Breaking High Confidence: Practical Face Impersonation under High-Security Thresholds

Changjin Kim, Seunghun Paik, Dongsoo Kim, and Jae Hong Seo*

Department of Mathematics & Research Institute for Natural Sciences
Hanyang University, Seoul 04763, Republic of Korea
{changjinkim,whiteseoonguh,frds37,jaehongseo}@hanyang.ac.kr

Abstract. Face recognition systems (FRSs) are increasingly deployed in critical real-world services for authentication, such as banking applications and airport identity checks, necessitating stringent security configurations. Consequently, the security vulnerabilities of FRSs have garnered significant attention. While existing studies have extensively explored FRS security, prior analyses have primarily focused on medium-security threshold settings, which are not directly applicable to FRSs operating under high-security constraints. In this paper, we propose the first successful impersonation attack against FRSs under high-security threshold settings. Among various threat models, we focus on a practical and challenging scenario: score-based impersonation attacks under strict rate limits. To precisely evaluate the feasibility of such attacks, we provide a principled mathematical analysis characterizing the gaps in each stage of the attack pipeline. Our method significantly enhances impersonation capabilities in score-based attacks, even under elevated decision thresholds. On the LFW benchmark, with a budget of only 100 confidence score queries per identity, our attack achieves an impersonation success rate exceeding 92% against Amazon Rekognition at a confidence score threshold of 99—recommended setting for law enforcement scenarios. We further observe consistently robust performance across multiple open-source FRSs evaluated at similarly stringent decision thresholds.

Keywords: High-Security · Impersonation Attack · Commercial APIs

1 Introduction

Biometric recognition has become a common way to verify identity in everyday life, and its convenience has driven rapid progress and widespread adoption across diverse applications. In particular, face recognition is now a widely used authentication method for identity verification across a broad range of services beyond personal devices to public infrastructure—for example, U.S. agencies deploy facial biometrics in airport identity checks [64]. It is also increasingly used in financial onboarding, such as remote account opening and eKYC verification [16].

* Corresponding author.

Table 1: Summary of prior attacks and settings. All attack success rate (ASR) values are computed using the LFW evaluation protocol. We denote by AWS(k) the AWS Rekognition confidence score threshold of k between 0-100.

Method	Access	Query budget	ASR@FMR (Mean Cos. Sim.)	Notes
Shahreza et al. [57]	white-box	–	94.90@10^{-3}	
Shahreza et al. [59]	white-box	–	96.40@10^{-3}	3D face reconstruction
Shahreza et al. [46]	white-box	–	96.40@10^{-3}	high-resolution
Shahreza et al. [58]	white-box	–	96.48@10^{-3}	
Template-based				
Mai et al. [41]	black-box	–	95.20@10^{-3}	
Dai et al. [8]	black-box	–	94.35@10^{-3}	high-resolution
Dong et al. [12]	black-box	–	98.00@10^{-3}	
Jung et al. [27]	black-box	–	73.04@10^{-4}	
Razzhigaev et al. [51]	black-box	300,000	(0.90)	
Score-based				
Razzhigaev et al. [52]	black-box	50,000	(0.98)	
Razzhigaev et al. [53]	black-box	20,000	(0.97)	
Park et al. [49]	black-box	4,000	96.25@10^{-3}	
Kim et al. [36]	black-box	100	99.60@0	13.9@AWS(90)
Ours	black-box	100	92.8@AWS(99)	

As deployments expand into high-stakes domains discussed above, reducing false matches becomes critical (often at the cost of rejecting more legitimate users). Thus, practical deployments in such settings frequently adopt conservative decision thresholds. In a face recognition system (FRS), decision thresholds are set based on a false match rate (FMR), which means the ratio of non-matching attempts that are incorrectly accepted as matches, and adversaries against FRSs that aim to induce false matches are referred to as impersonation attacks. As a representative benchmark for high-security decision thresholds, NIST FRTE (1:1 verification) reports performance using FMR = 10^{-6} [45]. In commercial API settings, AWS Rekognition's public-safety (law-enforcement) guidance recommends using a high confidence threshold (99% or higher) [4].

Impersonation attacks against FRSs have been widely studied. They are commonly categorized in terms of the feedback type from the target FRS through querying faces: template-based attacks [8, 12, 27, 41, 51] that obtain a feature template from queries, and score-based attacks [36, 49, 52, 53] that obtain similarity feedback (e.g., cosine similarity or confidence score) with the enrolled face. These attacks have demonstrated effectiveness against various FRSs, including even commercial APIs such as Face++ [17], Kairos [28], and AWS [1]. Typical commercial APIs return the confidence score in response to queries, without publicly disclosing any internal structure or formula for computing the confidence score. Under such settings, score-based attacks in the black-box setting—i.e., no internal knowledge of the target FRS is available to the adversary—have gained attention as a plausible threat against real-world FRS deployments.

Despite these advancements, however, we find that prior works have mainly analyzed medium-security settings (e.g., FMR = 10^{-3}), and the analysis for high-security regimes (e.g., FMR = 10^{-6}) has been less explored. Table 1 provides representative prior work in terms of feedback type, query budget, and

the attack success rate (ASR) at the given FMR. We can observe that all these works either require unrealistically large query budgets (4,000—300,000) for realistic deployments with rate limits, or the ASR sharply drops at high-security threshold settings. Such a landscape leaves a critical yet underexplored regime on the security analysis of FRSs in high-security settings, despite the increasing adoption of real-world FRS deployments under such scenarios.

1.1 Our Contributions

In this paper, we analyze the security of FRSs in high-security settings (i.e., FMR $= 10^{-6}$ or a confidence score threshold of 99) by proposing a black-box score-based attack, revealing their vulnerability for the first time under a practical query budget of at most 100. We first systematically analyze existing score-based attacks that operate under low query budgets, identifying multiple bottlenecks that limit ASRs in stringent threshold settings. Through a theoretical analysis of the geometry of the template space, we propose a series of techniques to address these bottlenecks, thereby achieving a non-trivial ASR in various open-source and commercial FRSs. Notably, using only 100 black-box score queries, our attack achieves an ASR of 92.8% against the AWS Rekognition API [1] at a threshold of 99 on the LFW dataset [23], where prior work failed to achieve even a non-negligible ASR under such a strict query budget. Our findings suggest that system-level high-security configurations alone may not suffice to ensure security in adversarial settings, highlighting the need for a better understanding of the inherent robustness of FRSs in such high-security regimes. We summarize our contributions as follows:

- We conduct the first security analysis of FRSs under high-security thresholds, focusing on practical black-box score-based impersonation.
- We revisit prior query-efficient score-based attacks, demonstrating why it fails under high-security settings through systematic analyses.
- By analyzing geometric properties of the template space, we propose multiple techniques to enable such attacks in high-security settings, whose effectiveness is supported by both theoretical and experimental analyses.
- We conduct extensive experimental analyses, revealing vulnerabilities of various open-source and commercial FRSs in high-security threshold settings using at most 100 queries. Our attack achieves high ASRs of 92.8% (LFW [23]), 95.24% (CFP-FP [55]), and 96.87% (AgeDB [44]) on AWS Rekognition at a confidence score threshold of 99. It also achieves 32.63%–89.27% on open-source FRSs at FMR $= 10^{-6}$ thresholds.

