GRANTFACE: WHO GETS TO RECOGNIZE YOUR FACE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial data submitted to a legitimate recognition service can be reused by unauthorized face recognition systems. Our intuition is that an image contains much more than a recognizer uses: variations it largely ignores may still matter to other recognizers. We turn this model-specific insensitivity into protection spaces that preserve authorized matching while disrupting unauthorized recognition. Our method, GRANTFACE, constructs these low-sensitivity spaces by aggregating the authorized recognizer's input-to-embedding sensitivities over synthetic probes. The protection space is constructed once and generalizes across identities and inputs without modifying the authorized recognizer or accessing unauthorized recognizers. We prove leading-order optimality for authorized release agreement on synthetic probes and give sufficient conditions for a selective protection interval. Across four benchmarks, GRANTFACE retains an average authorized TAR of 93.29% while reducing unauthorized TAR to 4.41%, with selectivity persisting for recognizers sharing the same backbone and after adaptive fine-tuning. By eliminating per-input optimization, GRANTFACE protects a new face in 0.26 ms.

1 INTRODUCTION

Face recognition (FR) is widely used for identity verification, access control, and other security-sensitive applications. Its growing adoption has also raised privacy concerns (Melzi et al., 2024). Users may provide facial data to a specific FR service for legitimate authentication, yet other recognition systems can later reuse the same data for purposes beyond those originally intended. A 2026 parliamentary briefing reported that UK police had used a passport database covering over 45 million people for FR searches, a use campaigners described as a "breach of the right to privacy" (Murphy & Bunn, 2026). Canadian privacy guidance likewise emphasizes limiting biometric use to the purposes for which the data were collected (Office of the Privacy Commissioner of Canada, 2025). Restricting access to the service or its model does not itself reduce the recognition utility of image copies obtained by third parties. Selective face protection acts on the images themselves, preserving the matching utility required by the service while reducing cross-model reuse.

Existing approaches protect facial data against unintended recognition in several ways. Recognition-suppression methods perturb facial images to prevent FR systems from reliably identifying them (Shan et al., 2020; Cherepanova et al., 2021), but this also sacrifices the recognition utility required by legitimate services. Privacy-preserving face recognition (PPFR) instead aims to preserve biometric utility while hiding sensitive facial information (Mi et al., 2022; 2023; 2024; Dai et al., 2025). These methods rely on dedicated protected representations, retrained recognition models, or modified recognition pipelines, tightly coupling the privacy mechanism to the underlying FR system. This coupling can limit compatibility with deployed recognition systems and hinder practical adoption (Sun & Liu, 2025). More recent methods attempt to preserve utility for a designated recognizer (Ding et al., 2022; Pu et al., 2025; Chow et al., 2024), but typically construct protection for individual samples or identities. They therefore incur per-sample optimization or per-identity setup as new inputs or users arrive. This raises a question: can the authorized recognizer's sensitivity structure define a protection space that remains selective across users and inputs?

To address this challenge, we propose GRANTFACE, which builds on a simple intuition: an image can vary in ways that matter little to a particular recognizer. Pruning studies illustrate redundancy within networks: many connections can be removed while retaining accuracy after fine-tuning (Han et al., 2015). We explore a complementary form of redundancy at the input: perturbations that barely change the authorized recognizer's identity embedding may affect other recognizers substan-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tially. While adversarial perturbations exploit highly sensitive directions to induce large embedding changes (Goodfellow et al., 2015), GRANTFACE constructs a protection space from the authorized recognizer's low-sensitivity directions. This model-specific sensitivity gap enables selective suppression of unauthorized recognition without accessing or querying unauthorized recognizers.

To construct this space, GRANTFACE measures how input perturbations affect the final identity embedding used for matching. Aggregating these sensitivities, measured by Jacobian Gram matrices, over identity-independent synthetic probes identifies a low-sensitivity protection space shared across inputs. Once constructed, this space is applied across faces without further optimization or modification to the authorized recognizer.

Experiments show that GRANTFACE retains an average authorized TAR of 93.29% while reducing unauthorized TAR to 4.41%. The same protection space generalizes to unseen identities and datasets, and selectivity persists for unauthorized recognizers sharing the same backbone and under fine-tuning and reconstruction attacks. Subspace ablations show that the protection space preserves authorized matching where alternatives with the same perturbation energy fail. Strength sweeps show that unauthorized recognition degrades before authorized recognition is substantially affected. This separation gives deployments flexibility to balance authorized utility and unauthorized suppression by tuning protection strength for each authorized recognizer. Protecting a new face requires only sampling a perturbation from the fixed space and adding it to the image, with negligible online overhead.

Our main contributions are:

- We introduce model-specific protection spaces that preserve authorized matching while disrupting recognition by other recognizers. GRANTFACE identifies a shared space through spectral analysis of sensitivities aggregated over synthetic probes.
- We establish a direct connection between model sensitivity and face verification, providing a theoretical foundation for selective protection. Our analysis proves optimal leading-order agreement on synthetic probes and explains model selectivity through a selective protection interval with explicit acceptance guarantees.
- Experiments across four benchmarks demonstrate a larger recognition gap than competing methods at comparable authorized utility. GRANTFACE generalizes across identities, retains selectivity under adaptive fine-tuning, and protects each face in 0.26 ms.

2 PROBLEM FORMULATION

2.1 SELECTIVE PROTECTION

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the image domain. A recognizer $F : \mathbb{R}^n \to \mathbb{S}^{d-1}$ maps an image to a unit-length feature vector, its *identity embedding*. Let F_A be the fixed authorized recognizer and $\mathcal{U}$ a specified family of unauthorized recognizers, with $F_U \in \mathcal{U}$; subscripts A and U refer to quantities of F_A and F_U. Face verification determines whether two images $x_1, x_2 \in \mathcal{X}$ depict the same person. A recognizer $F \in \{F_A\} \cup \mathcal{U}$ scores the pair by the cosine similarity $s_F(x_1, x_2) = \langle F(x_1), F(x_2) \rangle$ and accepts it as a match when $s_F \geq \tau_F$, with a fixed threshold $\tau_F \in (-1, 1)$ for each recognizer.

A shared randomized mechanism produces a protected face $x^p = \mathcal{P}_A(x; z)$ with random draw z. The mechanism is constructed once for F_A and applied across users and inputs without further optimization. Over the distribution $\mathcal{D}_+$ of same-person pairs, let $\bar{a}_F^0$ and $\bar{a}_F^p$ be the clean and protected acceptance probabilities at threshold τ_F; the latter also averages over independent draws for the two images. For $\rho \geq 0$ and $\gamma > 0$, we seek (ρ, γ)-selective protection satisfying

$$\bar{a}_A^p \geq \bar{a}_A^0 - \rho, \qquad \bar{a}_U^p \leq \bar{a}_A^p - \gamma \quad \text{for every } F_U \in \mathcal{U}. \tag{1}$$

Here ρ bounds authorized acceptance loss and γ sets the minimum protected recognition gap.

2.2 THREAT MODEL

The trusted service applies $\mathcal{P}_A$ internally, before F_A compares the protected images for user authentication. The protected images must therefore retain same-person matching utility. The service

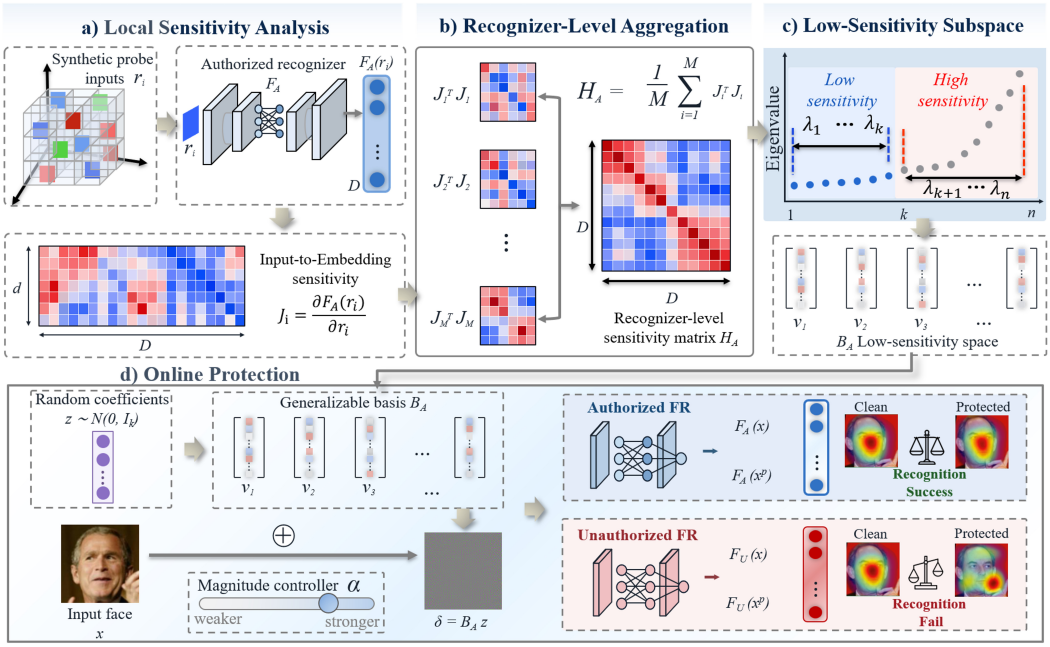


Figure 1: Overview of GRANTFACE. (a–c) Synthetic-probe sensitivities form H_A; its smallest-eigenvalue directions give B_A. (d) Each face is protected as $x^p = x + \alpha B_A z$, with random coefficients z and strength α.

has white-box access to F_A and constructs $\mathcal{P}_A$ using only this recognizer and identity-independent synthetic probes. The deployed recognizer and its verification pipeline remain fixed (Microsoft, 2026).

Protection is tied to the trained recognizer: anyone running F_A at the same threshold obtains the same decisions (Appendix A.1.5). GRANTFACE complements controls on access to this recognizer by reducing the utility of protected images to other recognizers.

