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

Wenqi Guo, Mohamed S. Shehata, and Shan Du

Department of Computer Science, Mathematics, Physics and Statistics

University of British Columbia, Kelowna, BC, Canada

{wg25r, mohamed.sami.shehata, shan.du}@ubc.ca

ABSTRACT

Privacy-preserving face recognition (PPFR) hides facial appearance while retaining machine recognition, whereas perception-preserving face de-identification (De-ID) retains human recognizability while blocking face recognition models. We show that both designs preserve identity signals that an adaptive attacker can extract. We introduce FaceLinkGen, a unified distillation attack that learns from paired protected and original images. Across MinusFace, PartialFace, and DecoyFace, FaceLinkGen regenerates faces accepted by Face++ at rates of 81.0–99.0% and by Amazon at rates of 74.9–99.2%. Against four De-ID methods, it raises mixed-gallery Recall@1 to 55.2–89.6% and remains effective with limited paired data. These results establish adaptive evaluation as essential for measuring facial privacy and show that resistance to fixed recognition or reconstruction models does not prevent identity leakage.

1 INTRODUCTION

Facial privacy research commonly considers two tasks: 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). Perception-preserving De-ID pursues an almost 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) and MinusFace (Mi et al., 2024), 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), the synthesis-based PerceptFace (Wang et al., 2025a), Protego (Wang et al., 2025b), and the diffusion-based *Weakening Diffusion Purification* (Salar et al., 2025), termed WDP in this paper.

Despite their opposite goals, both tasks face the same fundamental tension: privacy-sensitive 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 adaptive server with query access has a strong opportunity to learn how 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 from which an adaptive FR model may learn.

Based on this observation, we propose FaceLinkGen, a unified distillation-based attack framework for both PPFR and De-ID systems. Our results show that the vulnerability is not merely theoretical. For PPFR, FaceLinkGen regenerates faces with acceptance rates of 81.0–99.0% on Face++ and 74.9–99.2% on Amazon across MinusFace, PartialFace, and DecoyFace. For De-ID, FaceLinkGen achieves 48.4–89.6% Recall@1 when matching a protected query against an unprotected gallery across the evaluated methods, demonstrating that identity linkage remains practical after protection.

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 evaluate FaceLinkGen against representative methods from both families and show that substantial identity information survives their protection mechanisms.

2 RELATED WORK

2.1 PRIVACY-PRESERVING FACE RECOGNITION

Most previous PPFR methods operate in the frequency domain. Their main underlying assumption is that machines recognize faces using high-frequency information, whereas the visual information perceived by humans lies in the low-frequency domain. Thus, they argue that the original image cannot be reconstructed after the low-frequency information is removed and the frequency channels are disrupted. PartialFace first removes the low-frequency information and then shuffles the high-frequency information. MinusFace uses an autoencoder-style network to reconstruct the original image and computes the residual between the original and reconstructed images. This residual contains very little information, which is concentrated primarily in high frequencies, and it is shuffled again in the frequency domain and converted back to the spatial domain. FracFace disrupts the frequency space using a fractal-style transformation. In contrast, DecoyFace uses a decoy-based method that separates reconstruction-sensitive information from recognition-sensitive information and replaces the reconstruction-sensitive information with another person’s face. Consequently, when the reconstruction network is trained, it recovers a face from the decoy rather than the protected identity.

2.2 FACE ANONYMIZATION

There are two types of face anonymization/face de-identification (De-ID): utility-preserving and non-utility-preserving. The simplest non-utility-preserving method is to blur the user’s face. A more advanced non-utility-preserving face De-ID could be understood as simply a “face swap,” in which the pipeline uses a different face to replace the user’s existing face to completely protect the user’s privacy. This is useful for goals like dataset sharing and non-main-character anonymization (e.g., protecting passersby’s privacy). Utility-preserving De-ID methods usually aim to make the face still look like the user to human eyes while avoiding detection by FR models. One application is for people who want to post their photos on social media and remain recognizable to themselves and their families, but do not want their images to be scraped by bots to build a face search engine. We target the utility-preserving type in this paper, and it will be referred to directly as “face De-ID.”