2 Related Work

Score-based Attacks To impersonate the enrolled identity from the target system's score responses, several studies consider gradient-free optimization techniques, such as hill-climbing [19, 51, 52, 66], genetic algorithms [5, 12, 27, 61], or

gradient estimation [49, 53]. Recent studies exploit the generative model's latent space to recover high-fidelity facial images [5, 12, 27, 49, 66]. While these attacks recover high fidelity as the number of score queries increases, they require a huge number of queries (4,000–300,000). Furthermore, as demonstrated in several studies [37, 50], their adaptive query trajectories often leave structured footprints that are detectable by the target FRS service provider, enabling system-level defenses such as query filtering or request rejection.

Notably, a recent study [36] demonstrated that 100 non-adaptive score queries suffice to achieve a non-trivial ASR on various commercial APIs, which is based on the geometry of template space. However, their experimental setting is about scenarios with moderate security, e.g., a default threshold in AWS; for thresholds of high-stakes scenarios, their ASR drastically diminishes to 0.

Template-based Attacks From the success of the NbNet [41] in the black-box attack setting, several template inversion attacks have been proposed with various methodologies. They often assume that the adversary can query the chosen face to the target FRS, obtaining the template extracted from the face. Under this assumption, they train the inverse model that reconstructs faces from the templates. Earlier attacks tried to build such a model directly [15, 41, 65]. Recently, the advancement of generative models, e.g., StyleGANs [30, 31] or diffusion models [22], facilitates the attack, enabling the adversary to recover high-fidelity facial images from these models. Several studies attempted to leverage the latent space of generative models, e.g., training an adaptor network from the template to the latent vector [8, 46, 56]. However, they require the adversary to obtain compromised templates enrolled in the target FRS, e.g., through a data breach, which is considered a stronger assumption than score-based attacks.

3 Backgrounds and Problem Formulation

3.1 Face Recognition

Typical FRSs employ the feature extractor that returns a fixed-length feature vector—called *face template*—from the input face. The feature extractor serves as a distance-preserving mapping, i.e., faces from the same identity are mapped to face templates close to each other, or vice versa. FRS determines whether two facial images represent the same identity by computing the similarity score between templates; if it exceeds a threshold, they are recognized as the same person. Recent FRSs [7, 11, 33, 34, 43, 54, 72] represent face templates as a unit vector and measure their similarity scores through cosine similarity.

As the demand for face recognition-based applications increases, several commercial API services, such as AWS Rekognition [1], Tencent [62], Kairos [28], and Face++ [17], have been adopted in various FRS deployments [3, 63]. The user can enroll the face in these services, and then query the face to obtain the confidence score with the previously enrolled one. Confidence scores are used for determining whether the enrolled and queried faces are from the same person, and the formula for computing the confidence score is publicly unknown.

3.2 Score-based Face Impersonation Attack

We consider two entities, the adversary and the target FRS where the target face is enrolled. Following prior score-based impersonation attacks [12,36,51–53,66], the adversary can query its own face to the target FRS and obtain a confidence score. We consider the *black-box* scenario; the adversary does not know the internal information of the target FRS, including the parameters, architecture, and training data of the feature extractor, and the decision threshold. Any prior information on the enrolled identity is unknown to the adversary. In practical scenarios with commercial APIs, the number of score queries made by the adversary is limited due to API rate limits and financial costs [25, 36]. Hence, we assume a query-limited adversary with a small query budget, e.g., tens to hundreds. Under this setting, the adversary aims to craft a facial image that is recognized as the same identity.

4 Revisiting Prior Low-Query Score-based Attacks

Motivated by this limitation, we revisit a recent score-based attack under a limited query budget [36] as our baseline attack method. We systematically analyze several errors involved during the attack and identify why the attack becomes less effective when attacking FRSs under stringent threshold settings.

4.1 High-Level Idea of the Attack

The core insight of their attack is that well-trained face recognition models behave as almost isometries, i.e., their embedding spaces exhibit similar geometric structures in terms of pairwise similarity. They empirically verify this assumption across various open-source FRSs. They also observed that each score query reveals the target template's relative position within the embedding space. Thus, given a surrogate face recognition model, the adversary can estimate the target template's position in the surrogate model's embedding space with score queries.

Precisely, if we denote s_i as the score from i^{th} face image query, then one can write $s_i = \langle x_i, y \rangle + \epsilon_i$ for the templates $x_i, y \in \mathbb{S}^{d-1}$ of queried and enrolled faces, respectively. Here, ϵ_i indicates the error inherited from the mismatch between the target and surrogate models' embedding spaces, which is unknown to the adversary. That is, after Q queries, the adversary obtains an equation $Ay = s - \epsilon$, where A is a $Q \times d$ matrix whose i^{th} row is x_i and $s = (s_i)_{i=1}^{Q}, \epsilon = (\epsilon_i)_{i=1}^{Q} \in \mathbb{R}^Q$. Since this equation is underdetermined and the errors are unknown, the adversary utilizes the least square method to estimate y while ignoring errors, i.e., computing $\widehat{y} \leftarrow A^{\dagger}s$, where $A^{\dagger}$ denotes the pseudoinverse of A. Finally, the adversary can recover the candidate facial image of the enrolled identity by inverting $\widehat{y}$ through template inversion methods, i.e., inverse models, for the surrogate model [15,41,46,56,60]. A detailed algorithm is provided in Section A.1 of the supplementary material.

Key Techniques to Instantiate the Attack To realize the attack, they proposed two techniques. The first one is the orthogonal face set (OFS), a set of facial images whose templates are (almost) pairwise orthogonal in the embedding space. They showed that the orthogonality helps reduce the effect of errors $\epsilon_1, \ldots, \epsilon_Q$ after being multiplied by $A^\dagger$. In addition, since commercial APIs return confidence scores rather than cosine similarities, they propose an estimation method of the cosine similarity from the confidence score via numerical methods.

