# DiffPrivate: Facial Privacy Protection with Diffusion Models

Minh-Ha Le
Linköping University
Linköping, Sweden

Niklas Carlsson
Linköping University
Linköping, Sweden

## Abstract

The widespread use of facial recognition (FR) technology has heightened concerns about personal privacy. With surveillance systems becoming ubiquitous, the demand for effective privacy-enhancing technologies is growing urgent. In response to this challenge, we introduce DiffPrivate, a versatile technique designed to protect individuals from FR systems (FRS) through two distinct approaches: a Perturb-based and an Edit-based approach. The Perturb-based mode generates robust adversarial samples by manipulating the diffusion process of a latent diffusion model to alter identity-specific features, ensuring the preservation of visual fidelity to the original images. On the other hand, the Edit-based approach employs an additional DDPM model for fine-grain editing of attributes, allowing for more precise control over the appearance while subtly shifting the identity features to evade FRS. By leveraging the strengths of both modes, DiffPrivate effectively shields an individual's identity against advanced defense mechanisms like DiffPure, maintaining high image quality. Our experiments demonstrate that DiffPrivate achieves competitive attack performance in terms of success rates and transferability while producing more natural-looking adversarial images than state-of-the-art methods. Overall, DiffPrivate represents a significant step towards balancing personal privacy and image naturalness in the face of advancing FR technology.

## 1 Introduction

The combination of vast availability of photos on social media and the increasing integration of surveillance systems is redefining personal privacy. At the core of this transformation is the widespread adoption of facial recognition (FR) technology, facilitated by significant advancements in deep learning and neural networks [9, 18, 48]. FR systems (FRS) like Amazon Rekognition [1], Face++ [13], and Clearview.ai [19] already exhibit remarkable accuracy in identifying individuals from vast online image repositories. As this technology continues to advance and proliferate, the scrutiny of individuals—both online and offline—will reach unprecedented levels.

With vast amounts of facial data being harvested across various platforms, often without individual knowledge or explicit consent, this threat is quickly growing. As FRS are increasingly deployed, virtually every aspect of our daily life – from routine shopping excursions [36] to international travel [41] – will therefore quickly become susceptible to monitoring and tracking. The potential misuse of FR technology in contexts such as stalking [52], identity theft [8], and clandestine governmental surveillance [19, 38, 47] raises further serious concerns about personal security and autonomy

Although potential future regulations may address some concerns, the current trends underscore the need for privacy-enhancing technologies to safeguard individual privacy. Driven by this observation, there is a growing interest in designing privacy-enhancing solutions that utilize adversarial machine learning to create images that deceive various FRS. In this paper, we focus on adversarial attacks that fool automated FRS but still produce high-utility images.

**Adversarial Attacks as a Defense:** Recent advancements in countering FRS have primarily centered around the development of adversarial attacks to modify face images. Of these, early noise-based techniques [6, 66] often suffer from compromised visual quality, making their modifications glaringly evident. While Fawkes system achieved notable success in protecting privacy by adding subtle modifications, also this strategy struggled with image quality preservation [50]. Other strategies, including patch-based methods [29, 51] and image distortion techniques [33, 56, 64], anonymization [30, 32] though promising, were limited in practicality and effectiveness. Additionally, targeted attacks such as clean-label poison attacks [49, 70] have been explored, focusing on manipulating specific images using image classification models.

**GAN-Based Solutions:** Recent advancements have leveraged Generative Adversarial Networks (GANs) [14, 28] to balance privacy protection with retaining non-identifying features. Approaches like makeup transfer [21, 67] and facial attribute manipulation [25, 44] enable subtle image modifications. While many early GAN-based methods maintain some visual appeal, they exhibit relatively low attack success rates and require retraining for new targets. To address these limitations, StyleAdv [31] recently combined semantic editing with adversarial attacks, leveraging StyleGAN's disentangled latent space [28] for high-quality image reconstruction and improved success rates. Despite these advancements, there has been almost no work using diffusion models for adversarial attacks against FRS. In this work, we aim to design such solutions and explore to what extent such solutions may help address the GAN-based methods vulnerability against sophisticated defense techniques like DiffPure [40] or to increase the transferability of the attacks. Notably, most GAN-based methods are designed to attack a particular model (e.g., whitebox attacks that are not very transferable) or are ineffective against FRS using DiffPure.

**Our Diffusion-Based Solution Approach:** To address the above research gap, we introduce DiffPrivate. This method generates robust adversarial samples using diffusion models, improving upon the limitations of previous methods like StyleAdv. We originally chose diffusion models for their generative flexibility, which are ideal for complex, high-quality manipulations needed for privacy-preserving facial transformations. However, our results show that Stable Diffusion also outperforms architectures like StyleGAN in reconstructing finer details. Furthermore, the sequential nature of diffusion processes enables more controlled, gradual modifications, preserving the natural appearance while subtly altering


*Proceedings on Privacy Enhancing Technologies 2025(2), 1–17*

https://doi.org/XXXXXXX.XXXXXXX

identity-specific features. Combining these strengths of diffusion models allows us to enhance the effectiveness of privacy protection.

