FACELINKGEN: A RE-EVALUATION OF IDENTITY LEAKAGE IN PRIVACY-PRESERVING FACE RECOGNITION AND FACE ANONYMIZATION SYSTEMS USING SIMPLE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy-preserving face recognition (PPFR) and face anonymization have different goals, but both must retain some identity-related information for their intended use. We show that an adaptive attacker can learn this information. We propose FaceLinkGen, a simple distillation-based attack that trains a face recognition model to map protected inputs back to standard face embeddings. FaceLinkGen applies to keyless PPFR systems and perception-preserving face de-identification (De-ID) systems. For PPFR, the recovered embeddings can be used to regenerate faces that match the original person. Across MinusFace, PartialFace, and DecoyFace, the regenerated faces achieve acceptance rates of 81.0–99.4% on Face++ and 74.9–99.6% on Amazon. For De-ID, FaceLinkGen links protected faces to unprotected images of the same person, reaching Recall@1 values of 48.4–89.6% in the one-side-protected setting across the evaluated methods. FaceLinkGen exposes identity leakage across all evaluated methods, including DecoyFace and WDP, whose protection resists the tested U-Net attacks on face recovery and protected-to-unprotected linkage, respectively. The attack also remains effective when trained with limited paired data. Code is available at `https://anonymous.4open.science/r/FaceLinkGenRelease-08C8/`.

1 INTRODUCTION

Two common facial privacy tasks are privacy-preserving face recognition (PPFR) and perception-preserving face anonymization or de-identification (De-ID). PPFR aims to let a server verify a user's identity without exposing the user's face (Erkin et al., 2009), while perception-preserving De-ID has a nearly opposite goal: the protected face should remain recognizable to humans while becoming unrecognizable to face recognition (FR) models.

Representative PPFR methods include frequency-domain approaches such as PartialFace (Mi et al., 2023), MinusFace (Mi et al., 2024), and FracFace (Dai et al., 2025), as well as decoy-oriented approaches such as DecoyFace (Ren et al., 2026). The De-ID methods studied in this paper include the perturbation-based TIP-IM (Yang et al., 2021) and Protego (Wang et al., 2025b), synthesis-based PerceptFace (Wang et al., 2025a), and the diffusion-based *Weakening Diffusion Purification* (WDP) (Salar et al., 2025).

Both tasks face a common tension: privacy-coupled identity information must remain available to support their intended utility. In a keyless PPFR system, no user-held secret key controls access to the protected representation. If that representation remains useful for identity matching, an attacker with query access can learn to extract its identity signal and regenerate a visually recognizable face. Similarly, if a perception-preserving De-ID image must still look like the same person to humans, it necessarily retains identity cues that an FR model can learn to use.

We propose FaceLinkGen, a unified distillation-based attack framework for PPFR and De-ID systems. Its main advantage is coverage across all evaluated protection methods. Our results show that the vulnerability is not merely theoretical. For PPFR, FaceLinkGen regenerates faces with acceptance rates of 81.0–99.4% on Face++ and 74.9–99.6% on Amazon across MinusFace, Partial-

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.

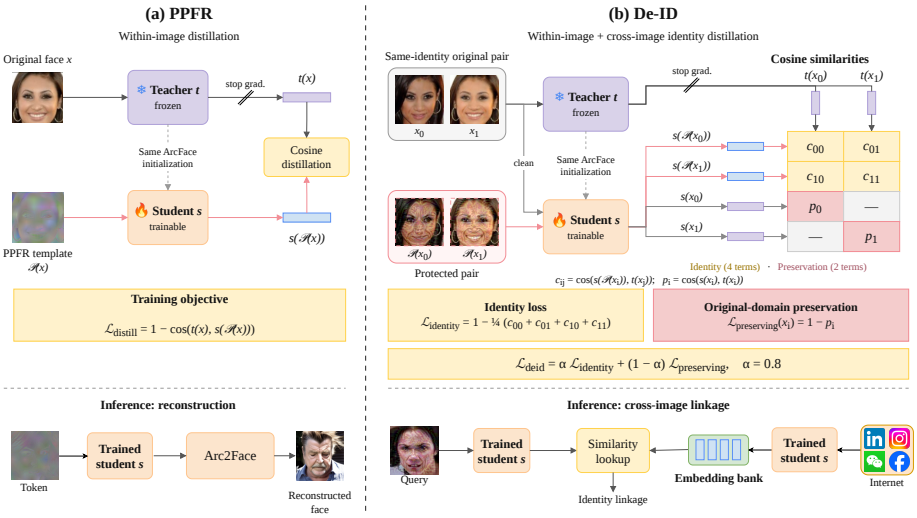

Figure 1: Method overview of FaceLinkGen. (a) For PPFR, training extracts the ground-truth identity embedding from the original face and encourages the student to produce the same embedding from the protected template by matching the teacher's output using a cosine-similarity loss. At inference time, the student extracts an embedding from an unseen template, and Arc2Face is used to synthesize a face image from that embedding. (b) For De-ID systems, we use two images of the same identity along with their protected counterparts. The goal is to maximize cosine similarity across the 2×2 matching grid between the two original images and their protected counterparts. To preserve the student's ability to recognize unprotected faces, we also feed the student an original image and enforce that its output for that image matches the teacher's.

Face, and DecoyFace. For De-ID, FaceLinkGen achieves 48.4–89.6% Recall@1 when matching a query against a gallery (with either the query or the gallery protected) across the evaluated methods, demonstrating that identity linkage remains practical after protection. Although an improved attack (U-Net) achieves a slightly higher attack success rate (ASR) on MinusFace, PartialFace, and TIP-IM, FaceLinkGen can recover identities from DecoyFace and link WDP-protected queries to unprotected images, where the tested U-Net attacks fail.

In summary, our contributions are as follows: (1) we identify a shared structural weakness underlying keyless PPFR and perception-preserving face de-identification, two tasks with seemingly opposite goals; (2) we propose FaceLinkGen, a unified distillation-based attack framework that exploits the identity signal retained for utility; and (3) we show that substantial identity information survives the protection mechanisms of representative methods from both families, with successful attacks across all evaluated methods, including methods that resist the tested U-Net attacks.

2 Related Work

Privacy-Preserving Face Recognition The aim of PPFR is to enable a server to confirm a user's identity without revealing the user's face. For example, a person could use facial authentication to approve a transaction on an online banking site, while ensuring that the bank does not receive the original image and cannot reconstruct the user's face from the provided template. The unprotected face never leaves the user's device, and the conversion happens locally. The server will only receive the protected template.