4.2 Investigating Errors from Approximations

While elegant, the aforementioned approach relies on several approximations. Specifically, for high-threshold settings, the attack's efficacy becomes much more sensitive to these approximations, as the face reconstructed by the adversary must be accurate enough to surpass the tight decision threshold. For this reason, we investigate and quantify the impact of these approximations on the precision of the recovered face and their consequences for the success of the attack.

Throughout the attack pipeline, the error occurs when (i) interpreting score queries as cosine similarities (**E1**), (ii) solving the system of linear equations through pseudoinverse (**E2**), and (iii) the reconstruction error from the template inversion (**E3**). To formalize this, let us denote z as the template of the final reconstructed template, and A, y, s, $\widehat{y} = A^\dagger s$, and ϵ follow the same notation as in Section 4.1. Recall that $Ay = s - \epsilon$. Then we have that

$$\|y - z\|_2 \leq \underbrace{\|y - A^\dagger(s-\epsilon)\|_2}_{(\mathbf{E2})} + \underbrace{\|A^\dagger(s-\epsilon) - \widehat{y}\|_2}_{(\mathbf{E1})} + \underbrace{\|\widehat{y} - z\|_2}_{(\mathbf{E3})}. \tag{1}$$

We now analyze each error as follows. First, since $\widehat{y} = A^\dagger s$, the error (**E1**) corresponds to $\|A^\dagger \epsilon\|_2$. If we denote the operator norm of $A^\dagger$, denoted by $\|A^\dagger\|$, then $\|A^\dagger\epsilon\|_2 \leq \|A^\dagger\| \cdot \|\epsilon\|_2 = \sigma_{\min}(A)^{-1} \cdot \|\epsilon\|_2$, where $\sigma_{\min}(A)$ denotes the smallest singular value of A. This observation is essentially identical to the analysis by [36] based on the condition number, which justifies the use of OFS. The error (**E2**) corresponds to the gap between y and its projection onto the row space of A. Finally, the error (**E3**) depends on the template inversion method.

Quantifying the Errors in [36] To quantify the errors (**E1**)–(**E3**), we design the following experiments that measure the cosine similarity between the target and reconstructed faces. We follow the same attack algorithm in [36].

- (Exp. 1) Black-box attack: errors (**E1**)-(**E3**) affect.
- (Exp. 2) White-box attack: errors (**E2**) and (**E3**) affect.
- (Exp. 3) White-box attack with the ideal inversion: only error (**E2**) affects.
- (Exp. 4) White-box attack with the ideal inversion + PCA subspace: error (**E2**) remains, but its amount is minimized due to the optimality of the PCA subspace in terms of (average) projection gap.

We use the same surrogate and inverse models as in [36], and for the black-box attack, we select the ViT-based FRS [34] as the target model[1], whose training

[1] The surrogate and target models correspond to F_5 and F_1 in Tab. 2, respectively.

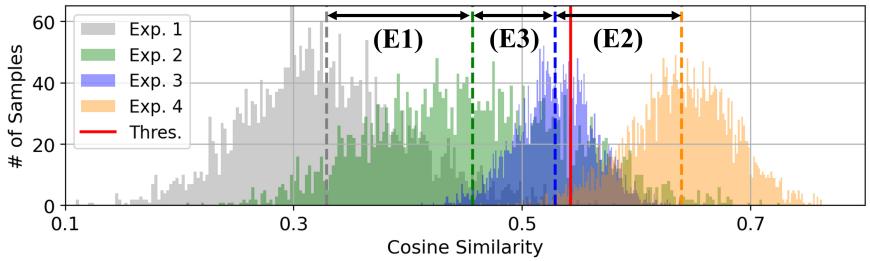


Fig. 1: The errors (**E1**)-(**E3**) in the attack by [36] measured through experiments (Exp. 1–4). The dashed lines indicate the average of each distribution. The solid red line denotes the estimated threshold when FMR $= 10^{-6}$. Best viewed in color.

dataset, architecture, and loss differ from the surrogate model. We run the PCA on the embeddings of the MS1MV3 dataset [10] extracted from the surrogate model and select the top 100 components with the highest singular values. We use the LFW dataset [23] as the target enrolled faces. To simulate the high-threshold scenario, we provide an estimate of the threshold for FMR $= 10^{-6}$.

The results are visualized in Fig. 1. For the baseline case (Exp. 1), almost all samples fail to exceed the threshold line of FMR $= 10^{-6}$ (red solid line) due to errors, except for only a small portion of samples ($\approx 0.2\%$). However, when we sequentially remove the effect of errors, the distribution of cosine similarity values gradually moves toward the right, and notably, for the rightmost case (Exp. 4), more than 97% of samples exceed the threshold. The averages of cosine similarity values for each experiment (Exp. 1–4) are 0.3289, 0.4567, 0.5289, and 0.6392, respectively. This result indicates that all these errors (**E1**)-(**E3**) in Eq. (1) intricately yet significantly affect the failure of [36] under high-security threshold settings. In the following section, we present our techniques to reduce each error, thus addressing the limitation.

5 Reducing Errors to Surpass High Thresholds

To overcome the errors (**E1**)-(**E3**) in Eq. (1), we revisit the techniques in [36] and show that those errors can be further reduced without additional queries.

5.1 (E1): Correction Matrix to Tame Metric Distortion

Motivated by [35], we utilize the *correction matrix* to reduce the error (**E1**). The correction matrix is defined as the inverse of the matrix $S = (s_{i,j})$, where $s_{i,j}$ denotes the cosine similarity between the i^{th} and j^{th} OFS elements obtained from score queries. With the correction matrix, the adversary computes $\widehat{y} \leftarrow A^T S^{-1} s$ instead of $A^\dagger s$. While [35] justified this idea through a somewhat exotic assumption, i.e., OFS serves as a universal basis over faces, we propose a theoretically principled interpretation of the correction matrix in terms of metric distortion.

Recall that $\widehat{y} \leftarrow A^\dagger s$ corresponds to solving the least squares over the surrogate embedding space, ignoring the metric distortion characterized by the additive error ϵ. Instead, if we apply a first-order approximation to such a distortion

by introducing a positive definite matrix Σ, we can incorporate the correction matrix with the (approximated) projection over the target template space. For templates x, y in the surrogate model's embedding space, let us assume that their cosine similarity in the target model can be approximated by $x^T \Sigma y$. Then we can write $S = A\Sigma A^T$, and $s = A\Sigma y$, and $\widehat{y} = A^T S^{-1} s = A^T (A\Sigma A^T)^{-1} (A\Sigma) s$.