DiffPrivate includes two main approaches: one based on DiffAE [43] (Edit-based) and another on Stable Diffusion [46] (Perturb-based). Although these approaches vary, both start by projecting real facial images into a diffusion model's latent space.

In the Perturb-based approach, DiffPrivate modifies the diffusion process by working with latent codes collected at different steps. It uses optimized text embeddings to guide the cloaking of faces, aiming to maintain the original's appearance. This approach identifies and obscures identity features through "cross-attention maps," targeting specific areas for gradual adjustment over multiple diffusion steps to preserve image quality. The Edit-based approach uses a specialized DiffAE model for detailed editing of facial attributes, allowing for precise adjustments that distance the identity from being recognized by FRS. This approach focuses on direct edits to maintain the natural look of the image.

Both approaches aim to create images that can evade FRS while preserving human likeness and identity resemblance. Our tests show that DiffPrivate effectively resists FRS and safeguard visual integrity, often enhancing the aesthetic quality similar to a "beauty filter". This dual functionality does not compromise the subject's identity, aligning with the vital importance of not altering facial features in ways that subjects may perceive negatively. Consequently, DiffPrivate signifies an advancement in reconciling the necessity for natural-looking images with privacy protections against FRS.

Our main contributions are summarized as follows:

- We introduce DiffPrivate, a strong adversarial attack against FRS that provides relatively strong protection, without sacrificing naturalness or image quality.
- While attacking the semantic latent space using a conditional diffusion model presents challenges, our series of targeted interventions in the diffusion process allows us to develop an effective new approach for face anonymization.
- We implement two complementing versions. Our DiffAE version achieves high success rate using attribute-based edits to push the identity toward a target, while our Stable Diffusion version performs more subtle perturbation-based edits.
- Through comparisons against other state-of-the-art attacks, DiffPrivate is shown to achieve competitive attack performance in terms of attack success rate and transferability.
- Our experiments demonstrate that DiffPrivate produces more natural-looking adversarial images than state-of-the-art methods and is considerably more robust against advanced FR defense methods such as DiffPure.

**Outline:** After giving an overview of related works (§2) and the models used (§3), we introduce DiffPrivate (§4), detail experiments (§5), and present results (§6-§9). We then provide an ablation study (§10), discuss limitations (§11), and conclude (§12).

## 2 Related Work

**Privacy Protection:** The development of strategies to undermine FRS has introduced both poisoning and evasion tactics. Poisoning attacks, as described by Shan et al. [50], involve altering images within the gallery sets of FRS. While effective, this method is impractical for individual users due to the complexity and access required. On the other hand, evasion attacks present a more user-friendly alternative, aiming to modify facial images to mislead FRS during their operational phase [6, 21, 29, 51, 66, 67]. Evasion methods, particularly those capable of blackbox attacks, are more accessible for protecting personal privacy [6, 21]. However, the effectiveness of current techniques varies, with some sacrificing visual quality for attack success or vice versa.

**Adversarial Attacks:** Traditional methods to generate adversarial examples have primarily focused on optimizing additive noise within the pixel domain. Goodfellow et al. [15] introduced the concept with the Fast Gradient Sign Method (FGSM), highlighting deep neural networks' vulnerability to perturbations. Madry et al. [35] enhanced model resilience through adversarial training with the Projected Gradient Descent (PGD) method, while Carlini and Wagner [3] developed a more sophisticated optimization-based approach, the CW attack, for effective perturbations. Further advancements included integrating momentum into adversarial example generation for overcoming local optima [10]. In general, Fawkes [50], LowKey [6], and other pixel perturbation methods like Ulixes [7], Face-Off [4] and FoggySight [12] introduce strategic noise to images, adding noise patterns to faces that, while effective, can be perceptually noticeable and thus not ideal for all scenarios. Others have created semantic adversarial examples that subtly alter image attributes to fool binary classifiers, offering insights into vulnerabilities of classifiers less complex than the methods used in FRS (like those we attack here) [26]. Recently, the focus has shifted to using generative models like Variational Autoencoders (VAEs) and GANs to create more realistic and semantically consistent adversarial examples. This new direction, as explored by Wong et al.[63], Xiao et al.[65], and Qiu et al. [44], significantly enhances the sophistication and effectiveness of adversarial attacks, showcasing the potential of generative models in this domain.

**GAN-Based Approaches:** GANs offer a promising approach for creating realistic adversarial images [21, 67]. These methods strike a balance between maintaining image quality and evading detection, making them an attractive option for evading FRS. Nonetheless, as FR technologies evolve, so too must these evasion techniques to ensure continued effectiveness.