There are two main types of such De-ID systems: adversarial perturbation-based methods and perception-preserving synthesis-based methods. Adversarial perturbation-based methods optimize the input such that it finds a weak corner of the FR model and thus bypasses its recognition. However, these methods are usually weak and optimized for a specific model, and they also add unnatural artifacts to the image. Perception-preserving synthesis-based methods like PerceptFace generate a different identity but aim to keep the changes minimal such that it still looks like the same person to a human.

2.3 PREVIOUS ATTACKS

One important work that presents an argument similar to ours is by Calmon et al. (2015). It shows that perfect privacy is possible only if the information used for utility is decoupled from private information. This is closely related to our attack, which relies on the fact that human perception and machine recognition use the same or heavily shared fundamental information.

Recent work has also begun to challenge the reconstruction-centric perspective in non-facial domains by shifting attention toward semantic-level inversion. In such settings, the attacker aims not to recover the original image itself, but to regenerate information that is semantically consistent with the original identity. This goal is both easier to achieve and better aligned with realistic attack objectives. For instance, Yue et al. (2023) proposes a semantic recovery framework that leverages generative models to synthesize semantically consistent images without pixel-level similarity in federated

learning attacks. Complementary findings further suggest that traditional reconstruction metrics fail to capture how humans perceive privacy leakage (Sun et al., 2023).

3 THREAT MODEL

PPFR. For PPFR systems, the adversary is the service provider itself, similar to backdoor attacks. This follows the original motivation of PPFR (Erkin et al., 2009): the user wants the server to verify their identity without ever seeing their real face. We model the server as honest-but-curious (HBC), as in DecoyFace (Ren et al., 2026) and the original PPFR paper (Erkin et al., 2009): it follows the protocol faithfully but tries to learn as much as it can from the templates it receives. The adversary has black-box oracle access to the protection function: it can submit an original image and obtain the corresponding protected template, but it has no knowledge of the method’s internals, and any internal randomness of the method is freshly sampled per query and unknown to the 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 full setting, the adversary can query the oracle without limit and build a large paired dataset of original images and templates. This could happen if the protection method is developed by the system provider itself or is open-source. In the few-sample setting, the adversary can obtain far fewer identities (256–8192), which tests how well the attack works when the paired data are scarce in settings in which the protection method is held by a third-party protection service or is a closed-source standardized protection method.

De-ID. For De-ID systems, the adversary is a web scraper that collects online images, extracts their embeddings into a database, and lets people search the database for similar faces. The adversary’s goal is to link a face to other images of the same person, whether the images are protected or not. We used only 2,048 training images from 1,024 identities for De-ID systems, as it is more time-consuming to generate more samples for attackers because many methods require per-sample optimization. We also tested the few-sample setting for De-ID with only 32–1,024 identities.

Both of these settings match Kerckhoffs’s principle, where the security of the system should depend on the protection of the key (the random component in the system) and not the algorithm itself. Even without official access to the system, the attacker can implement a proxy system themselves that uses the same method as the deployed system, only without the secret randomness. Additionally, in realistic settings, the attacker is actually stronger than our threat model assumes: since both PPFR and De-ID systems likely run locally on the user’s device for the purpose of privacy, an attacker can query these systems locally without limits.

4 FACELINKGEN

4.1 BACKGROUND

The simplicity of our method is intentional. We show that even strong protection methods fail under a standard distillation process, which indicates that the vulnerability lies in the representation itself rather than in any particular implementation.

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 or cannot be reconstructed back to x . De-ID methods aim for the opposite: to remove or hide z_m while retaining z_p , so that the image still looks natural and 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, they are strongly correlated, and this correlation is what our method exploits.