Most existing PPFR approaches work in the frequency domain. Their main assumption is that machines recognize faces using high-frequency information, whereas the visual information perceived by humans lies in the low-frequency domain. On this basis, they argue that the original image cannot be reconstructed after the low-frequency information is removed and the frequency channels are disrupted. Among them, PartialFace first removes the low-frequency information and then shuffles the high-frequency information, whereas MinusFace uses an autoencoder-style network to

reconstruct the original image and computes the residual between the original and reconstructed images to remove low-frequency information. The residual thus contains very little information. MinusFace then shuffles it again in the frequency domain and converts it back to the spatial domain. Another method, FracFace, uses a fractal-style transformation to disrupt the frequency channels. In contrast, DecoyFace separates reconstruction-sensitive information from recognition-sensitive information and replaces the reconstruction-sensitive information with another person's face. A trained reconstruction network therefore recovers the decoy face rather than the protected identity.

Face Anonymization There are two types of face anonymization or de-identification (De-ID): utility-preserving and non-utility-preserving. The simplest non-utility-preserving method is blurring the face. A more advanced non-utility-preserving method is a form of "face swap," where a different face replaces the user's face to protect privacy. This is useful for dataset sharing and anonymizing people other than the main subject (e.g., protecting passersby's privacy). Utility-preserving De-ID methods aim to keep the face recognizable to humans while avoiding recognition by FR models. For example, users may want to share photos on social media that friends and family recognize, while preventing web scrapers from indexing their faces to build a face search engine that links their photos across platforms. In this paper, we target utility-preserving methods, referred to simply as "face De-ID" in the following text.

There are two main types of such De-ID systems: adversarial perturbation-based methods and perception-preserving synthesis-based methods. Adversarial perturbation-based methods add noise to exploit model weaknesses and avoid recognition. However, these perturbations are often weak and model-specific (Wang et al., 2025a), and they introduce visible artifacts. Perception-preserving synthesis-based methods like PerceptFace generate a different identity while keeping visual changes minimal, so the face still appears to belong to the same person to human observers.

Previous Attacks Prior work by Calmon et al. (2015) shows that perfect privacy is possible only if the information needed for utility is decoupled from private information. This principle underlies our attack: human perception and machine recognition rely on largely shared facial information, making full separation difficult, if not impossible.

Recent work in non-facial domains has also begun shifting attention from pixel-level reconstruction to semantic-level inversion. In such settings, the attacker aims to regenerate information that is semantically consistent with the original identity without recovering the original image itself. This goal is easier to achieve and better aligned with realistic attack objectives. For instance, Yue et al. (2023) propose a semantic recovery framework that uses generative models to synthesize semantically consistent images without pixel-level similarity in federated learning attacks. Other findings suggest that traditional reconstruction metrics fail to capture how humans perceive privacy leakage (Sun et al., 2023).

3 Threat Model

Threat Model for PPFR For PPFR systems, the adversary is the service provider itself. The service provider does not need to be the algorithm developer, which can be a third party or a standardization agency. This follows the original threat model of PPFR (Erkin et al., 2009), where the user wants the server to verify identity without seeing the real face. We model the server as honest-but-curious (HBC) following prior work (Erkin et al., 2009; Ren et al., 2026): it faithfully follows the protocol (without compromising any systems) but attempts to recover the user's facial image from the templates it receives. We restrict the adversary to black-box oracle access to the protection function: it can submit an original image and obtain the corresponding protected template, but cannot inspect internal states or random seeds; any internal randomness in the method is freshly sampled for each query and is unknown to the adversary.

This constraint includes two scenarios: first, when the service provider did not build the protection mechanism and lacks knowledge of the system; and second, when an insider, such as an employee, can access the data and invoke the protection function but cannot view the system design. It also extends to an external attacker who can run the PPFR method locally on their own device to obtain the pairs, although this attacker is not our primary adversary.

This matches the black-box setting of MinusFace (Mi et al., 2024) and is stricter than the "minimal assumptions" setting of Zhang et al. (2024), which requires thousands of server queries per target and models the adversary more like an outside attacker than a service provider.

We consider two regimes. In the first, the HBC service provider can query the protection method without rate limits, which applies when the protection method is an open standard or developed by the service provider itself. In the second, the service provider has limited access to the protection method, so the adversary obtains templates for far fewer identities (256–8192). This tests attack performance when paired data are limited, such as when the protection method is held by a third-party service or is closed-source.

Threat Model for De-ID For De-ID systems, the adversary is a web scraper that collects online images, stores their embeddings in a database, and searches for matching faces. The adversary's goal is to link a query face to other images of the same person, whether those images are protected or not. Because several De-ID methods require per-sample optimization, making large-scale data generation costly for an attacker, we evaluate training with 2,048 images from 1,024 identities, as well as a few-sample setting with 32–1,024 identities. We also assume a black-box setting in which the attacker can acquire these pairs by directly querying the De-ID tools using public face datasets such as MS1M without knowledge of system internals, since these tools must be made publicly available for users to anonymize their faces.

We justify our threat models by Kerckhoffs's principle, which states that security should not rely on keeping an algorithm secret; rather, it should rely on keeping the key secret. Thus, the adversaries do not have access to the random components in the systems but do have access to an implementation of the algorithm itself.

4 FACELINKGEN

Background For a face image x and its protected image $\tilde{x}$, we assume that they contain model-identifiable information z_m and human-perceptual information z_p. PPFR methods aim to remove z_p while retaining z_m, so that the image remains usable for recognition but reveals nothing to a human observer and cannot be reconstructed back to x. De-ID methods aim for the opposite: to hide or alter z_m while retaining z_p, so that the image is still identifiable to a human but cannot be matched by a recognition system. Both directions assume that z_m and z_p can be separated; however, in natural face images, we hypothesize that they are strongly correlated, such that retaining one leaves enough identity signal for an adaptive model to learn a mapping back to the other. We tested this hypothesis using our proposed FaceLinkGen attack framework.

Consider PPFR first. Methods such as Arc2Face (Papantoniou et al., 2024) and FaceID IP-Adapter (Ye et al., 2023) can regenerate a face from z_m (the face embedding) alone. The fact that this works demonstrates the core issue: z_m carries enough identity information to render a face recognizable to humans. Given that PPFR retains z_m, it is reasonable to construct an attack that extracts Arc2Face-compatible z_m and generates a new image of the person, without directly extracting z_p.