Recent developments like AnonFACES [32] and StyleID [30] offer new ways to protect privacy. AnonFACES aims to keep images looking natural while hiding identities, using groups of similar images to maintain some level of uniqueness without revealing too much. StyleID uses GANs to change faces in a way that the person cannot be recognized but keeps important features. Adding to these ideas, StyleAdv [31] uses StyleGAN's advanced features to make high-quality images that are hard for FRS to identify. It keeps the naturalness of protected photos and is easy to use for editing and protecting privacy. However, its effectiveness partly relies on how well the StyleGAN encoder can recreate faces, which is still a hurdle. These steps forward highlight the ongoing work to find a balance between keeping images useful and high-quality while protecting privacy against more advanced recognition systems. In Sec. 6, we compare our performance against two state-of-the-art GAN-based approaches (StyleAdv [31], AMT-GAN [21]) and three perturbation-based approaches (CW [3], PGD [35], Fawkes [50]).

**Diffusion Models:** Diffusion models, a class of generative models, have gained attention for their ability to generate high-quality,

realistic images by learning to reverse a noise-adding process [20, 39, 53]. Initially introduced by Sohl-Dickstein et al. [53], these models generate images by gradually removing noise, simulating the reverse of a diffusion process. Ho et al. [20] further advanced diffusion models by simplifying their training and improving the quality of generated samples. Nichol and Dhariwal [39] extended these improvements, refining model architectures and training methods to produce even more realistic images. In addition to image synthesis, diffusion models have shown promise in adversarial machine learning, particularly in improving adversarial robustness [2, 16, 40, 45, 60]. Studies by Gowal et al. [16], Rebuffi et al. [45], and Wang et al. [60] have leveraged diffusion models to generate synthetic data for adversarial training, enhancing model resilience. Moreover, Nie et al. [40] and Carlini et al. [2] have explored the use of pretrained diffusion models for purifying input images from adversarial noise, providing both empirical and certified defenses. Despite these advancements, the potential of diffusion models in creating adversarial attacks has not been fully explored. This work aims to investigate this untapped area, contributing to the understanding of diffusion models' capabilities in adversarial contexts.

**Non-Digital Attacks:** While outside the scope of this paper, others have shown that adversarial examples also can be created in the real world. This includes patch-based methods such as Adv-Hat [29] and Adv-Glasses [51]. While such approaches are not always practical and visible patches attached on faces often are easily detected, others have shown that the use of sunglasses impair face identity recognition more significantly than face masks [42] but do not target protection against FRS.

# 3 Background

## 3.1 Diffusion Model

Diffusion models, a type of generative model, involve two key stages: (1) a *forward process* that transforms an input image $\mathbf{x}_0$ into a purely noisy state $\mathbf{x}_T$ over $T$ forward steps, and (2) a *reverse process* that reconstructs $\mathbf{x}_0$ from $\mathbf{x}_T$ through $T$ reverse steps.

During the forward process, Gaussian noise is incrementally added at each step $t$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\gamma_t\mathbf{x}_{t-1}, \delta_t\mathbf{I}), \tag{1}$$

where $\gamma_t$ is a scaling factor that diminishes the image's intensity, $\mathbf{I}$ is the identity matrix, and $\delta_t\mathbf{I}$ is the variance of the Gaussian noise added at step $t$. Combining the noise over the first $t$ steps, the distribution of the noisy image $\mathbf{x}_t$ at any step $t$ can then be expressed relative the original image $\mathbf{x}_0$ as follows:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\rho_t\mathbf{x}_0, (1-\rho_t)\mathbf{I}), \tag{2}$$

where $\rho_t = \prod_{i=1}^{t}(1-\delta_i)$. The diffusion model aims to learn the reverse distribution $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$, needed for recovering the original image from its noisy state. For small differences between steps, this reverse distribution can be approximated as:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(v_\phi(\mathbf{x}_t, t), \tau_t^2\mathbf{I}), \tag{3}$$

where $v_\phi(\mathbf{x}_t, t)$, often a neural network, predicts the mean of the reverse Gaussian distribution at each step, and $\tau_t^2$ is the variance. The neural network is trained using a loss function that compares the actual noise added at each step to the predicted noise:

$$\mathcal{L} = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0,\eta_t}\left[\|\eta_t - \eta_\phi(\mathbf{x}_t, t)\|^2\right], \tag{4}$$

where $\eta_t$ is the actual noise added to $\mathbf{x}_0$ to produce $\mathbf{x}_t$, and $\eta_\phi(\mathbf{x}_t, t)$ is the noise predicted by the model.

**DDIM Framework:** The Denoising Diffusion Implicit Model (DDIM) framework [54] introduced a deterministic reverse process. Here, the forward process is defined as:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\rho_{t-1}\mathbf{x}_0 + \sqrt{1-\rho_{t-1}}\frac{\mathbf{x}_t - \rho_t\mathbf{x}_0}{\sqrt{1-\rho_t}}, \mathbf{0}\right), \tag{5}$$

which dictates the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$, given the original image $\mathbf{x}_0$. Then, in the reverse process, DDIM first estimates $\mathbf{x}_0$ from the noisy image $\mathbf{x}_t$:

$$g_\phi(\mathbf{x}_t, t) = \frac{\mathbf{x}_t - \sqrt{1-\rho_t}\cdot\eta_\phi(\mathbf{x}_t, t)}{\sqrt{\rho_t}}, \tag{6}$$

and then defines the reverse transition using this estimate:

$$\mathbf{x}_{t-1} = \sqrt{\rho_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\rho_t}\eta_\phi(\mathbf{x}_t, t)}{\sqrt{\rho_t}}\right) + \sqrt{1-\rho_{t-1}}\eta_\phi(\mathbf{x}_t, t). \tag{7}$$

Importantly for our purposes, DDIM can thus act both as an encoder, generating a latent noise representation $\mathbf{x}_T$ from $\mathbf{x}_0$, and as a decoder, reconstructing $\mathbf{x}_0$ from $\mathbf{x}_T$.