Consider PPFR first. Methods such as Arc2Face (Papantoniou et al., 2024) and FaceID IP-Adapter (Ye et al., 2023) regenerate a face from z_m (face embedding) alone. The fact that this path works at all is the point: z_m carries enough identity information to render a face that a human can recognize. The PPFR transformation therefore did not discard z_p ; it moved the identity into a domain that an off-the-shelf FR model no longer reads, while leaving it recoverable. Even in a

hypothetical and unlikely situation where it is possible to extract another set of z_m (non-standard facial embeddings) for verification that does not contain any visual information, the mapping is still 1-to-1 and could likely be learned in reverse.

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, which is similar to adversarial training.

The second route (PerceptFace) applies a transformation that replaces the z_m in x with someone else’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 from z_p , then z_p retains enough of the original identity for a model to learn the mapping back to z_m . The perceptual fidelity that makes De-ID practical is the same property that leaves the identity signal in the image.

In all three cases, the protection holds only with respect to a fixed model. Nothing about the transformation removes identity from the image; it only places identity where the current recognition pipeline cannot follow. Once the model is allowed to adapt, the correlation between z_m and z_p is sufficient to undo the protection.

4.2 METHODS

Our attack uses a pre-trained ArcFace face embedding model to initialize a student model s and a teacher model t . They are initialized with the same weights. The teacher model remains frozen throughout training, while only the student model is trained. To make the student extract the same identity information as the teacher, whose ArcFace embedding serves as the ground truth, we use a simple cosine-similarity loss, $\mathcal{L}_{distill}$. Here, x is the original face and $\mathcal{P}$ denotes any PPFR/De-ID protection method.

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

This is enough to attack PPFR systems because the goal is to extract the information from the image to reconstruct it, which depends on a single image. However, for De-ID systems, we also need a cross-image loss. The facial embedding $t(x)$ contains not only identity-related information but also information about lighting, pose, and background. To prevent the student from learning shortcuts based on such information and degrading cross-image linkage, we add a cross-image distillation loss. We combine the within-image and cross-image distillation terms into a single identity loss, $\mathcal{L}_{identity}$. This loss matches each protected image in a pair with both teacher embeddings.

$$\mathcal{L}_{identity} = 1 - \frac{1}{4} \sum_{x \in \{x_0, x_1\}} \sum_{x' \in \{x_0, x_1\}} \cos(s(\mathcal{P}(x)), t(x'))$$

For De-ID systems, the attacker also encounters a mixture of protected and unprotected images. Therefore, the student model must also recognize original faces. To achieve this, we add an original-domain-preserving loss, $\mathcal{L}_{preserving}$, only for De-ID attacks. Given a clean, unprotected image, the student model is trained to produce the same embedding as the teacher.

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

The final training loss for PPFR is $\mathcal{L}_{distill}$, and the final training loss for De-ID systems combines $\mathcal{L}_{identity}$ and $\mathcal{L}_{preserving}$ with a simple sum, $\mathcal{L}_{deid}$. Since the preserving loss only acts as a regularization term to maintain the model’s existing capabilities, we give it a lower weight. We set α to 0.8 in this paper.

$$\mathcal{L}_{deid} = \alpha \cdot \mathcal{L}_{identity} + (1 - \alpha) \cdot \mathcal{L}_{preserving}$$

4.3 METRICS

For PPFR systems, since the goal is to recover the original human-perceivable information, we used Arc2Face (Papantoniou et al., 2024) to convert the distilled embeddings back to original faces. To

validate the generated faces, we verified them against the original faces at the strictest threshold using both Face++ ($\text{FAR}@1e-5$) and the Amazon ($\text{Conf} > 80$) Face API, two APIs popular in the face privacy research community. We also report the AdaFace cosine similarity between the original and regenerated faces. We also present *unselected* examples as qualitative results.