We can observe that $\widehat{y}$ becomes the projection of y onto the row space of A in the metric determined by Σ, namely[2], $\widehat{y} = \arg\min_{z \in R(A)} (y-z)^T \Sigma (y-z)$, where $R(A)$ denotes the vector space spanned by the rows of A. That is, $\widehat{y}$ provides a better estimate in the target embedding space in terms of the linear combination of OFSs. While this interpretation relies on approximating the metric distortion via Σ, we empirically verify that the correction matrix substantially reduces the error (**E1**), especially when attacking the commercial FRS.

5.2 (E2): OFS with an Optimal Projection Gap

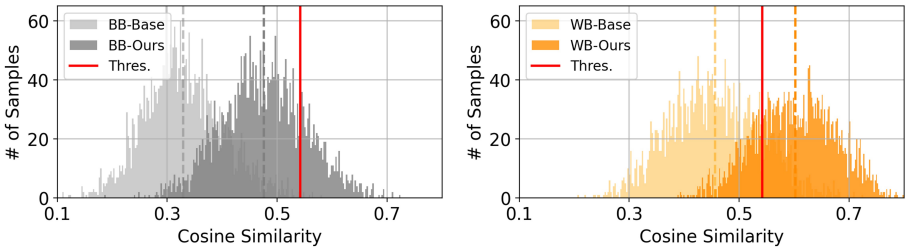
While the error (**E2**) stems from the subspace projection, the baseline OFS by [36] enforces orthogonality only. Instead, we select the OFS obtained from the PCA, which provides the optimal subspace to minimize the projection gap. More precisely, for a dataset $X \subset \mathbb{R}^d$, the PCA with k principal components finds the row orthogonal matrix $W \in \mathbb{R}^{k \times d}$ which minimizes $\frac{1}{|X|} \sum_{y \in X} \|W^T W y - y\|_2^2$. Hence, with a sufficiently large dataset for X, e.g., large-scale public training datasets, the adversary can effectively reduce the error (**E2**) by setting $A = W$, i.e., using the faces corresponding to the principal components as the OFS.

5.3 (E3): Improved Inverse Model

To reduce the error (**E3**), we propose a new inverse model called the *SPNet*. Our design starts from the NbNet [41], the first black-box inverse model. Recently, [47] improved the NbNet by adopting the style mapping network of StyleGAN [30] with the adaptive instance normalization (AdaIN) [24], regarding an input face template as a style. They replaced each batch normalization [26] with AdaIN. In contrast, [58] pointed out that the NbNet is prone to producing blurry output, proposing a new architecture, called DSCasConv, based on the skip connection. We find that merging these techniques can achieve the best of both worlds, obtaining a more accurate inverse model.

In addition, we made the following micro-level modifications, including (i) removing `tanh` at the last layer to obtain a sharper facial image, (ii) replacing `ReLU` with `GELU`, and (iii) employing pixel-wise normalization (PNorm) [29] for the deconvolution layer block. Moreover, by following [47]'s observation, we directly use the MS1MV3 training dataset [10]. Though our design choices are empirical, we empirically verify that they substantially improve the inversion precision, as well as the attack success rate. We provide the detailed architecture configuration in Section C of the supplementary material.

[2] The proof is given in Section B of the supplementary material.



(a) Black-Box Attacks **(b)** White-Box Attacks

Fig. 2: The error analysis with our techniques—correction matrix (**E1**), PCA-based OFS (**E2**), and SPNet (**E3**)—with a comparison to the baseline [36] in the black-box (BB) and white-box (WB) attacks. In the BB attack of ours, all the techniques are enabled, whereas only PCA-based OFS and SPNet are enabled in the WB attack.

Training Losses To train the proposed inverse model, we employ the following losses that are widely used for training the inverse model.

- (Pixel Loss) it minimizes the absolute error between the original face x and reconstructed ones $\widehat{x}$, namely, $\mathcal{L}_{\text{Pixel}}(x, \widehat{x}) = \|x - \widehat{x}\|_1$.
- (Identity Loss) it maximizes the cosine similarity between templates of the original face x and reconstructed ones $\widehat{x}$ extracted from various FRSs. More precisely, for FRSs $F_1, \ldots, F_k$, we define $\mathcal{L}_{\text{ID}}(x, \widehat{x}) = \sum_{i=1}^{k}(1 - \langle F_i(x), F_i(\widehat{x})\rangle)$, assuming that each embedding is normalized.

Since the adversary trains the inverse model of its own surrogate FRS, unlike black-box inverse models [15,41], we also include the surrogate model to compute $\mathcal{L}_{\text{ID}}$. The final loss is $\mathcal{L}_{\text{Train}} = \lambda_1 \mathcal{L}_{\text{Pixel}} + \lambda_2 \mathcal{L}_{\text{ID}}$ for hyperparameters λ_1, λ_2.

5.4 Effect of the Proposed Techniques on the Errors

To show the validity of the proposed techniques, we show how they affect the cosine similarity of original and recovered faces in the black-box and white-box settings. In the black-box attack setting, we enable all the techniques (correction matrix, PCA-based OFS, and SPNet) to reduce all the errors (**E1**-**E3**). In contrast, since the error (**E1**) is not engaged in the white-box attack setting, we only enable the PCA-based OFS and SPNet. All the remaining experimental settings, e.g., choice of the surrogate/target FRSs and the dataset to run the PCA, are equivalent to those of Fig. 1 of Section 4.2. The results are visualized in Fig. 2a and 2b, respectively. We can observe that our techniques substantially improve the cosine similarities in both settings, showing their validity. Notably, we can observe that these techniques finally enable the similarity values to surpass a high-security threshold; 20% and 78.07% of samples exceed the red line in black-box and white-box attacks, respectively.

6 Experimental Analysis

To verify the effectiveness of the proposed techniques, we conduct extensive experimental analyses on various open-source and commercial FRSs.

Table 2: Summary of target open-source and commercial FRSs. Thres. denotes estimated threshold when FMR $= 10^{-6}$ for each open-source FRS.

FRS	Architecture	Loss	Train Dataset	Thres.
F_S [11]	ResNet-100 [21]	ArcFace	Glint360K [6]	0.5422
F_1 [34]	ViT-Base [13]	AdaFace [33]+KPRPE	WebFace12M [73]	0.5195
F_2 [9]	ResNet-100	CosFace [69]+TopoFR	MS1MV2 [11]	0.5216
F_3 [39]	ResNet-100	SphereFace-R	MS1MV2	0.5537
F_A [1]	Not publicly disclosed			

6.1 Experimental Settings

Attack Experimental Setting We select the following open-source FRSs as the target systems: ViT-KPRPE [34] (F_1), TopoFR [9] (F_2), SphereFace-R [39] (F_3). Their pre-trained parameters are publicly available in their public GitHub repositories, e.g., `CVLFace` [32] and `OpenSphere` [38]. For the surrogate model (F_S) held by the adversary, we use the ArcFace [11] from the `InsightFace` library [20]. We select the AWS CompareFace API (F_A) as the target commercial system. We summarize the specification of each FRS in Tab. 2.

