Scores Tell Everything about Bob: Non-adaptive Face Reconstruction on Face Recognition Systems

Sunpill Kim[†‡], Yong Kiam Tan[‡], Bora Jeong[†‡], Soumik Mondal[‡], Khin Mi Mi Aung[‡] and Jae Hong Seo[†*]
[†]*Department of Mathematics, Research Institute for Natural Sciences, Hanyang University, Seoul, Republic of Korea*
[‡]*Institute for Infocomm Research (I²R), A*STAR, Singapore*
{*ksp0352, jbr915, jaehongseo*}*@hanyang.ac.kr,* {*tanyk1, soumikm, mmaung*}*@i2r.a-star.edu.sg,* [*]*Corresponding author*

***Abstract*—Face recognition systems (FRSs) typically store databases of discriminative real-valued *template* vectors, which are extracted from each enrolled user's facial image(s). Such template databases must be carefully protected for user privacy—indeed, the dangers of template leakages have been widely reported in the literature. In contrast, the *similarity scores* between queried images and enrolled users is often unprotected and can be readily queried through typical FRS APIs. Such scores provide a potential avenue of adversarial attack on FRSs, but recently proposed score-based attacks remain largely impractical because they essentially rely on *trial-and-error* strategies that use an enormous number of *adaptive* queries (>50K) for face reconstruction.**

We present the first practical score-based face reconstruction and impersonation attack against three commercial FRS APIs: AWS CompareFaces, FACE++, and KAIROS, as well as five commonly used pre-trained open-source FRSs. Our attack is carried out in the black-box FRS model, where the adversary has no knowledge of the FRS (underlying models, parameters, template databases, etc.), except for the ability to make a limited number of similarity score queries. Notably, the attack is straightforward to implement, requires no trial-and-error guessing, and uses a *small* number of *non-adaptive* score queries. We motivate the attack by analyzing the topological meaning of similarity scores and then present our novel method using *orthogonal face sets*: a precomputed approximate basis set of human-like face images that enables us to get meaningful similarity scores from a small number of non-adaptive queries. Our approach successfully reconstructs human-like impersonation images with $>20\%$ (resp. $>96\%$) success rates across three test datasets when directly attacking the AWS CompareFaces API (resp. open-source CosFace FRS) using only 100 queries—up to two orders of magnitude fewer queries than previous approaches. We provide evidence that personally identifiable biometric features are captured in our reconstructions by evaluating our approach in transfer-like attack settings and through other image similarity metrics.

1. Introduction

Biometric authentication systems are widely used in practice, e.g., eye/iris and face recognition at border controls and fingerprint recognition on mobile devices. Face recognition systems (FRSs) have gained significant popularity as a means of biometric user authentication because they are non-intrusive, highly accurate, and convenient to deploy. Recent advances in FRS accuracy are due, in large part, to the successful use of deep learning (DL)-based neural networks as biometric template extractors that can map from user facial images to short but highly discriminative real-valued template vectors. These templates are stored by FRSs, and used to identify users by comparing the template similarity of their facial images against the stored database.

In all biometric authentication systems, user privacy is a key priority that must never be compromised. The standard security requirement for biometric template protection is well established in ISO/IEC 24745 [1], and there is a significant body of research studying the dangers of template leakages [2], [3], [4], [5], as well as corresponding template protection methods [6], [7] that aim to securely store templates against adversaries that can compromise the database while allowing efficient computation of the desired scores. However, there is no rigorous regulation or guideline for privacy-preserving handling of scores, e.g., commercial FRS services [8], [9], [10] typically protect enrolled templates but allow clients to query for similarity scores against enrolled identities freely. Notably, a feasible attack on scores can be used to bypass template protection methods.

Recently, several *score-based* (SB) attack methods have been proposed in the literature [11], [12], [13], [14]. These SB attacks target a black-box FRS model, where the adversary has no knowledge of the target FRS, such as the structure of its underlying template extractors, model parameters, or template databases—in other words, the adversary can only glean limited information by making a number m of similarity score queries. At a high-level, existing attacks all use a *trial-and-error* hill climbing strategy to generate new images and select ones with higher scores iteratively; new image iterations are generated either through Gaussian distributions [11], EigenFace [12], or StyleGAN [13], [14]. However, such approaches have three key drawbacks:

1) they require an *enormous* number of similarity score queries to reconstruct face images ($m > 50$K) or even for impersonation ($m > 1$K);
2) personally identifiable features of the targeted identity are not well captured in reconstructed images;
3) their attack queries are *adaptive* for a fixed target user,

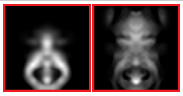
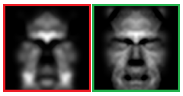
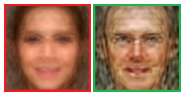
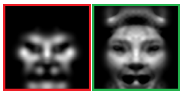
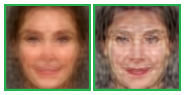
Target FRS (Direct)	T_1 (CosFace)						T_{AWS}	T_{KAIROS}
Target FRS (Transfer)	$T_{\mathsf{FACE++}}$							
Method / Target	Razzhigaev et al. [11]		Razzhigaev et al. [12]		Vendrow et al. [13]		Ours		
									
Queries	22400	200000	15600	200000	2300	200000	100	100	100
LPIPS	0.4998	0.3928	0.3087	0.3287	0.4485	0.4500	0.2714	0.2111	0.2154
									
Queries	48400	200000	5800	200000	2150	200000	100	100	100
LPIPS	0.5933	0.4750	0.3573	0.2537	0.4180	0.3877	0.2321	0.2332	0.2661
									
Queries	19200	200000	7200	200000	1100	200000	100	100	100
LPIPS	0.5352	0.3538	0.2829	0.2204	0.3781	0.3140	0.2270	0.2216	0.2656

TABLE 1: **(Direct attack)** Reconstructed face images using various attack approaches against the targeted FRSs; for prior approaches [11], [12], [13]: (left) face images at successful impersonation, (right) attempted full face reconstruction using a large number of queries. **(Transfer attack)** A Type-1 (resp. Type-2) attack is successful if the transfer target FRS considers the image similar to the original target image (resp. another image of the same identity); green box: both Type-1 and Type-2 transfer attack success; yellow box: success on Type-1 only; red box: failure on both types. **(LPIPS scores)** Lower is better, i.e., generated image is more similar to the original target image according to the LPIPS metric.

i.e., each new query is sequentially determined based on preceding score queries.

These drawbacks limit their applicability against commercial FRSs, e.g., making many sequential incorrect queries can have significant monetary costs.

1.1. Our Contribution

We propose a new *non-adaptive* SB attack for face reconstruction, which requires up to two orders of magnitude fewer queries ($m = 100$) than previous SB attacks (with $m > 50$K, see Section 2). Non-adaptiveness means that all of our attack queries against an enrolled identity can be made independently and simultaneously, without waiting to learn from the results of previous queries. We demonstrate the effectiveness of our attack on five open-source FRS [15], [16], [17], [18], [19] and three commercial FRS [8], [9], [10] with three widely-used test datasets, LFW [20], AgeDB [21], and CFP-FP [22]. In fact, for the three commercial FRS, we do not have any public information about their underlying model architecture, loss functions, training datasets, or even their internal methods used to calculate scores; thus, our attack is carried out in a real-world black-box FRS setting.

Table 1 shows various sample face reconstructions from our experiments using different methods and targets (full details in Section 5). The small number of queries used in our approach makes it feasible and cheap to apply against commercial FRS targets (T_{AWS}, $T_{\mathsf{F++}}$, and T_{KAI}), with significant success rates: up to 29.6%, 19.9%, and 12.2% when targeting the AgeDB dataset for the three commercial FRSs, respectively. Our reconstructed images can be used for transfer-like attacks, i.e., using the results from attacking target T_A against another target T_B, or another image of the same identity. They also outperform prior work in classical image similarity metrics such as the Learned Perceptual Image Patch Similarity (LPIPS) [23] and Deep Image Structure and Texture Similarity (DISTS) [24] (latter score not shown here). This indicates that our attack can successfully recover some personally identifiable biometric face features of the targeted identities using only $m = 100$ score queries.

1.2. A New Topological Approach Using Orthogonal Face Sets

Although similarity scores reveal partial information about enrolled templates, one can hardly design a non-adaptive SB attack without a clear mathematical intuition and understanding of the processes used to generate similarity scores. We motivate our attack by analyzing the topological meaning of similarity scores generated by DL-based FRS models trained by so-called *metric learning* [15], [16], [17], [18], [19], [25], [26]. In metric learning, a feature extractor's parameters are trained using a loss function which is designed to make the model's latent space mimic a metric space. Many biometric recognition systems are trained by metric learning with angular loss as it has been shown to successfully extract discriminative features from various types of biometric information [27], [28], [29], [30].

To design our attack, we assume that the FRS feature extractor is well-trained via metric learning. More precisely,

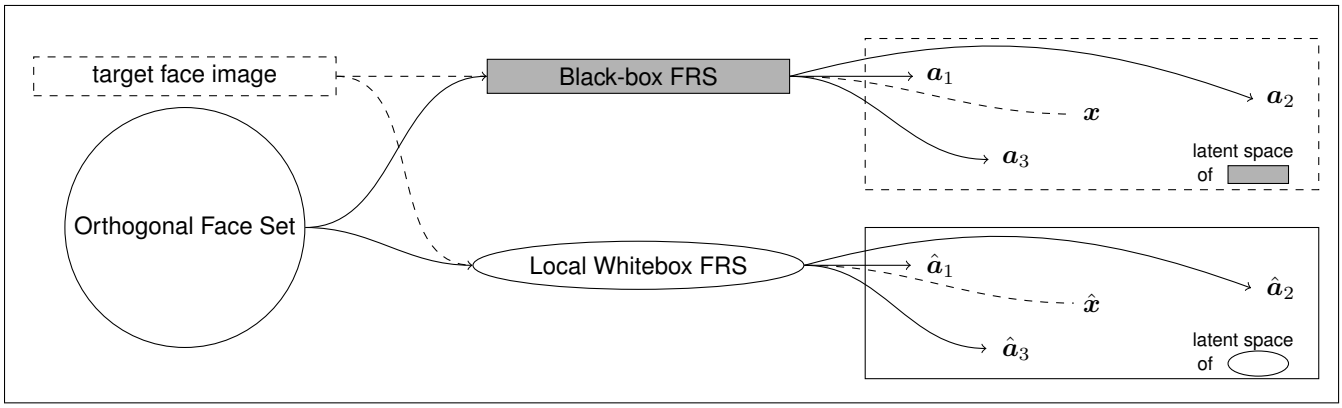


Figure 1: Conjecture 1 says that $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \approx \langle \hat{\boldsymbol{a}}_i, \hat{\boldsymbol{x}} \rangle$, where $\langle \cdot, \cdot \rangle$ is the standard inner product, i.e., the similarity scores of a target FRS are approximately the same as those of a local FRS. Conjecture 2 says that the OFS and similarity scores from the local FRS can be used to determine $\hat{\boldsymbol{x}}$. If one has $\hat{\boldsymbol{x}}$, the local FRS can be inverted to reconstruct a face image that maps (close) to $\hat{\boldsymbol{x}}$ via the local FRS and (close) to $\boldsymbol{x}$ via the target FRS.

we present two intuitive conjectures and analyze their consequences. Our first conjecture is as follows:

Conjecture 1. *Given two DL-based FRS independently trained by metric learning, their latent spaces V and $\widehat{V}$ are almost isometric via face images. If the feature vectors $\boldsymbol{x}, \boldsymbol{y} \in V$ of two face images are close (resp. far), the feature vectors $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \widehat{V}$ in the other latent (metric) space of the same face images are also close (resp. far).*

Several previous template-based attacks [2], [3], [4] use the above conjecture implicitly to circumvent the difficulties of the black-box FRS model. For example, consider training an inverse model of a target FRS to reconstruct face images using perceptual loss [31]. Naively, the parameters of the target FRS would be required to train the reconstructor, but this is not available in the black-box FRS model. Instead, another local (white-box) FRS is used to compute substitute gradients [2]. Conjecture 1 can be used to help explain the success of using a local white-box FRS to replace gradients from the target black-box FRS.

Although a local FRS can be used to substitute outputs from the target black-box FRS, more is needed for achieving high-accuracy attacks with low resources. For example, the base attack success rate is low [2], and additional machine learning techniques, such as knowledge distillation [32], are needed to mimic further the target black-box FRS [3], [4]. These additional techniques drastically improve the accuracy of template-based attacks, but they require many queries. We present a second conjecture to help design a new method using only a small number of non-adaptive queries.