## 3.2 Latent Diffusion Model

Latent Diffusion Model [46] is an advanced variant of diffusion models that employs a Variational Autoencoder (VAE) to encode high-dimensional data, such as images, into a more manageable, lower-dimensional latent space. This simplification enables the diffusion process to operate in the compressed latent space instead of the original high-dimensional image space.

The VAE comprises two main components: an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder function maps an image $I$ into a latent representation $z = \mathcal{E}(I)$, of the image, which the decoder later can use to reconstruct the image $\hat{I} = \mathcal{D}(z)$. In the context of diffusion models, the forward diffusion process is applied to the latent representation $z$. Similar to described in the previous subsection, this process involves incrementally adding Gaussian noise to the latent representation over $T$ steps:

$$q(z_t|z_{t-1}) = \mathcal{N}(\gamma_t z_{t-1}, \delta_t\mathbf{I}), \tag{8}$$

where $z_0$ is the initial latent representation, and $z_T$ is the fully noised latent representation at the final step.

The reverse process first learns the reverse distribution $p(z_{t-1}|z_t)$, uses it to iteratively obtained the denoised latent representation $z_0$, which is then decoded back into the image space: $\hat{I} = \mathcal{D}(z_0)$. This latent diffusion approach, exemplified in Stable Diffusion, reduces computational complexity while enabling efficient training and high-quality image generation. It effectively combines the VAE's efficiency in encoding images with the diffusion model's capability to generate diverse, high-fidelity outputs. In this work, we integrate targeted manipulation of the above processes within Stable Diffusion to achieve our goals using a novel Perturb-based approach.

## 3.3 Diffusion Autoencoders

Diffusion Autoencoders (DiffAEs) [43] integrate the powerful image generation capabilities of diffusion models with the semantic understanding of autoencoders. Unlike standard diffusion models, which lack interpretable latent codes, DiffAEs combine a learnable encoder with a diffusion decoder. This two-pronged approach extracts both high-level meaning and fine-grained details from an

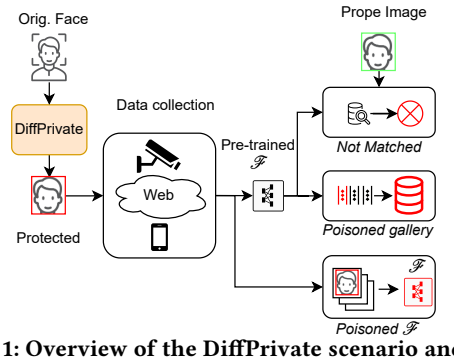



**Figure 1: Overview of the DiffPrivate scenario and use cases**

image, creating a code that is not only reconstructs the original image but also allows for manipulation of its semantic attributes. These capabilities enable advanced applications such as feature-based image editing, denoising, and conditioned sampling.

**Semantic Encoder:** The semantic encoder is designed to map an input face image $I$ into a semantic latent code $\mathbf{z} = \text{Enc}(I)$ that encapsulates high-level semantics of the face. By manipulating $\mathbf{z}$, we can induce semantic changes in the image, affecting attributes like expression, age, or gender in case of facial images.

**Conditional DDIM:** In the Conditional DDIM framework, as proposed in DiffAE, the DDIM model is conditioned on the semantic code $\mathbf{z}$ and includes a noise prediction network $\eta_\phi(\mathbf{x}_t, t, \mathbf{z})$, with $\mathbf{z}$ serving as an additional input. During the decoding phase, the reconstructed image $I = \mathbf{x}_0$ is obtained by executing a deterministic generative process, described as:

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z}) = \begin{cases} \mathcal{N}(f_\phi(\mathbf{x}_1, 1, \mathbf{z}), \mathbf{0}), & \text{if } t = 1, \\ q(\mathbf{x}_{t-1}|\mathbf{x}_t, f_\phi(\mathbf{x}_t, t, \mathbf{z})), & \text{otherwise,} \end{cases} \tag{9}$$

where $f_\phi(\mathbf{x}_t, t, \mathbf{z}) = (\mathbf{x}_t - \sqrt{1-\rho_t} \cdot \eta_\phi(\mathbf{x}_t, t, \mathbf{z}))/\sqrt{\rho_t}$. Here, $q(\cdot|\cdot,\cdot)$ is defined similarly to the DDIM process described earlier.