We evaluate the attack success rate (ASR) by measuring how many reconstructed faces are authenticated as the target face. The thresholds for open-source FRSs are set on the FMR of 10^{-6}. Since the AWS CompareFace returns the confidence score between 0 and 100, we select 80, 90, and 99 as the decision thresholds; note that 80 is the default threshold [2] and 99 is recommended for high-stakes use-cases [4]. For evaluation, we select LFW [23], CFP-FP [55], and AgeDB [44] datasets, which are widely used for benchmarking the accuracy of FRSs. These datasets consist of pairs of facial images from the same or different identities; we enroll the first face of the pairs from the same identity and measure the ASR. Details on these datasets are provided in Section D of the supplementary material. To craft the PCA-based OFS, we select 100 principal components of templates of the MS1MV3 dataset [36] extracted from F_S. When attacking F_A, we use the same method to convert confidence scores into cosine similarities as [36], whose details are provided in Section A.2 of the supplementary material.

Threshold Estimation for FMR $= 10^{-6}$ In LFW, CFP-FP, and AgeDB, low FMRs like 10^{-6} cannot be directly measured because of insufficient numbers of false pairs. While the IJB-C benchmark [42] provides low FMR settings, e.g., 10^{-6}, the resulting threshold is not aligned with our evaluation datasets because, unlike them, the templates in the IJB-C protocol are the average of embeddings from multiple faces of the same identity. Thus, we estimate the threshold for FMR $= 10^{-6}$ by using large-scale face datasets. Since F_S, F_2, and F_3 are trained from MS1M-derived datasets (MS1MV2, Glint360K), to avoid potential overlap, we select the CASIA-WebFace dataset [71] that is known not to be derived from the MS1M lineage. Similar to the NIST-FRTE protocol [45], we randomly sample 4,096 identities, sample one image per identity, and compute the false-pair cosine similarities from all cross-identities to find the candidate threshold for FMR $= 10^{-6}$. Tab. 2 reports the average over 1,000 repetitions. Though

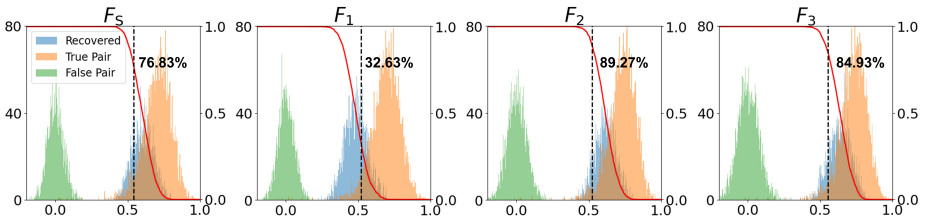


Fig. 3: Attack results of ours on various open-source FRSs in the LFW dataset. x-axis: cosine similarity. y-axis: the number of samples (left), ASR (right). Red solid lines denote the ASR for each threshold, and black dashed lines denote the threshold at FMR $= 10^{-6}$. We also report the ASR at that threshold. Best viewed in color.

CASIA-WebFace may introduce distributional bias, the resulting ASRs are likely conservative: at directly measurable FMRs 10^{-2} and 10^{-3}, CASIA-WebFace often yields comparable or tighter thresholds than the benchmark-specific ones. This suggests that the resulting ASRs based on CASIA-WebFace thresholds are not optimistic estimates, and details are provided in Tab. 2 and 3 of Section D in the supplementary material.

Training Inverse Model We train SPNet with the MS1MV3 dataset [10]. We use the AdamW optimizer [40] with an initial learning rate of 0.001, weight decay of 0.0005, and a batch size of 64 for 81,000 total steps (approximately one epoch for the MS1MV3 dataset). A linear warm-up is applied during the first 5% of training steps, followed by cosine annealing decay to 1% of the initial learning rate. The loss coefficients are set to $\lambda_1 = \lambda_2 = 10$. For the identity loss, we employ an ArcFace ResNet-100 trained on MS1MV3 with ElasticFace-Cos+ [7].

6.2 Attack Results

Open-source FRSs Fig. 3 visualizes cosine similarity scores of true/false pairs (orange/green histograms) of the LFW dataset and those between the target and reconstructed faces from our attack. Our attack achieves non-trivial ASRs in all target FRSs, achieving 76.83% in the white-box setting (F_S) and 32.63%–89.27% in the black-box setting (F_1–F_3). Since all training settings for F_1 differ from F_S, it exhibits the lowest ASR, though it still achieves a non-negligible ASR. For F_2 and F_3, our attack achieves much higher ASR, even exceeding the result in the white-box setting. This stems from the extent of non-uniformity in the face template distributions, which is discussed in Section E.1 of the supplementary material. We provide the results for the CFP-FP and AgeDB datasets in Section E.2 of the supplementary material.

AWS CompareFace Tab. 3 shows that the baseline attack exhibits an ASR of at most 0.60% in the tight threshold setting (99), though it achieves a non-trivial ASR at moderate thresholds. When our techniques for (**E1**) and (**E2**) are enabled, the attack achieves an ASR of up to 63.50%. Such a gain is also observed when we replace NbNet with Arc2Face [48], a diffusion-based inverse

Table 3: ASRs (%) on benchmark datasets against F_A. Baseline corresponds to "NbNet" without our techniques. C: correction matrix. P: PCA-based OFS.

Used Tools/Tech.	LFW			CFP-FP			AgeDB		
	80	90	99	80	90	99	80	90	99
Baseline* [36]	27.00	13.90	–	20.60	9.50	–	29.60	17.70	–
Reproduced	25.48	12.54	0.60	23.78	11.17	0.57	26.97	13.63	0.37
NbNet+C	70.89	48.98	2.83	64.14	41.43	2.46	61.17	40.27	2.10
NbNet+P	99.60	97.37	46.68	99.89	98.40	51.26	99.73	98.47	55.67
NbNet+CP	99.67	98.20	56.10	99.63	98.63	60.60	99.93	99.23	63.50
Arc2Face+CP	99.43	97.83	53.92	99.26	97.71	53.03	99.13	96.53	44.83
SPNet+CP	99.90	99.83	**92.80**	99.87	99.84	**95.24**	100.00	100.00	**96.87**

*These results are reported in their original paper.

model that produces high-resolution (512×512) faces. After replacing NbNet with SPNet, the ASR exceeds 92.80% even in a high-stakes scenario threshold. For comparison with [36], we re-implement its attack and evaluate it under the same AWS CompareFace API version as ours. The reproduced ASR slightly differs from the originally reported ASR. We guess that this discrepancy may stem from differences in the API/model version. We access the AWS CompareFace API with `FaceModelVersion: v7` (announced: 2023-12-05; accessed: 2026-03-05), whereas the baseline likely used v6 (announced: 2022-05-06), given the camera-ready deadline (2023-08-18) of the IEEE S&P 2024 summer cycle.