Unauthorized parties obtain protected faces and may know their identity labels. They control their own recognizers and seek to reuse these images for identity matching outside the service's authentication workflow. Their recognizers may be developed after deployment. We consider two adversaries: *direct adversaries* apply a fixed F_U, including models with the same backbone but different training data or objectives; *adaptive adversaries* fine-tune F_U on protected training faces before recognizing new inputs.

The protection algorithm is public. The basis B_A of the protection space (Section 3.2) and per-release random draws remain internal to the trusted service.[1] We study individual protected releases, assuming the protection space has not been recovered. Direct and fine-tuning evaluations use images as released; Appendix A.3 additionally evaluates reconstruction preprocessing.

3 METHODOLOGY

To pursue the selective protection objective in Equation 1, GRANTFACE uses the authorized recognizer's sensitivity structure: how input perturbations change its identity embeddings. As shown in Figure 1, we identify a shared low-sensitivity protection space and sample perturbations within it. We prove leading-order optimality for release agreement on synthetic probes and give sufficient conditions for a selective interval of protection strengths.

[1]For the additive mechanism in Equation 5, projecting onto the complement of the protection space gives $(I - B_A B_A^\top)x^p = (I - B_A B_A^\top)x$. This removes the noise and the clean image component in that subspace.

3.1 RECOGNIZER SENSITIVITY

A network contains much more than it needs. An image contains much more than a recognizer uses. Pruning studies illustrate the first observation: substantial numbers of connections can be removed while retaining accuracy after fine-tuning (Han et al., 2015). Our intuition is to explore the second. A face recognizer acts as a high-dimensional filter, responding strongly to some input variations while largely ignoring others. These directions offer a natural space for perturbations with little effect on its identity embedding.

To understand how this filtering happens, we start with the first linear operator $W = W^{(1)}$ acting on the vectorized face. Its right singular vectors associated with the smallest singular values identify directions with weak first-layer responses without backpropagation. However, the first layer tells only part of the story: a perturbation that produces a noticeable response here may still have little effect on the final embedding, because subsequent layers can suppress that response. A first-layer criterion can therefore overlook directions with low sensitivity at the output (Proposition 1, Appendix A.1). We therefore measure sensitivity through the full input-to-embedding (end-to-end) map (Figure 1(a)). To quantify it, we linearize the recognizer: for an input perturbation $\delta \in \mathbb{R}^n$, Lipschitz continuity of each Jacobian $J_F(x) = \partial F(x)/\partial x \in \mathbb{R}^{d\times n}$, with constant L_F, gives

$$F_A(x+\delta) = F_A(x) + J_A(x)\delta + q_A, \qquad \|q_A\|_2 \le \frac{L_A}{2}\|\delta\|_2^2 \tag{2}$$

(Appendix A.1.1), so up to the remainder q_A, the squared response $\|J_A(x)\delta\|_2^2$ captures the sensitivity of F_A along δ.

For three of the four recognizers, the first-layer basis is 42–$253\times$ more sensitive at the embedding than the protection space constructed below (Table 3).

3.2 CONSTRUCTING THE LOW-SENSITIVITY SPACE

A recognizer-level protection space requires sensitivity structure that persists across inputs. Accordingly, we aggregate local sensitivities over identity-independent synthetic probes $r_1, \dots, r_M \in \mathcal{X}$, random noise images in our experiments (Figure 1(b)), keeping the construction independent of protected users and evaluation faces. With $J_i = J_A(r_i)$, we define the mean sensitivity matrix as

$$H_A = \frac{1}{M}\sum_{i=1}^{M} J_i^\top J_i \in \mathbb{R}^{n\times n}. \tag{3}$$

To compare candidate subspaces, consider a k-dimensional subspace with orthonormal basis $B \in \mathbb{R}^{n\times k}$ ($B^\top B = I_k$). Sampling $z \sim \mathcal{N}(0, I_k)$ at strength $\alpha > 0$ gives the mean squared linear response

$$\frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_z\|J_i(\alpha B z)\|_2^2 = \alpha^2 \operatorname{Tr}(B^\top H_A B), \tag{4}$$

since $\mathbb{E}[zz^\top] = I_k$. All such subspaces carry the same energy $\mathbb{E}_z\|\alpha B z\|_2^2 = \alpha^2 k$, so this response ranks them by sensitivity alone. With eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ of H_A and orthonormal eigenvectors $v_1, \dots, v_n$, the optimal subspace is $B_A = [v_1, \dots, v_k]$, the k directions along which probe embeddings move least, attaining $\alpha^2 \sum_{j=1}^{k} \lambda_j$ (Figure 1(c); Appendices A.1.2 and A.1.6). Table 3 confirms its advantage at equal energy.

3.3 PROTECTING SHARED FACES

As illustrated in Figure 1(d), the resulting basis B_A is computed once and shared across users and inputs. Each new face is protected by

$$z \sim \mathcal{N}(0, I_k), \qquad \delta = \alpha B_A z, \qquad x^p = \mathcal{P}_A(x; z) = x + \delta. \tag{5}$$

All Jacobian computation happens offline; online protection is a Gaussian draw, one matrix–vector product, and an addition. The strength α is the only deployment parameter and can be adjusted without rebuilding B_A or modifying the recognizer (Section 4.4).

3.4 VERIFICATION ANALYSIS

To connect embedding sensitivity to verification decisions, we first analyze agreement between protected probe releases. We then derive a provable selective interval of protection strengths that preserves authorized matching while suppressing unauthorized matching.

When two independently protected copies of a probe are verified against each other, the expected score loss is the release disagreement: for independent $z, z' \sim \mathcal{N}(0, I_k)$,

$$D_A(B, \alpha) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{z,z'} \big[1 - \langle F_A(r_i + \alpha B z), F_A(r_i + \alpha B z') \rangle\big]. \tag{6}$$

Theorem 1 (Leading-order release disagreement) *For any B with $B^\top B = I_k$, $D_A(B, \alpha) = \alpha^2 \operatorname{Tr}(B^\top H_A B) + O(\alpha^3)$ as $\alpha \to 0$, and B_A minimizes the leading coefficient, with minimum $\sum_{j=1}^{k} \lambda_j$.*

Thus, to leading order, no subspace preserves release agreement better than B_A (Appendix A.1.2). Verification, however, compares two different faces, so acceptance also depends on their clean separation and the matching threshold.

Pairwise verification. For a face pair, protection moves the difference between the two embeddings, and the pair remains a match as long as this difference stays within a matching radius. Formally, fix a face pair (x_1, x_2) and $F \in \{F_A, F_U\}$, and stack the independent draws as $Z = (z_1^\top, z_2^\top)^\top \sim \mathcal{N}(0, I_{2k})$. Since embeddings are unit vectors, $s_F \geq \tau_F$ exactly when the embedding distance is at most the matching radius $h_F = \sqrt{2(1 - \tau_F)}$. With the clean separation and linear response

$$d_F = F(x_1) - F(x_2), \qquad C_F = [J_F(x_1) B_A, \, -J_F(x_2) B_A], \tag{7}$$

the protected pair is accepted exactly when $\|d_F + \alpha C_F Z + e_F\|_2 \leq h_F$, where e_F is the difference of the two remainders in Equation 2; a_F^p denotes the probability of this event over Z. This distance form stays accurate even when the two clean embeddings coincide (Appendix A.1.4).

To identify a selective interval, we first set the remainder e_F aside. As α grows, each draw Z moves the embedding difference along the ray $\alpha \mapsto d_F + \alpha C_F Z$, which starts inside the ball of radius $h > \|d_F\|_2$ and leaves it once, at a crossing strength $\kappa_F(Z; h) \in (0, \infty]$. A recognizer that responds strongly to the protection leaves the ball early, whereas an authorized pair with a small response C_A leaves it late. Quantiles of κ_F over draws, written $Q_p[\cdot]$, turn this picture into strength thresholds, and the following theorem adjusts the radii to account for the remainder.

Theorem 2 (Selective protection interval) *Fix a strength cap $\bar{\alpha} > 0$ and $t > 0$, and set $\epsilon_F = L_F \bar{\alpha}^2 (\sqrt{2k} + t)^2/2$ and $\eta = e^{-t^2/2}$. For a pair with $\|d_A\|_2 < h_A - \epsilon_A$ and $\|d_U\|_2 < h_U$ and for $\beta \in (0, 1)$, let $\alpha_A = Q_\beta[\kappa_A(Z; h_A - \epsilon_A)]$ and $\alpha_U = Q_{1-\beta}[\kappa_U(Z; h_U + \epsilon_U)]$. Every $\alpha \in [\alpha_U, \min\{\alpha_A, \bar{\alpha}\}]$ satisfies $a_A^p \geq 1 - \beta - \eta$ and $a_U^p \leq \beta + \eta$.*

Here ϵ_F shrinks or enlarges the matching radius to absorb the remainder, a bound that holds with probability at least $1 - \eta$, and β sets the confidence of each threshold. A selective interval thus exists whenever $\alpha_U \leq \min\{\alpha_A, \bar{\alpha}\}$, and every strength in it provably preserves authorized matching while suppressing unauthorized matching (Appendix A.1.3). Figure 2 illustrates this separation empirically for each authorized recognizer, using sampled strengths and mean unauthorized TAR.

To make the role of the spectral construction explicit, Markov's inequality also bounds authorized acceptance directly by sensitivity: with margin $m_A = h_A - \epsilon_A - \|d_A\|_2 > 0$, Frobenius norm $\|\cdot\|_F$, and $0 < \alpha \leq \bar{\alpha}$,

$$a_A^p \geq 1 - \frac{\alpha^2 \|C_A\|_F^2}{m_A^2} - \eta, \qquad \|C_A\|_F^2 = \sum_{j=1}^{2} \|J_A(x_j) B_A\|_F^2. \tag{8}$$

Authorized acceptance thus stays high whenever the pair's sensitivity $\|C_A\|_F^2$ is small relative to its margin; averaged over probe pairs, this sensitivity is $2 \operatorname{Tr}(B_A^\top H_A B_A) = 2 \sum_{i=1}^{k} \lambda_i$, the quantity B_A minimizes. On the unauthorized side, selectivity requires F_U to respond to the protection

differently from F_A: Appendix A.1.5 bounds the acceptance gap by the difference between their clean separations and responses, each rescaled by the matching radius. Even recognizers sharing the backbone of F_A differ enough to be suppressed (Section 4.4).