During the encoding process, the stochastic code $\mathbf{x}_T$ of the input image is obtained using the Conditional DDIM encoder, $\mathbf{x}T = \text{DDIMenc}(I, \mathbf{z})$, by reversing the deterministic generative process:

$$\mathbf{x}_{t+1} = \sqrt{\rho_{t+1}} f_\phi(\mathbf{x}_t, t, \mathbf{z}) + \sqrt{1-\rho_{t+1}} \eta_\phi(\mathbf{x}_t, t, \mathbf{z}). \tag{10}$$

In this process, $\mathbf{x}_T$ is encouraged to encode primarily the information that is not captured by $\mathbf{z}$, essentially focusing on stochastic or variable details that are not represented in the semantic code.

In this work, we incorporate DiffAE in our Edit-based approach.

# 4 DiffPrivate Framework

## 4.1 Use Case Scenario and General Approach

We consider a scenario where social media users aim to prevent their profiles from being linked to query photos submitted to a FR service. The scenario is motivated by the growing practice of entities like companies and law enforcement agencies scraping public photos from social media and then using FR technology to link query photos to identities and accounts. In this quickly developing scenario, many users may seek to enhance their privacy by uploading slightly altered images that either (1) cause the wrong identity to be suggested, (2) introduce ambiguity in query databases, or (3) increase the likelihood that any FRS model $\mathcal{F}$ trained on such data returns the wrong identity, helping users avoid detection. Fig. 1 illustrates the DiffPrivate approach and these use cases.

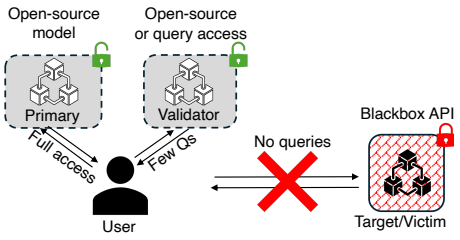



**Figure 2: Our blackbox security model**

To protect user privacy from FRS, we develop methods that allow users to generate images with adversarial perturbations. The methods are tunable, allowing users adjusting the desired alteration level to balance privacy protection and visual resemblance. In addition to increased control, this flexibility enables future use cases, including the use of different protection levels for primary subjects and bystanders. We expect the desired privacy protection to be individual, but consider user studies determining people's current interest, expectation, and desirable properties of such a system as interesting future work. Although our solutions are defensive, we refer to them as "attacks" on FRS to align with established terminology.

## 4.2 Attacker Model

We assume that individuals do not control any of the query photos submitted for identification, their public appearance, or their physical appearance when in public. Instead, they can alter and control the photos they upload to social media. For example, in the above scenario the user would use DiffPrivate to create a "protected" version of the original image, recognizing that this image if shared on social media may be collected and used either as a "probe" image or be added (with labels from the social media) to the gallery set $\mathcal{G}$, or even used for training purposes of the FRS model $\mathcal{F}$.

Furthermore, we assume a "blackbox" model where we do not have access to the FRS models $\mathcal{F}$ that we are attacking (and hence defending the user against) but have access to at one or more other FRS models that we can use to execute our adversarial attack. This assumption is realistic as there are many open-source models and APIs available, but the users may not always know what FRS they need to protect themselves against. However, it also places an additional weight on the need for a highly transferable solution and/or the user generating enough samples to pollute the FRS.

We employ a "blackbox" model assumption unlike the security model used in [31], which permits a limited number of queries to a victim model (we refer to their model as "semi-blackbox") or models that have allowed even more API queries [11]. In our blackbox scenario (Fig. 2), we pragmatically assume access only to an open-source model, referred to as the "primary" model, and some limited queries to a "validator" model (both differ from the "victim" model), and refer to the model that we try to protect the user from as the "victim/target" model. Our approach, with its stricter security model, demonstrates enhanced robustness compared to existing methods. Sec. 5.2 outlines and further motivates our evaluation in both blackbox and semi-blackbox settings.

## 4.3 Attacker Goals and Variations

In this paper, we consider two distinct goals targeting each of the two desirable properties described above:

- **Attribute-based Edits:** In this scenario, we aim to edit the image enough that we can shift the identity to a target identity, helping the user create images that potentially could be used to pollute the database of the FRS and/or damage a model trained on such polluted dataset. Here, we think of an attacker that may even apply strong defenses (e.g., DiffPure) but that may use publicly collected images for their database (and training of their FRS model). For this reason, we want the ability to create images that the FRS map to a different targeted identity with some significant probability.
- **Perturbation-based Alterations with Wide Protection:** In this scenario, we aim to perturb the image just enough to fool FRS enough that they would return the wrong identity, while people would still recognize the person. As the users typically do not know who the attacker is or what model they are attacking, it is important that this attack is highly transferable and protects against a wide range of FRS.

Furthermore, to ensure high utility, the generated images should be of high quality, have a high degree of naturalness, and in the second case closely resemble the person in the photo (even though the FRS is fooled to believe otherwise).