De-ID follows two routes, and both fail for related reasons. The first adds adversarial perturbations that mask z_m, so that an off-the-shelf model $\mathcal{F}$ produces $\mathcal{F}(\tilde{x}) \neq z_m$. Here z_m is still present in the image, and only the model's reading of it has been disrupted. Training with the objective $\arg\min_{\mathcal{F}} \mathcal{L}(\mathcal{F}(\tilde{x}), z_m)$ recovers the mapping. This is similar to adversarial training, where models trained on adversarial examples can defend against adversarial attacks, at least for the specific methods they were trained on (Goodfellow et al., 2015).

The second route, exemplified by PerceptFace, applies a transformation that replaces z_m in x with another person's identity, so z_m is genuinely changed rather than hidden. This appears stronger, but it carries a structural constraint: for the method to be useful, the result must still look like the same person to a human observer, which means z_p must be preserved. If a human can still recognize the person using z_p, then z_p retains enough information for re-identification. We can therefore train a model to learn a mapping from z_p back to z_m embeddings and use them for linkage.

Our Method Figure 1 summarizes FaceLinkGen. We initialize a student model s and a teacher model t from the same pre-trained ArcFace feature extractor with identical weights. To maintain a unified framework and ensure a fair comparison, we use the Arc2Face-compatible model even for De-ID tasks, which do not require regeneration. The teacher t remains frozen throughout training, while only the parameters of s are updated.

Given a public face dataset $\mathcal{X}$ with an example image $x \in \mathcal{X}$ and a protection transformation $\mathcal{P}$, the protected input is $\tilde{x} = \mathcal{P}(x)$. The central objective is to train s to recover from $\tilde{x}$ the identity representation produced by t on the corresponding original image. We use cosine distance as the distillation objective:

$$\mathcal{L}_{\text{distill}} = 1 - \cos\big(s(\mathcal{P}(x)),\, \text{sg}(t(x))\big),$$

where sg denotes the stop-gradient operator.

For PPFR, single-image distillation is sufficient because the attack objective is to recover an identity representation from an individual protected template. The student therefore learns the mapping

$$\mathcal{P}(x) \xrightarrow{s} t(x)$$

.

For De-ID systems, however, single-image supervision alone is insufficient. The teacher embedding $t(x)$ may encode image-specific factors such as pose, illumination, and expression, which the protected image retains, unlike in PPFR. If the student is trained only to reproduce the teacher embedding of the exact corresponding image, it may exploit instance-specific correlations rather than learning an identity-consistent representation. Additionally, since the attacker's goal here is to link the image with *another* image of the same person in the database, cross-image supervision is also better aligned with the inference setting.

We therefore introduce cross-image supervision. Given two images (x_0, x_1) of the same identity, each protected image is matched against the teacher embeddings of both original images:

$$\mathcal{L}_{\text{identity}} = 1 - \frac{1}{4}\sum_{i=0}^{1}\sum_{j=0}^{1} \cos\big(s(\mathcal{P}(x_i)),\, \text{sg}(t(x_j))\big).$$

The diagonal terms $(i = j)$ provide within-image distillation, while the off-diagonal terms $(i \neq j)$ explicitly encourage identity consistency across different images of the same person.

A realistic De-ID adversary may encounter a mixture of protected and unprotected images. The student should therefore recover identity from protected inputs without degrading its representation of ordinary faces. To preserve performance on the original image domain, we introduce an original-domain-preserving loss:

$$\mathcal{L}_{\text{preserving}} = 1 - \cos\big(s(x),\, \text{sg}(t(x))\big).$$

The final PPFR training objective is $\mathcal{L}_{\text{PPFR}} = \mathcal{L}_{\text{distill}}$. For De-ID, we combine identity recovery and original-domain preservation. Since the preserving loss only acts as a regularization term to maintain the model's existing capabilities, we assign it a lower weight:

$$\mathcal{L}_{\text{DeID}} = \alpha\mathcal{L}_{\text{identity}} + (1-\alpha)\mathcal{L}_{\text{preserving}}, \qquad \alpha = 0.8.$$

At inference, FaceLinkGen supports different downstream attacks depending on the protection setting. For PPFR, suppose an HBC service provider receives an unknown victim's protected template T, where $T \notin \{\mathcal{P}(x_i) : x_i \in \mathcal{X}\}$. The trained student first maps T into the ArcFace identity space as $e_T = s(T)$. The recovered embedding can then condition Arc2Face to synthesize a face corresponding to the victim:

$$F = \text{Arc2Face}(e_T, \epsilon), \qquad \epsilon \sim p(\epsilon),$$

where ϵ is the diffusion initialization noise that controls identity-independent attributes not uniquely specified by the face embedding, such as pose and expression.

For De-ID, the attacker instead uses the recovered representation for identity linkage. Let $\mathcal{C}$ denote a large-scale corpus of public face images, which may contain either protected or unprotected images. The attacker first constructs an embedding gallery: $\mathcal{G} = \{(s(x),\, x) : x \in \mathcal{C}\}$.

Given a protected query face F, its recovered embedding is $e_F = s(F)$. The attacker then identifies the gallery image whose embedding has the highest cosine similarity to the query:

$$x^* = \arg\max_{x \in \mathcal{C}} \cos\big(e_F,\, s(x)\big).$$

This links the protected face to an identity represented in the public image corpus.

Because the recovered embedding is task-independent, the De-ID linkage attack also applies to PPFR, and the PPFR reconstruction attack also applies to De-ID. These cross-applications provide limited practical advantage. PPFR is not intended to conceal identity from recognition systems, since identity information must be preserved for authorization. Conversely, reconstructing a visually recognizable face from a De-ID representation is of limited value because the protected face already preserves the human-perceptual identity information of the original face. These cross-applications could nevertheless be useful when a PPFR system is intended only for within-system verification and should not enable cross-dataset linkage, or when an image regenerated from a protected De-ID image could serve as a token for linkage in a database with a different embedding space. These cases are outside the scope of this paper.