Conjecture 2. *For a well-trained FRS with angular loss, there is a set $\mathcal{O}$ of m face images such that for any vector $\hat{\boldsymbol{x}}$ in the FRS's latent space, the set of feature vectors $\hat{\boldsymbol{a}}_i$ of $\mathcal{O}$ and the set of cosine similarity values $\{\langle \hat{\boldsymbol{a}}_i, \hat{\boldsymbol{x}} \rangle + e_i\}_{i \in [m]}$ with some errors e_i can be used to approximate $\hat{\boldsymbol{x}}$. We call the set $\mathcal{O}$ an orthogonal face set (OFS) as it is similar to an orthogonal basis for human face images.*

We design a new *non-adaptive* SB attack, where the intuitive use of Conjectures 1 and 2 is shown in Figure 1. Briefly, our attack consists of a pre-processing and a reconstruction step. In the pre-processing step, we generate an OFS that is expected to satisfy Conjecture 2. This step is performed using a pre-trained local feature extractor F_L and its inverse model F_L^{-1} regardless of the target FRS T. In the reconstruction step, we reformulate the problem by using a small number of online queries to T to find a template $\hat{\boldsymbol{x}}$ of the local model instead of directly solving the problem against the target (unknown) FRS. Finally, we reconstruct a face image using the inverse model F_L^{-1} and use it to fool T. In Section 3, we provide analysis and empirical evidence for our conjectures. We provide an algorithm for finding an OFS and discuss other attack details in Section 4.

2. Related Work

Many adversarial attacks on DL-based FRS have been proposed with the aim of compromising the FRS verification/identification process under various scenarios. Sharif et al. [33] developed physically realizable and inconspicuous impersonation attacks on FRS; they further demonstrate that their approach can be adapted to black-box scenarios by attacking FACE++. In their impersonation scenario, images of the targeted identity are available, and the goal is to fool the FRS into misclassifying a source identity as the target. In contrast, we consider *face reconstruction* (and subsequent impersonation) with no information about the enrolled identity other than the similarity score queries. There is a similar attack scenario called *model inversion* [34], [35], [36], which aims to find pre-images of identities used in the training process (so-called closed-set scenario, Definition 1 in Duong et al. [3]), as opposed to reconstruction of arbitrary input face images (so-called open-set scenario, Definition 2 in Duong et al. [3]). Our work lies within the broader umbrella of black-box adversarial [37] and privacy [34] attacks on image recognition systems, but we focus on

Method	Basis	Input	Strategy	Pre-processing	Target Model	Number of Queries: Reconstruction	Number of Queries: Impersonation
[2]	Template-based (learning)	templates from queried image and target image	training the inverse model of target model	VGG-19	FaceNet	≥1M	N/A
[3]				ResNet-50	FaceNet, ArcFace, ...		
[4]							
[5]				StyleGAN	ArcFace		
[11]	Score-based (adaptive)	cosine similarity score between queried image and target image	genetic algorithm	Gaussian Blob	FaceNet, ArcFace	300K	≈25K
[12]				EigenFace		50K	≈4K
[13]				StyleGAN	FaceNet	≤160K	≈1K
[14]					ArcFace	≤256K	N/A
Ours	Score-based (non-adaptive)	confidence score (same as above)	linear equation	ResNet-50 and NbNet	FaceNet, ArcFace, ...	100	100
			weighted sum		CompareFaces (AWS)		
					FACE++, KAIROS		

TABLE 2: A summary of various FRS attack methods. The number of queries for full face reconstruction for each attack is taken from the literature. We also evaluate open source implementations (where available [11], [12], [13]) of previous attacks in our experiments to provide the number of queries required to carry out impersonation.

exploiting features specific to FRSs such as the fact that their underlying models are trained with metric learning.

To the best of our knowledge, most recent reconstruction attacks on FRS can be categorized into two groups: template-based or score-based attacks. Table 2 summarizes related approaches and our attack with respect to the capabilities of the adversary, target model, and the number of required queries.

2.1. Template-based Attack

Template-based attacks use leaked templates from queried images to find an inverse model (from templates to images). In this attack scenario, there is an assumption that the adversary can steal templates enrolled in the database of a target FRS in order to train an inverse model. Using the inverse model, the adversary can compromise user privacy by inverting new templates output by the target FRS.

There are several lines of research deploying template-based attacks [2], [3], [4], [5]. Mai et al. [2] proposed NbNet, an inverse model of FaceNet [19], consisting of neighborly de-convolutional blocks using a DenseNet architecture to address the issues of repeated, noisy, and insufficient channel details. When training NbNet, they reconstruct face images similar to the original image of the template using pixel loss and perceptual loss. The perceptual loss proposed by Johnson et al. [31] is particularly important because it trains the reconstructed image to be recognized as the same identity by different FRSs; Mai et al. [2] use the L2-norm of the intermediate feature map of Simonyan and Zisserman [38] to calculate the perceptual loss. Duong et al. [3] and Truong et al. [4] target the ArcFace FRS [17] by training GAN-based inverse models DiBiGAN and DAB-GAN which have the structure of ProGAN [39] and TransGAN [40], respectively. The adversarial loss is added for training each GAN and the bijective loss is also added for a one-to-one correspondence between the latent space and the face image space in the structure of the GAN. Therefore, the total loss term in each approach consists of 4 components, including pixel loss and perceptual loss, the latter is calculated using the feature map of He et al. [41].

Dong et al. [5] use a pre-trained StyleGAN2 [42], [43] to reconstruct face images. They train a multi-layer perceptron that takes templates of target FRS ArcFace as inputs and returns input vectors of StyleGAN2 using mean squared error as a loss function. This approach using StyleGAN2 has advantages in terms of generating high-resolution and human-like face images compared to previous methods [2], [3], [4]. Nevertheless, the ability to steal target templates and make many template queries to the target feature extractor is a very strong limiting assumption.

2.2. Score-based (SB) Attack

Score-based (SB) attacks are the focus of this paper; recall that in this black-box attack scenario, the adversary has no access to the FRS. In particular, it cannot steal templates enrolled or generated by the target FRS. Instead, the adversary can only obtain similarity scores by querying the target FRS. We additionally focus on the number of FRS queries m as an important metric for this paper to model the cost of real-world attack scenarios. Adversarial SB attacks typically run an optimization algorithm to find an image with the highest similarity to a given target. There are several score-based attacks [11], [12], [13], [14] that use genetic algorithm-type approaches, with respective target FRSs as shown in Table 2. The approach by Razzhigaev et al. [11] starts with 4480 Gaussian Blobs, which are Gaussian noise images in the shape of components of human faces such as the eyes and nose. The adversary queries the target FRS for similarity scores corresponding to the 4480 images and then sets the most similar image as the initial image. Next, the adversary adds a batch of noise to the current image and sends this batch of new images to the target FRS, updating the current image to the most similar image among the new images. The noise is generated from Gaussians with randomly sampled parameters at each iteration. This method uses up to $m = 300000$ similarity queries in total, including the 4480 queries made during initialization. A subsequent work [12], reduced the similarity queries needed to around $m = 50000$ by replacing Gaussian Blobs with 1024 images generated by EigenFace [44], a method for generating major components in human faces. Vendrow et al. [13] and Dong

et al. [14], propose optimization attacks using a pre-trained StyleGAN2 [42], [43] to improve the quality of generated face images. They run genetic algorithms on the input vector of a StyleGAN2, where the input vector is initialized to zero or a random vector and random noise is added to the vector at each iteration. These methods aim to find an input vector that generates a face image with highest similarity score. The methods use up to $m = 161600$ and 256000 queries [13], [14], respectively.

Tao et al. [36] target two commercial FRSs: Microsoft Azure and Clarifai. We note that their attack is designed for closed-set attack scenarios but can be modified to apply to open-set scenarios. Nevertheless, their approach also uses a trial-and-error strategy and requires many online queries (100K for initialization). In this paper, we focus on drastically reducing m to make it cheap and feasible to apply against FRSs, including commercial systems.

3. Background and Analysis of Conjectures

3.1. Notation

An n-dimensional real vector in $\mathbb{R}^n$ is denoted by boldface lowercase letters, e.g., $\boldsymbol{x}$, and often considered as an $n \times 1$ column matrix. The standard inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. Matrices are denoted by boldface uppercase letters, e.g., $\boldsymbol{A}$. A superscript $(\cdot)^T$ denotes the vector/matrix transpose operation. For example, $\boldsymbol{x}^T \in \mathbb{R}^{1\times n}$ is the transpose of the column vector $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$. We denote the L2-norm of a vector/matrix as $||\cdot||$ and the set of natural numbers less than or equal to $m \in \mathbb{N}$ as $[m]$.

3.2. Errors for Linear Systems

When solving a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ for $\boldsymbol{x}$, we often do not know $\boldsymbol{b}$ exactly, but instead have $\boldsymbol{b}' = \boldsymbol{b} + \boldsymbol{e}$, which contains a small unknown error term $\boldsymbol{e}$. The best thing we can do is to solve $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{b}'$ exactly, giving us a different solution vector $\boldsymbol{x}'$. The error sensitivity in the linear system can be measured using the condition number of matrix $\boldsymbol{A}$, defined as $\mathrm{cond}(\boldsymbol{A}) = ||\boldsymbol{A}||||\boldsymbol{A}^+|| \geq ||\boldsymbol{A}\boldsymbol{A}^+|| = 1$, where $\boldsymbol{A}^+$ is the pseudoinverse of $\boldsymbol{A}$. Note that for a square non-singular matrix $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$. From the condition number, we can bound how the relative error $||\boldsymbol{b}' - \boldsymbol{b}||/||\boldsymbol{b}||$ in $\boldsymbol{b}$ influences the relative error $||\boldsymbol{x}' - \boldsymbol{x}||/||\boldsymbol{x}||$ in $\boldsymbol{x}$ by:

$$\frac{||\boldsymbol{x}' - \boldsymbol{x}||}{||\boldsymbol{x}||} \leq \mathrm{cond}\,(\boldsymbol{A}) \frac{||\boldsymbol{b}' - \boldsymbol{b}||}{||\boldsymbol{b}||}. \tag{1}$$

A system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with large $\mathrm{cond}\,(\boldsymbol{A})$ is called an ill-conditioned problem. Informally, for a small change (error $\boldsymbol{e}$) in the inputs (vector $\boldsymbol{b}$), there is a large change in the approximated solution (vector $\boldsymbol{x}'$). This means that the original solution (vector $\boldsymbol{x}$) to the linear system becomes hard to find. Note that the matrix norm of any orthonormal matrix is 1, which means the condition number of an orthonormal matrix is always 1 (the minimum condition number).

3.3. Face Recognition System and Threat Models

Modern FRSs typically use a deep convolutional neural network as a deep feature extractor $F : \mathcal{I} \to \mathbb{R}^n$ to map a human face image to the feature vector or *template* space. Note that images in $\mathcal{I}$ may have been refined by appropriate image processing algorithms such as face detection and face alignment. For a target FRS T, we write F_T for its associated feature extractor. For simplicity, we assume that templates are always normalized and following recent studies [15], [16], [17], [25], [26], we focus our discussion on the cosine similarity $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ between templates $\boldsymbol{x}$ and $\boldsymbol{y}$.

To motivate our threat model, note that if the feature extractor F takes images $i \in \mathcal{I}$ as input and outputs a template $F(i) \in \mathbb{R}^n$ and there is no limit to the number of queries, then the adversary can train an inverse model F^{-1} such that $F^{-1}(F(i)) \approx i$ and reconstruct the image $F^{-1}(\boldsymbol{x})$ from any compromised template $\boldsymbol{x} \in \mathbb{R}^n$. However, the output of an FRS is usually *not* the raw template.

For a black-box FRS T, fix an enrolled image $\mathsf{img}_{\mathsf{target}}$. The goal of the adversary is to reconstruct $\mathsf{img}_{\mathsf{target}}$ through the query interface of T, thus uncovering private information about the targeted identity. We assume that the adversary can only obtain an output score between the templates of the enrolled image and a queried image $\mathsf{img}_{\mathsf{query}} \in \mathcal{I}$. No other access to the feature extractor F_T is available. We further consider two types of FRS output scores:

- **Cosine similarity.** Target FRS $T : \mathcal{I} \to [-1, 1]$ is the function $T(\mathsf{img}_{\mathsf{query}}) = \langle F_T(\mathsf{img}_{\mathsf{target}}), F_T(\mathsf{img}_{\mathsf{query}}) \rangle$
- **Confidence.** Target FRS $T : \mathcal{I} \to [0, 1]$ is the function $T(\mathsf{img}_{\mathsf{query}}) = h(F_T(\mathsf{img}_{\mathsf{target}}), F_T(\mathsf{img}_{\mathsf{query}}))$ where h is an unknown function mapping the output of the FRS's feature extractor F_T to confidence probabilities.

There is a difference between these two types of scores in terms of the information revealed about the latent space. If the cosine similarity score is provided, the adversary can directly solve a suitable linear equation to recover target templates. For the confidence score, the adversary does not have any information about how the final score between two templates is calculated, which makes matters more challenging. The cosine similarity score is often used by open-source FRSs, while many commercial FRSs output a confidence estimate for their prediction as a probability.