## 4.4 Formal Goals and High-Level Approach

To achieve the above goals, we present two variations of our attack, both built using the same high-level approach, but incorporating slightly different loss functions and diffusion models. Specifically, we present one somewhat simpler solution using DiffAE for attribute-based edits and one somewhat more complex variation that applies perturbation-based alterations with Stable Diffusion.

In both cases, we attack a FRS assumed to use an unknown machine learning model $\mathcal{M}$ and embedding function $\mathcal{F}$, which have been trained on a dataset $\mathcal{D}$ consisting of paired data points $(\vec{X}_i, \vec{y}_i)$, where $\vec{X}_i$ represents an input with specific dimensions (height, width, channels), $\vec{y}_i$ represents the corresponding ground-truth label (potentially poisoned) with $K$ possible categories.

**Adversarial Attack Strategy:** As attackers, our objective is to manipulate input data $\vec{X}_i^{\text{orig}}$ into adversarial examples $\vec{X}_i^{\text{adv}}$. These manipulated inputs aim to deceive the model into producing either a desired target label $\vec{y}_i^{\text{tgt}}$ or any label other than the true label $\vec{y}_i$.

**Recent GAN-based Progress:** Systems like StyleAdv [31] and StyleID [30] have successfully manipulated the identity using a latent-space approach leveraging StyleGAN [28]. In addition to StyleGAN being excellent for sampling high-quality images [69], much of the above works' success also came from leveraging StyleGANs highly disentangled latent space for image editing.

While these works have proven successful and there are multiple works proposing different methods for the StyleGAN encoder, there are always deviations, especially when it comes to the facial domain. Motivated by the strong and flexible encoders provided by diffusion models, in this work, we set out to find and incorporate similar models like StyleGAN for the diffusion context. Specifically, we look closer at to what extent similar solutions can be achieved using manipulations of the latent spaces of a diffusion model. After some exploration, we have found two approach variations to achieve the above goals: (1) an Edit-based version using DiffAE and (2) a Perturb-based version using Stable Diffusion. Despite differing mechanisms, both approaches share a fundamental goal of manipulating facial images to thwart FRS while preserving visual appeal.

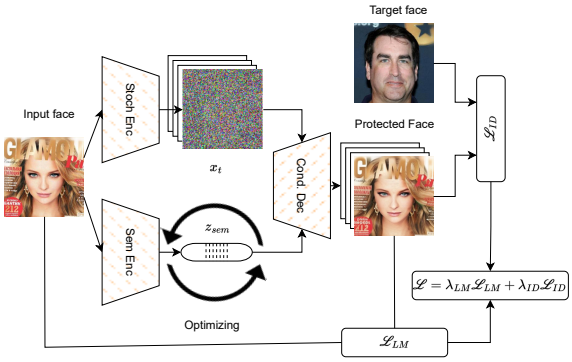



**Figure 3: Our Edit-based approach utilizes a diffusion autoencoder [43], takes the input face, and calculates the input's stochastic encoding $x_t$ and semantic latent code $z_{sem}$. The goal is to optimize $z_{sem}$ by minimizing a loss function that pushes the identity toward a target identity.**

**High-level Similarities and Differences:** Both approaches use optimization algorithms and diffusion models to iteratively refine modifications that subtly alter facial images, impeding FRS identification. However, they differ somewhat in how this is achieved. In particular, our Edit-based version using DiffAE models focuses on manipulating semantic latent codes within the diffusion model framework, while the perturbation-based version uses Stable Diffusion to intervene in the diffusion process. Furthermore, our Edit-based version prioritizes maintaining high visual quality and preserving facial features, whereas the perturbation-based edits emphasize introducing subtle modifications deep into the diffusion process to elude recognition by FRS. These core differences underscore the flexibility of our solution approach to achieve the overarching goal of enhancing facial privacy protection using diffusion models.

## 4.5 Edit-Based Version using DiffAE

When looking for a good replacement to StyleGAN, we found that DiffAE seems to tick all the boxes that StyleGAN offers and more: (1) bi-directional encoding and decoding, (2) highly disentangled latent codes for effective real image editing, and (3) a diffusion-based approach with a deep computational graph, suitable for robust adversarial attacks. Our initial strategy therefore involves using DiffAE to create adversarial samples. Specifically, echoing the objectives of StyleAdv, we seek to find a semantic latent code $z_{sem}$ that can generate an adversarial sample that ideally is indistinguishable to human observers, yet capable of tricking FR technology.

Taking this simple approach, as depicted in Fig. 3, we first input an image[1] into DiffAE, where it is encoded into a stochastic code $x_t$ and a semantic latent code $z_{sem}$. We then enter an optimization loop with respect to $z_{sem}$, seeking a modified version $\hat{z}_{sem}$ that fulfills our loss function criteria.

Our goal is to produce an adversarial image $\mathbf{I}_{adv}$ that effectively bypasses FRS. We achieve this by optimizing the semantic code $\mathbf{z}_{sem}$ of the input image $\mathbf{I}$, resulting in an adversarial semantic code

[1] Our choice of imagery in Figs. 3 and 4 are illustrative of DiffPrivate preserving complex backgrounds and keeping other visual elements intact, a task that prior methods like StyleID and StyleAdv find challenging.