Visualization of Attack Result Fig. 4 provides examples of the attack results; the 1st row and 2nd rows are the faces of the true pairs in the LFW dataset. The 3rd row shows the reconstruction of target faces in the 1st row. We also provide the confidence score from F_A. Although the resulting faces are visually dissimilar to the targets, the confidence score is sufficiently high to surpass the high-stakes scenario threshold, even surpassing the score of true pairs in some cases. This shows the discrepancy between the human's view and the embedding-level similarity, as our attack explicitly exploits the embedding space's geometry.

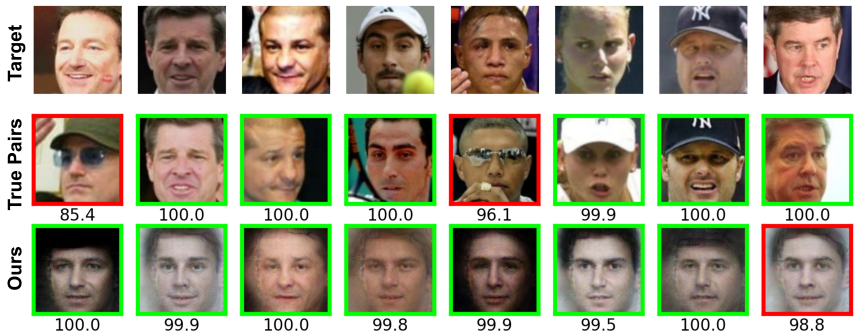
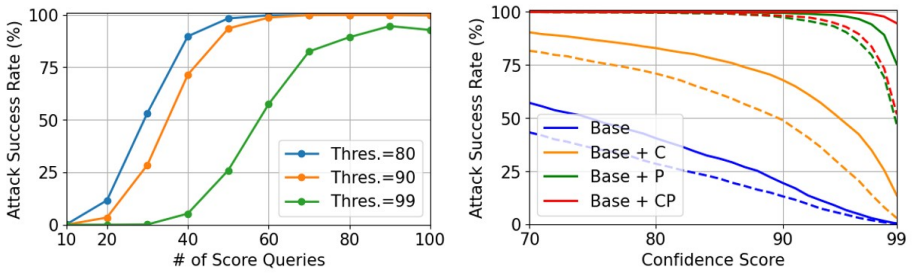


Fig. 4: True pairs of the LFW dataset (1st and 2nd rows) and the reconstructed faces from our attack (3rd row) targeting those in the 1st row. Green/red boxes denote whether the confidence score with the 1st row's faces exceeds 99 or not, respectively.



(a) Query Budget Analysis **(b)** Effect of Techniques

Fig. 5: Ablation studies against F_{A} in the LFW dataset with respect to the effect of query budget and the effect of the correction matrix (C) and PCA-based OFS (P). In (b), dashed lines and solid lines denote the ASR results from the baseline NbNet in [36] and the proposed SPNet. Best viewed in color.

Comparison with Additional Baselines We further examine whether three score-based attacks [49, 52, 53] remain competitive in the 100-query regime. For [52], we use the query-wise trend reported in [36]; for the other two attacks, we adapt their optimization procedures to the 100-query setting. As these attacks rely on many iterative score queries, ranging from 4,000 to 300,000, their ASRs remain limited under this budget, reaching only about 1% even at FMR$= 10^{-4}$. In contrast, our attack achieves non-trivial ASRs under the same query budget and high-security thresholds. Detailed experimental settings and results are provided in Section E.5 of the supplementary material.

6.3 Ablation Study

We conduct ablation studies on our techniques against F_{A}; the results for F_1–F_3 are provided in Section E.3 of the supplementary material.

Query Budget Analysis Fig. 5a shows ASR on LFW as the number of OFS queries increases from 10 to 100 under various decision thresholds. At thresholds 80 and 90, ASR increases rapidly and approaches 100% with 60 queries. On the other hand, at the high-security setting (99), ASR remains low for small budgets, rising sharply after approximately 50 queries to achieve non-trivial ASR. Overall, 50 queries suffice to achieve a non-trivial ASR in all threshold regimes we tested.

Effect of Proposed Techniques on ASR Fig. 5b depicts the effect of the proposed techniques to handle errors (**E1**-**E3**) on the ASR for F_{A}. We can observe that each technique (correction matrix, PCA-based OFS, and SPNet) consistently and independently improves the ASR, even for both the baseline NbNet and the proposed SPNet. This demonstrates how our techniques can overcome limitations of [36] observed in Section 4.2 in a strict threshold setting.

Inverse Model We analyze the impact of our modifications to design SPNet. We train the baseline NbNet of [47] for F_S, and then add micro-level modifications and DSCasConv [58], fixing the loss and the training recipe. Measuring the inversion precision on LFW and the ASR of our attack, Tab. 4 shows that our design improves both. We further assess the robustness of both inversion precision and ASR to training randomness over five random seeds; the ASR on F_A exceeds 90% across all five seeds, while the inversion precision remains nearly unchanged, varying by less than 10^{-4}.

Table 4: Ablation study on the inverse model

Model	Inv. Prec.	ASR on F_A
NbNet [47]	0.8809	56.10
+ Tuning/Micro	0.9687	91.41 ± 1.75
+ DSCasConv [58]	**0.9782**	**91.88 ± 1.60**

7 Discussion

Potential Mitigation of the Attack To defend against the proposed attack, a straightforward solution is to hide the confidence score. However, this may not be the fundamental solution because of insider adversaries who can access the raw API responses, e.g., service or API developers. Instead, one may consider designing an FRS model having a stronger metric distortion—amplifying the error (**E1**)—with respect to the open-source FRSs, e.g., FRSs that maintain high accuracy under tighter thresholds suggested by [35]. To reduce the gain from the PCA, one may design a loss function that facilitates spreading the template distribution, such as some loss functions studied in the literature [14].