3.4. Analysis of Conjectures 1 & 2

3.4.1. Conjecture 1. The main goal of deep (metric) learning-based feature extractors in FRS is to achieve strong intra-class compactness and inter-class separability, i.e., images from the same identity are close while images from different identities are distant. If two feature extractors F_L and F_T are well-trained, then FRSs using the respective extractors should output similar cosine similarities for the same input image pairs with $\langle F_L(i), F_L(i') \rangle \approx \langle F_T(i), F_T(i') \rangle$.

In Figure 2, we plot the distribution of cosine similarities for all pairs from the LFW dataset; the dataset consists of both true pairs (same identity) and false pairs (different identities). The X-axis indicates the cosine similarity

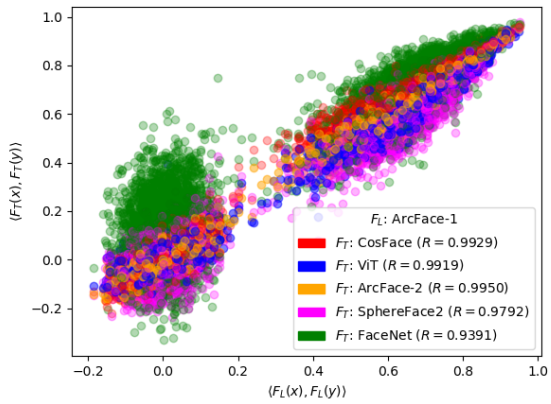


Figure 2: Scatter plots of the cosine similarity values output by feature extractors for all pairs from the LFW dataset; R is the correlation coefficient between scores output by F_L and different extractors F_T.

between two images from a local feature extractor F_L, and the Y-axis indicates the cosine similarity between those same images from five target feature extractors. The plot shows that the output cosine similarity for each image pair is nearly identical regardless of the feature extractor, which means that Conjecture 1 is empirically satisfied.

3.4.2. Conjecture 2. We can consider Conjecture 2 as solving a linear system with an error as in Section 3.2. More precisely, the equation is defined as follows:

$$\widehat{\boldsymbol{A}}\hat{\boldsymbol{x}} = \hat{\boldsymbol{b}}, \hat{\boldsymbol{b}'} = \hat{\boldsymbol{b}} + \boldsymbol{e} \tag{2}$$

where each row in $\widehat{\boldsymbol{A}} \in \mathbb{R}^{m\times n}$ is $\hat{\boldsymbol{a}}_i^T \in \mathbb{R}^{1\times n}$ and we only know $\hat{\boldsymbol{b}'}$ for some unknown error terms $\boldsymbol{e} = [e_1, \cdots, e_m]^T$.

Although solving this system is not a trivial problem in general, two things can help us approximate $\hat{\boldsymbol{x}}$ well. First, we can choose $\widehat{\boldsymbol{A}}$ arbitrarily.[1] The second is that we expect the norm of $\hat{\boldsymbol{x}}$ to be close to 1. This is natural because cosine similarity is used in metric learning-based FRSs.

To exploit these two advantages, we consider choosing a pre-processed image set $\mathcal{O}$ such that $\widehat{\boldsymbol{A}}$ forms an orthogonal basis with condition number 1. This not only reduces the relative error of the solution of the linear system but also helps to better approximate $\hat{\boldsymbol{x}}$. Briefly, by using the pseudoinverse matrix $\widehat{\boldsymbol{A}}^+$ and finding $\boldsymbol{z} = \widehat{\boldsymbol{A}}^+(\hat{\boldsymbol{b}} + \boldsymbol{e})$, we obtain $\boldsymbol{z}$ as an approximation of $\hat{\boldsymbol{x}}$, where its norm is $||\boldsymbol{z}|| = ||\hat{\boldsymbol{b}} + \boldsymbol{e}||$ from $(\widehat{\boldsymbol{A}}^+)^T\widehat{\boldsymbol{A}}^+ = \boldsymbol{I}$. Here, $\boldsymbol{z}$ is a least square solution such that $||\widehat{\boldsymbol{A}}\boldsymbol{z}' - \hat{\boldsymbol{b}'}|| \geq ||\widehat{\boldsymbol{A}}\boldsymbol{z} - \hat{\boldsymbol{b}'}||$ for all $\boldsymbol{z}' \in \mathbb{R}^n$, and if $||\hat{\boldsymbol{b}} + \boldsymbol{e}||$ is almost 1, our approximation also has a norm of almost 1.

From Conjecture 1, we can consider the error term e_i as $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - \langle \hat{\boldsymbol{a}}_i, \hat{\boldsymbol{x}}\rangle$, which implies e_i is not large, where $\boldsymbol{a}_i$ and $\hat{\boldsymbol{a}}_i$ (resp. $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$) are template of the pre-processed image set (resp. target image) using feature extractors F_T and F_L. We can reformulate the problem using a local FRS with small errors instead of the target FRS.

[1] We consider the minimum number of queries m, which means $m < n$ where n is the FRS template dimension, usually 512 in practice.

	F_{T1}	F_{T2}	F_{T3}	F_{T4}	F_{T5}
ϵ (upper bound)	19.01	22.81	20.61	31.38	58.02
average	4.75	5.41	4.84	8.42	23.41

TABLE 3: Experimental ϵ of ϵ-isomtery $F_T \circ F_L^{-1}$ and the average of the difference between two angles in each space.

3.4.3. Additional Evidence. Even if template $\hat{\boldsymbol{x}}$ is perfectly recovered in the above manner using the local FRS F_L, we still require a local inverse model F_L^{-1} in order to reconstruct an *image* from $\hat{\boldsymbol{x}}$. The question remains how close $F_T(F_L^{-1}(\hat{\boldsymbol{x}}))$ is to the template $\boldsymbol{x}$, where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are $F_T(\mathsf{img}_{\mathsf{target}})$ and $F_L(\mathsf{img}_{\mathsf{target}})$, respectively. To investigate this, we define the following new notion of ϵ-isometry to show how much the angles are preserved.

Definition 1 (Almost Isometry). *Given a positive real number ϵ, an ϵ-isometry or almost isometry is a map $M : V \to W$ between metric spaces V and W such that for $\boldsymbol{v}, \boldsymbol{v}' \in V$ one has $|d_W(M(\boldsymbol{v}), M(\boldsymbol{v}')) - d_V(\boldsymbol{v}, \boldsymbol{v}')| < \epsilon$, and for any point $\boldsymbol{w} \in W$ there exists a point $\boldsymbol{v} \in V$ with $d_W(\boldsymbol{w}, M(\boldsymbol{v})) < \epsilon$, where d_V and d_W are the metrics of V and W, respectively.*

For given pre-trained local inverse model F_L^{-1} and target feature extractor F_T, we can consider the transformation $M : \mathbb{R}^n \to \mathbb{R}^n$ as $F_T \circ F_L^{-1}$ with $d_{\mathbb{R}^n}$ being angular distance (recall that templates are normalized). The smaller ϵ is, the more similar template to $\boldsymbol{x}$ can be obtained using M. In Table 3, we provide an experimental upper bound ϵ and the average value of the difference between the angles in each target system based on 10000 randomly sampled template vectors $\boldsymbol{v}, \boldsymbol{v}'$. Table 3 tells us that if we can approximate the local template $\hat{\boldsymbol{x}}$ well using $\boldsymbol{z}$, then by selecting a good local inverse model F_L^{-1}, we can also reconstruct the image $F_L^{-1}(\boldsymbol{z})$ such that the angle between corresponding template $F_T(F_L^{-1}(\boldsymbol{z}))$ and $\boldsymbol{x}$ in the target FRS will be small.

4. Proposed Method

In this section, we present our new non-adaptive SB attack on black-box FRSs. We start by describing our attack algorithm under the setting where the adversary can obtain cosine similarity scores. Then, we provide the GenOFS algorithm to generate an Orthogonal Face Set (OFS) of human-like face images, which is the main ingredient of our attack. Finally, we provide two additional techniques to attack commercial FRSs by circumventing the problem that these FRSs do not output cosine similarity scores directly. An illustrated overview of our attack is provided in Figure 3.

4.1. Proposed Attack Algorithm

Fix a target FRS T and let $\boldsymbol{x} = F_T(\mathsf{img})$ be the enrolled template for a targeted identity. Suppose we have a well-

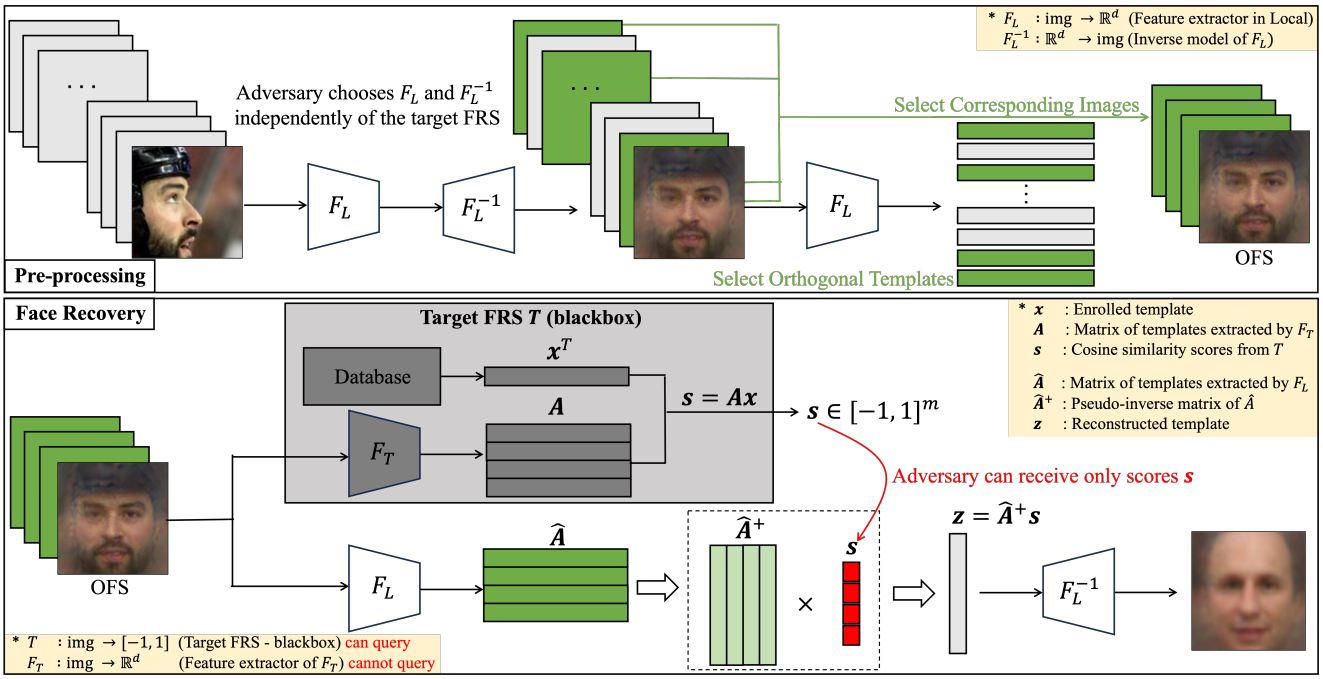


Figure 3: An overview of our attack which consists of a pre-processing and a reconstruction step. In pre-processing, templates of the OFS are chosen to be (nearly) orthogonal to each other in order to obtain as much information as possible from a small number of queries. Face reconstruction is carried out by reformulating the linear system for target template $\boldsymbol{x}$ into a linear system for $\hat{\boldsymbol{x}}$ consisting of known local FRS templates, followed by applying a local inverse model.

trained local feature extractor F_L, its local inverse model, F_L^{-1}, and a face image set $\mathcal{O} = \{o_i | i \in [m]\}$.

Querying T with o_i yields the m similarity scores $\langle \boldsymbol{x}, F_T(o_i) \rangle$. From Conjecture 1, we have $\langle \boldsymbol{x}, F_T(o_i) \rangle \approx \langle \hat{\boldsymbol{x}}, F_L(o_i) \rangle$, so we can reformulate the problem to find a template $\hat{\boldsymbol{x}} = F_L(\mathsf{img})$ for F_L instead of directly solving the problem against the unknown feature extractor F_T of target FRS T. More precisely, the output scores is represented by the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{s}$ where $\boldsymbol{A}$ is a $(m \times n)$-matrix with rows $[\boldsymbol{A}]_i = F_T(o_i)$. Based on Conjecture 1, the similarity scores $\boldsymbol{s}$ from the target FRS serve as a good proxy for the similarity scores $\hat{\boldsymbol{s}}$ that we expect to get from F_L, thus:

$$\boldsymbol{A}F_T(\mathsf{img}) = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{s} = \underbrace{\hat{\boldsymbol{s}} + \boldsymbol{e} \approx \hat{\boldsymbol{s}}}_{\text{Conjecture 1}} = \widehat{\boldsymbol{A}}\hat{\boldsymbol{x}} = \widehat{\boldsymbol{A}}F_L(\mathsf{img})$$

where $\widehat{\boldsymbol{A}}$ is a $(m \times n)$-matrix with rows $[\widehat{\boldsymbol{A}}]_i = F_L(o_i)$.