For De-ID systems, since the goal is to prevent faces from being searched for, we tested the retrieval performance on a closed-set face database with two settings: both the gallery and the queries are protected (the setting that is harder for the defender per (Wang et al., 2025b)) and only one random side is protected (the setting that is harder for the attacker per (Wang et al., 2025b)). We report the Recall@1 and Recall@5 metrics.

Metric	Reproduced Result	Original Paper Result
Cosine similarity ($\downarrow$)	0.081	0.027
Identity Leakage Ratio ($\downarrow$)	2.61%	2.93%
Identity Redirection Ratio ($\uparrow$)	97.39%	97.07%
Face Validity Ratio ($\uparrow$)	99.99%	99.88%

Table 1: Reproduction results compared with the original paper for DecoyFace.

4.4 SETTINGS AND BASELINE

In this paper, we evaluated PPFR methods including MinusFace (Mi et al., 2024), PartialFace (Mi et al., 2023), and DecoyFace (Ren et al., 2026). MinusFace and PartialFace are frequency-based, and DecoyFace is decoy-based. We excluded FracFace (Dai et al., 2025) because there are some inconsistencies between the code and the paper, and neither their code nor our reproduced code can replicate its results. For both MinusFace and PartialFace, we used the official implementation from Tencent/TFace. For DecoyFace, since no code has been published, we reproduced and verified it using Paper-Replication (Hans & Bilonis, 2026) against the paper. We verified the key metrics of our reproduction against the paper; the verification results are in Table 1. All of our metrics are very close to the original numbers reported in the paper. Our implementation is better in 3 out of 4 metrics, and all metrics remain in the region of successful protection. For De-ID systems, we evaluated the perturbation-based methods TIP-IM (Yang et al., 2021) and Protego (Wang et al., 2025b), the perception-based method PerceptFace (Wang et al., 2025a), and the diffusion-based method Weakening Diffusion Purification (WDP) (Salar et al., 2025). We did not use many recent methods like CLIP2Protect (Shamshad et al., 2023), as they conflicted with the goal of dodging impersonation. These methods can successfully trick the FR model into recognizing the “protected” image as someone else, but they cannot prevent the FR model from still recognizing the person as themselves. This offers little privacy protection. We also evaluated a method called CanFG (Wang et al., 2024) that sits between PPFR and De-ID methods. It claims to reach both goals by converting someone’s physical identity to a virtual identity. Because it is difficult to deduplicate identities across datasets, we used a single dataset, MS1M (Guo et al., 2016), for training, validation, and testing by splitting it within the dataset based on identity labels. For PPFR, we used 30,000 identities for training, 2,000 identities for validation, and 2,000 identities for testing. For De-ID systems, since many methods use per-sample optimization and it is expensive to collect a large number of samples (both in our experiment and for real attackers), we used only 2,048 training images from 1,024 identities, 128 validation images, and 384 testing images. For the recall metrics for De-ID systems, we used closed-set retrieval with 192 query images and 192 target images, along with another 107,484 images from the unprotected image set as distractors. This simulates the real Internet environment, where the images of only a small number of people are protected.



Figure 1: U-Net attack results for PartialFace (top row) and MinusFace (bottom row). 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 for longer, and AdamW+IN training.

As a baseline, we used the U-Net (Ronneberger et al., 2015) to convert the protected image back into the original image, essentially fitting the U-Net to be an inverse function of the protection method.

This is the attack tested on PPFR methods like MinusFace, PartialFace, FracFace, and DecoyFace. We tested it on both PPFR and De-ID systems. The U-Net is trained using $\mathcal{L}_{ppfr-unet}$, where $\mathcal{U}$ is the U-Net.