Finally, averaging over same-person pairs gives $\bar{a}_F^p = \mathbb{E}_{\mathcal{D}_+}[a_F^p]$. If one strength lies in the interval of Theorem 2 for almost every pair in $\mathcal{D}_+$ and every $F_U \in \mathcal{U}$, with common β and t, GRANTFACE therefore achieves Equation 1 with $(\rho, \gamma) = (\beta + \eta,\ 1 - 2(\beta + \eta))$ for $\beta + \eta < 1/2$. Table 1 yields its empirical counterpart, $(\hat{\rho}, \hat{\gamma}) = (0.1064, 0.5420)$.

We next evaluate selective face protection empirically across datasets and recognizers.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. To assess whether protection generalizes across datasets, we evaluate it on four widely used face recognition benchmarks: LFW (Huang et al., 2008), AgeDB-30 (Moschoglou et al., 2017), CALFW (Zheng et al., 2017), and CFP-FP (Sengupta et al., 2016). LFW evaluates face recognition under unconstrained real-world variations, while AgeDB-30 and CALFW introduce substantial age variations between matched facial images. CFP-FP evaluates robustness to large pose changes through frontal-to-profile face matching. We follow the official verification protocols of each dataset, consisting of 6,000 pairs for LFW, AgeDB-30, and CALFW, and 7,000 pairs for CFP-FP. None of these facial images or identities is used to construct the recognizer-level protection space; all protection spaces are derived solely from identity-independent synthetic probes and are fixed before evaluation.

Evaluation metrics. We report verification accuracy (ACC) and true accept rate (TAR) at thresholds calibrated to a false accept rate (FAR) of 1%. In Table 1, each recognizer uses the same threshold for clean and protected inputs; FAR=1% denotes the calibration point. Clean and protected TAR estimate $\bar{a}_F^0$ and $\bar{a}_F^p$, respectively. Authorized TAR loss and the protected recognition gap correspond to ρ and γ in Equation 1.

Implementation details. All face images are aligned and resized to 112×112 pixels. The recognizers in the main comparison are pretrained with the ArcFace objective on the MS1M dataset. Each protection space is built from $M = 128$ random-noise probes with dimension $k = 512$. All experiments are implemented in PyTorch 2.6 with CUDA 11.8 and conducted on NVIDIA RTX A6000 GPUs.

4.2 SELECTIVE PROTECTION EFFECTIVENESS

Table 1 evaluates direct adversaries from Section 2.2. Authorized ACC averages 94.47%, compared with 97.81% on clean inputs, while unauthorized ACC averages 56.19%. Authorized TAR remains at 93.29% versus 96.52% on clean inputs, whereas unauthorized TAR falls to 4.41%. Across all settings in Table 1, the largest authorized TAR loss is 10.64 percentage points and the smallest protected recognition gap is 54.20 points. These give $(\hat{\rho}, \hat{\gamma}) = (0.1064, 0.5420)$ as the empirical counterpart of Equation 1.

To assess whether selectivity depends on the operating threshold, we also report verification AUC across all four benchmarks (Appendix A.6). The results show that protection weakens unauthorized pair discrimination while preserving authorized discrimination.

Against Targeted Anonymization (Ding et al., 2022) and PMT/FDM (Pu et al., 2025), GRANTFACE achieves a substantially larger recognition gap at comparable authorized performance and cuts online latency from over one second to 0.26 ms per image (Table 8, Appendix A.5).

Table 1: Protection at $\alpha = 1.0$. We report ACC / TAR@FAR=1% (%). Bold marks the authorized recognizer; gray marks clean inputs. Darker cells show higher ACC on a common scale.

Dataset	Authorized	IRSE-100	DenseNet-201	SwinMLP-B	TransFace ViT-B
LFW	Clean	99.15 / 98.80	99.38 / 99.27	98.85 / 98.50	98.78 / 98.27
	IRSE-100	**96.58 / 95.80**	66.30 / 18.70	62.50 / 1.00	62.02 / 0.37
	DenseNet-201	53.72 / 2.00	**98.87 / 99.00**	58.85 / 1.43	62.72 / 0.77
	SwinMLP-B	56.35 / 0.67	54.73 / 0.43	**96.48 / 95.70**	62.85 / 1.20
	TransFace ViT-B	53.28 / 0.40	56.08 / 0.13	54.28 / 0.40	**93.55 / 87.63**
AgeDB-30	Clean	97.45 / 96.13	97.98 / 97.07	98.13 / 97.10	98.08 / 96.93
	IRSE-100	**93.13 / 87.80**	63.72 / 33.60	57.22 / 1.27	52.62 / 0.37
	DenseNet-201	57.55 / 3.73	**97.10 / 96.47**	59.97 / 13.33	57.45 / 1.20
	SwinMLP-B	52.45 / 0.00	54.30 / 1.33	**96.05 / 94.67**	59.55 / 2.37
	TransFace ViT-B	50.07 / 0.03	51.12 / 0.10	51.98 / 0.00	**93.28 / 90.67**
CALFW	Clean	95.58 / 91.97	95.67 / 92.17	95.77 / 92.50	95.72 / 92.47
	IRSE-100	**92.65 / 87.63**	63.85 / 23.20	54.90 / 0.40	53.72 / 0.03
	DenseNet-201	53.85 / 3.43	**94.92 / 91.30**	56.87 / 8.90	56.10 / 0.67
	SwinMLP-B	51.23 / 0.00	51.65 / 0.17	**94.27 / 91.07**	56.43 / 1.80
	TransFace ViT-B	48.57 / 0.00	49.15 / 0.00	50.80 / 0.00	**92.30 / 88.60**
CFP-FP	Clean	98.40 / 97.91	98.59 / 98.29	99.03 / 98.91	98.43 / 97.97
	IRSE-100	**90.37 / 95.00**	62.97 / 38.97	57.21 / 2.89	61.34 / 1.71
	DenseNet-201	51.13 / 7.23	**97.46 / 97.80**	58.09 / 16.49	59.27 / 6.00
	SwinMLP-B	55.34 / 0.94	51.40 / 3.49	**95.40 / 97.94**	61.06 / 8.77
	TransFace ViT-B	51.69 / 0.63	56.81 / 0.86	52.19 / 0.29	**89.07 / 95.57**

Table 3: Subspace selection at equal perturbation energy ($\alpha = 1.0$). Top: authorized ACC / TAR@FAR=1% (%). Bottom: end-to-end sensitivity $10^7 \operatorname{Tr}(B^\top H_A B)$ of the first-layer basis, i.e., the right singular vectors of $W^{(1)}$ with the k smallest singular values (ratio to B_A in parentheses), and of B_A; Table 5 in the appendix gives additional spectral statistics.

Method	IRSE-100	DenseNet-201	SwinMLP-B	TransFace	Mean
Gaussian Noise	52.17 / 0.50	56.93 / 0.10	53.63 / 0.27	51.53 / 0.30	53.57 / 0.29
Random Subspace	53.04 / 0.50	56.04 / 0.09	53.07 / 0.26	51.15 / 0.26	53.33 / 0.28
High-Sensitivity	50.43 / 0.17	52.45 / 0.30	52.73 / 0.23	50.70 / 0.23	51.58 / 0.23
Low-Sensitivity (GRANTFACE)	**96.58 / 95.80**	**98.87 / 99.00**	**96.48 / 95.70**	**93.55 / 87.63**	**96.37 / 94.53**
First-Layer ($W^{(1)}$)	254 (49×)	256 (42×)	1010 (253×)	0	–
Low-Sensitivity (GRANTFACE)	**5.19**	**6.16**	**3.99**	**0**	–

4.3 EFFICIENCY AND SCALABILITY

The one-time construction of the protection space takes 450.23 s on average (Table 2), after which protection requires only 0.261 ms per image, corresponding to a throughput of approximately 3,829 images per second. Since the same protection space is shared across all subsequent users and inputs, this offline cost is amortized over deployment. Unlike per-image or per-identity protection methods, protecting a new user requires neither additional optimization nor user-specific enrollment images, so the per-image online cost remains constant as the number of protected inputs grows.

Table 2: Protection efficiency.

Recognizer	Offline (s)	Latency (ms/image)	Throughput (images/s)
IRSE-100	384.21	0.260	3841
DenseNet-201	534.98	0.259	3861
SwinMLP-B	496.16	0.259	3858
TransFace ViT-B	385.55	0.266	3759
Mean	**450.23**	**0.261**	**3829**

4.4 ABLATION AND ANALYSIS

Subspace selection. We compare B_A with isotropic Gaussian noise, a random k-dimensional subspace, and the high-sensitivity subspace spanned by the k leading eigenvectors of H_A, all with the same perturbation energy. Gaussian noise and random subspaces reduce mean authorized ACC to

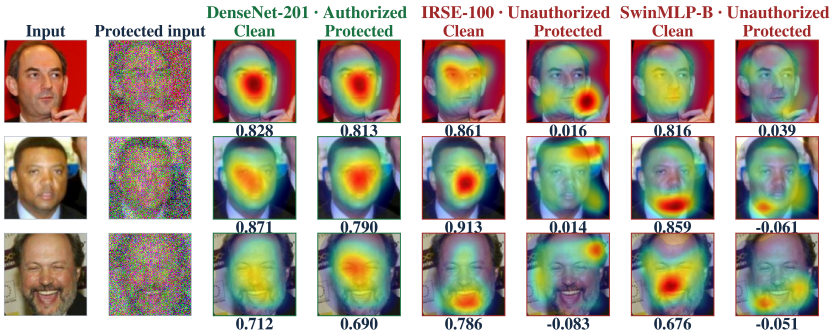


Figure 3: Attention maps and cosine similarities before and after protection at $\alpha = 1.0$. Each map shows how strongly each spatial position of the feature map affects the cosine similarity.