Note that the terms highlighted in red are unavailable to the attacker. Nevertheless, we can attempt to reconstruct the target image img using two further approximations:

$$\underbrace{\mathsf{img} \approx \overbrace{F_L^{-1}(\hat{\boldsymbol{x}})}^{\text{Conjecture 2}} \approx F_L^{-1}(\boldsymbol{z})}_{\text{Almost isometry property}} = F_L^{-1}(\widehat{\boldsymbol{A}}^+ \boldsymbol{s}) \qquad (3)$$

Here, Conjecture 2 says that if $\widehat{\boldsymbol{A}}$ is a well-chosen basis, then z is a good approximation of $\hat{\boldsymbol{x}}$, and the almost isometry property of $F_T \circ F_L^{-1}$ means it closely preserves cosine similarity values. That means $\langle \hat{\boldsymbol{x}}, \boldsymbol{z} \rangle \approx \langle \boldsymbol{x}, F_T(F_L^{-1}(\boldsymbol{z})) \rangle$, i.e., our reconstructed image $F_L^{-1}(\boldsymbol{z})$ will have a template that is close to $\boldsymbol{x}$ for the target FRS. The detailed description of attack algorithm is presented in Algorithm 1.

Algorithm 1 AttackCosSim

Require: $F_L : \mathcal{I} \to \mathbb{R}^n$, $F_L^{-1} : \mathbb{R}^n \to \mathcal{I}$, target FRS $T : \mathcal{I} \to [-1, 1]$ and face image set $\mathcal{O} = \{o_1, \cdots, o_m\} \subset \mathcal{I}$
Ensure: Reconstructed face image $\mathsf{img}' \in \mathcal{I}$

```
for i ∈ [m] do
    â_i ← F_L(o_i)
    s_i ← T(o_i)                          // Query the score
end for
Â ← [â_1, ··· , â_m]^T and s ← [s_1, ··· , s_m]^T
z ← Â^+ s                                 // Solve a linear system
img' ← F_L^{-1}(z)                        // Reconstruct the image
Return img'
```

It is important to note that the solving and reconstruction are done using local models F_L, F_L^{-1} without needing any knowledge about target FRS except for the cosine similarity queries (one for each element of the set $\mathcal{O}$). There are two remaining problems to be tackled:

1) we need to generate the set $\mathcal{O}$ independently of the target feature extractor, and

2) for target FRSs that output confidence scores, we do not have direct access to similarities, so the above reformulation for $\hat{\boldsymbol{x}}$ does not work directly.

4.2. Orthogonal Face Set (Pre-processing)

In this subsection, we provide an algorithm to generate an OFS inspired by Conjecture 2. Our main idea is use a well-trained local feature extractor $F_L : \mathcal{I} \to \mathbb{R}^n$ and its inverse model $F_L^{-1} : \mathbb{R}^n \to \mathcal{I}$ (the same models used in attacking T). We emphasize that F_L and F_L^{-1} are local models which can be arbitrarily pre-trained by the adversary, and the adversary can make an unlimited number of (local) queries to these models.

From F_L, we can compute the cosine similarity between the templates of each image pair in any given face image dataset $\mathcal{Z}$. We can then choose a subset $\mathcal{O} \subseteq \mathcal{Z}$ such that their templates extracted by F_L are pairwise orthogonal to each other. However, finding a large set of face images with templates that are exactly pairwise orthogonal is challenging. Therefore, we used a relaxed version[2] whose templates are δ-orthogonal to each other, as in the following definition.

Definition 2 (Almost Orthogonal Set). *For $\delta \geq 0$, a set of unit vectors V is an δ-orthogonal set or almost-orthogonal set if the cosine similarity between any pair of two distinct vectors $v, v' \in V$ is smaller than δ, i.e., $|\langle v, v' \rangle| \leq \delta$.*

For a pre-trained local feature extractor F_L and image set $\mathcal{Z} = \{z_1, \cdots, z_l\}$, a naïve approach to generate an OFS is to let $\boldsymbol{v}_i = F_L(z_i)$ for $i \in [l]$. Then, we can select the subset $\mathcal{Z}' = \{z_i | i \in J\} \subseteq \mathcal{Z}$ as OFS using indices J such that $\max_{j,k \in J} |\langle \boldsymbol{v}_j, \boldsymbol{v}_k \rangle| \leq \delta$. This naïve approach can generate δ-orthogonal face image set correctly, but for reconstruction, it is not the best way to generate OFS.

We observe that the inverse model F_L^{-1} is typically trained by perceptual loss, which measures image similarities more robustly than pixel loss [2]. Thus, facial image templates generated by F_L and inverted by F_L^{-1} should result in images that capture personally identifiable biometric features of user identities from input images while excluding information irrelevant to the identity. Accordingly, we expect that output images from the inverse model will be more useful as part of the OFS as compared to raw input face images. For example, the images shown in Figure 4 are human-like face images generated by an inverse model. Note that each image is front facing and has less identity-irrelevant information such as background details. For this reason, we use images $F_L^{-1}(F_L(z_i))$ as the starting point to find a set of face image set whose templates are almost-orthogonal to each other. The proposed GenOFS algorithm and an example OFS are provided in Algorithm 2 and Figure 4, respectively.

[2]Note that if all templates in OFS are pairwise orthogonal, the condition number of the corresponding matrix is 1. In our experiments, we set $\delta = \cos 85^\circ \approx 0.0872$ and the resulting condition number is 2.6.

Algorithm 2 GenOFS

Require: $F_L : \mathcal{I} \to \mathbb{R}^n$, $F_L^{-1} : \mathbb{R}^n \to \mathcal{I}$, face image set $\mathcal{Z} = \{z_1, \cdots, z_l\} \subset \mathcal{I}$, and pre-defined parameter δ
Ensure: Orthogonal Face Set $\mathcal{O}$
1: **for** $i \in [l]$ **do**
2: $\quad \boldsymbol{v}_i \leftarrow F_L(F_L^{-1}(F_L(z_i)))$
3: **end for**
4: Set $I \subseteq [l]$ such that $\max_{j,k \in I} |\langle \boldsymbol{v}_j, \boldsymbol{v}_k \rangle| \leq \delta$
5: $\mathcal{O} \leftarrow \{o_i \in \mathcal{I} | o_i = F_L^{-1}(F_L(z_i)) \wedge i \in I\}$
6: **Return** $\mathcal{O}$

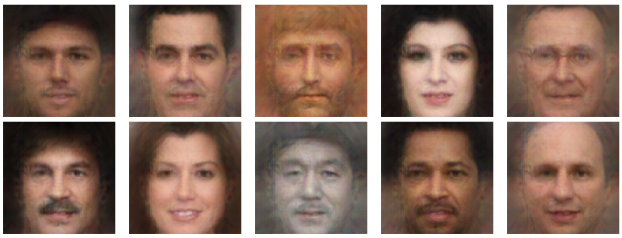


Figure 4: Examples of the Orthogonal Face Set

4.3. Reformulation using Confidence Score

We now turn to the setting where an adversary can only obtain confidence scores between an enrolled image and queried images. Note that we do not have information about the exact metrics used in commercial FRSs [8], [9], [10], nor the functions used to convert from (assumed) cosine similarities to confidence probabilities. To attack such systems, we provide two additional heuristic techniques.

Our first technique is a method to convert confidence scores back into cosine similarity scores, which we can use with Algorithm 1. Recent work by Knoche et al. [45] introduced a way to compute the confidence score for FRSs based on cosine similarity between two input images. They use the DOGBOX algorithm [46] to determine the coefficients of a logistic sigmoid function $f(d)$, defined as follows:

$$f(d) = \frac{L}{1 + e^{-k \cdot (d - d_0)}} + b \tag{4}$$

where L, d_0, k, and b are coefficients.

Using a local feature extractor F_L, we follow their algorithm to fit a logistic sigmoid function $f(\cdot)$ and then use the inverse function $f^{-1}(\cdot)$ as a proxy to invert confidence scores into cosine similarity scores. The adversary can obtain the coefficients of the logistic sigmoid function using any dataset consisting of true-and-false image pairs. Moreover, the function $f^{-1}(\cdot)$ can be computed as part of the pre-processing step using a local model F_L and does not affect the number of online queries that have to be made to a target FRS in our attack.

In Figure 5, we plot the confidence scores output by two commercial FRSs AWS and FACE++ for the CFP-FP dataset and the logistic sigmoid determined by the DOGBOX algorithm on that dataset. The plot shows that the estimation

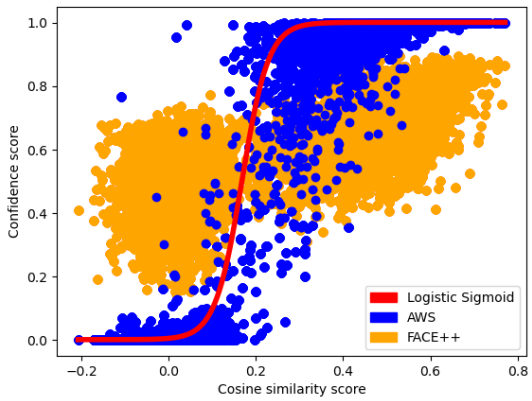


Figure 5: Comparison between predicted pairs (cosine similarity, confidence) and genuine pairs for CFP-FP

Algorithm 3 AttackConfidence

Require: F_L, F_L^{-1}, target FRS $T : \mathcal{I} \to [0, 1]$, face image set $\mathcal{O}$, and auxiliary information aux (description of function $f^{-1} : [0, 1] \to [-1, 1]$ or parameter $l \in [m]$)
Ensure: Reconstructed face image $\text{img}' \in \mathcal{I}$

for $i \in [m]$ **do**
  $\hat{\boldsymbol{a}}_i \leftarrow F_L(o_i)$
  $c_i \leftarrow T(o_i)$ // Query the score
end for
if $l \notin$ aux **then** // First Technique
  $s_i \leftarrow f^{-1}(c_i)$ for $\forall i \in [m]$
  $\hat{\boldsymbol{A}} \leftarrow [\hat{\boldsymbol{a}}_1, \cdots, \hat{\boldsymbol{a}}_m]^T$ and $\boldsymbol{s} \leftarrow [s_1, \cdots, s_m]^T$
  $\boldsymbol{z} \leftarrow \hat{\boldsymbol{A}}^+ \boldsymbol{s}$
else // Second Technique
  Set $I \subseteq [m]$ such that $\min_{j \in I} c_j \geq \max_{k \notin I} c_k$ and $|I| = l$
  $\boldsymbol{z} \leftarrow \sum_{i \in I} c_i \hat{\boldsymbol{a}}_i$
end if
$\text{img}' \leftarrow F_L^{-1}(\boldsymbol{z})$ // Reconstruct the image
Return img'

of the conversion algorithm is fairly accurate in the case of AWS but less so in FACE++.

Our second heuristic technique is a direct weighted sum that can be applied to approximate $\hat{\boldsymbol{x}}$ instead of solving the linear system. We observe that a higher confidence score indicates that the target identity looks more like the corresponding image in the OFS. Therefore, we estimate the template $\hat{\boldsymbol{x}}$ using a weighted sum of each OFS template, using confidence scores as coefficients in a linear combination. In addition, we observe that if we used all templates from OFS, the output linear combination $\hat{\boldsymbol{x}}$ would contain the templates of non-similar face images. Therefore, we restrict the number of templates used in the linear combination to the top l highest confidence scores. A description of the confidence score attack algorithm using our techniques is presented in Algorithm 3.

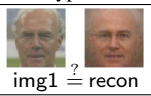
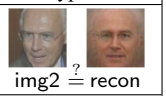
	FRS		Evaluation on Target FRS	
Attack	Source	Target	Type-1	Type-2
Direct	T_A	T_A	 img1 $\stackrel{?}{=}$ recon	 img2 $\stackrel{?}{=}$ recon
Transfer	T_A	T_B	 img1 $\stackrel{?}{=}$ recon	 img2 $\stackrel{?}{=}$ recon

TABLE 4: Classification by attack scenario; recon is the reconstructed face image using an attack against source FRS T_A and target image img1; img2 is a different face image of same identity as img1; T_B is a different FRS from T_A. Note that recon is *not* re-tuned for img2 nor T_B.

5. Experiments

In this section, we provide implementation details and experimental results for our proposed attack on open-source and commercial FRSs and compare our approach against previous methods [11], [12], [13] using their respective open source implementations. The source code for our attack implementation is publicly available.

5.1. Face Datasets and Face Recognition Systems

For target face images, we select three datasets, LFW [20], AgeDB-30 [21], and CFP-FP [22], which are widely used benchmarks for unconstrained FRS face verification. The LFW and AgeDB datasets each contain 3000 positive and 3000 negative image pairs, while CFP-FP contains 3500 positive and 3500 negative pairs. We use positive pair images that are aligned and cropped to 112×112 image from all target datasets. In particular, one image from each true pair is used as the main target for reconstruction (Type-1 attack) and the second image of the same identity pair for a transfer-like attack (Type-2 attack). To avoid confusion of terminology, Table 4 shows our classification of various attack scenarios used in the experiments.