$$\mathcal{L}_{ppfr-unet} = \|x - \mathcal{U}(\mathcal{P}(x))\|_1$$

For methods besides DecoyFace, we used the same U-Net that the authors used in the MinusFace attack, as well as a version with instance normalization (Ulyanov et al., 2017). For PPFR, we trained the U-Net both using our hyperparameter-tuned settings and using the original MinusFace attack settings (a learning rate of $1e-3$ and weight decay of $1e-4$ with the SGD optimizer); these settings are used to reproduce the original “protected” results and the results under better-tuned attacker models. For De-ID systems, we also included a preserving loss for the same reason that the attacker faces a mixture of original and protected images. The final loss for the De-ID attack using U-Net is $\mathcal{L}_{deid-unet}$.

$$\mathcal{L}_{deid-unet} = \beta \cdot \|x - \mathcal{U}(\mathcal{P}(x))\|_1 + (1 - \beta) \cdot \|x - \mathcal{U}(x)\|_1$$

We used $\beta = 0.8$ in this paper.

4.5 SOFT BIOMETRIC EXTRACTION

Beyond hard identity recognition, the 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 applies only to PPFR, as these systems are meant to allow the system to identify the person without compromising their privacy. De-ID systems purposely maintain human perception, and thus retaining soft biometrics is intended. Privacy frameworks like the Canadian Privacy Act (Branch, 2025) explicitly protect race and age. A robust privacy-preserving system must therefore prevent the recovery of these attributes from its templates.

Our empirical results for generated images demonstrate that these soft biometrics remain visible in the regenerated faces. This matches previous research, which indicates 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 it 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. We also trained the same model architecture using an untrained teacher model as the base model on raw images to determine the baseline for extracting this soft biometric information from raw images. The results are reported in Table 2.

4.6 IMPLEMENTATION DETAILS

Each protection method produces a different form of output. All De-ID systems produce regular RGB images, whereas PPFR methods usually produce arrays with different shapes. MinusFace, FracFace, and PartialFace all 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 use a single convolutional adapter to adapt the channels and bilinear interpolation to adapt the spatial dimensions. In our setting, since the student model and the DecoyFace model use the same Res-100 (He et al., 2015) architecture, they have the same shape, and this layer is only for compatibility if the model changes. Note that even if they have the same architecture, they are from different model families (AdaFace (Kim et al., 2023) vs. ArcFace (Deng et al., 2022)) to avoid circularity and an artificial boost in attack performance. 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 curves

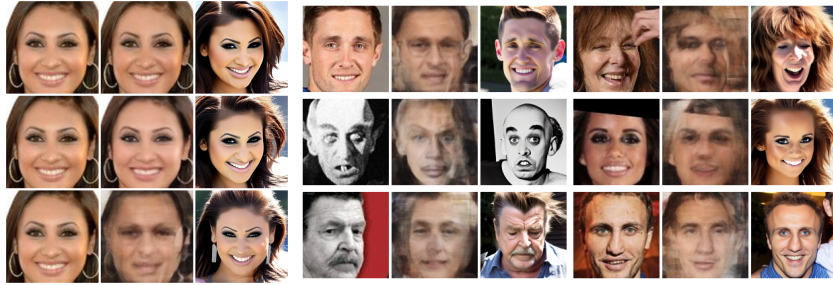


Figure 2: Left: Reconstructions produced by each attack. Rows correspond to PartialFace, MinusFace, and DecoyFace; columns show the original image, the U-Net attack, and the distillation attack. Right: More examples of reconstruction comparisons between U-Net and distillation for DecoyFace.

back (indicating overfitting), and the checkpoints with the lowest validation loss are selected due to the smaller sample size and more unstable training.