Metrics For PPFR systems, where the goal is to recover human-recognizable faces, we use Arc2Face (Papantoniou et al., 2024) to reconstruct images from distilled embeddings. To evaluate reconstruction fidelity, we verify generated faces against originals using both Face++ (`FAR@1e-5`) and Amazon (binary pass/fail output) verification APIs, two APIs popular in the face privacy research community, under their strictest operational thresholds. We also report the AdaFace cosine similarity between original and reconstructed faces, and present uncurated qualitative examples. Note that Face++, the Amazon API, and AdaFace are *not* our attack targets; they are used as proxies for human visual judgment to verify that the generated and original faces depict the same person.

For De-ID systems, since the goal is to prevent face-based retrieval, we evaluate retrieval performance on a closed-set face gallery under two settings: both gallery and query images are protected (the setting that is harder for the defender per Wang et al. (2025b)), and only one randomly selected side (either query or gallery) is protected (the setting that is harder for the attacker per Wang et al. (2025b)). We report Recall@1, Recall@5, and mean reciprocal rank (MRR).



Figure 2: U-Net attack results for PartialFace (top row) and MinusFace (bottom row). IN denotes instance normalization. From left to right, the columns show the original image, SGD training without IN, SGD training with IN at an earlier training step, SGD training with IN at a later training step, and AdamW+IN training.

Targets and Baseline We evaluate three PPFR methods: the frequency-based MinusFace (Mi et al., 2024) and PartialFace (Mi et al., 2023), and the decoy-based DecoyFace (Ren et al., 2026). For MinusFace and PartialFace, we use official implementations from `Tencent/TFace`. Because public DecoyFace code is unavailable, we reproduced it using Paper-Replication (Hans & Bilionis, 2026) and verified the key metrics against the original results (Table 3 in the Appendix). Our reproduced values are close to those reported in the original paper. We omit FracFace (Dai et al., 2025) because there are inconsistencies between the code and the paper, and neither their code nor our reproduced code can replicate its results. For De-ID systems, we evaluate the perturbation-based TIP-IM (Yang et al., 2021) and Protego (Wang et al., 2025b), the perception-based PerceptFace (Wang et al., 2025a), and the diffusion-based Weakening Diffusion Purification (WDP) (Salar et al., 2025). We focus on dodging methods rather than impersonation methods such as CLIP2Protect (Shamshad et al., 2023). Impersonation methods can make an FR model identify a "protected" image as someone else, but they do not prevent the model from also recognizing the person's actual identity, offering little privacy protection.

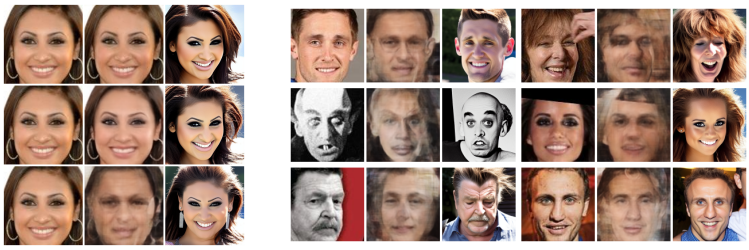

Figure 3: Left: Reconstructions produced by each attack. Rows correspond to PartialFace, MinusFace, and DecoyFace; columns show the original image, the U-Net attack, and the distillation attack. Right: More uncurated examples of reconstruction comparisons between U-Net and distillation for DecoyFace.

Given that most facial datasets are collected from public images and often contain images of the same people across datasets, and it is difficult to deduplicate identities across datasets, using different datasets for training and testing actually has a higher risk of data leakage than using a single dataset and splitting the images by person ID. Thus, we use only MS1M (Guo et al., 2016) as the real image dataset in our paper. To further validate how well our method transfers to a different distribution, we use TPDNE (`TLeonidas/this-person-does-not-exist`) as a synthetic dataset for PPFR testing in Table 6 in the appendix. Since TPDNE does not have paired samples, we cannot use it for De-ID evaluation.

For PPFR, we use 30,000 identities for training, 2,000 for validation, and 2,000 for testing. For De-ID systems, where per-sample optimization makes sample generation computationally expensive (both in our experiments and for real attackers), we use 2,048 training images across 1,024 identities, 128 validation identities, and 384 test identities. Both settings split the dataset based on identity labels and not image indices. Closed-set retrieval evaluation uses 192 query images and 192 target images, mixed with 107,484 unprotected distractor images, reflecting realistic web-scale retrieval where only a small minority of photos are protected.

As a baseline, we evaluate U-Net (Ronneberger et al., 2015) inversion, which trains a network to directly reconstruct clean faces from protected inputs using pixel-level loss. This architecture represents the standard reconstruction attack used in prior evaluations of MinusFace, PartialFace, FracFace, and DecoyFace. We apply it to both PPFR and De-ID systems. For PPFR, the U-Net is trained with an ℓ_1 reconstruction loss: $\mathcal{L}_{\text{ppfr-unet}} = \|x - \mathcal{U}(\mathcal{P}(x))\|_1$. For MinusFace and PartialFace, we evaluate both the original architecture from Mi et al. (2024) (without normalization) and an improved version incorporating instance normalization (Ulyanov et al., 2017). The former reproduces the results reported in the original paper, in which the authors claimed that the protection method was successful, and the latter gives results under a better-tuned attacker. For DecoyFace, following Ren et al. (2026), we use an asymmetric U-Net that maps intermediate feature representations to RGB images.

For De-ID systems, since the attacker also faces a mixture of original and protected images, the U-Net also includes an identity-preserving loss similar to that used in our method, with $\beta = 0.8$: $\mathcal{L}_{\text{deid-unet}} = \beta \cdot \|x - \mathcal{U}(\mathcal{P}(x))\|_1 + (1 - \beta) \cdot \|x - \mathcal{U}(x)\|_1$.

5 Results

5.1 PPFR Systems

Revisiting the U-Net Attack We first re-examine U-Net inversion against PartialFace and MinusFace. Although MinusFace and PartialFace reported that these representations resist U-Net inversion, our better-trained U-Net successfully recovers clean faces (Figure 2, rightmost column). Our experiments attribute this discrepancy to architectural and optimization choices. Training the original MinusFace U-Net (without normalization layers) using SGD (Sutskever et al., 2013) produces nearly uniform images where inversion fails. This differs from the blurry faces reported in the original papers, but in both cases the inversion clearly fails. Adding instance normalization improves feature propagation: early in training, coarse facial structures emerge while fine details

Table 1: PPFR method results under U-Net and distillation attacks. Both AdamW+IN U-Net and distillation successfully invert MinusFace and PartialFace, achieving high acceptance rates on Face++ and Amazon and exceeding the AdaFace similarity threshold (0.19, selected to maximize F1 on LFW). Distillation is the only evaluated attack that succeeds against DecoyFace.