$\hat{z}_{sem}$. This modified code, along with $\mathbf{x}_t$, is then processed through the DDIM decoding method to generate $\mathbf{I}_{adv}$:

$$\mathbf{I}_{adv} = \text{DDIM}_{dec}(\mathbf{x}_t, \hat{z}_{sem}). \tag{11}$$

In practical terms, particularly for targeted privacy protection, our objective is encapsulated in the following optimization problem:

$$\min_{\hat{z}_{sem}} \mathcal{L}_{ID}(\mathbf{I}_{adv}) = \mathcal{D}(\mathcal{F}(\mathbf{I}_{adv}), \mathcal{F}(\mathbf{I}_{tgt})), \tag{12}$$

where $\mathbf{I}_{tgt}$ represents the facial image of the target identity, $\mathcal{D}$ is the cosine distance, and $\mathcal{F}$ is the FR model. It is noteworthy that in real-world scenarios, access to $\mathcal{F}$ may not be available.

For better-preserving features of the original image, such as a facial landmark, we integrate a landmark loss, which calculates L2 distance of the facial landmark of input face $\mathbf{I}$ and adversarial face $\mathbf{I}_{adv}$: $\mathcal{L}_{LM} = \|\text{LM}(\mathbf{I}) - \text{LM}(\mathbf{I}_{adv})\|$, where LM is a facial landmark detection model. The optimization problem now becomes:

$$\min_{\hat{z}_{sem}} \mathcal{L}_{ID}(\mathbf{I}_{adv}) = \lambda_{ID}\mathcal{L}_{ID} + \lambda_{LM}\mathcal{L}_{LM}, \tag{13}$$

where $\lambda_{ID}$ and $\lambda_{LM}$ are hyperparameters balancing the loss terms.

Algorithm 1 presents our Edit-based method (based on DiffAE) for generating adversarial images that alter the appearance of the input image $\mathbf{I}$ with the goal of misleading FRS. The inputs consist of the original image $\mathbf{I}$, iteration count $N$, mixing coefficient $\gamma$, learning rate $\eta$, identity loss weight $\lambda_{ID}$, landmark stability weight $\lambda_{LM}$, encoder function $Enc$, original latent code $z_{org}$, target latent code $z_{tgt}$, and validation threshold $\theta$. The outputs are the adversarial image $\mathbf{I}_{adv}$ and the optimized mixing parameter $\alpha$.

The algorithm starts by encoding the input image into a latent representation $\mathbf{z}$ and transforming it using $DDIM_{enc}$ to $\mathbf{x}_T$. The methodology initializes a vector $\alpha^{(0)}$ as zero, representing no initial mixing between $z_{org}$ and $z_{tgt}$. This zero vector allows the generation of a semantic latent code $\hat{z}_{sem}$, which is then processed with $DDIM_{dec}$ to render the initial adversarial image $\mathbf{I}_{adv}^{(0)}$.

The algorithm employs an iterative optimization loop using the Adam optimizer. Here, gradient updates to $\alpha$ minimize a loss function that balances identity and landmark stability costs, weighted by $\lambda_{ID}$ and $\lambda_{LM}$. This continuous adjustment of $\alpha$ aims at finding an optimal blend of $z_{org}$ and $z_{tgt}$ that makes the resulting adversarial image deviate sufficiently from the original to outwit FRS while still conforming to the provided validator criteria $\theta$.

The loop terminates when the adversarial image satisfies the validator condition or reaches the iteration limit $N$, finalizing with $\alpha_{opt}$ and $\hat{z}_{sem}$ used to decode the final adversarial image $\mathbf{I}_{adv}$. This procedure offers a systematic approach to editing the image's latent space for desired adversarial effects.