We test our attack against ResNet-100 models with these defense methods applied. With the fixed surrogate F_S, our attack achieves ASRs of 41.33% for [14] at FMR= 10^{-6}, which is roughly 47.94% and 43.60% drops compared to the same ResNet-100 target models F_2 and F_3. In Section E.1 of the supplementary material, we further show that [14] yields a less concentrated PCA singular-value spectrum, thus increasing the error (**E2**). In contrast, the ASR for [35] is 0.13% at FMR= 10^{-6}, substantially weakening the attack efficacy. We believe that these empirical results may help better understand the inherent robustness of FRS, leaving their further investigation as an interesting future direction.

Image Quality of the Reconstructed Faces The proposed inverse model crafts relatively low-resolution faces compared to recent inversion models; in fact, our inverse model produces 128×128 resolution images, while recent models can handle 512×512 or 1024×1024 resolution [46, 48, 56]. Nevertheless, all the reconstructed faces pass the face detection algorithm of the AWS CompareFace API without any additional treatments, and as we already observed in Tab. 3, even reducing (**E1**) and (**E2**) alone already suffices to achieve a non-trivial ASR. When we combine the proposed attack with Arc2Face [48], which produces faces of 512×512 resolution, our attack still maintains 53.92% of ASR when the threshold is set to 99, i.e., a high-security setting[3]. Improving high-resolution inverse models would further enhance the ASR.

[3] We provide the example recovered faces in Section E.4 of the supplementary material.

On the Feasibility of Real-World Physical Attacks While we showed the validity of the proposed attack against the commercial API, our evaluation considers a digital setting by directly submitting faces to the API. However, in the real-world physical attack, the reconstructed faces may deform during presentation, e.g., printing or occlusion, and sensor-level defenses such as presentation attack detection [18, 67, 68] can be employed. Nevertheless, we emphasize that our attack exposes the vulnerability in the feature extraction stage, which is orthogonal to the sensor-level countermeasures mentioned above. This suggests that the attack could potentially be realized in physical settings when combined with off-the-shelf spoofing attacks, leaving it as an interesting future direction.

The Effect of Surrogate (Mis-)Alignment Surrogate-target misalignment (**E1**) plays a significant role in attack efficiency, as we observed in our analyses against F_1–F_3 and the defense method by [35]. Since our primary goal is to analyze a commercial FRS F_{A}, we do not exhaustively explore this aspect. Nevertheless, its systematic characterization across diverse surrogate-target pairs in terms of architecture, training datasets, and losses would help us better understand the FRS vulnerability, and we leave it as important future work.

8 Conclusion

Despite the increasing adoption of FRSs in high-stakes scenarios, the integrity of FRSs under such scenarios has not yet been thoroughly explored. In this paper, we demonstrate the vulnerability of various open-source and commercial FRSs even in low-FMR (e.g., 10^{-6}) settings under black-box score-based attack scenarios with a practical query budget. Our attack is enabled by a geometric interpretation of score queries over the embedding space, together with a novel inverse model that achieves high inversion precision. Beyond the attack itself, the proposed inverse model may be utilized for other applications, such as identity-preserving synthetic face generation [48,70]. We hope that our findings contribute to improving the security and robustness of future FRS designs.

Ethical Statement Our goal is to understand the potential vulnerability of the FRSs against score-based attacks in strict security settings, not to facilitate misuse. All the experiments were done in a strictly controlled setting, and **we did not launch attacks against real-world FRS deployments**.

Acknowledgements

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) (RS-2021-II210727), and in part by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00332210).

References

1. Amazon Web Services: Amazon rekognition, https://aws.amazon.com/rekognition/, accessed: 2026-02-27
2. Amazon Web Services: Comparefaces - amazon rekognition, https://docs.aws.amazon.com/rekognition/latest/APIReference/API_CompareFaces.html
3. Amazon Web Services: Identity verification using amazon rekognition, https://aws.amazon.com/rekognition/identity-verification/, accessed: 2026-03-05
4. Amazon Web Services: Use cases that involve public safety, https://docs.aws.amazon.com/rekognition/latest/dg/considerations-public-safety-use-cases.html
5. An, S., Yao, Y., Xu, Q., Ma, S., Tao, G., Cheng, S., Zhang, K., Liu, Y., Shen, G., Kelk, I., et al.: Imu: Physical impersonating attack for face recognition system with natural style changes. In: 2023 IEEE Symposium on Security and Privacy. pp. 899–916. IEEE (2023)
6. An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T.: Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4042–4051 (2022)
7. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1578–1587 (2022)
8. Dai, L., Shen, Z., Zhou, Z., Yu, P., Xia, Z.: Clip-fti: Fine-grained face template inversion via clip-driven attribute conditioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3479–3487 (2026)
9. Dan, J., Liu, Y., Deng, J., Xie, H., Li, S., Sun, B., Luo, S.: Topofr: A closer look at topology alignment on face recognition. Advances in Neural Information Processing Systems **37**, 37213–37240 (2024)
10. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5203–5212 (2020)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
12. Dong, X., Miao, Z., Ma, L., Shen, J., Jin, Z., Guo, Z., Teoh, A.B.J.: Reconstruct face from features based on genetic algorithm using gan generator as a distribution constraint. Computers & Security **125**, 103026 (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Duan, Y., Lu, J., Zhou, J.: Uniformface: Learning deep equidistributed representation for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3415–3424 (2019)
15. Duong, C.N., Truong, T.D., Luu, K., Quach, K.G., Bui, H., Roy, K.: Vec2face: Unveil human faces from their blackbox features in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6132–6141 (2020)
16. European Banking Authority (EBA): Guidelines on the use of remote customer onboarding solutions under article 13(1) of directive (eu) 2015/849 (eba/gl/2022/15),