53.57% and 53.33%, with TAR approaching zero (Table 3). High-sensitivity directions yield 51.58% ACC and 0.23% TAR, whereas the protection space retains 96.37% ACC and 94.53% TAR. The protection space preserves authorized matching where these equal-energy alternatives fail, consistent with Equation 8 when low probe sensitivity transfers to evaluation faces. For the first-layer basis of Section 3.1, Proposition 1 predicts excess sensitivity, and the bottom of Table 3 confirms it for three of the four recognizers: this basis is 42–253× more sensitive at the embedding than B_A, even though their $W^{(1)}$ has no null space (Table 5, Appendix A.1).

Flexible protection. Figure 2 shows a clear separation between authorized and mean unauthorized TAR from about $\alpha = 0.7$ onward within the plotted range. This separation persists across multiple sampled strengths, giving deployments flexibility to balance authorized utility and unauthorized suppression by adjusting α. The location and width of the useful range depend on the authorized recognizer, consistent with the mechanism in Theorem 2. Appendix A.2 gives the full $\alpha \in [0, 2.5]$ sweep and sampled ranges under explicit TAR criteria.

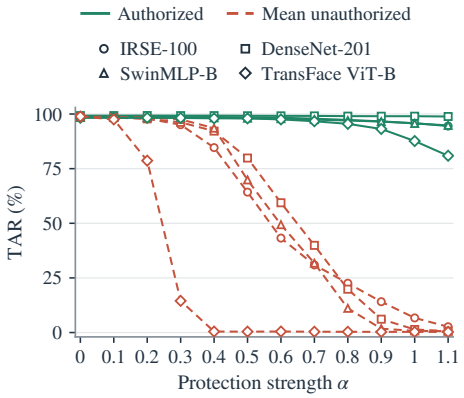


Figure 2: Authorized and mean unauthorized TAR across protection strengths.

Qualitative analysis. In Figure 3, the authorized DenseNet-201 keeps its attention and score after protection, while unauthorized recognizers shift attention off the face and their scores fall to near zero or below. Appendix A.7 traces this contrast to the direction in which protection moves each embedding.

Deployment choices. The same protection space supports different strength settings, so tuning α requires no additional offline construction. The protected inputs in Figure 3 are hard to read by design: they are intermediate inputs consumed by F_A (Section 2.2). Larger α can strengthen suppression, while smaller α reduces perturbation magnitude when human-viewable releases are desired. Sufficiently large perturbations also degrade authorized matching (Appendix A.2), so the strength should be chosen for the authorized recognizer and the deployment's needs.

Models with the same backbone. A direct adversary may share the backbone of F_A while using different training data or a different recognition objective. Unauthorized TAR still degrades earlier as protection strength increases, while authorized recognition remains effective over a wider range (Appendix A.4). Training alone thus makes these recognizers respond to the protection differently enough from F_A to be suppressed (Appendix A.1.5).

Table 4: Unauthorized recognition on LFW before and after fine-tuning.

Attacker	Clean	Protected Zero-shot	Protected Adaptive
MobileFaceNet	97.71 / 96.67	57.88 / 3.96	62.82 / 6.65
iResNet100	99.44 / 99.58	54.90 / 2.17	50.61 / 0.81
SwinMLP-B	99.05 / 98.95	62.22 / 7.91	50.30 / 1.16
Mean	98.73 / 98.40	58.33 / 4.68	54.58 / 2.87

Adaptive attacks. We instantiate the adaptive adversary from Section 2.2 with MobileFaceNet (Chen et al., 2018), iResNet100 (Deng et al., 2022), and SwinMLP-B. Each attacker fine-tunes its own recognizer on protected samples before evaluation. Unauthorized recognition remains low after fine-tuning (Table 4). For iResNet100 and SwinMLP-B, TAR decreases from 2.17% to 0.81% and from 7.91% to 1.16%, respectively. The mean TAR decreases from 4.68% under zero-shot evaluation to 2.87% after adaptation, remaining far below the 98.40% clean performance.

Robustness against reconstruction attacks. As shown in Appendix A.3, representative reconstruction attacks fail to recover unauthorized recognition utility, with GFPGAN (Wang et al., 2021) and DiffBIR (Lin et al., 2024) yielding unauthorized TARs of 6.78% and 0.62%, respectively.

5 Related Work

5.1 Facial Privacy Protection

Prior work on facial privacy broadly follows two directions. Recognition-suppression methods, such as Fawkes and OPOM, perturb facial images to reduce their utility to face recognition systems, but do not preserve recognition for a designated legitimate service (Shan et al., 2020; Zhong & Deng, 2023). Privacy-preserving face recognition (PPFR) instead retains identity utility while concealing sensitive facial information through frequency-domain representations, feature transformations, obfuscation, or protected templates (Mi et al., 2023; 2024; Yuan et al., 2024; Jin et al., 2024; Zhong et al., 2025; Dai et al., 2025). These methods achieve better privacy–utility trade-offs, but typically rely on dedicated representations, adapted recognition models, transformed templates, or modified inference pipelines. Their primary objective is to protect facial information while retaining recognition utility, rather than to control which existing recognizer can use the protected input.

5.2 Selective Face Recognition Protection

Recent methods seek to preserve recognition for a designated recognizer while limiting recognition by others. Targeted Anonymization generates adversarial faces that remain matchable by an authorized recognizer while transferring recognition errors to other models (Ding et al., 2022). Chameleon learns a personalized protection mask for each identity and enables authorized recognition through a de-obfuscation mechanism (Chow et al., 2024). More recently, Facial Data Minimization optimizes obfuscated inputs to preserve the representation of an authorized service (Pu et al., 2025), while ErasableMask restores legitimate recognition by removing the protective perturbation when authorization is required (Shen et al., 2026).

These methods achieve selective protection through sample- or identity-specific construction, or through an additional recovery procedure. GrantFace constructs one protection space for the authorized recognizer and applies it across users and inputs. This space is obtained by aggregating Jacobian Gram matrices of normalized identity embeddings over synthetic probes. The authorized recognizer matches protected faces, without modification. Our analysis relates this construction to acceptance probabilities for independently protected face pairs at fixed verification thresholds.

6 Conclusion

We present GrantFace, a recognizer-level approach to selective face protection that exploits the end-to-end sensitivity structure of the authorized recognizer. GrantFace constructs a generalizable low-sensitivity protection space that preserves authorized recognition while suppressing unauthorized recognition, without modifying the deployed recognizer or requiring access to unauthorized recognizers. Our analysis proves leading-order optimality for release agreement on synthetic probes and gives conditions for a selective protection interval. Experiments demonstrate cross-model separation, robustness to adaptive attackers, and low online overhead. These results highlight recognizer-specific sensitivity as a basis for scalable selective protection: a single space serves new users and inputs without further optimization. Future work will incorporate perceptual constraints and extend the formulation beyond additive perturbations.

AI USE STATEMENT

Generative AI tools were used for language polishing and literature retrieval to improve writing quality and related-work coverage. All retrieved references were independently verified. The intellectual and technical contributions are solely those of the authors.

ETHICS STATEMENT

This work investigates selective protection of facial data against unauthorized recognition. All experiments are conducted on established public face recognition benchmarks, and no new biometric data are collected.

REPRODUCIBILITY STATEMENT

The proposed method, theoretical analysis, experimental protocol, and implementation details are described in the main paper and appendix. All datasets and models used are publicly available. Code and a demonstration will also be released to support reproducibility and will be available at: https://anonymous.4open.science/r/PAIRS-7EEC.

REFERENCES

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition*, pp. 428–438, Cham, 2018. Springer International Publishing. ISBN 978-3-319-97909-0.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P. Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=hJmtwocEqzc.

Ka-Ho Chow, Sihao Hu, Tiansheng Huang, and Ling Liu. Personalized privacy protection mask against unauthorized facial recognition. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXII*, pp. 434–450, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73006-1. doi: 10.1007/978-3-031-73007-8_25. URL https://doi.org/10.1007/978-3-031-73007-8_25.

Wanying Dai, Beibei Li, Naipeng Dong, Guangdong Bai, and Jin Song Dong. Fracface: Breaking the visual clues - fractal-based privacy-preserving face recognition. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/02272b67360899b10238109072571ca5-Abstract-Conference.html.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3087709. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3087709.

Kangyi Ding, Teng Hu, Xiaolei Liu, Weina Niu, Yanping Wang, and Xiaosong Zhang. Targeted anonymization: A face image anonymization method for unauthorized models. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022. doi: 10.1109/ICME52920.2022.9859898.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning*

Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1412.6572.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, pp. 1135–1143. MIT Press, 2015.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, October 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie. URL https://inria.hal.science/inria-00321923.

Shuaifan Jin, He Wang, Zhibo Wang, Feng Xiao, Jiahui Hu, Yuan He, Wenwen Zhang, Zhongjie Ba, Weijie Fang, Shuhong Yuan, and Kui Ren. FaceObfuscator: Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 6849–6866, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIX*, pp. 430–448, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73201-0. doi: 10.1007/978-3-031-73202-7_25. URL https://doi.org/10.1007/978-3-031-73202-7_25.

Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. An overview of privacy-enhancing technologies in biometric recognition. *ACM Computing Surveys*, 56(12), October 2024. ISSN 0360-0300. doi: 10.1145/3664596. URL https://doi.org/10.1145/3664596.

Yuxi Mi, Yuge Huang, Jiazhen Ji, Hongquan Liu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (eds.), *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 6755–6764. ACM, 2022. doi: 10.1145/3503161.3548303. URL https://doi.org/10.1145/3503161.3548303.

Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using random frequency components. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19616–19627, 2023. doi: 10.1109/ICCV51070.2023.01802.

Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 297–307. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00036. URL https://doi.org/10.1109/CVPR52733.2024.00036.

Microsoft. Specify a face recognition model. https://learn.microsoft.com/en-us/azure/ai-services/face/how-to/specify-recognition-model, 2026. Microsoft Learn, last updated March 19, 2026, accessed September 2, 2026.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1997–2005. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.250. URL https://doi.org/10.1109/CVPRW.2017.250.