Attack Method	Metric	MinusFace	PartialFace	DecoyFace
AdamW+IN U-Net	Face++	0.998	1.000	0.014
	Amazon	0.998	1.000	0.000
	AdaFace Sim	0.803	0.940	0.074
SGD U-Net	Face++	0.000	0.008	–
	Amazon	0.000	0.064	–
	AdaFace Sim	0.009	0.268	–
Distillation (Ours)	Face++	0.980	0.994	0.810
	Amazon	0.994	0.996	0.749
	AdaFace Sim	0.629	0.696	0.457

remain blurry, matching reported behavior in the prior papers; with continued training, structures become clearer, but the details are still hard to recognize.

Switching from SGD to AdamW alongside instance normalization enables complete reconstruction. This indicates that the apparent resilience of prior defenses reflected evaluation against under-converged U-Nets rather than inherent resistance to inversion. On DecoyFace, however, U-Net reconstructions depict the decoy identity rather than the original subject, confirming that decoy substitution successfully misleads direct pixel inversion. This matches the results reported in the DecoyFace paper.

FaceLinkGen Results Figure 3 (left) compares reconstructions from the tuned U-Net (AdamW+IN) and FaceLinkGen. Table 1 reports the attack success rates (ASR) verified by commercial Face++ and Amazon APIs, as well as the AdaFace cosine similarity between original and reconstructed faces. Under the original MinusFace training setup (SGD without normalization), U-Net inversion fails across both MinusFace and PartialFace. With AdamW and instance normalization, however, U-Net inversion achieves near-100% verification rates on MinusFace and PartialFace. FaceLinkGen achieves Face++ acceptance rates of 98.0% and 99.4% on them as well, comparable to 99.8% and 100.0% for AdamW+IN U-Net. Its AdaFace similarity is slightly lower because FaceLinkGen performs semantic reconstruction, generating another image of the same person, unlike U-Net inversion, which recovers the original image. However, an AdaFace similarity of 0.6–0.7 is well within same person differnt image range. FaceLinkGen's main advantage is that it also breaks DecoyFace (81.0% on Face++, 74.9% on Amazon), which completely resists U-Net inversion.

On DecoyFace, even AdamW-trained U-Nets cannot pass face verification: although the reconstructions resemble faces, they depict the decoy subject, which is the intent of DecoyFace. In contrast, FaceLinkGen achieves an 81.0% verification rate on Face++ and 74.9% on Amazon, recovering most of the identities. Uncurated visual examples in Figure 3 (right) further support this: although U-Net can reconstruct a face, it does not match the original person's identity, whereas FaceLinkGen can generate another face of the same individual. This demonstrates that while decoy substitution diverts U-Net reconstruction, the template retains sufficient identity information for recognition, which distillation reliably extracts and uses to regenerate faces.

To verify embedding quality directly, we compare cosine similarity between student and teacher embeddings against the empirical intra-identity similarity distribution (Figure 5 in Appendix). The student-teacher similarity exceeds the similarity between different photos of the same individual, confirming that distilled embeddings faithfully preserve the identity signal more than another image of the same person would.

5.2 De-ID Systems

Table 2 reports the De-ID attack results. We use the one-side-protected setting as our primary measure of attack effectiveness because in the both-sides-protected setting, retrieving one protected image using another protected image of the same individual is already highly effective before any attack. For example, PerceptFace achieves 0.917 Recall@1 before attack. This high baseline recall

Table 2: Attack results on the De-ID systems: recall on the 192-ID test pool (gallery size 107,676). "Average" is the arithmetic mean of the scores from the one-side and both-sides settings, reflecting a realistic scenario where protected and unprotected images coexist in the gallery.

		U-Net			Distillation (Ours)			Before Attack		
		Recall@1	Recall@5	MRR	Recall@1	Recall@5	MRR	Recall@1	Recall@5	MRR
TIP-IM	One side	0.969	0.974	0.973	0.896	0.958	0.923	0.031	0.052	0.045
	Both sides	0.974	0.979	0.975	0.885	0.922	0.905	0.255	0.391	0.317
	Average	**0.971**	**0.977**	**0.974**	0.891	0.940	0.914	0.143	0.222	0.181
WDP	One side	0.031	0.042	0.036	0.484	0.703	0.578	0.005	0.021	0.011
	Both sides	0.490	0.667	0.577	0.620	0.745	0.673	0.682	0.839	0.762
	Average	0.260	0.354	0.306	**0.552**	**0.724**	**0.625**	0.344	0.430	0.387
PerceptFace	One side	0.255	0.438	0.342	0.823	0.911	0.858	0.036	0.062	0.052
	Both sides	0.891	0.932	0.911	0.911	0.943	0.925	0.917	0.974	0.938
	Average	0.573	0.685	0.627	**0.867**	**0.927**	**0.891**	0.477	0.518	0.495
Protego	One side	0.542	0.677	0.607	0.880	0.932	0.908	0.026	0.026	0.027
	Both sides	0.349	0.474	0.415	0.911	0.943	0.926	0.026	0.047	0.039
	Average	0.445	0.576	0.511	**0.896**	**0.938**	**0.917**	0.026	0.036	0.033

stems from two factors. First, because the gallery consists predominantly of unprotected photos (as on the real Internet, where only a small minority of users protect their photos), protection shifts embeddings into distinct regions rather than hiding them in the unprotected crowd, making protected images stand out and cluster together; t-SNE visualizations (Figure 4 in Appendix A) confirm that all evaluated methods except Protego form isolated embedding clusters, which corresponds to their high both-sides-protected recall. Second, several methods deterministically transform facial features rather than removing them, effectively assigning each person a consistent pseudonym across photos. This mirrors classic linkage flaws, such as using a virtual but stable MAC address to avoid device tracking, static pseudo-identifiers, or unsalted hashes: while the original value is masked, deterministic substitution leaves records linkable across sessions, even though the true identity is never recovered.