https://www.eba.europa.eu/sites/default/files/document_library/Publications/Guidelines/2022/EBA-GL-2022-15%20GL%20on%20remote%20customer%20onboarding/1043884/Guidelines%20on%20the%20use%20of%20Remote%20Customer%20Onboarding%20Solutions.pdf
17. Face++: Face++, https://www.faceplusplus.com/, accessed: 2026-02-27
18. Fang, H., Liu, A., Wan, J., Escalera, S., Escalante, H.J., Lei, Z.: Surveillance face presentation attack detection challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6361–6371 (2023)
19. Galbally, J., McCool, C., Fierrez, J., Marcel, S., Ortega-Garcia, J.: On the vulnerability of face verification systems to hill-climbing attacks. Pattern Recognition **43**(3), 1027–1038 (2010)
20. Guo, J., Deng, J., An, X., Yu, J., Gecer, B.: Insightface: 2d and 3d face analysis project, https://github.com/deepinsight/insightface, accessed: 2026-03-03
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
23. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: 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)
24. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1501–1510 (2017)
25. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning. pp. 2137–2146. PMLR (2018)
26. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. PMLR (2015)
27. Jung, Y.G., Park, J., Dong, X., Park, H., Teoh, A.B.J., Camps, O.: Face reconstruction transfer attack as out-of-distribution generalization. In: European Conference on Computer Vision. pp. 396–413. Springer (2024)
28. Kairos AR Inc.: Kairos face recognition, https://face.kairos.com/, accessed: 2026-02-27
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
30. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
31. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
32. Kim, M.: Cvlface: High-performance face recognition all-in-one toolkit, https://github.com/mk-minchul/CVLface, accessed: 2026-03-03
33. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18750–18759 (2022)
34. Kim, M., Su, Y., Liu, F., Jain, A., Liu, X.: Keypoint relative position encoding for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 244–255 (2024)

35. Kim, S., Paik, S., Hwang, C., Kim, M., Seo, J.H.: Non-adaptive adversarial face generation. Advances in Neural Information Processing Systems **38**, 37315–37354 (2025)
36. Kim, S., Tan, Y.K., Jeong, B., Mondal, S., Aung, K.M.M., Seo, J.H.: Scores tell everything about bob: Non-adaptive face reconstruction on face recognition systems. In: 2024 IEEE Symposium on Security and Privacy. pp. 1684–1702. IEEE (2024)
37. Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., Zhao, B.Y.: Blacklight: Scalable defense for neural networks against query-based black-box attacks. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 2117–2134 (2022)
38. Liu, W., Wen, Y.: Opensphere, https://github.com/ydwen/opensphere, accessed: 2026-03-03
39. Liu, W., Wen, Y., Raj, B., Singh, R., Weller, A.: Sphereface revived: Unifying hyperspherical face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2458–2474 (2022)
40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
41. Mai, G., Cao, K., Yuen, P.C., Jain, A.K.: On the reconstruction of face images from deep face templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(5), 1188–1202 (2018)
42. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics. pp. 158–165. IEEE (2018)
43. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14234 (2021)
44. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
45. National Institute of Standards and Technology (NIST): Face recognition technology evaluation (frte) 1:1 verification, https://pages.nist.gov/frvt/html/frvt11.html, accessed: 2026-03-03
46. Otroshi Shahreza, H., Marcel, S.: Face reconstruction from facial templates by learning latent space of a generator network. Advances in Neural Information Processing Systems **36**, 12703–12720 (2023)
47. Paik, S., Hwang, C., Kim, S., Seo, J.H.: On the reversibility of locality-sensitive hashing-based biometric template protections. IEEE Transactions on Dependable and Secure Computing (2025)
48. Papantoniou, F.P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., Zafeiriou, S.: Arc2face: A foundation model for id-consistent human faces. In: European Conference on Computer Vision. pp. 241–261. Springer (2024)
49. Park, H., Park, J., Dong, X., Teoh, A.B.J.: Towards query efficient and generalizable black-box face reconstruction attack. In: 2023 IEEE International Conference on Image Processing. pp. 1060–1064. IEEE (2023)
50. Park, J., McLaughlin, N., Alouani, I.: Mind the gap: Detecting black-box adversarial attacks in the making through query update analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10235–10243 (2025)

51. Razzhigaev, A., Kireev, K., Kaziakhmedov, E., Tursynbek, N., Petiushko, A.: Black-box face recovery from identity features. In: European Conference on Computer Vision. pp. 462–475. Springer (2020)
52. Razzhigaev, A., Kireev, K., Udovichenko, I., Petiushko, A.: Darker than black-box: Face reconstruction from similarity queries. arXiv preprint arXiv:2106.14290 (2021)
53. Razzhigaev, A., Mikhalchuk, M., Kireev, K., Udovichenko, I., Kuznetsov, A., Petiushko, A.: Inverting black-box face recognition systems via zero-order optimization in eigenface space. arXiv preprint arXiv:2506.09777 (2025)
54. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–823 (2015)
55. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1–9. IEEE (2016)
56. Shahreza, H.O., George, A., Marcel, S.: Face reconstruction from face embeddings using adapter to a face foundation model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5584–5593 (2025)
57. Shahreza, H.O., Hahn, V.K., Marcel, S.: Face reconstruction from deep facial embeddings using a convolutional neural network. In: 2022 IEEE International Conference on Image Processing. pp. 1211–1215. IEEE (2022)
58. Shahreza, H.O., Hahn, V.K., Marcel, S.: Vulnerability of state-of-the-art face recognition models to template inversion attack. IEEE Transactions on Information Forensics and Security **19**, 4585–4600 (2024)
59. Shahreza, H.O., Marcel, S.: Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3d face reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 14248–14265 (2023)
60. Shahreza, H.O., Marcel, S.: Template inversion attack using synthetic face images against real face recognition systems. IEEE Transactions on Biometrics, Behavior, and Identity Science **6**(3), 374–384 (2024)
61. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 1528–1540 (2016)
62. Tencent: Tencent cloud face recognition, https://www.tencentcloud.com/products/facerecognition, accessed: 2026-02-27
63. Tencent: Tencent face recognition use-cases, https://www.tencentcloud.com/document/product/1059/37345, accessed: 2026-03-05
64. Transportation Security Administration: Facial comparison technology, https://www.tsa.gov/news/press/factsheets/facial-comparison-technology
65. Truong, T.D., Duong, C.N., Le, N., Savvides, M., Luu, K.: Vec2face-v2: Unveil human faces from their blackbox features via attention-based network in face recognition. arXiv preprint arXiv:2209.04920 (2022)
66. Vendrow, E., Vendrow, J.: Realistic face reconstruction from deep embeddings. In: NeurIPS 2021 Workshop Privacy in Machine Learning (2021)
67. Wang, C.Y., Lu, Y.D., Yang, S.T., Lai, S.H.: Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20281–20290 (2022)
68. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6678–6687 (2020)
69. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
70. Wu, H., Singh, J., Tian, S., Zheng, L., Bowyer, K.: Vec2face: Scaling face dataset generation with loosely constrained vectors. In: International Conference on Learning Representations. vol. 2025, pp. 23490–23503 (2025)
71. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
72. You, J., Li, S., Sun, Y., Wei, J., Guo, M., Feng, C., Ran, J.: Lvface: Progressive cluster optimization for large vision models in face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11840–11849 (2025)
73. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10492–10502 (2021)