5 RESULTS

5.1 PPFR SYSTEMS

Revisiting the U-Net Attack We first revisit the U-Net attack against PartialFace and MinusFace. Although both papers report that their protected representations resist U-Net inversion, our U-Net recovers clean faces from them (the rightmost subplot in Figure 1). Figure 2 shows reconstruction comparisons between U-Net and distillation for DecoyFace. We looked into this gap and found that small implementation choices strongly affect how well the attack works. With the original U-Net used in MinusFace (no normalization layers) and the SGD optimizer (Sutskever et al., 2013), the attack produces a uniform color image. This differs from the blurry faces reported in the original papers, but in both cases the inversion clearly fails. Adding instance normalization improves training considerably: at earlier training steps, coarse facial structure appears, though the faces remain blurry and unrecognizable, closely matching the reported results, and at later training steps, the face is clearer but the details are still hard to recognize. Switching to AdamW on top of instance normalization makes the reconstruction cleanly successful. This suggests that prior work may not have trained the attacking U-Net to convergence, giving a false sense of security against U-Net inversion. For DecoyFace, however, although a U-Net can produce some blurry images of someone, that person has a different identity from the person in the original face image. This matches the reported results in the paper.

The full visual comparison of the better U-Net (AdamW + IN) and the distillation attack is shown on the left in Figure 2, with the DecoyFace comparison against the U-Net attack shown on the right. In Table 3, we show the attack success rate (ASR) verified

Table 2: Soft biometric recovery rate from the protected template without regeneration. 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

by Face++ and Amazon, as well as the similarity of the generated face to the original face. Note that Face++ and Amazon are not our attack targets; they merely serve as verifiers of face similarity. From the table, we can see that the U-Net attack failed on all methods when trained using the setting in MinusFace (SGD optimizer with no normalization). However, after switching the optimizer to AdamW and adding instance normalization, the attack is almost perfect on MinusFace and PartialFace. The distillation attack on both MinusFace and PartialFace is also highly successful; its ASR is slightly lower than that of the U-Net attack because the distillation attack is a semantic reconstruction that generates another image of the same person, unlike the U-Net attack, which recovers the original image. The semantic reconstruction adds more variation and noise, just as using another

Table 3: PPFR method results under U-Net and distillation attacks.

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++	0.972	0.990	0.810
	Amazon	0.960	0.992	0.749
	AdaFace Sim	0.538	0.663	0.457

image of the same person yields lower face similarity and pass rate. This does not mean the attack is poorer; the difference is merely due to the introduced variation.

On DecoyFace, both U-Net attacks failed to pass the face verification test. Looking at the examples, we can see that the U-Nets for DecoyFace do generate fairly realistic faces, but they do not depict the person in the original image, which is the intent of DecoyFace. However, under our distillation attack, the attack rate rose to around 70–80%, and most of the identities were recovered. This shows that even though the decoy can mislead the U-Net reconstruction, it preserves enough identity information for verification that it can also be extracted and used to regenerate faces.

5.2 DE-ID SYSTEMS

The main attack results against the De-ID systems are reported in Table 4. Beyond these, we observe an interesting and somewhat ironic effect: using one protected image to retrieve *another* protected image of the same person is highly effective—for PerceptFace, its Recall@1 is above 90%. We attribute this to two factors. First, because the gallery is dominated by unprotected images (as on the real Internet, where only a small minority protect their photos), protecting an image pushes its embedding toward a distinctive region rather than hiding it in the unprotected crowd, making protected images stand out and cluster together. We plotted (Figure 4 in Appendix) the t-SNE of embeddings of unprotected images and protected images and observed that protection methods besides Protego yield clustered embeddings. This corresponds to the high both-sides-protected recall for all methods besides Protego. Second, many of these methods merely transform identity rather than destroy it, effectively assigning each person a stable pseudonym that remains consistent across their images and is therefore matchable. This mirrors classic linkage vulnerabilities: using a virtual but stable MAC address to avoid device tracking, or applying unsalted hashing to “protect” linkable fields. In each case, the original value is hidden from the attacker, yet the substituted value is deterministic, so records remain linkable even though the true identity is never recovered.

For the attack results shown in Table 4, 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. On certain methods such as WDP and Protego, however, the U-Net fails to recover the identity while distillation still succeeds. This holds in both the one-side-protected and both-sides-protected settings; notably, the U-Net attack even makes the originally vulnerable both-sides-protected scenario harder to attack. We also report a “Mixed” score that averages the two settings, since in practice an attacker’s gallery will contain an unpredictable mixture of protected and unprotected images of each person.