For the attack results shown in Table 2, both the U-Net attack and our distillation attack achieve meaningful ASR against TIP-IM and PerceptFace by retrieving the target person from the gallery. FaceLinkGen extends attack coverage to WDP in the one-side-protected setting, where U-Net reaches only 0.031 Recall@1 and distillation reaches 0.484. Even though the number 0.484 might seem low, it means that around half of the tested users' identities are compromised, which indicates substantial leakage. Distillation also achieves substantially higher recall than U-Net on PerceptFace (0.823 vs. 0.255) and Protego (0.880 vs. 0.542), although U-Net achieves higher Recall@1 than distillation on TIP-IM in both settings (0.969/0.974 vs. 0.896/0.885), likely because the weak protection of TIP-IM causes ASR for both methods to saturate around 0.9–1.0. We also show in the Appendix that even with only a small set of samples, our method can attack both types of systems.

6 Conclusion

We introduced FaceLinkGen, a unified distillation attack targeting both keyless PPFR and perception-preserving face De-ID systems. FaceLinkGen reconstructs human-recognizable faces from PPFR templates and can link De-ID queries to unprotected images. In many scenarios, FaceLinkGen matches U-Net's ASR and surpasses it in others. It also remains effective in cases where U-Net fails entirely (DecoyFace). Overall, these results indicate that existing keyless PPFR approaches and De-ID systems are vulnerable to distillation attacks. We therefore call on future developers to carefully consider the robustness of privacy-preserving face systems against distillation attacks before deployment. We have discussed more potential defense methods in the Appendix.

7 AI Usage Disclosure

LLMs were used to search for related work, brainstorm ideas, proofread, and implement ideas by writing code and monitoring experiments. A human was always in the loop to monitor the LLMs, exercise critical judgment, and make the final decisions about their outputs.

REFERENCES

Wei Ao and Vishnu Naresh Boddeti. CryptoFace: End-to-End Encrypted Face Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19197–19206, 2025.

Flavio P. Calmon, Ali Makhdoumi, and Muriel Médard. Fundamental limits of perfect privacy. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1796–1800, June 2015. doi: 10.1109/ISIT.2015.7282765.

Wanying Dai, Beibei Li, Naipeng Dong, Guangdong Bai, and Jin Song Dong. FracFace: Breaking The Visual Clues—Fractal-Based Privacy-Preserving Face Recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

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 and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3087709.

Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-Preserving Face Recognition. In Ian Goldberg and Mikhail J. Atallah (eds.), *Privacy Enhancing Technologies*, pp. 235–253, Berlin, Heidelberg, 2009. Springer. ISBN 978-3-642-03168-7. doi: 10.1007/978-3-642-03168-7_14.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. https://arxiv.org/abs/1607.08221v1, July 2016.

Atharva Hans and Ilias Bilionis. Coding-agents can replicate scientific machine learning papers, July 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015.

Arun Kumar Jindal, Imtiyazuddin Shaik, Vasudha Vasudha, Srinivasa Rao Chalamala, Rajan Ma, and Sachin Lodha. Secure and Privacy Preserving Method for Biometric Template Protection using Fully Homomorphic Encryption. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1127–1134, December 2020. doi: 10.1109/TrustCom50675.2020.00149.

Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age, August 2019.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality Adaptive Margin for Face Recognition, February 2023.

Legislative Services Branch. Consolidated federal laws of Canada, Privacy Act, June 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019.

Pietro Melzi, Hatef Otroshi Shahreza, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Sébastien Marcel, and Christoph Busch. Multi-IVE: Privacy enhancement of multiple soft-biometrics in face embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 323–331, 2023. URL https://openaccess.thecvf.com/content/WACV2023W/DVPBA/html/Melzi_Multi-IVE_Privacy_Enhancement_of_Multiple_Soft-Biometrics_in_Face_Embeddings_WACVW_2023_paper.html.

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, August 2023.

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, March 2024.

Daile Osorio-Roig, Paul A. Gerlitz, Christian Rathgeb, and Christoph Busch. Reversing Deep Face Embeddings with Probable Privacy Protection, October 2023.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2Face: A Foundation Model for ID-Consistent Human Faces, August 2024.

Zhihan Ren, Lijun He, Xinyao Wang, Xinzhu Fu, and Fan Li. DecoyFace: Beyond Obfuscation via Controllable and Imperceptible Identity Misdirection for Privacy-Preserving Face Recognition, July 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015.

Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. Enhancing facial privacy protection via weakening diffusion purification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8235–8244, 2025.

Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. CLIP2Protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20595–20605, 2023.

Xiaoxiao Sun, Nidham Gazagnadou, Vivek Sharma, Lingjuan Lyu, Hongdong Li, and Liang Zheng. Privacy Assessment on Reconstructed Images: Are Existing Evaluation Metrics Faithful to Human Perception?, October 2023.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1139–1147. PMLR, May 2013.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization, November 2017.

Tao Wang, Yushu Zhang, Xiangli Xiao, Kun Xu, Lin Yuan, Wenying Wen, and Yuming Fang. Make Identity Unextractable yet Perceptible: Synthesis-Based Privacy Protection for Subject Faces in Photos, September 2025a.

Ziling Wang, Shuya Yang, Jialin Lu, and Ka-Ho Chow. Protego: User-Centric Pose-Invariant Privacy Protection Against Face Recognition-Induced Digital Footprint Exposure, August 2025b.

Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3897–3907, 2021.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, August 2023.

Zhuowen Yuan, Zhengxin You, Sheng Li, Zhenxing Qian, Xinpeng Zhang, and Alex Kot. On Generating Identifiable Virtual Faces. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1465–1473, Lisboa Portugal, October 2022. ACM. ISBN 978-1-4503-9203-7. doi: 10.1145/3503161.3548110.

Kai Yue, Richeng Jin, Chau-Wai Wong, Dror Baron, and Huaiyu Dai. Gradient Obfuscation Gives a False Sense of Security in Federated Learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 6381–6398, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe Jin, and Xuejun Li. Validating Privacy-Preserving Face Recognition Under a Minimum Assumption. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12205–12214, June 2024. doi: 10.1109/CVPR52733.2024.01160.

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, December 2024.