In our attack implementation, a pre-trained ArcFace [17] ResNet-50 model and NbNet [2] are used as a local feature extractor $F_L : \mathcal{I} \to \mathbb{R}^{512}$ and its inverse model $F_L^{-1} : \mathbb{R}^{512} \to \mathcal{I}$, respectively. Note that many FRSs use a default template dimension of 512, which we adopt in this paper. The local pre-trained feature extractor and inverse model are trained on the same dataset MS1MV3 [47]; since these local models can be selected at will, unlike the target FRSs, using the same training dataset for both models does not bias our attack results.

To carry out a comprehensive evaluation of the effectiveness of our attack, we selected five open-source FRSs and three commercial FRSs as targets.

For the five open-source FRSs, we assume a black-box setting where the adversary can obtain cosine similarity scores. Table 5 gives a detailed description of each target open-source FRS and our local models. Our model choices were motivated not just by performance (TAR@FAR) but also by model diversity, e.g., being different from the local

	Open-source FRS and Inverse Model			Training Dataset	TAR@FAR(%)		
	Name	Architecture	Loss	Name	LFW	AgeDB	CFP-FP
Local	F_L	ResNet-50 [41]	ArcFace [17]	MS1MV3 [47]	99.60@0.03	96.93@0.73	97.54@0.49
	F_L^{-1}	NbNet-B [2]	Perceptual [2]	MS1MV3 [47]	N/A	N/A	N/A
Target	T_1	ResNet-100 [41]	CosFace [15]	Glint360K [48]	99.60@0	97.10@0.53	98.63@0.14
	T_2	Vision Transformer [16]	CosFace [15]	MS1MV3 [47]	99.60@0	96.63@0.53	97.89@0.57
	T_3	ResNet-50 [41]	ArcFace [17]	WebFace12M [49]	99.60@0	96.67@0.70	98.66@0.26
	T_4	SFNet-20 [18]	SphereFace2 [26]	VGGFace2 [50]	99.16@0.13	91.63@4.30	95.49@2.69
	T_5	Inception-ResNet [51]	Triplet [19]	CASIA-WebFace [52]	98.10@1.23	89.43@11.10	91.17@6.23

TABLE 5: Description of each open-source FRS and the local (inverse) feature extractors

Dataset	L	T_1	T_2	T_3	T_4	T_5
LFW	$78.2°$	$72.5°$	$77.3°$	$74.3°$	$75.8°$	$63.6°$
AgeDB	$80.8°$	$76.4°$	$80.5°$	$79.1°$	$78.7°$	$70.4°$
CFP-FP	$81.1°$	$77.3°$	$81.1°$	$79.6°$	$81.1°$	$73.7°$

TABLE 6: Thresholds derived using the cross validation scheme from Huang et al. [20] for each open-source FRS

model in terms of architecture, loss function, and/or training dataset. To the best of our knowledge, T_1 and T_2 are among the best in accuracy for ResNet-based and non-ResNet-based architectures, respectively; T_3 is similar to the local model, differing only in the training dataset; T_4 is a lightweight model which differs from the others in terms of architecture, loss, and training dataset; and T_5 is a classical model that has been widely used as a target in previous studies. The baseline performance of T_4 and T_5 are relatively weaker compared to F_L, T_1–T_3, but we include them to test the applicability of our attack on a range of models.

We follow the leave-one-out cross validation scheme from Huang et al. [20] (average of thresholds with highest accuracy across 10 training-validation splits) to set default acceptance thresholds for each dataset and target FRS before applying any of our attacks. The selected thresholds are shown in Table 6, and the TAR@FAR value for each threshold is shown in the rightmost columns of Table 5.

The three target commercial FRSs do not provide any information about model architecture, loss function, training datasets, and function used to compute the confidence score, i.e., our attacks on these FRSs are in a real-world black-box setting. Nevertheless, we show the performance of each commercial FRS TAR@FAR across our three datasets and several score acceptance thresholds in Table 7. All three commercial FRSs provide a default threshold setting, which is directly applicable to many verification applications. However, users can change the threshold to suit their application. The FACE++ FRS provides additional guidelines to achieve desired FAR 0.1%, 0.01%, or 0.001% using the acceptance thresholds 0.7039, 0.7656, or 0.8088, respectively; we use all three recommended thresholds in our experiments. However, the other commercial FRSs, AWS CompareFaces and KAIROS, only provide default threshold recommendations. For our evaluation, we follow their default thresholds and we arbitrarily choose two additional thresholds set slightly lower or slightly higher than the original thresholds.

Name	Thresh.	LFW	AgeDB	CFP-FP
T_{AWS}	0.7	99.33@0	87.97@0	93.17@0
	$0.8^{\dagger}$	99.17@0	83.90@0	90.21@0
	0.9	98.63@0	75.23@0	84.33@0
$T_{\mathsf{FACE++}}$	0.7039^{*}	99.10@0.07	85.96@1.87	39.39@0.46
	$0.7656^{\dagger}$	97.22@0	64.23@0.30	21.07@0
	0.8088^{*}	92.63@0	39.60@0.07	9.95@0
T_{KAI}	0.5	95.92@0.14	54.06@1.09	74.16@0.04
	$0.6^{\dagger}$	84.58@0.07	21.77@0.07	39.53@0
	0.7	58.76@0	4.28@0	10.13@0

TABLE 7: TAR@FAR performance (%) of commercial FRSs. "†" indicates the default recommended threshold in each FRS. "*" indicates another threshold setting provided by the corresponding FRS.

	LFW		AgeDB		CFP-FP	
Target	1	2	1	2	1	2
T_1	96.17	47.87	99.60	46.37	99.89	41.03
T_2	95.36	40.46	99.73	43.86	99.62	29.71
T_3	95.83	38.70	96.03	33.80	95.54	29.03
T_4	88.07	43.03	88.63	26.17	89.14	23.31
T_5	80.57	55.83	94.47	61.80	97.54	77.71

TABLE 8: Direct ASR (%) on the open-source FRSs

5.2. Direct Attack Result

We start by evaluating the performance of our direct attack on the target FRSs. We generate an OFS $\mathcal{O}$ using the local feature extractor F_L and its inverse model F_L^{-1} based on the MS1MV3 dataset as an image set $\mathcal{Z}$ and set $\delta = \cos 85° \approx 0.0872$ for the GenOFS algorithm. We set the size of $\mathcal{O}$ to 99 images, so that the total number of online queries made to the target FRS by our attack is always $m = 100$, including one last query for the reconstructed image. The same OFS is used throughout all subsequent experiments unless otherwise stated. For a visual overview of this section, we refer readers to Figure 6 which plots score histograms for our direct attack against all targeted FRSs on the LFW dataset (quantitative discussion below).

5.2.1. Cosine Similarity Score Setting. We apply Algorithm 1 with $\mathcal{O}$ against the open-source FRSs and record the attack success rates (ASR) for both Type-1 and Type-2 direct attacks in Table 8. The relative success of our attacks can be seen in Figure 6 (for LFW).

We defer discussion of T_5; for T_1–T_4, our direct ASR generally correlates well with the baseline performance of each model, e.g., T_4 being a lightweight model has the lowest Type-1 ASR, while the Type-1 ASR of T_1–T_3 is at least

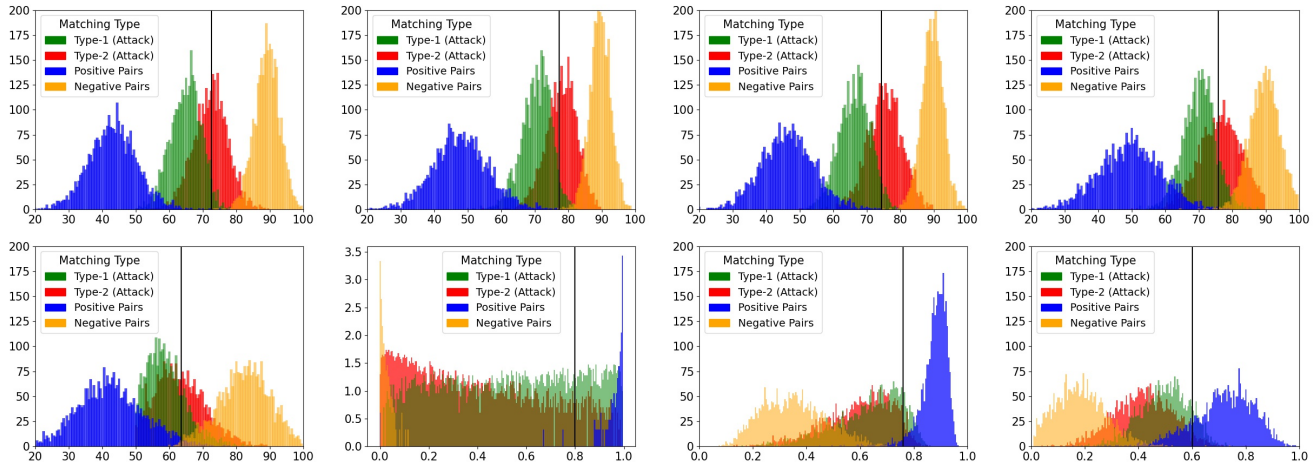


Figure 6: Score histograms for the reconstructed face images in Type-1 and Type-2 attack for each target FRS on the LFW dataset, from left to right: (top row) T_1, T_2, T_3, T_4, (bottom row) T_5, T_{AWS}, $T_{\mathsf{FACE++}}$, T_{KAI}; scores for the positive/negative pairs of the dataset are also plotted for comparison. The horizontal axis is the angle between images for T_1–T_5 and the respective confidence scores for the commercial FRSs; the vertical axis is the number of images in each bin, where T_{AWS} is shown in logarithmic scale because its outputs are bimodal (matching a sigmoid output, see Figure 5); the default acceptance thresholds are shown as black vertical lines. For the open-source FRSs, an image is accepted if its angle is below the threshold, whereas for the commercial FRSs, an image is accepted when its confidence score is above the threshold.

95% across all target datasets. As explained in Section 3, a key intuition used throughout the design of our attack is that well-trained local and target models should be almost isometric for face images (Conjecture 1), so that we can use reconstructions from the local (inverse) model to attack the target model, and vice-versa for similarity scores. We can apply this intuition in the converse direction, i.e., if the local and target similarity scores are *less* correlated (see T_4 in Figure 2), then our local approximation in Equation 3 may incur a worse error term, ultimately leading to a poorer face reconstruction. Thus, perhaps counter-intuitively, a weaker FRS like T_4 may be *harder* to attack using our approach, as its output scores are less well-correlated with those of our strong local model F_L.

In the more difficult Type-2 setting, Table 8 shows that our attack has some success, although with lower rates than the Type-1 setting. This indicates that our attack can reconstruct face images that not only resemble a single target image, but also contains personally identifiable biometric features of the target identity that transfer across images of the same identity. We further investigate the transferability of our attack in the next section. We also observe that stronger target FRSs, such as T_1, are more susceptible to Type-2 attacks compared to T_2, T_3, T_4. This indicates that FRSs that capture more discriminative features of each identity into a template may more vulnerable to our attack.

Turning to the results for T_5, we see it is an outlier in ASR. One reason is our standardized threshold selection method (Table 6)—the thresholds for T_5 are less stringent compared to the others because it has poor inter-class separation, giving it a high baseline FAR and making it an easier target for attacks (see Figure 6). Nevertheless, the results

		LFW		AgeDB		CFP-FP	
Target	Thresh.	1	2	1	2	1	2
T_{AWS}	0.7	38.2	10.5	39.9	4.1	29.2	2.6
	0.8	27.0	6.4	29.6	2.2	20.6	1.2
	0.9	13.9	2.4	17.7	0.7	9.5	0.4
$T_{\mathsf{FACE++}}$	0.7039	50.7	22.9	62.2	12.9	41.1	10.2
	0.7656	14.5	3.5	19.9	1.2	8.9	0.8
	0.8088	2.1	0.3	3.0	0.1	1.3	0
T_{KAI}	0.5	49.5	22.6	55.5	8.4	36.5	5.0
	0.6	11.9	3.5	12.2	0.4	6.1	0.3
	0.7	0.4	0.2	0.3	0	0.1	0

TABLE 9: Direct ASR (%) on the commercial FRSs

show that our attack works against T_5, even though its training loss function (triplet loss) differs significantly from the local model (angular loss). Future work could design experimental protocols to investigate this gap more closely.

5.2.2. Confidence Score Setting. In Table 9, we present our direct attack result against the three commercial FRSs, T_{AWS}, $T_{\mathsf{FACE++}}$, and T_{KAI}, using Algorithm 3. We use an inverse sigmoid function f^{-1} for T_{AWS} (as computed and shown in Figure 5) and parameter $l = 15$ as auxiliary attack information and for the other two FRSs, respectively.