Joe Murphy and Sarah Bunn. Facial recognition technology in policing. Rapid response, Parliamentary Office of Science and Technology, UK Parliament, April 2026. URL https://doi.org/10.58248/RR94.

Office of the Privacy Commissioner of Canada. Guidance for processing biometrics – for businesses. Office of the Privacy Commissioner of Canada, 2025. URL https://www.priv.gc.ca/en/privacy-topics/health-information-genetics-biometrics/biometrics/gd_bio_org-final/. Last modified 11 August 2025; accessed 25 August 2026.

Yuwen Pu, Jiahao Chen, Jiayu Pan, Diqun Yan, Xuhong Zhang, and Shouling Ji. Facial data minimization: Shallow model as your privacy filter. *IEEE Transactions on Dependable and Secure Computing*, 22(5):5186–5202, 2025. doi: 10.1109/TDSC.2025.3563477.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pp. 1–9. IEEE Computer Society, 2016. doi: 10.1109/WACV.2016.7477558. URL https://doi.org/10.1109/WACV.2016.7477558.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. ISBN 978-1-939133-17-5.

Sipeng Shen, Yunming Zhang, Dengpan Ye, Xiuwen Shi, Long Tang, Haoran Duan, Yueyun Shang, and Zhihong Tian. Erasablemask: A robust and erasable privacy protection scheme against black-box face recognition models. *IEEE Trans. Multim.*, 28:1342–1351, 2026. doi: 10.1109/TMM.2025.3639912. URL https://doi.org/10.1109/TMM.2025.3639912.

Zhifang Sun and Zhe Liu. Ensuring privacy in face recognition: A survey on data generation, inference and storage. *Discover Applied Sciences*, 7, 2025. doi: 10.1007/s42452-025-06987-2.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9164–9174, 2021. doi: 10.1109/CVPR46437.2021.00905.

Lin Yuan, Wu Chen, Xiao Pu, Yan Zhang, Hongbo Li, Yushu Zhang, Xinbo Gao, and Touradj Ebrahimi. Pro-face c: Privacy-preserving recognition of obfuscated face via feature compensation. *Trans. Info. For. Sec.*, 19:4930–4944, January 2024. ISSN 1556-6013. doi: 10.1109/TIFS.2024.3388976. URL https://doi.org/10.1109/TIFS.2024.3388976.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. URL http://arxiv.org/abs/1708.08197.

Yaoyao Zhong and Weihong Deng. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3590–3603, 2023. doi: 10.1109/TPAMI.2022.3175602.

Zhizhou Zhong, Yuxi Mi, Yuge Huang, Jianqing Xu, Guodong Mu, Shouhong Ding, Jingyun Zhang, Rizen Guo, Yunsheng Wu, and Shuigeng Zhou. Slerpface: face template protection via spherical linear interpolation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i10.33162. URL https://doi.org/10.1609/aaai.v39i10.33162.

A Appendix

A.1 Proofs

A.1.1 First-Order Expansion with Remainder Bound

For any recognizer F, the fundamental theorem of calculus gives

$$F(x+\delta) - F(x) - J_F(x)\delta = \int_0^1 \big(J_F(x+t\delta) - J_F(x)\big)\delta \, dt,$$

whose norm is at most $\int_0^1 L_F t\|\delta\|_2^2 \, dt = L_F\|\delta\|_2^2/2$. Taking $F = F_A$ gives Equation 2. The remainder bounds assume this Lipschitz condition holds along the perturbation segments. Regionwise smoothness of a piecewise-affine backbone alone does not ensure this condition for the normalized recognizer.

A.1.2 Proof of Theorem 1

Equation 4 follows from $\mathbb{E}_z\|J_i(\alpha Bz)\|_2^2 = \alpha^2 \operatorname{Tr}(B^\top J_i^\top J_i B\, \mathbb{E}[zz^\top]) = \alpha^2 \operatorname{Tr}(B^\top J_i^\top J_i B)$.

Expansion. Fix a probe r_i, let $z, z' \sim \mathcal{N}(0, I_k)$ be independent, and set $W = (z, z')$. For unit vectors, $1 - \langle u, v\rangle = \frac{1}{2}\|u - v\|_2^2$, and Equation 2 gives $F_A(r_i + \alpha Bz) - F_A(r_i + \alpha Bz') = \alpha J_i B(z - z') + e_i$ with $\|e_i\|_2 \le L_A\alpha^2\|W\|_2^2/2$. Expanding the square,

$$\Big|1 - \langle F_A(r_i + \alpha Bz), F_A(r_i + \alpha Bz')\rangle - \tfrac{\alpha^2}{2}\|J_i B(z - z')\|_2^2\Big| \le K_i\big(\alpha^3\|W\|_2^3 + \alpha^4\|W\|_2^4\big)$$

for a constant K_i independent of α. Since Gaussian moments are finite and $z - z' \sim \mathcal{N}(0, 2I_k)$, the expected disagreement at r_i is $\alpha^2\|J_iB\|_F^2 + O(\alpha^3)$. Averaging over probes gives $D_A(B, \alpha) = \alpha^2 \operatorname{Tr}(B^\top H_A B) + O(\alpha^3)$, whose leading term is the mean squared linear response in Equation 4.

Minimization. Write $H_A = V \operatorname{diag}(\lambda_1, \dots, \lambda_n)V^\top$ and let w_i be the squared norm of row i of $V^\top B$. The diagonal of the rank-k projection $V^\top BB^\top V$ gives $0 \le w_i \le 1$ and $\sum_i w_i = k$, so

$$\operatorname{Tr}(B^\top H_A B) - \sum_{i\le k}\lambda_i = \sum_{i>k}\lambda_i w_i - \sum_{i\le k}\lambda_i(1 - w_i) \ge \lambda_k\Big(\sum_{i>k} w_i - \sum_{i\le k}(1 - w_i)\Big) = 0,$$

with equality at B_A. If $\lambda_k < \lambda_{k+1}$, equality forces $w_i = 1$ for $i \le k$, so the minimizing subspace is unique; a tie $\lambda_k = \lambda_{k+1}$ admits distinct minimizers. □

A.1.3 Proof of Theorem 2

Remainders. Let $E = \{\|Z\|_2 \le \sqrt{2k} + t\}$, so $\Pr(E^c) \le \eta$ by Gaussian concentration (Hsu et al., 2012). With $\delta_j = \alpha B_A z_j$, Equation 2 writes the protected embedding difference as $d_F + \alpha C_F Z + e_F$ with $\|e_F\|_2 \le L_F(\|\delta_1\|_2^2 + \|\delta_2\|_2^2)/2 = L_F\alpha^2\|Z\|_2^2/2$, which is at most ϵ_F on E whenever $\alpha \le \bar{\alpha}$.

Acceptance bounds. Acceptance means $\|d_F + \alpha C_F Z + e_F\|_2 \le h_F$. With $\Psi_F(a) = \Pr(\|d_F + \alpha C_F Z\|_2 \le a)$, the triangle inequality on E gives, for any pair,

$$\Psi_F(h_F - \epsilon_F) - \eta \le a_F^p \le \Psi_F(h_F + \epsilon_F) + \eta. \tag{9}$$

Crossing strength. For $h > \|d_F\|_2$ and $w = C_F Z \ne 0$, the quadratic $\|d_F + aw\|_2^2 - h^2$ in a has a negative constant term, so its roots have opposite signs and, for $\alpha \ge 0$, $\|d_F + \alpha w\|_2 \le h$ exactly when $\alpha \le \kappa_F(Z; h)$, the positive root

$$\kappa_F(Z; h) = \frac{\sqrt{\langle d_F, w\rangle^2 + (h^2 - \|d_F\|_2^2)\|w\|_2^2} - \langle d_F, w\rangle}{\|w\|_2^2}; \tag{10}$$

if $w = 0$, $\kappa_F = \infty$. Hence $\Psi_F(h) = \Pr(\kappa_F(Z; h) \ge \alpha)$.

Quantiles. For $X \in (0, \infty]$, $Q_p[X] = \inf\{a \ge 0 : \Pr(X \le a) \ge p\}$ with $\inf\emptyset = \infty$, so $\Pr(X < \alpha) \le p$ whenever $\alpha \le Q_p[X]$. For $\alpha \le \alpha_A$, this gives $\Psi_A(h_A - \epsilon_A) \ge 1 - \beta$. For each finite a, the set $\{Z : \|d_U + aC_U Z\|_2 = h_U + \epsilon_U\}$ is Lebesgue-null, so κ_U has no atoms in $(0, \infty)$ and $\Pr(\kappa_U < \alpha_U) = \Pr(\kappa_U \le \alpha_U) \ge 1 - \beta$ when $\alpha_U < \infty$. For $\alpha \ge \alpha_U$, this gives $\Psi_U(h_U + \epsilon_U) \le \beta$. Substituting both bounds into Equation 9 proves the theorem. For the same pair, a union bound further gives the probability that F_A accepts and F_U rejects the protected pair is at least $1 - 2\beta - \eta$. □

Authorized acceptance. On E, $\alpha\|C_A Z\|_2 \le m_A$ implies authorized acceptance by the triangle inequality. Since $\mathbb{E}\|C_A Z\|_2^2 = \|C_A\|_F^2$, Markov's inequality and $\Pr(E^c) \le \eta$ give Equation 8. For any admissible B, uniformly sampled probe pairs give $M^{-2}\sum_{i,\ell}\|[J_i B, -J_\ell B]\|_F^2 = 2\,\mathrm{Tr}(B^\top H_A B)$, which Theorem 1 minimizes at B_A.

Gaussian distance distribution. The quantiles α_A and α_U depend on the clean pair only through the law of the affine distance. With a thin singular value decomposition $C_F = U\,\mathrm{diag}(\sigma_1,\dots,\sigma_q)V^\top$, $q = \mathrm{rank}(C_F)$, columns u_ℓ of U, and $d_\perp = (I - UU^\top)d_F$, rotational invariance gives

$$\|d_F + \alpha C_F Z\|_2^2 \overset{d}{=} \|d_\perp\|_2^2 + \sum_{\ell=1}^{q}(u_\ell^\top d_F + \alpha\sigma_\ell\xi_\ell)^2, \qquad \xi_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{11}$$

a weighted sum of noncentral chi-square variables plus a constant. Averaging Equation 9 over genuine (same-person) pairs bounds $\bar{a}_F^p$; averaging over impostor (different-person) pairs bounds the FAR at the fixed threshold.