A Embedding Visualization

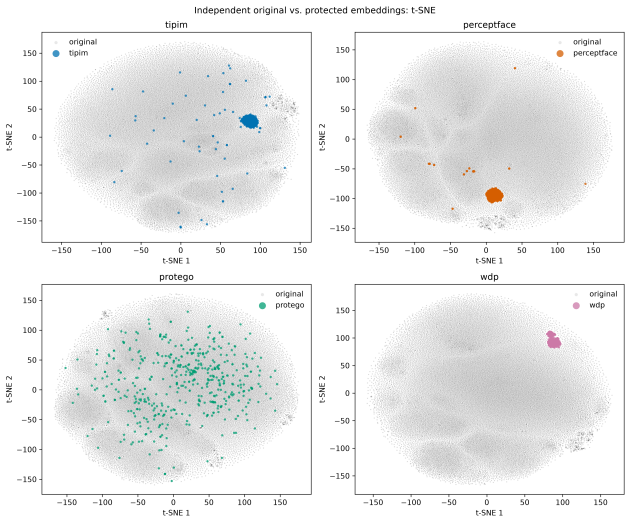

Figure 4: t-SNE plot of each method's embeddings compared to the original embeddings. Gray dots represent embeddings of unprotected distractor images.

B Implementation Details

The protection methods produce different output formats. All De-ID systems produce RGB images, whereas the evaluated PPFR methods produce arrays of different shapes. MinusFace and PartialFace both produce $C \times H \times W$ tensors with different values of C, while H and W remain the original image height and width. We replace the first convolutional layer of the student model with one that accepts C input channels rather than three, without changing the remaining architecture. DecoyFace produces intermediate CNN feature cubes (by default, $64 \times 56 \times 56$). To adapt to this output, we remove layers from the student model until the input shape of the next layer is similar to that of DecoyFace, then adjust the channel count with a single convolutional adapter and the spatial dimensions with bilinear interpolation. Because the student and DecoyFace models use the same Res-100 (He et al., 2015) architecture, the relevant feature tensors have the same shape, and the convolutional adapter is only for compatibility if the model changes. Although they share an architecture, the models come from different families (AdaFace (Kim et al., 2023) vs. ArcFace (Deng et al., 2022)) to avoid circularity and an artificial boost in attack performance. In the full-data setting, all distillation methods are trained using a learning rate of `1e-4` and weight decay of 0.01 with AdamW (Loshchilov & Hutter, 2019). All runs are executed on an NVIDIA DGX B200. Training on PPFR methods is stopped when it plateaus, while De-ID training is stopped when the validation loss rises again (indicating overfitting), and the checkpoints with the lowest validation loss are selected due to the smaller sample size and greater training instability.

Table 3: Reproduction results compared with those reported in the original DecoyFace paper. Our reproduction performs better on three of the four metrics, with a lower identity leakage ratio, a higher identity redirection ratio, and a higher face validity ratio. It has slightly higher cosine similarity with the original face (0.081), but the value remains within the near-zero range of similarities between different people, and thus the Identity Leakage Ratio remains low.

Metric	Reproduced Result	Original Paper Result
Cosine similarity (↓)	0.081	**0.027**
Identity Leakage Ratio (↓)	**2.61**%	2.93%
Identity Redirection Ratio (↑)	**97.39**%	97.07%
Face Validity Ratio (↑)	**99.99**%	99.88%

C Soft Biometric Extraction

Beyond identity recognition, exposure of soft biometric attributes presents a significant privacy risk. Characteristics such as skin color, age, and gender are sensitive personal data that enable unauthorized profiling and algorithmic discrimination. This threat model applies only to PPFR, which aims to identify a person while preserving their privacy. De-ID systems deliberately preserve human-recognizable appearance, so retaining soft biometrics is intentional. Privacy frameworks like the Canadian Privacy Act (Legislative Services Branch, 2025) explicitly protect race and age. A robust privacy-preserving system must therefore prevent the recovery of these attributes from its templates.

The regenerated faces show that these soft biometrics remain visible. This agrees with prior findings that ArcFace embeddings retain such information (Melzi et al., 2023; Osorio-Roig et al., 2023). Because our extracted embeddings closely resemble the original facial embeddings, we hypothesize that a model can learn a direct mapping from the template to these attributes without facial reconstruction.

To assess this, we appended a lightweight probe to the trained student model and trained the resulting model for 10 epochs on the FairFace dataset (Kärkkäinen & Joo, 2019) to predict age, gender, and race (7 categories). The probe consists of a shared hidden layer with ReLU activation, followed by three task-specific linear output heads. The model was trained with cross-entropy loss, a learning rate of `2e-5`, and weight decay of 0.01 using AdamW. To establish a baseline for soft biometric extraction, we also trained the same probe architecture on raw images, using the undistilled teacher model as the base model. The results are reported in Table 5.

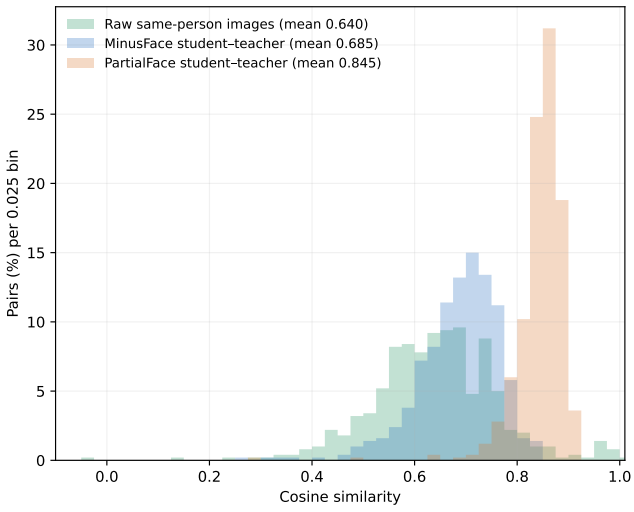

Figure 5: Distributions of cosine similarity for raw same-person image pairs and student-teacher embedding pairs from MinusFace and PartialFace.

Table 4: PPFR method results under U-Net and distillation attacks, evaluated against a different image from the same identity directory. MinusFace and PartialFace distillation use the best checkpoints after continued training. The teacher baseline uses embeddings of the original unprotected inputs. The teacher baseline is shown in separate rows because it is not associated with any protection method.

Attack Method	Metric	MinusFace	PartialFace	DecoyFace
AdamW+IN U-Net	Amazon	0.962	0.986	0.000
	AdaFace Sim	0.519	0.600	0.041
SGD U-Net	Amazon	0.000	0.030	–
	AdaFace Sim	0.007	0.163	–
Distillation (Ours)	Amazon	0.784	0.890	0.419
	AdaFace Sim	0.425	0.470	0.307
Teacher baseline	Amazon		0.926	
	AdaFace Sim		0.493	

Table 5: Soft biometric recovery performance from the protected template without regeneration. Acc denotes accuracy, and MAE denotes mean absolute error in years. Age MAE is calculated using the midpoint of each age group, with ages greater than 70 marked as 75.