Our attack is able to achieve non-negligible Type-1 and Type-2 ASR against the targets, especially for the lower and default threshold settings. This is significant because it indicates that our attack (using only 100 queries) is viable against commercial systems. For example, on the LFW dataset, our attack against T_{AWS} with default threshold achieves ASR 27.0% and 6.4% in Type-1 and 2 attacks, respectively. Since there is no false acceptance with default threshold setting in T_{AWS}, this result shows that there is room

Source	T_1		T_2		T_3		T_4		T_5	
Target \ Type	1	2	1	2	1	2	1	2	1	2
T_1	N/A		61.70	18.60	89.13	36.70	3.90	2.07	0.83	0.60
T_2	84.47	48.93	N/A		82.47	40.23	8.63	3.27	1.23	1.17
T_3	83.10	37.70	58.67	17.13	N/A		5.37	1.90	0.83	0.77
T_4	69.70	52.27	51.33	29.57	67.03	42.50	N/A		35.90	25.40
T_5	55.57	46.43	41.67	31.97	54.67	40.73	60.73	44.27	N/A	

TABLE 10: Transfer ASR (%) on the open-source FRSs with the LFW dataset

for leakage of face image biometric information in score-based attacks. Similarly, for $T_{\mathsf{FACE++}}$, when the threshold is set to the default recommended value, it is claimed that the FAR is about 0.01%, but our Type-1 ASR is 1000 times higher than this FAR; it is also about 2000 times higher at the threshold 0.8088, where the FAR is stated to be 0.001%. In the case of the T_{KAI}, the ASR is lower than the other two systems, but given the baseline performance shown in Table 7, our result is not negligible.

The ASR for these commercial FRSs is lower than those of open-source FRSs. This is expected because we can only approximately recover similarity scores from confidence scores in our approach. We also suspect, but cannot confirm, that the commercial systems may deploy certain template- or score-based defenses. Lastly, we observe that the default thresholds for commercial FRSs are set much more stringently than those for our open-source FRSs, at the cost of rejecting some true samples; this is visualized in Figure 6 for LFW. However, for usability, many commercial FRSs allow users to set their own acceptance thresholds in their APIs. Our attack indicates that an absence of threshold lower limits could pose dangers to user privacy, and that further standardization or guidelines may be necessary.

5.3. Transfer Attack

In this subsection, we experiment with several transfer-like attacks (Table 4 bottom row) on the target FRSs to investigate the transferability of our attack's reconstructed face images on the LFW dataset.

5.3.1. Cosine Similarity Score Setting. We test whether reconstructed images from our attack method against each open-source FRS (T_1–T_5) can be identified as the same identity at different target FRSs for both Type-1 and Type-2 attacks, results are shown in Table 10.

Attacking a strong source FRS (such as T_1) leads to reconstructed faces that transfer well against other targets, e.g., using generated images from source T_1 provides the best transfer Type-2 ASR against all targets, the most challenging attack setting in this paper. Conversely, images from weaker source models T_4 and T_5 have much lower transfer ASR (against T_1–T_3), which is to be expected since our attack may not adequately recover information from such models (see Section 5.2.1).

5.3.2. Confidence Score Setting. We follow the same experimental configurations as in Section 5.2.2, except that the source and target FRSs are different; results are in Table 11.

Source	T_{AWS}		$T_{\mathsf{FACE++}}$		T_{KAI}	
Tar. \ Type	1	2	1	2	1	2
T_1	32.00	11.53	2.57	1.43	1.23	1.53
T_2	39.37	21.23	8.10	4.13	4.90	3.50
T_3	27.93	11.23	3.07	1.47	1.77	1.17
T_4	55.10	42.97	43.00	30.70	50.73	37.30
T_5	50.57	41.90	58.37	47.90	70.53	58.30
T_{AWS}	N/A		0.03	0	0.37	0.30
$T_{\mathsf{FACE++}}$	13.57	5.25	N/A		2.59	1.33

TABLE 11: Transfer ASR (%) on the commercial FRSs with the LFW dataset

Our approach's ASR is much higher when using T_{AWS} as the source FRS for all transfer attack settings (except T_5). This difference may be partly explained by the different conversion techniques used to target the commercial FRSs. The first technique using an inverse sigmoid for T_{AWS} may enable our attack to better recover personally identifiable face features from confidence scores, as opposed to the second technique for $T_{\mathsf{FACE++}}, T_{\mathsf{KAI}}$.

5.4. Comparison Against Prior Approaches

Our attack significantly improves over previous approaches [11], [12], [13] in terms of the number of queries for score-based impersonation and face reconstruction. To the best of our knowledge, there is no prior attack method that can succeed within 100 queries. In this section, we provide a qualitative comparison against prior approaches using the LFW dataset, including performance on transfer-like attacks and on other similarity metrics.

We apply our approach and approaches from previous studies [11], [12], [13] to attack FRS T_1 on the LFW dataset. For our *non-adaptive* attack, we test increasing sizes for OFS $\mathcal{O}$, up to $m = 100$ queries. For all prior attacks, due to their *adaptive* nature, we run them until they generate an image that passes the acceptance threshold.

In Figure 7, for the sake of effectively showing the difference between each method, we use a logarithmic scale for the number of queries. Our attack achieves a similar ASR with two orders of magnitude fewer queries for a Type-1 direct attack. We also observe that at a fixed Type-1 ASR (96.17%) our attack yields almost 7x higher ASR (47.87%) for Type-2 direct attack than previous methods ($<7.00\%$).

We further investigate the transferability of our attack in Table 12; the attack based on Gaussian blobs [11] is excluded because the shape of images generated are not human-like (see Table 1 for examples) and are not detected

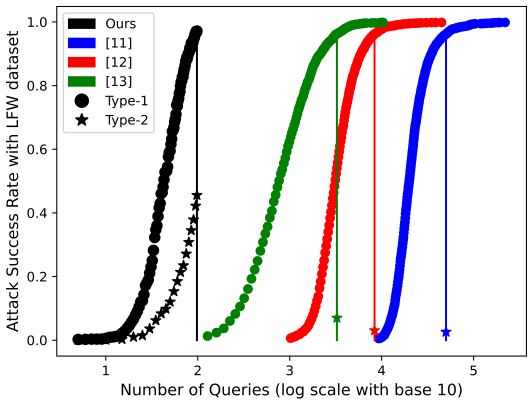


Figure 7: Number of queries for Type-1 direct attack success against T_1 with the LFW dataset (horizontal axis $\log_{10}$ scale). The vertical lines mark the number of queries at which methods achieve the same ASR (96.2%) as our attack.

(Method,Source)	([12],T_1)		([13],T_1)		(Ours,T_1)	
Target \ Type	1	2	1	2	1	2
T_2	4.67	1.37	5.53	2.67	84.47	48.93
T_3	5.23	0.97	2.90	0.93	83.10	37.70
T_4	12.47	6.77	14.10	10.50	69.70	52.27
T_5	5.93	4.23	16.90	13.60	55.57	46.43
T_{AWS}	0.13	0	0.27	0.17	27.47	10.10
$T_{\mathsf{FACE++}}$	0.30	0.03	3.94	1.87	34.89	14.05

TABLE 12: Transfer ASR (%) for various attack methods

by the face detection pre-processing steps of commercial FRSs (we were unable to obtain a score from queries).

For all target FRSs, our approach achieves significantly higher ASR for both Type-1 and Type-2 transfer attacks compared to previous approaches. Notably, for commercial FRS targets, the transfer ASR of prior approaches is zero or close to zero, unlike our approach.

In Table 13, we measure the similarity between the target image and reconstructed image for each method, using five commonly used metrics: $\text{LPIPS}_{\text{Alex}}$/$\text{LPIPS}_{\text{VGG}}$ [23], DISTS [24], Structural Similarity Index Measure (SSIM) [53], and L2-norm. As a simple comparison, the last column shows the similarity score of the other image of the same identity (true pair) in the LFW dataset. Our attack achieves better scores (more similar) on all metrics. Interestingly, we observe that using our attack on T_{AWS} achieves slightly better scores than against T_1, except $\text{LPIPS}_{\text{VGG}}$. A detailed description of each metric is relegated to Appendix A.

	Other Methods			Ours		LFW Pair
	[11]	[12]	[13]	T_1	T_{AWS}	
$\text{LPIPS}_{\text{Alex}}$	0.531	0.338	0.394	0.228	**0.228**	0.200
$\text{LPIPS}_{\text{VGG}}$	0.475	0.391	0.472	**0.322**	0.323	0.302
DISTS	0.394	0.285	0.316	0.243	**0.240**	0.205
SSIM	0.589	0.721	0.523	0.781	**0.787**	0.771
L2-norm	72.38	39.57	48.44	30.79	**29.32**	23.94

TABLE 13: Comparison of image similarities average on LFW dataset (lower is better, except SSIM, best in **bold**)

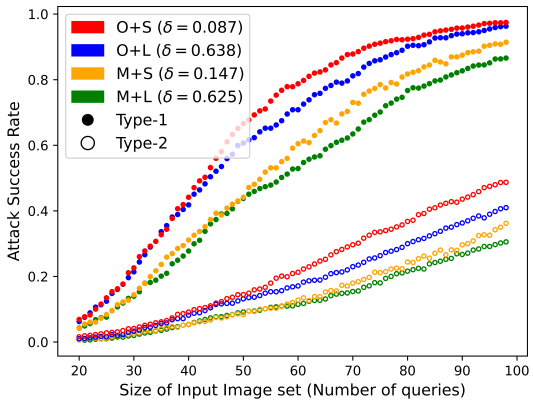


Figure 8: Comparison of ASR using various face image sets

5.5. Ablation Study (Orthogonal Face Set)

The main tuning knob in our attack is the choice of OFS. We investigate the ASR of our attack using various alternative face image sets as input of Algorithm 1.

5.5.1. Tweaking OFS Generation. As explained in Section 4.2, our method in Algorithm 2 tries to generate an OFS from images that have identity-irrelevant information removed (label "O"). An alternative is to use raw images from the face dataset MS1MV3 (label "M"). Further, we select an OFS with small δ value as explained in Definition 2 (label "S"); it is also possible to randomly select an OFS, resulting in larger δ values (label "L"). Finally, we can also modify the size of the OFS.

In Figure 8, we plot the direct ASR against T_1 on the LFW dataset using combinations of the above changes to OFS generation (up to size 100). First, we observe that "O" leads to stronger ASR compared to "M"; removing identity-irrelevant information is clearly helpful for our attack. Second, we note that "S" performs better than "L" when the OFS size is large, but with a crossover point at smaller OFS sizes; it may be worth swapping between "S" and "L" depending on the desired query size. Finally, we observe (as expected) that a larger OFS leads to higher ASR; nevertheless, the figure also shows that our attack remains somewhat feasible with fewer queries, e.g., $>50\%$ ASR can be achieved using only 50 queries.

5.5.2. Large Random OFS. It is difficult to find a large OFS from the MS1MV3 dataset satisfying the almost orthogonality requirements for small δ values. To test OFS sizes beyond 100, we use randomly selected face image sets as the OFS. As explained in Section 3.4, there are four important points for our attack to be successful, briefly:

1) $\boldsymbol{e}$: the error $\boldsymbol{e}$ from the Conjecture 1 should be small;
2) $\widehat{\boldsymbol{A}}$: the condition number of $\widehat{\boldsymbol{A}}$ should be small, hence the use of almost-orthogonal templates in GenOFS;
3) ϵ: the ϵ of almost-isometry is empirically small for highly accurate FRS such as T_1, T_2, and T_3;

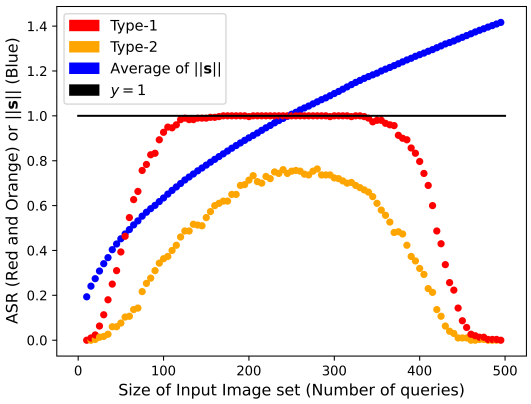


Figure 9: Comparison of ASR and average norm of score vector $||\boldsymbol{s}||$ using various sizes of input face image sets

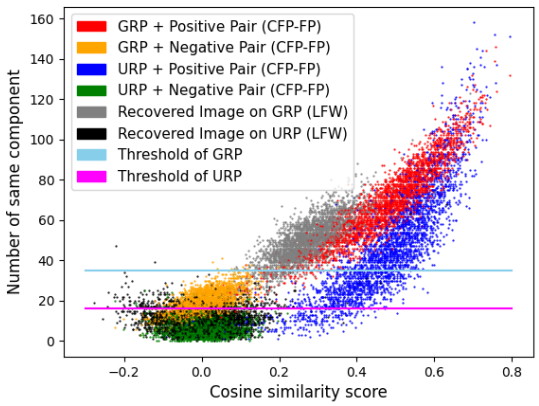


Figure 10: Distribution of score pair on GRP and URP-IoM

4) $||\boldsymbol{s}||$: if $\widehat{\boldsymbol{A}}$ is an orthogonal matrix, $\boldsymbol{z} = \widehat{\boldsymbol{A}}^{+}\boldsymbol{s}$ is a least squares solution whose norm is equal to $\boldsymbol{s}$.