6 FEW-SAMPLE SETTINGS

For both PPFR and De-ID methods, we also tested a few-sample setting. This is used to simulate a setting where the attacker can obtain only a small number of training pairs. This could be a result of the protection method being held by a third-party protection service or being a closed-source standardized protection method. The few-sample setting is also tested because PerceptFace (Wang et al., 2025a) and CanFG (Wang et al., 2024) acknowledged that reversion is possible with

Table 4: Attack results on the De-ID systems: recall on the 192-ID test pool (gallery size 107,676). “Mixed” averages the one-side and both-sides settings, reflecting a realistic scenario where protected and unprotected images co-exist 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
	Mixed	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
	Mixed	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
	Mixed	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
	Mixed	0.445	0.576	0.511	0.896	0.938	0.917	0.026	0.036	0.033

a large number of paired samples and proposed to use rate limiting to mitigate this. We thus test whether the attack would be successful with only a limited number of samples. In this section, we used a slightly lower learning rate of $5e-5$ for both distillation and U-Net. We tested both PPFR and De-ID systems against a limited number, N , of identities and 2 images per identity, where $N \in \{256, 512, 1024, 2048, 4096, 8192\}$ for PPFR systems and $N \in \{32, 64, 128, 256, 512, 1024\}$ for De-ID systems. These numbers are chosen because the main adversary in the PPFR systems is still the service provider, which could have high but limited access to the method, while the adversaries of De-ID systems are usually outsiders whose access could be further limited.

Since the student model is pre-trained to recognize faces, it would be unfair to compare a U-Net trained from scratch in few-sample settings. Thus, we pre-train the U-Net with an autoencoder-style training objective, $\mathcal{L}$, using regular clean faces.

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

The skip connections of the U-Net are removed to avoid the model learning shortcuts. The U-Net is trained with the full MS1M dataset, as it does not use protected data pairs, with the AdamW optimizer and a learning rate of $1e-3$ until the validation loss plateaus. Then, the skip connections are added back, and it is trained until the loss plateaus again. The same pre-trained U-Net is used for all downstream tasks.

In this setting, since both U-Net and distillation are fine-tuned, we used a learning rate of $5e-5$ for both. Figure 3 plots the attack success rate against sample size. 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 0.4. TIP-IM is the only exception, for which U-Net outperforms distillation. This is because TIP-IM’s protection is too weak, allowing U-Net to easily purify the perturbations. Notably, distillation on TIP-IM follows roughly the same curve as on other methods. For PPFR methods, both U-Net and distillation can penetrate PartialFace with relatively high ASR even at low sample counts. For MinusFace, however, distillation maintains a much higher ASR than the U-Net does until both are nearly saturated at around 2048 samples.

7 CONCLUSION

We presented FaceLinkGen, a unified distillation attack against keyless PPFR and perception-preserving face De-ID systems. The attack exploits the identity signal retained to support recognition

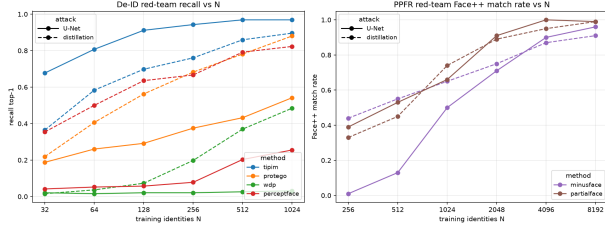


Figure 3: Few-sample attack success rate across N . The De-ID ASR is defined as Recall@1, and the PPFR attack success rate is defined as the rate of successful verification by Face++. N refers to the number of identities.

or visual fidelity and maps protected inputs to standard identity embeddings. Across the evaluated PPFR systems, FaceLinkGen regenerated faces with high acceptance rates, including DecoyFace, where direct U-Net reconstruction produced the decoy identity. Across four De-ID systems, the adapted recognizer recovered strong linkage performance in galleries containing protected and unprotected images. The few-sample results further show that this leakage can be learned from limited paired data. These findings make adaptive attacks a necessary part of privacy evaluation: protection against a fixed recognizer or reconstruction network is insufficient when the protected representation continues to encode identity.