A.1.4 Score Decay at Aligned Embeddings

Write $s_F^p = s_F(x_1^p, x_2^p)$ for the protected score. Differentiating $s_F = 1 - \|d_F\|_2^2/2$ along the protection directions gives the first-order score response $v_F := \sum_{j=1}^{2}\|B_A^\top \nabla_{x_j} s_F\|_2^2 = \|C_F^\top d_F\|_2^2$, which vanishes at aligned embeddings, $F(x_1) = F(x_2)$, even when $C_F \neq 0$. The distance form captures the resulting quadratic decay: with $d_F = 0$ and $\|e_F\|_2 \le L_F\alpha^2\|Z\|_2^2/2$,

$$\left|1 - s_F^p - \tfrac{\alpha^2}{2}\|C_F Z\|_2^2\right| \le \alpha\|C_F Z\|_2\|e_F\|_2 + \tfrac{1}{2}\|e_F\|_2^2 = O(\alpha^3),$$

so, almost surely as $\alpha \to 0$,

$$\frac{1 - s_F^p}{\alpha^2} \to \frac{1}{2}\|C_F Z\|_2^2 \overset{d}{=} \frac{1}{2}\sum_{\ell=1}^{q}\sigma_\ell^2\xi_\ell^2, \tag{12}$$

with convergence in expectation as in Appendix A.1.2. For two copies of one input, $C_F = [J_F(x)B_A, -J_F(x)B_A]$, and the leading expected score loss $\alpha^2\|J_F(x)B_A\|_F^2$ matches the per-probe term of Theorem 1. For $d_F \neq 0$, the linear term $-\alpha d_F^\top C_F Z$ has variance $\alpha^2 v_F$, while $-d_F^\top e_F$ contributes at order α^2 and is absorbed by the adjusted radii of Theorem 2.

A.1.5 Acceptance under Similar Verification Geometry

Theorem 2 separates the recognizers only if F_U loses acceptance at smaller strengths than F_A. We bound how far their acceptance probabilities can differ. Scaling acceptance radii to one, let $\mu_F = d_F/h_F$, $T_F = C_F/h_F$, and $Y_F = \mu_F + \alpha T_F Z$, so the affine acceptance probability is $a_F^{\text{aff}} = \Pr(\|Y_F\|_2 \le 1)$. Up to rotations of the embedding space and of the Gaussian noise, the normalized geometries differ by

$$\Delta_\alpha^2 = \min_{Q\in\mathrm{O}(d),\, O\in\mathrm{O}(2k)}\left\{\|\mu_U - Q\mu_A\|_2^2 + \alpha^2\|T_U - QT_A O\|_F^2\right\}. \tag{13}$$

For every $\omega > 0$,

$$\left|a_U^{\text{aff}} - a_A^{\text{aff}}\right| \le \Pr\left(\left|\|Y_A\|_2 - 1\right| \le \omega\right) + \frac{\Delta_\alpha^2}{\omega^2}. \tag{14}$$

Indeed, for a minimizing (Q, O), couple Y_U with $\widetilde{Y}_A = Q(\mu_A + \alpha T_A O Z)$, which satisfies $\|\widetilde{Y}_A\|_2 \overset{d}{=} \|Y_A\|_2$ and $\mathbb{E}\|Y_U - \widetilde{Y}_A\|_2^2 = \Delta_\alpha^2$. If the decisions differ, either $\|\widetilde{Y}_A\|_2$ lies within ω of 1 or $\|Y_U - \widetilde{Y}_A\|_2 > \omega$; Markov's inequality completes the bound. For any $\alpha > 0$, $\Delta_\alpha = 0$, equivalently $(\mu_F, T_F T_F^\top)$ agree up to a common output rotation, gives equal affine acceptance at every scale; if $F_U = QF_A$ with orthogonal Q and equal thresholds, the recognizers make identical decisions for every protected pair. With remainder coverage $\Pr(\|e_F\|_2 > \epsilon_F) \le \eta_F$ and $\bar{\epsilon}_F = \epsilon_F/h_F$, the same coupling and a union bound give

$$|a_U^p - a_A^p| \le \Pr\left(\left|\|Y_A\|_2 - 1\right| \le \omega + \bar{\epsilon}_A + \bar{\epsilon}_U\right) + \frac{\Delta_\alpha^2}{\omega^2} + \eta_A + \eta_U. \tag{15}$$

Small geometric differences therefore limit selectivity unless substantial authorized probability lies near the threshold or the remainders are large. Appendix A.4 tests recognizers that share the backbone of F_A.

Table 5: First-layer and end-to-end sensitivity for the same k and probes: $b = \sum_{j=1}^{k} \lambda_j(H_A)$, $c = \|\bar{G}\|_2 \sum_{j=1}^{k} \sigma_j(W)^2$, and $e_W = \mathrm{Tr}(B_W^\top H_A B_W)$. Spectra are ordered increasingly; B_W contains the right singular vectors associated with the k smallest singular values of W.

Recognizer	$\sigma_{\min}(W)$	b	c	e_W/b	e_W
IRSE-100	2.44×10^{-1}	5.19×10^{-7}	6.87×10^{-2}	48.92	2.54×10^{-5}
DenseNet-201	4.46×10^{-2}	6.16×10^{-7}	7.67×10^{-3}	41.53	2.56×10^{-5}
SwinMLP-B	5.12×10^{-2}	3.99×10^{-7}	1.31	253.26	1.01×10^{-4}
TransFace ViT-B	0	0	0	–	0

FIRST-LAYER AND END-TO-END SENSITIVITY

Proposition 1 (First-layer sensitivity as a sufficient condition) *Write $F_A = \Phi \circ W$ with first linear operator $W = W^{(1)} \in \mathbb{R}^{m \times n}$, let $G_i = \partial\Phi(W r_i)$ and $\bar{G} = M^{-1} \sum_{i=1}^{M} G_i^\top G_i$, and order $\lambda_j(H_A)$ and $\sigma_j(W)^2 = \lambda_j(W^\top W)$, $j = 1, \ldots, n$, increasingly. Then:*

(i) $H_A = W^\top \bar{G} W$, *so* $\ker W \subseteq \ker H_A$;

(ii) $\delta^\top H_A \delta \le \|\bar{G}\|_2 \|W\delta\|_2^2$ *for every* $\delta \in \mathbb{R}^n$;

(iii) $\sum_{j=1}^{k} \lambda_j(H_A) \le \|\bar{G}\|_2 \sum_{j=1}^{k} \sigma_j(W)^2$ *for every* $1 \le k \le n$.

Proof. The chain rule gives $J_i = G_i W$, hence (i), and (ii) follows from $\delta^\top H_A \delta = \|\bar{G}^{1/2} W\delta\|_2^2$. Let B_W contain the right singular vectors of W for its k smallest singular values. The minimization in Appendix A.1.2 and (ii) give $\sum_{j \le k} \lambda_j(H_A) \le \mathrm{Tr}(B_W^\top H_A B_W) \le \|\bar{G}\|_2 \sum_{j \le k} \sigma_j(W)^2$. □

By (ii), $\{\delta : \|\bar{G}\|_2 \|W\delta\|_2^2 \le \tau\} \subseteq \{\delta : \delta^\top H_A \delta \le \tau\}$ for every $\tau > 0$, with strict inclusion unless $H_A = \|\bar{G}\|_2 W^\top W$. First-layer control is thus sufficient for low mean sensitivity but generally restrictive.

A.1.6 RANK OF THE EMPIRICAL SENSITIVITY MATRIX

Differentiating $\|F_A(r_i)\|_2^2 = 1$ gives $F_A(r_i)^\top J_i = 0$, so $\mathrm{rank}(J_i) \le d - 1$. Subadditivity of rank and Proposition 1(i) give

$$\mathrm{rank}(H_A) \le \min\big\{n, \mathrm{rank}(W), M(d-1)\big\}. \tag{16}$$

Hence $\dim\ker H_A \ge n - M(d-1)$. When $k \le \dim\ker H_A$, every k-dimensional subspace of $\ker H_A$ has zero mean squared linear response, and the minimizer is not unique if $k < \dim\ker H_A$.

FIRST-LAYER SPECTRAL COMPARISON

Table 5 extends the summary in Table 3 with $\sigma_{\min}(W)$ and the first-layer spectral bound c, alongside the optimal end-to-end sensitivity b and the end-to-end sensitivity e_W of the first-layer basis. The proof of Proposition 1(iii) gives $b \le e_W \le c$: the first gap measures the effect of subspace selection, and the second measures the looseness of the bound. For IRSE-100, DenseNet-201, and SwinMLP-B, the first-layer basis has $41.53\times$–$253.26\times$ the end-to-end sensitivity of B_A. This gap motivates constructing the protection space from the full input-to-embedding sensitivity. TransFace ViT-B represents a special case: its k smallest singular values are zero, so the corresponding first-layer basis lies in $\ker W$. By Proposition 1(i), these directions also lie in $\ker H_A$, yielding $b = e_W = c = 0$.

A.2 FULL PROTECTION-STRENGTH SWEEP

Figure 4 extends the focused view of Figure 2 to the full $\alpha \in [0, 2.5]$ sweep in separate panels. Figure 5 provides an alternative view of the lower-strength portion of the sweep. The criteria in Table 6 select consecutive tested strengths of 1.0–1.1 for IRSE-100, 0.9–1.9 for DenseNet-201, 0.9–1.1 for SwinMLP-B, and 0.4–0.8 for TransFace ViT-B. These are empirical ranges on the sampled grid. The unauthorized metric is an average over recognizers, so the displayed threshold does not bound

Table 6: Selective ranges on the tested $\Delta\alpha = 0.1$ grid. Each range satisfies authorized TAR loss from clean inputs ≤ 5 percentage points and mean unauthorized TAR $\leq 10\%$. Entries give the minimum authorized TAR (A-TAR) and maximum mean unauthorized TAR (U-TAR) (%, FAR=1%) within each range.