We showed that 1) and 3) are experimentally satisfied and we expect 2) to be satisfied provided we use the GenOFS algorithm. For point 4), we expect template vectors $\boldsymbol{z}$ to have norm 1 (for most FRSs). Thus, we hypothesize that it would be ideal if $||\boldsymbol{s}||$ is also close to 1. In Figure 9, we plot the average norm of $\boldsymbol{s}$ and corresponding ASR when directly attacking T_1 on the LFW dataset, and using randomly chosen input images as the OFS. Interestingly, we observe a peak in ASR almost precisely where the returned similarity score vectors have average norm close to 1.

A direction for future work is to generate a large OFS with small δ artificially, e.g., by traversing the input space ($\mathbb{R}^{512}$) of the local inverse model F_L^{-1} with gradient descent.

6. Discussion

We discuss various alternative attack scenarios and the implications of our results. Detailed descriptions are deferred to Appendix B.

6.1. Defense Methods

6.1.1. Hiding score outputs. The key takeaway from our attack is that a small number of similarity score queries ($m = 100$) can leak significant information in modern FRSs. This leakage can be exploited to an extent *even if* raw similarity information is not directly available but can be estimated, e.g., from the confidence scores output by commercial FRSs (Section 4.3). Consequently, as a first defense, developers building on FRS APIs (commercial or open-source) should take care never to reveal raw score values to untrusted end-users. Further upstream, commercial FRS services could provide more restrictive APIs that only return binary values (accept or deny) based on pre-defined, non-user-configurable thresholds. Restricting threshold access is important, for example, if the score is between 0 to 100, binary search can pinpoint the score using 6 to 7 queries.

6.1.2. Template protection. Many template protection methods have been proposed [6], [7], [54], [55] based on Homomorphic Encryption (HE) and Locality Sensitive Hashing (LSH). These approaches aim to protect template security and privacy while allowing efficient computation of the desired scores. We focus our discussion on recently proposed methods [6], [7] as representative examples. Engelsma et al. [6] use end-to-end HE to ensure that templates are never stored nor revealed as plaintext to adversaries. However, our attack bypasses such protections as long as we can get similarity score outputs from the target FRS.

Jin et al. [7] propose two LSH constructions, GRP-IoM and URP-IoM, to transform the template $\boldsymbol{x} \in \mathbb{R}^n$ into a protected template $\boldsymbol{x}' \in [a]^b$ using public randomness R for some pre-defined parameters a and b. Since the LSH consists of non-linear functions such as $\max$, the authors argue that the adversary cannot obtain the target template $\boldsymbol{x}$ from $\boldsymbol{x}'$, even if the latter is compromised and the adversary knows R, a, b. However, we observe that the LSH approach is also vulnerable to score-based attacks. Similar to our first technique in Section 4.3, given R, a, b, we can use an existing dataset (CFP-FP) and local model F_L to approximate an inverse function for converting component-wise LSH matching score into similarity scores. In Figure 10, we plot the score pairs for the CFP-FP dataset and score pairs of reconstructed images using our adapted attack on the LFW dataset. For GRP-IoM (resp. URP-IoM), our approach achieves ASR 96.33% and 48.50% (resp. 16.47% and 7.20%) in Type-1 and Type-2 attacks, respectively.

6.1.3. Exact matching. There is another class of cryptographic template protection methods [56], [57], [58] based on Fuzzy Commitment [59] or Fuzzy Vault [60]. These methods use *exact matching* on templates without calculating scores, so they are naturally secure against score-based attacks. At a high level, these methods employ error-correcting codes on input user image templates together with a matching process based on cryptographically secure hashing to check the corrected templates against template databases. However, state-of-the-art exact matching meth-

ods suffer from performance degradation when applied to FRSs, e.g., in Dong et al. [56] and Rathgeb et al. [57], template values are binarized in order to apply classical error-correcting codes, leading to performance degradation due to quantization. Kim et al. [58] propose a new error-correcting code that applies directly to real-valued templates, but also with performance degradation due to low correction capacity. Nevertheless, improving exact matching methods may be a promising direction of future work to protect against the vulnerabilities identified in this paper.

6.2. Identification Scenario

In the open-set *identification* scenario, multiple identities are enrolled and the FRS returns top-k scores for a queried image among its enrolled identities. A queried image is identified as the identity with the highest score (above a threshold σ). Our attack can easily be adapted to this setting (Appendix B). For $k = 5$ and target FRS T_1 with $\sigma = \cos 72.5^\circ \approx 0.3$, our attack achieves ASR 89.06% on the IJB-C dataset [61], which supports 1-to-N identification.

6.3. Other Biometric Authentication

There are many DL-based biometric recognition systems such as voice [27], [28], fingerprint [29], and iris [30] trained by metric learning on cosine similarities to achieve strong intra-class compactness and inter-class separability for user identities. Since our attack primarily relies on information leaked through *similarity scores*, we conjecture that our approach can be readily adapted to attack various other kinds of biometric recognition systems beyond facial images.

6.4. Choice of Inverse Model

In our experiments, we fixed NbNet [2] as a local inverse model F_L^{-1}. Notably, we did not fully exploit the fact that our local models can be arbitrarily chosen and trained in white-box settings, e.g., NbNet was trained in a black-box setting [2]. This raises several possible directions for future work in strengthening our attack, e.g.:

- Choose a strong generative local inverse model, e.g., based on StyleGan, to reconstruct high-resolution and realistic face images from (local) templates.
- Use multiple local inverse models; currently, F_L^{-1} is used twice when generating OFS or reconstructing images from template $\boldsymbol{z}$ using F_L^{-1}. However, the purpose is different: for the OFS, we expect output images from F_L^{-1} to contain characteristic information of each identity without irrelevant information; on the other hand, when reconstructing images, we want F_L^{-1} to produce a good pre-image from a template. There is room to fine-tune inverse models for both purposes.

7. Conclusion

This work presents a novel, non-adaptive score-based attack on FRSs, requiring up to two orders of magnitude fewer queries than prior approaches. The attack is feasible to apply against a range of open-source and commercial FRS APIs (AWS CompareFaces, Face++, KAIROS), achieving significant attack success rates against the targeted systems. Our approach also works well in transfer-like attacks, i.e., compromising an identity for one system could lead to the compromise of the same identity for another (unattacked) system. Thus, our attack reveals potential security and privacy issues from score leakages in state-of-the-art FRSs and highlights the need for further study and regulation of the outputs of biometric authentication systems in order for them to be securely deployed.

Acknowledgment

We thank the anonymous shepherd and reviewers for their helpful feedback on this paper. This work was supported in part by the Institute of Information and Communication Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (A Study on Cryptographic Primitives for SNARK, 50%) under Grant 2021000727, and in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT), 50%, under Grant 2020R1C1C1A01006968. This work is also supported by A*STAR under its RIE2020 Advanced Manufacturing and Engineering (AME) Programmatic Programme (Award A19E3b0099) and by scholarships from A*STAR Graduate Academy.

References

[1] "Information security, cybersecurity and privacy protection – biometric information protection," International Organization for Standardization, Standard ISO/IEC 24745:2022, 2022. [Online]. Available: https://www.iso.org/standard/75302.html

[2] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1188–1202, 2019.

[3] C. N. Duong, T. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy, "Vec2Face: Unveil human faces from their blackbox features in face recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 6131–6140.

[4] T. Truong, C. N. Duong, N. Le, M. Savvides, and K. Luu, "Vec2face-v2: Unveil human faces from their blackbox features via attention-based network in face recognition," *CoRR*, vol. abs/2209.04920, 2022.

[5] X. Dong, Z. Jin, Z. Guo, and A. B. J. Teoh, "Towards generating high definition face images from deep templates," in *BIOSIG*, ser. LNI, A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. B. Raja, C. Rathgeb, A. F. Sequeira, and A. Uhl, Eds., vol. P-315. Gesellschaft für Informatik e.V., 2021, pp. 71–80.

[6] J. J. Engelsma, A. K. Jain, and V. N. Boddeti, "HERS: homomorphically encrypted representation search," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 4, no. 3, pp. 349–360, 2022.

[7] Z. Jin, J. Y. Hwang, Y. Lai, S. Kim, and A. B. J. Teoh, "Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 2, pp. 393–407, 2018.

[8] Amazon. CompareFaces API. [Online]. Available: https://docs.aws.amazon.com/rekognition/

[9] Megvii. Compare API. [Online]. Available: https://console.faceplusplus.com/documents/5679308

[10] Kairos. Compare API. [Online]. Available: https://www.kairos.com/docs/api

[11] A. Razzhigaev, K. Kireev, E. Kaziakhmedov, N. Tursynbek, and A. Petiushko, "Black-box face recovery from identity features," in *ECCV*, ser. LNCS, A. Bartoli and A. Fusiello, Eds., vol. 12539. Springer, 2020, pp. 462–475.

[12] A. Razzhigaev, K. Kireev, I. Udovichenko, and A. Petiushko, "Darker than black-box: Face reconstruction from similarity queries," *CoRR*, vol. abs/2106.14290, 2021.

[13] E. Vendrow and J. Vendrow, "Realistic face reconstruction from deep embeddings," in *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[14] X. Dong, Z. Miao, L. Ma, J. Shen, Z. Jin, Z. Guo, and A. B. J. Teoh, "Reconstruct face from features based on genetic algorithm using GAN generator as a distribution constraint," *Comput. Secur.*, vol. 125, p. 103026, 2023.

[15] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5265–5274.

[16] Y. Zhong and W. Deng, "Face transformer for recognition," *CoRR*, vol. abs/2103.14803, 2021.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[18] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller, "SphereFace revived: Unifying hyperspherical face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2458–2474, 2023.

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

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

[21] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *CVPR Workshops*. IEEE Computer Society, 2017, pp. 1997–2005.

[22] S. Sengupta, J. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *WACV*. IEEE Computer Society, 2016, pp. 1–9.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 586–595.

[24] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, 2022.

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *CVPR*. IEEE Computer Society, 2017, pp. 6738–6746.

[26] Y. Wen, W. Liu, A. Weller, B. Raj, and R. Singh, "SphereFace2: Binary classification is all you need for deep face recognition," 2022.

[27] Z. Li, M. Mak, and H. M. Meng, "Discriminative speaker representation via contrastive learning with class-aware attention in angular space," in *ICASSP*. IEEE, 2023, pp. 1–5.

[28] B. Han, Z. Chen, and Y. Qian, "Exploring binary classification loss for speaker verification," in *ICASSP*. IEEE, 2023, pp. 1–5.

[29] Y. Zhang, R. Zhao, Z. Zhao, N. Ramakrishnan, M. Aggarwal, G. Medioni, and Q. Ji, "Robust partial fingerprint recognition," in *CVPR Workshops*. IEEE, 2023, pp. 1011–1020.

[30] J. Wei, Y. Wang, H. Huang, R. He, Z. Sun, and X. Gao, "Contextual measures for iris recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 57–70, 2023.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, ser. LNCS, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711.

[32] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[33] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *CCS*, E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, Eds. ACM, 2016, pp. 1528–1540.

[34] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *CCS*, I. Ray, N. Li, and C. Kruegel, Eds. ACM, 2015, pp. 1322–1333.

[35] Z. Yang, J. Zhang, E. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *CCS*, L. Cavallaro, J. Kinder, X. Wang, and J. Katz, Eds. ACM, 2019, pp. 225–240.

[36] G. Tao, Q. Xu, Y. Liu, G. Shen, S. An, J. Xu, X. Zhang, and Y. Yao, "MIRROR: model inversion for deep learning network with high fidelity," in *NDSS*. The Internet Society, 2022.

[37] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *AsiaCCS*, R. Karri, O. Sinanoglu, A. Sadeghi, and X. Yi, Eds. ACM, 2017, pp. 506–519.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[39] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2018.

[40] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," pp. 14 745–14 758, 2021.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE Computer Society, 2016, pp. 770–778.

[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[43] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 8107–8116.

[44] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[45] M. Knoche, T. Teepe, S. Hörmann, and G. Rigoll, "Explainable model-agnostic similarity and confidence in face verification," in *WACV*. IEEE, 2023, pp. 1–8.

[46] C. Voglis and I. Lagaris, "A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization," in *WSEAS International Conference on Applied Mathematics*, vol. 7, 2004, pp. 9 780 429 081 385–138.

[47] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *ICCV Workshops*. IEEE, 2019, pp. 2638–2646.

[48] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: training 10 million identities on a single machine," in *ICCVW*. IEEE, 2021, pp. 1445–1449.

[49] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, D. Du, J. Lu, and J. Zhou, "WebFace260M: A benchmark for million-scale deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2627–2644, 2023.

[50] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *FG*. IEEE Computer Society, 2018, pp. 67–74.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *AAAI*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 4278–4284.

[52] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014.