	Gender Acc	Race Acc	Age Group Acc	Age MAE
MinusFace	92.52%	64.97%	54.43%	5.28
PartialFace	93.49%	68.64%	56.55%	4.92
DecoyFace	89.39%	62.20%	53.43%	5.66
Raw Images	94.29%	71.28%	60.30%	4.40

D Few-Sample Settings

We also evaluate attacks in a few-sample regime, where the adversary can acquire only a limited number of training pairs. Prior studies such as PerceptFace (Wang et al., 2025a) observe that inversion is possible with large paired datasets and propose rate limiting as a countermeasure. We therefore investigate whether attacks succeed under strict data budgets. We train both distillation and U-Net models with a learning rate of $5e{-}5$, since both are fine-tuned, across N identities with 2 images per identity, evaluating $N \in \{256, 512, 1024, 2048, 4096, 8192\}$ for PPFR and $N \in \{32, 64, 128, 256, 512, 1024\}$ for De-ID. These identity counts reflect realistic constraints: PPFR service providers typically have broader but still limited query access, whereas De-ID adversaries are usually outsiders operating under tighter query budgets.

Because the distillation student is pre-trained for face recognition, comparing it against a U-Net trained from scratch would disadvantage the U-Net in data-limited regimes. To ensure a balanced comparison, we pre-train the U-Net on clean faces using an autoencoder reconstruction loss:

$$\mathcal{L} = \|x - \mathcal{U}(x)\|_1.$$

We temporarily omit skip connections during initial pre-training to prevent trivial identity shortcuts. The U-Net is trained on the full MS1M dataset, as this pre-training does not use protected data pairs, using AdamW with a learning rate of $1e{-}3$ until validation loss plateaus, after which skip connections are re-enabled and training continues until convergence. This pre-trained model initializes all few-sample U-Net experiments.

Figure 6 plots the attack success rate against the number of training identities. For most De-ID methods, the distillation attack achieves substantially higher ASR than the U-Net does. On WDP, U-Net yields an ASR of nearly zero, whereas distillation reaches 0.40. TIP-IM is the only exception, for which U-Net outperforms distillation, as TIP-IM's protection is weak: its simple additive perturbations are readily removed by standard denoising. Notably, distillation on TIP-IM follows roughly the same curve as on other methods. For PPFR methods, both U-Net and distillation can successfully attack PartialFace with relatively high ASR even with few training identities. For Mi-

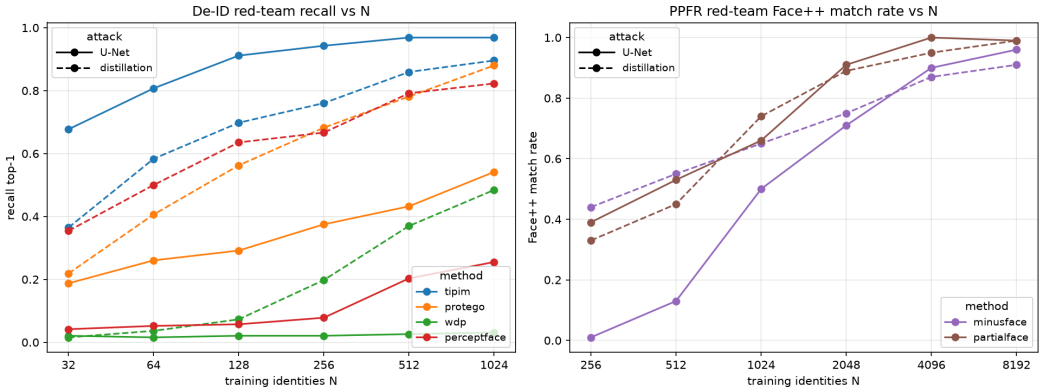

Figure 6: Few-sample attack success rate as a function of N. The De-ID ASR is the average Recall@1 across the one-side-protected and both-sides-protected settings, and the PPFR attack success rate is defined as the rate of successful verification by Face++. N refers to the number of identities.

Table 6: PPFR attacks on 200 aligned synthetic faces from the TPDNE dataset. Each reconstructed image is compared with the original cropped input from which the protected representation was generated. Amazon and Face++ report acceptance rates; AdaFace reports mean cosine similarity.

Attack	PPFR method	Amazon	Face++	AdaFace Sim
AdamW+IN U-Net	MinusFace	1.000	1.000	0.722
AdamW+IN U-Net	PartialFace	1.000	1.000	0.908
AdamW+IN U-Net	DecoyFace	0.005	0.015	0.060
Distillation	MinusFace	0.985	0.980	0.515
Distillation	PartialFace	1.000	1.000	0.638
Distillation	DecoyFace	0.415	0.660	0.333

nusFace, however, distillation maintains a much higher ASR than the U-Net does until both are nearly saturated at around 2,048 identities.

E Future Defense Directions

We suggest several potential directions for future PPFR design and evaluation, focusing mainly on stronger defense mechanisms and broader vulnerability assessments.

One rigorous approach is to incorporate secret keys into the conversion process, as in Yuan et al. (2022). This works as a multi-factor authentication system that requires both biometrics and a key, and it prevents attackers, including our method, from converting a face without the secret. A method could transmit the face embedding instead of the face itself, since there are many ways to incorporate keys into embeddings, such as rotation and dropout Zhong et al. (2024).

Alternatively, systems may return to formal cryptographic methods such as Ao & Boddeti (2025). These methods are traditionally viewed as computationally expensive, but modern computing resources make this trade-off acceptable. For instance, Jindal et al. (2020) reports a processing time of only 2.83 ms per face pair. Importantly, these computational costs can act as an effective client-side constraint against brute-force attacks, similar in concept to slow hashing. This enhances privacy while remaining imperceptible to regular users.

For De-ID tasks, although our evidence is not strong enough to show that retaining information for human perception while preventing model identity extraction is impossible, our results suggest that this is a likely possibility, and future work should weigh this trade-off carefully. Possible workarounds include using a similar key-based structure in which a person's face is altered to a dif-

ferent identity when posting on social media by a reversible transformation, and only trusted viewers can restore it to the original user's face when viewing.