Authorized	α range	Min. A-TAR	Max. mean U-TAR
IRSE-100	1.0–1.1	94.50	6.69
DenseNet-201	0.9–1.9	95.47	6.11
SwinMLP-B	0.9–1.1	94.80	1.76
TransFace ViT-B	0.4–0.8	95.50	0.47

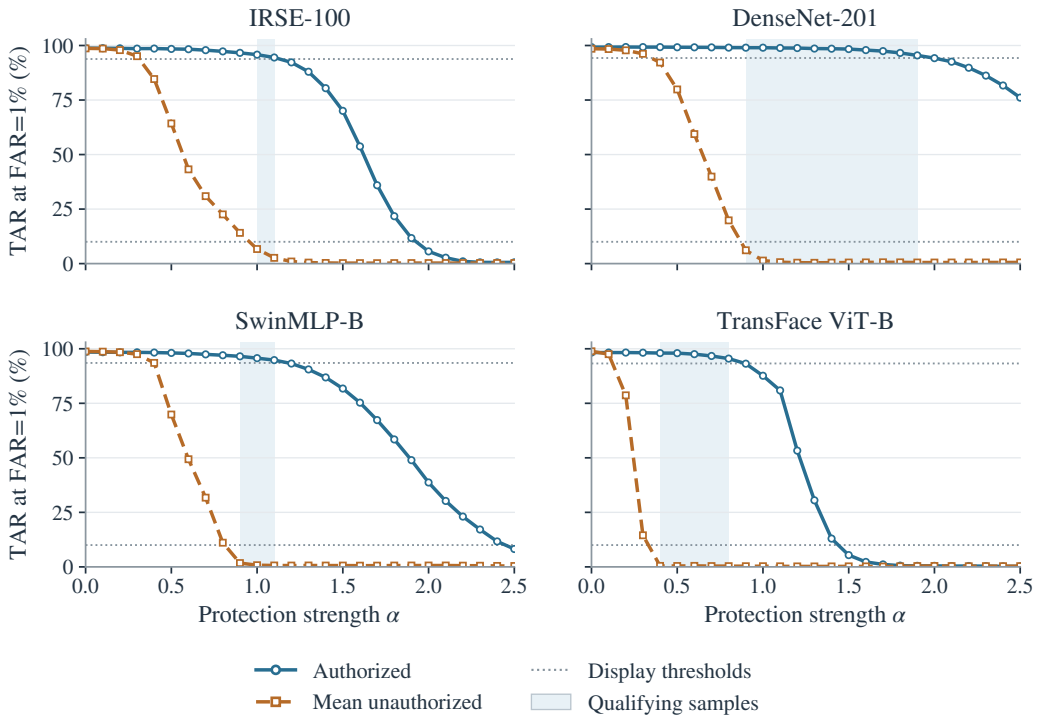


Figure 4: Unauthorized TAR falls before authorized TAR deteriorates, but large perturbations eventually impair authorized matching. Each panel shows the full strength sweep for one authorized recognizer; solid circles denote authorized TAR and dashed squares denote mean unauthorized TAR. Dotted lines mark the display thresholds of Table 6: authorized TAR within 5 percentage points of its clean value and mean unauthorized TAR at most 10%. Shading spans consecutive qualifying samples on the measured $\alpha = 0, 0.1, \ldots, 2.5$ grid; no threshold crossing is interpolated.

every unauthorized recognizer individually. The sweep also shows why protection strength must be chosen for the authorized recognizer: increasing α beyond its useful range eventually reduces authorized TAR.

A.3 Robustness against Reconstruction Attacks

We further evaluate whether reconstruction can recover recognition utility for unauthorized recognizers. Table 7 reports results for two representative reconstruction approaches, GFPGAN (Wang et al., 2021) and DiffBIR (Lin et al., 2024), covering GAN- and diffusion-based reconstruction. Both methods largely preserve recognition performance on clean images, indicating that reconstruction itself does not substantially degrade facial identity information. However, neither effectively removes

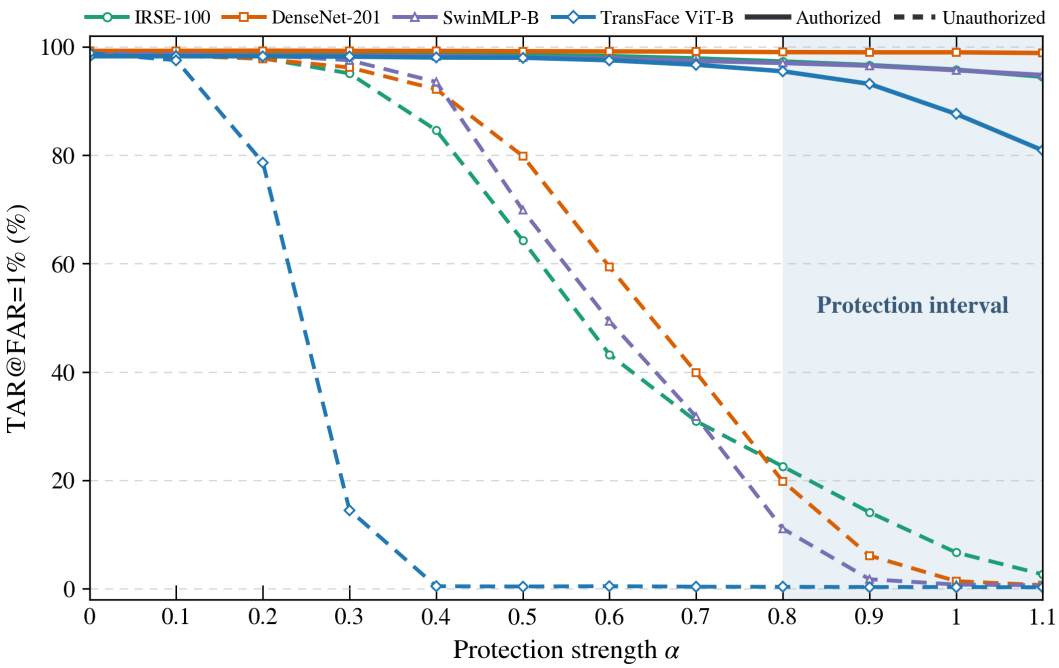


Figure 5: Alternative view of the strength sweep over $\alpha \in [0, 1.1]$. Colors identify the authorized recognizer; solid and dashed curves denote authorized and mean unauthorized TAR, respectively. Shading highlights $\alpha \in [0.8, 1.1]$ for visual comparison; it does not indicate a common qualifying range for all models. Ranges under the common performance criteria appear in Table 6.

Table 7: Reconstruction attacks on LFW. Entries show ACC / TAR@FAR=1% (%); U- marks unauthorized recognizers. Clean results are averaged over four recognizers.

Method	Family	Clean ACC / TAR ↑	Protected U-ACC / U-TAR ↓
Baseline	No reconstruction	99.04 / 98.71	58.64 / 2.29
GFPGAN	GAN	98.62 / 98.47	63.02 / 6.78
DiffBIR	Diffusion	98.51 / 98.42	58.30 / 0.62

the protection introduced by GRANTFACE. After reconstruction, unauthorized recognition remains strongly suppressed: GFPGAN achieves 63.02% U-ACC and 6.78% U-TAR, while DiffBIR yields 58.30% U-ACC and 0.62% U-TAR.

A.4 MODELS WITH THE SAME BACKBONE

By Appendix A.1.5, selectivity is limited by how far the normalized verification geometries of F_U and F_A differ. To test whether training alone induces such differences, we vary the training data or recognition objective while holding the backbone fixed. GRANTFACE remains selective in both settings.

Training data. Figure 6 considers recognizers with the same backbone and recognition objective, differing only in the training dataset. The MS1M model serves as the authorized recognizer, while models trained on WebFace4M and Glint360K serve as unauthorized recognizers. Although all three models exhibit comparable clean recognition performance, their responses diverge as the protection strength increases. Unauthorized TAR decreases substantially earlier, whereas the authorized recognizer remains effective over a much wider range of α. This separation shows that the protection remains selective even when the recognizers differ only in the data used for training.

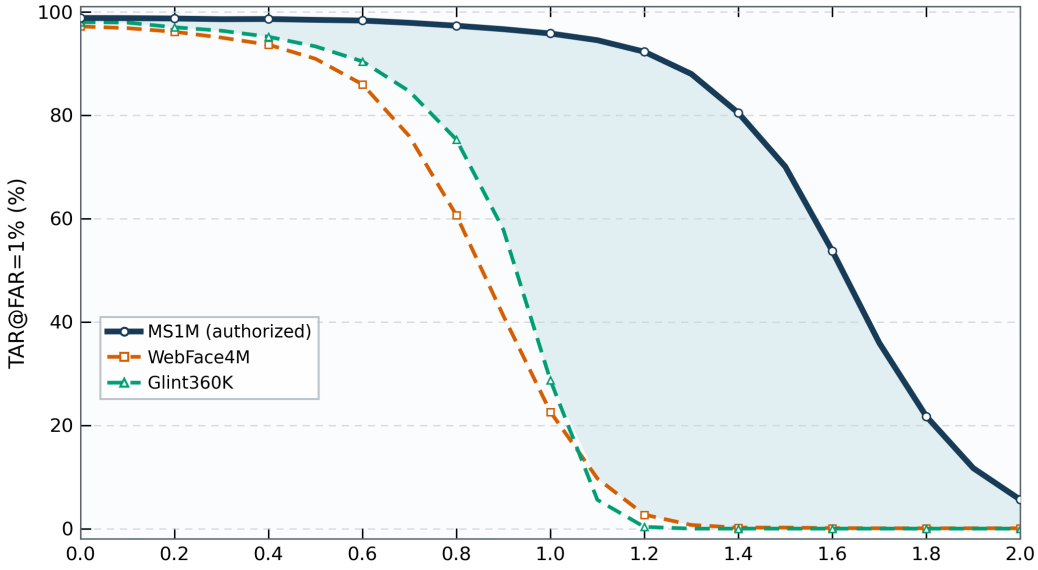


Figure 6: TAR versus protection strength α for models differing only in training data.