REFERENCES

- Legislative Services Branch. Consolidated federal laws of Canada, Privacy Act, June 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.
- 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 Bilonis. 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.
- 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.
- Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019.
- Pietro Melzi, Hatem Othrosi 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, Lin Yuan, Zhihua Xia, and Jian Weng. Make Privacy Renewable! Generating Privacy-Preserving Faces Supporting Cancelable Biometric Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10268–10276, Melbourne VIC Australia, October 2024. ACM. ISBN 979-8-4007-0686-8. doi: 10.1145/3664647.3680704.
- 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, Sibio Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, August 2023.
- 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.

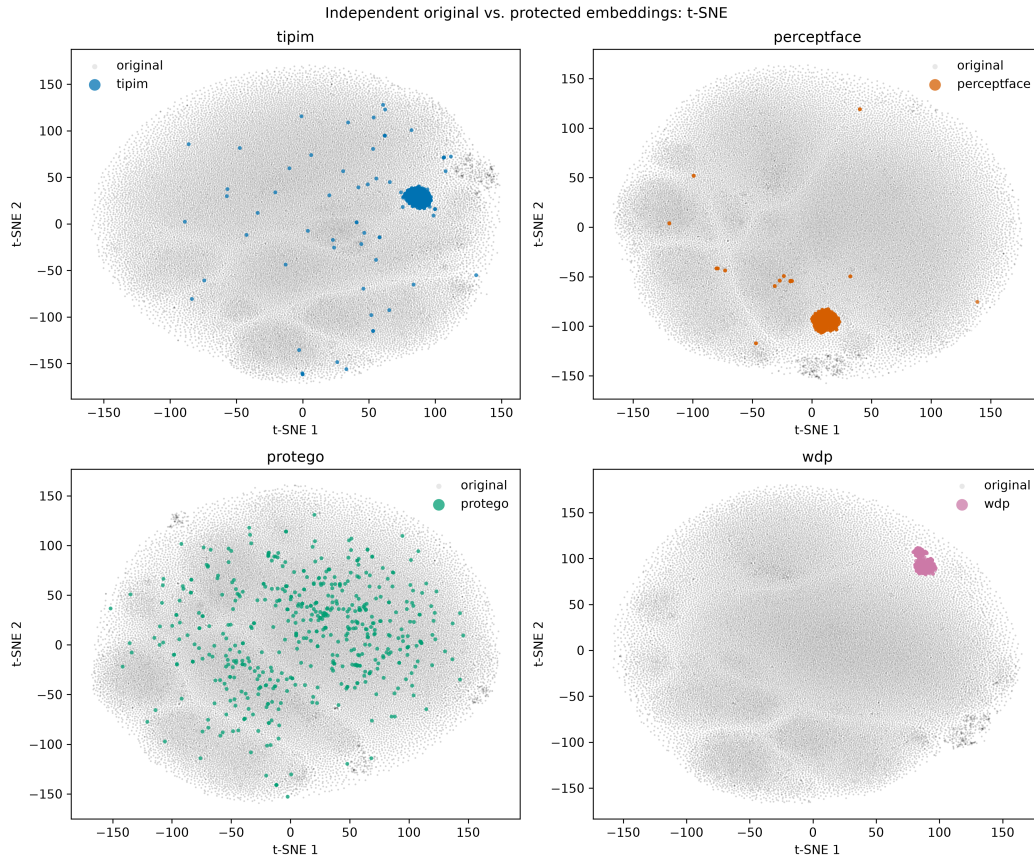


Figure 4: t-SNE plot of each method's embeddings compared to the original embeddings.