[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[54] X. Dong, K. Wong, Z. Jin, and J. Dugelay, "A cancellable face template scheme based on nonlinear multi-dimension spectral hashing," in *IWBF*. IEEE, 2019, pp. 1–6.

[55] A. K. Jindal, I. Shaik, Vasudha, S. R. Chalamala, R. M. A, and S. Lodha, "Secure and privacy preserving method for biometric template protection using fully homomorphic encryption," in *TrustCom*, G. Wang, R. K. L. Ko, M. Z. A. Bhuiyan, and Y. Pan, Eds. IEEE, 2020, pp. 1127–1134.

[56] X. Dong, S. Kim, Z. Jin, J. Y. Hwang, S. Cho, and A. B. J. Teoh, "Secure chaff-less fuzzy vault for face identification systems," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 17, no. 3, pp. 79:1–79:22, 2021.

[57] C. Rathgeb, J. Merkle, J. Scholz, B. Tams, and V. Nesterowicz, "Deep face fuzzy vault: Implementation and performance," *Comput. Secur.*, vol. 113, p. 102539, 2022.

[58] S. Kim, Y. Jeong, J. Kim, J. Kim, H. T. Lee, and J. H. Seo, "IronMask: Modular architecture for protecting deep face template," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 16 125–16 134.

[59] A. Juels and M. Wattenberg, "A fuzzy commitment scheme," in *CCS*, J. Motiwalla and G. Tsudik, Eds. ACM, 1999, pp. 28–36.

[60] A. Juels and M. Sudan, "A fuzzy vault scheme," *Des. Codes Cryptogr.*, vol. 38, no. 2, pp. 237–257, 2006.

[61] B. Maze, J. C. Adams, J. A. Duncan, N. D. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus benchmark-C: face dataset and protocol," in *ICB*. IEEE, 2018, pp. 158–165.

[62] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <1MB model size," *CoRR*, vol. abs/1602.07360, 2016.

[63] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *CoRR*, vol. abs/1404.5997, 2014.

Appendix A. Image Similarity Metrics

The following metrics are used to experimentally measure image similarities in Table 13.

Learned Perceptual Image Patch Similarity (LPIPS) LPIPS is an image similarity metric based on exploiting feature representations of deep learning models [23]. It outperforms traditional algorithms over a wide range of distortions. Similar to Perceptual Loss [31], LPIPS computes distances between feature representations at multiple layers of a model, based on the assumption that different layers represent different structures in an image. Zhang et al. [23] used SqueezeNet [62], AlexNet [63], and VGG [38] as deep learning models. This paper uses two LPIPS ($\text{LPIPS}_{\text{Alex}}$ and $\text{LPIPS}_{\text{VGG}}$) values computed using AlexNet and VGG[3], respectively.

Deep Image Structure and Texture Similarity (DISTS) DISTS [24] is an image similarity metric based on deep learning models following the LPIPS design. In DISTS, representations are extracted from VGG [38] and combined using structure and texture similarity measurements.

Structural Similarity Index Measure (SSIM) SSIM [53] is an image similarity metric designed to evaluate human visual quality differences rather than numerical errors. The core idea of SSIM is that the degree of distortion of structural information has a great effect on perception because human vision is specialized in deriving structural information from images. Therefore, similarity is evaluated in terms of luminance, contrast, and structural properties.

L2-norm This is equivalent to the mean squared error between the target and reconstructed images.

Appendix B. Alternative Attack Scenario Details

B.1. Protected Templates

We give concrete descriptions of the target LSHs: GRP-IoM and URP-IoM from [7] used in Figure 10.

B.1.1. GRP-IoM. For some three positive integers, n, a, and b, GRP-IoM consists of two algorithms, Setup and Hashing. Setup is a probabilistic algorithm that outputs a set of matrices $R = \left\{\boldsymbol{R}^{(1)}, \ldots, \boldsymbol{R}^{(b)}\right\}$ such that $\boldsymbol{R}^{(i)} \in \mathbb{R}^{a \times n}$ for $i \in [b]$. In addition, row vectors $\left\{\boldsymbol{r}_1^{(i)}, \boldsymbol{r}_2^{(i)}, \cdots, \boldsymbol{r}_a^{(i)}\right\}$ of $\boldsymbol{R}^{(i)}$ are independently sampled from $\mathcal{N}(\vec{0}, I_n)$. Hashing is a deterministic algorithm that takes a real-valued vector $\boldsymbol{x} \in \mathbb{R}^n$ and R. This algorithm outputs an integer vector $\boldsymbol{x'} = (x'_1, \cdots, x'_b) \in [a]^b$, where $x'_i = \text{argmax}_{k \in [a]} \boldsymbol{r}_k^{(i)} \cdot \boldsymbol{x}$ and $\cdot$ is element-wise multiplication. The formal description is provided in Algorithm 4.

B.1.2. URP-IoM. For some positive integers, n, a, b, and c, URP-IoM also consists of two algorithms, Setup and Hashing. Setup is a probabilistic algorithm that outputs a set of uniformly sampled permutations $\Sigma = \left\{\sigma^{(i,j)} : i \in [c], j \in [a]\right\}$ from $[n]$ to $[n]$. Hashing is a deterministic algorithm that takes a real-valued vector $\boldsymbol{x} \in \mathbb{R}^n$ and Σ. This algorithm outputs an integer vector $\boldsymbol{x'} = (x'_1, \cdots, x'_n) \in [b]^n$, where $x'_i = \text{argmax}_{k \in [b]} \left(\prod_{j=1}^{a} \sigma^{(i,j)}(\boldsymbol{x})\right)_k$. The formal description is provided in Algorithm 5.

[3] https://github.com/richzhang/PerceptualSimilarity

Algorithm 4 GRP-IoM

Require: $\boldsymbol{x} \in \mathbb{R}^n, a \in \mathbb{N}$, and $b \in \mathbb{N}$
1: **for** $i \in [b]$ **do**
2: Set $\boldsymbol{r}_j^{(i)} \leftarrow \mathcal{N}(\vec{0}, I_n)$ for $j \in [a]$
3: Set $x'_i \leftarrow \text{argmax}_{j \in [a]} \boldsymbol{r}_j^{(i)} \cdot \boldsymbol{x}$ for $j \in [a]$
4: Set $\boldsymbol{R}^{(i)} \leftarrow \left[\mathbf{r}_1^{(i)} \| \cdots \| \mathbf{r}_a^{(i)}\right]^T$
5: **end for**
6: Set $R \leftarrow \left\{\mathbf{R}^{(1)}, \mathbf{R}^{(2)}, \cdots, \mathbf{R}^{(b)}\right\}$
7: Set $\boldsymbol{x}' \leftarrow (x'_1, \cdots, x'_b)$
8: **return** $(\boldsymbol{x}', R)$

Algorithm 5 URP-IoM

Require: $\boldsymbol{x} \in \mathbb{R}^n, a \in \mathbb{N}, b \in \mathbb{N}$, and $c \in \mathbb{N}$
1: Set $\Sigma \leftarrow \{\sigma : [n] \to [n] | \sigma \text{ is bijective}\}$
2: **for** $i \in [c]$ **do**
3: Set $\sigma^{(i,j)} \xleftarrow{\$} \Sigma$ for $j \in [a]$
4: Set $x'_i \leftarrow \text{argmax}_{k \in [b]} \left(\prod_{j=1}^{a} \sigma^{(i,j)}(\boldsymbol{x})\right)_k$
5: Set $R^{(i)} \leftarrow \left\{\sigma^{(i,1)}, \cdots, \sigma^{(i,a)}\right\}$
6: **end for**
7: Set $R \leftarrow \left\{R^{(1)}, R^{(2)}, \cdots, R^{(n)}\right\}$
8: Set $\boldsymbol{x}' \leftarrow (x'_1, \cdots, x'_n)$
9: **return** $(\boldsymbol{x}', R)$

B.2. Identification Scenario

Biometric authentication systems are typically used in two settings: verification and identification. This paper focused on attacking the 1-to-1 verification setting. In the 1-to-N identification setting, when the user queries an image, FRSs must identify the closest identity among all the enrolled identities. For example, many commercial FRS services allow clients to query for top-k[4] similarity scores among the enrolled identities, i.e., if we make m queries, we will obtain the $k \times m$ scores from the database. For identification, a queried image is identified as the identity whose score is highest and with the corresponding score above a threshold $\sigma \geq 0$. The goal of an identification attack is to generate a query image that passes the score threshold, i.e., it is identified as one of the enrolled identities.

B.2.1. Adapting Our Attack For Identification. Our attack can be straightforwardly adapted as follows. For each image o_i in the OFS of size m, query the target FRS to obtain the top-k identities and their similarity scores. Observe that the top-k identities returned by the FRS may be different for each o_i. Nevertheless, after querying the whole OFS, we obtain $m \times k$ scores, where we observe that some identities will appear more than once. Accordingly, we select the identity I for which we know the most number of similarity scores (tie-breaking arbitrarily) and carry out the rest of our attack against this identity.

[4]Note that the target commercial FRSs [8], [9], [10] in our paper allow clients to set k up to 4096, 5, and 10, respectively.

Threshold	$72.54°$	$76.41°$	$77.29°$	$^{\dagger}80.15°$
ASR	89.06%	97.72%	98.06%	99.83%

TABLE 14: Direct ASR using various thresholds. "†" indicates the best threshold whose true positive identification rate is best. The other thresholds are from the Table 6.

	$\lvert\mathcal{O}\rvert$	$\text{cond}(\widehat{\boldsymbol{A}})$	$\lvert\lvert\boldsymbol{s}\rvert\rvert$	$\lvert\boldsymbol{s}\rvert$	ASR
Verification	45	1.6758	0.4223	45	53.10%
	75	2.0750	0.5465	75	89.27%
Identification	99	1.0740	0.4243	$\approx$3	89.06%

TABLE 15: Comparison between verification and identification scenario in terms of the norm of score vector $\boldsymbol{s}$ and ASR. Note that the operation $|\cdot|$ outputs a dimension of the input vector. In our case, the dimension indicates the number of cosine similarities, which is used for linear equation $\widehat{\boldsymbol{A}}\hat{\boldsymbol{x}} = \boldsymbol{s}$.

Compared to the verification scenario, we obtain far fewer similarity scores for I (typically around 3 in our experiments). However, these scores provide us with much more useful similarity information because they were returned in a top-k setting. Experimentally, we observe that the condition number of matrix $\widehat{\boldsymbol{A}}$ consisting of templates from the OFS is very small.

B.2.2. Experiment. To evaluate our adapted attack, we fix the open-source target FRS T_1 and the IARPA Janus Benchmark–C (IJB-C) face dataset [61]. We follow the IJB-C dataset evaluation protocol which supports 1-to-N identification attack scenarios. Briefly, we first extract 19593 templates from 3531 subject identities using T_1 (subjects have different numbers of images in the dataset). We then repeat the following evaluation process for 10000 iterations and take the average accuracy result:

1) Randomly select one face image from each identity and set the 3531 corresponding templates as enrolled templates in T_1.
2) Apply our attack against T_1 to reconstruct a face image for target I (as explained above).
3) Check if the reconstructed image is accepted by T_1.

Table 14 shows the attack success rate of our adapted attack for various acceptance thresholds. We note that our adapted attack works well in the identification scenario.

As remarked in Section 5.5, some of the key factors affecting the accuracy of our attack are the condition number of the template matrix $\widehat{\boldsymbol{A}}$ and norm of $\boldsymbol{s}$. In Table 15, we show experimental values for some OFS sizes at the acceptance threshold of $72.54°$. In the identification scenario, the condition number of $\widehat{\boldsymbol{A}}$ is very small, whilst $||\boldsymbol{s}||$ is relatively large despite only having 3 entries. We leave the investigation of these parameters to future work.

Appendix C.
Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary

This paper reports on an approach to reconstruct faces through queries to face recognition systems. By using the similarity scores revealed by face recognition systems in response to a query, the authors devise an approach based on a small set of non-adaptive queries to reconstruct faces from the returned similarity scores.

C.2. Scientific Contributions

Provides a Valuable Step Forward in an Established Field

C.3. Reasons for Acceptance

This work provides a valuable step forward in techniques of face reconstruction. The proposed approach can reconstruct face images with much fewer queries compared to prior techniques that use similarity scores to reconstruct faces. Whereas previous approaches used adaptive techniques in the query process, those approaches needed approximately 50,000 image queries. In contrast, the proposed approach uses only 100 queries. Achieving this vast reduction in the number of queries through a non-adaptive approach is a significant contribution.

To achieve such significant speedups, the proposed technique builds a model (using metric learning) with faces to interpret the similarity scores for face recognition queries. Leveraging the topological meaning of similarity scores, the proposed work significantly reduces the number of queries needed for face reconstruction. The approach significantly improves the attack practicality (e.g., reducing time and cost to launch such an attack).

The approach is evaluated against three commercial face recognition services with images from three datasets. Because the proposed solution relies only on the similarity scores, it has the potential to be adapted to other types of biometric authentication deep learning models.