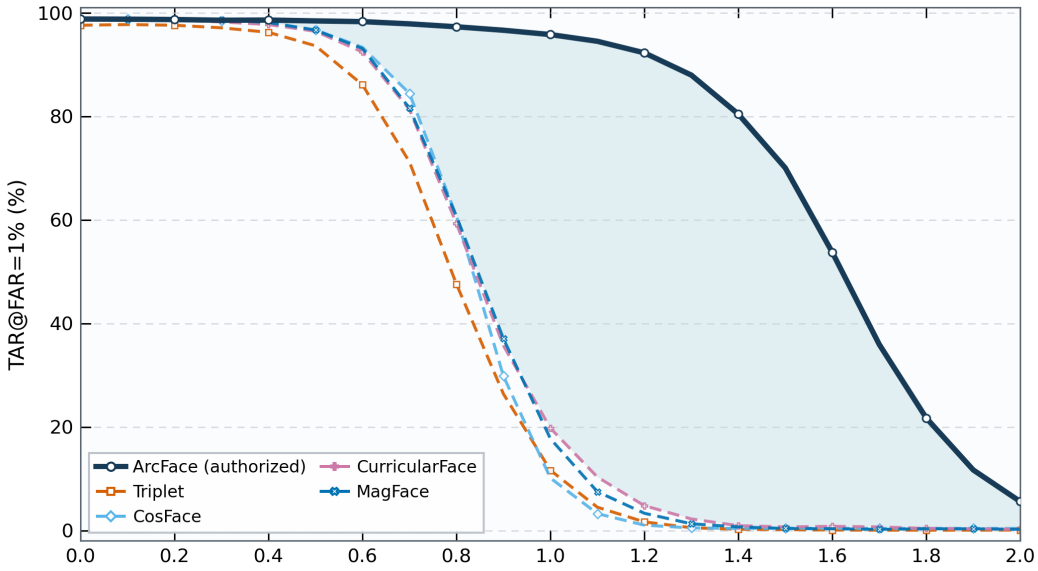


Figure 7: TAR versus protection strength α for models differing only in training objective.

Training objectives. The same pattern holds when the recognition objective varies. Figure 7 compares models sharing the same backbone and training data, with ArcFace serving as the authorized recognizer and Triplet, CosFace, CurricularFace, and MagFace as unauthorized recognizers. Despite the shared architecture and training data, the unauthorized recognizers lose recognition capability at substantially smaller protection strengths, while the authorized ArcFace model remains effective over a broader range. These results show that selectivity persists across training objectives even when architecture and training data are fixed.

Table 8: Comparison with selective face-protection methods: authorized/unauthorized TAR (%), mean recognition gap (p.p.), latency (ms/image), and throughput (images/s).

Method	LFW	AgeDB-30	CALFW	CFP-FP	Gap ↑	Lat. ↓	FPS ↑
Targeted Anon.	97.52 / 67.63	95.38 / 41.17	91.79 / 33.66	97.32 / 59.10	45.11	1207.07	0.83
PMT / FDM	97.63 / 82.30	95.43 / 81.54	91.21 / 78.47	**98.18** / 88.89	12.81	2019.04	0.50
GRANTFACE	**97.78 / 25.73**	**95.48 / 31.72**	**91.90 / 31.09**	97.60 / **39.48**	**63.69**	**0.26**	**3829.47**

Table 9: Verification AUC (%) across four benchmarks at $\alpha = 1.0$. Bold marks the authorized recognizer.

Dataset	Authorized	IRSE-100	DenseNet-201	SwinMLP-B	TransFace ViT-B
LFW	Clean	99.60	99.78	99.59	99.51
	IRSE-100	**98.87**	66.37	64.62	63.76
	DenseNet-201	55.21	**99.74**	61.97	65.37
	SwinMLP-B	57.97	56.75	**98.96**	65.24
	TransFace ViT-B	53.59	57.86	55.92	**98.04**
AgeDB-30	Clean	99.01	99.05	99.10	99.12
	IRSE-100	**97.54**	66.44	59.79	54.91
	DenseNet-201	59.66	**99.01**	63.26	60.59
	SwinMLP-B	52.75	56.24	**98.62**	63.55
	TransFace ViT-B	46.73	52.51	52.85	**97.71**
CALFW	Clean	97.06	97.20	97.19	97.22
	IRSE-100	**96.23**	67.95	56.11	54.41
	DenseNet-201	54.05	**97.05**	60.48	58.60
	SwinMLP-B	50.86	51.09	**97.03**	59.39
	TransFace ViT-B	49.21	49.34	52.46	**96.39**
CFP-FP	Clean	99.65	99.64	99.79	99.61
	IRSE-100	**95.31**	65.94	60.49	64.94
	DenseNet-201	52.33	**99.26**	61.46	63.30
	SwinMLP-B	57.91	53.75	**98.75**	65.25
	TransFace ViT-B	53.50	60.41	52.83	**95.20**

A.5 COMPARISON WITH EXISTING AUTHORIZED FACE PROTECTION METHODS

Table 8 shows that GRANTFACE achieves the largest average recognition gap among the compared methods: 63.69 percentage points (p.p.), compared with 45.11 p.p. for Targeted Anonymization (Ding et al., 2022) and 12.81 p.p. for PMT/FDM (Pu et al., 2025). It maintains high authorized TAR while reducing unauthorized TAR across all four datasets. Online protection takes 0.26 ms per image, with throughput above 3,800 images/s, whereas the baselines require more than one second per image.

A.6 THRESHOLD-INDEPENDENT VERIFICATION

The main evaluation uses the clean-calibrated verification threshold for both clean and protected inputs. To examine performance independently of this operating threshold, Table 9 reports verification AUC for the same settings. Across all datasets, the authorized recognizer retains AUC close to its clean performance, whereas unauthorized recognizers show substantially lower AUC, with many results approaching chance-level discrimination.

A.7 DIRECTION OF EMBEDDING DISPLACEMENT

Section 3.4 ties acceptance to how protection moves the pair's embedding difference. We visualize this movement directly. For each genuine pair, the reference embedding z_r is kept clean and only

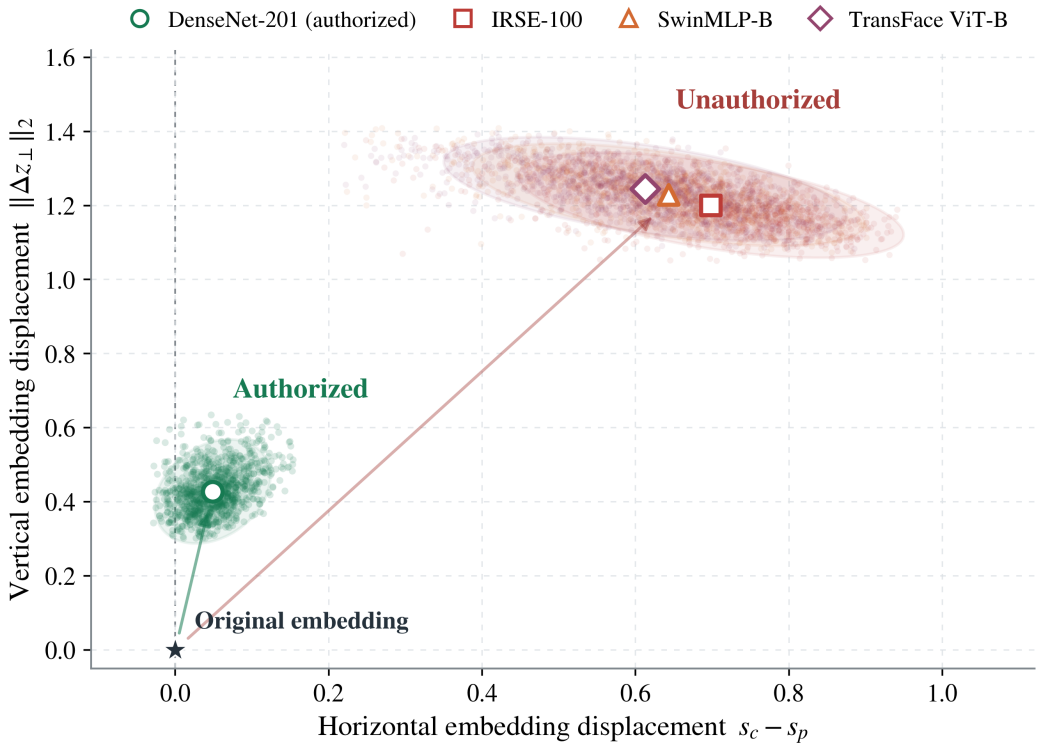


Figure 8: Embedding displacement relative to a clean reference z_r. Horizontal: cosine similarity decrease $s_c - s_p = -\Delta z^\top z_r$. Vertical: orthogonal displacement $\|\Delta z - (\Delta z^\top z_r) z_r\|_2$.

the query is protected, with displacement $\Delta z = z_q^p - z_q^c$. Since the score is $s = z_q^\top z_r$, the horizontal component $s_c - s_p = -\Delta z^\top z_r$ is exactly the score decrease, while the vertical component $\|\Delta z - (\Delta z^\top z_r) z_r\|_2$ is orthogonal to the reference and leaves the score unchanged.

Figure 8 shows distinct displacement patterns for the authorized and unauthorized recognizers. For the authorized recognizer, the displacement is concentrated predominantly in the vertical direction, while the horizontal component remains limited. Consequently, most of the embedding change has little effect on the cosine similarity with the clean reference, allowing same-person matching to be retained after protection. In contrast, the unauthorized recognizers exhibit substantially larger horizontal components. These changes directly reduce the cosine similarity between the protected query and the clean reference, leading to lower matching scores and suppressed unauthorized matching.