**Algorithm 1** Edit approach based on DiffAE

**Input:** $\mathbf{I}$, $N$, $\gamma$, $\eta$, $\lambda_{ID}$, $\lambda_{LM}$, $Enc$, $z_{org}$, $z_{tgt}$, $\theta$
**Output:** adversarial image $\mathbf{I}_{adv}$, optimized $\alpha$
▷ Image Encoding
$\mathbf{z} = Enc(\mathbf{I})$, $\mathbf{x}_T = \text{DDIM}_{enc}(\mathbf{I}, \mathbf{z})$
▷ Initialization for Optimized $\alpha$
$\alpha^{(0)} \leftarrow \mathbf{0}$, where $\mathbf{0}$ is a zero vector of the same size as $z_{org}$
$\hat{z}_{sem}^{(0)} \leftarrow z_{org} \cdot \alpha^{(0)} + z_{tgt} \cdot (1 - \alpha^{(0)})$
$\mathbf{I}_{adv}^{(0)} = \text{DDIM}_{dec}(\mathbf{x}_T, \hat{z}_{sem}^{(0)})$
Initialize Adam parameters: $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
Initialize $i \leftarrow 0$
**while** Validator$(\mathbf{I}, \mathbf{I}_{adv}^{(i)}) < \theta$ and $i < N$ **do**
    Update $\hat{z}_{sem}^{(i+1)} \leftarrow z_{org} \cdot \alpha^{(i+1)} + z_{tgt} \cdot (1 - \alpha^{(i+1)})$
    $\mathbf{I}_{adv}^{(i)} = \text{DDIM}_{dec}(\mathbf{x}_T, \hat{z}_{sem}^{(i)})$
    $\mathcal{L}(\mathbf{I}_{adv}^{(i)}) = \lambda_{ID} \cdot \mathcal{L}_{ID}(\mathbf{I}_{adv}^{(i)}) + \lambda_{LM} \cdot \mathcal{L}_{LM}(\mathbf{I}_{adv}^{(i)})$
    $t \leftarrow t + 1$
    $g \leftarrow \nabla_{\alpha^{(i)}} \mathcal{L}(\mathbf{I}_{adv}^{(i)})$
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g$
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g^2$
    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
    $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
    $\alpha^{(i+1)} \leftarrow \alpha^{(i)} - \eta \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
    $i \leftarrow i + 1$
**end while**
$\alpha_{opt} \leftarrow \alpha^{(i)}$, $\hat{z}_{sem} \leftarrow \hat{z}_{sem}^{(i)}$, $\mathbf{I}_{adv} = \text{DDIM}_{dec}(\mathbf{x}_T, \hat{z}_{sem})$

## 4.6 Perturb-Based Version w. Stable Diffusion

While DiffAE and the above approach provide an intuitive method for crafting the adversarial sample, as we will later show, this method tends to push the identity toward the target identity. This is actually the design intention of the DiffAE model, in which the decoupled latent space $z_{sem}$ is highly disentangled, with small changes in this latent space reflecting the visible change in pixel space. However, to achieve imperceptible adversarial perturbations, we adopt a more integral approach, as described next.

Our intuition here is that diffusion models, with their depth computational graph, would cancel out the modification in feature space which does not match pixel space. To fool the models into helping us craft adversarial samples carrying hidden features that do not match their visual representation, we need to make the modification deep into individuals passing of the diffusion process.

While crafting adversarial samples for diffusion models through diffusion process intervention often presents challenges, we highlight two hurdles specific to this approach: (1) the need for an accurate inversion method like StyleGAN's encoder or DiffAE's semantic encoder to project real images onto the latent space; and (2) memory efficiency, as the optimization process requiring interaction with the diffusion process can be memory intensive. Given these challenges, we opt for Stable Diffusion as our base model. This has several advantages. Most importantly, as a latent diffusion model, where the diffusion occurs in a compressed VAE latent space, Stable Diffusion significantly reduces memory costs and, crucially, delivers high-quality images, as supported by the research literature and the wide availability of supporting work around the model. This allows us to leverage existing solutions and tailor them to our specific needs. For inversion, we choose the advanced Null-text inversion technique proposed by Mokady et al. [37].

To integrate our solution into the inversion process, we first employ DDIM inversion to obtain a sequence of latent codes $z$ at various timesteps within the diffusion process. As illustrated in Fig. 4, the process commences with the original image's direct latent codes $z_0$ and progressively introduces noise until reaching $z_T$. Without optimization, initiating the denoising process from $z_T$ would merely yield a reconstructed version $\hat{z}_0$ that deviates substantially from the original $z_0$. Subsequently, the inversion process pivots towards optimizing the null-text, or unconditional prompt, that serves as input to Stable Diffusion. This optimization effectively brings the inverted codes closer to their originals while preserving

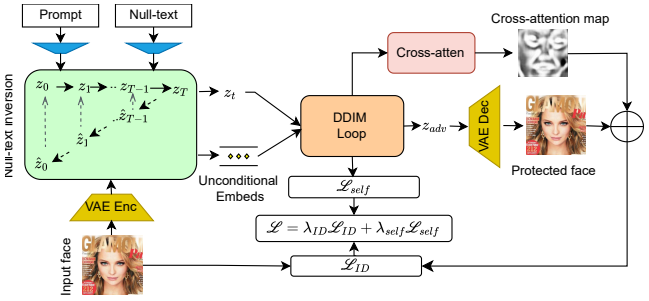



**Figure 4: Our Perturb-based approach utilizes a Stable Diffusion model, which goes through an inversion process that optimizes the null-text embedding, allowing accurate reconstruction of the original identity, then goes through a diffusion loop where cross attention maps are extracted, facilitating an optimization process pushing the identity features gradually away from the original while keeping the identity resemblance intact.**

**Algorithm 2** Perturbation approach based on Stable Diffusion

**Input:** $\mathbf{I}$, $N$, $\eta$, $\lambda_{ID}$, $\lambda_{self}$, $\text{Invert}^{Null}$, $\theta$
**Output:** adversarial image $\mathbf{I}_{adv}$
▷ Null-text inversion
$z_T, \{\varnothing_t\}_{t=1}^{T} = \text{Invert}^{Null}(\{\mathbf{z}_t\}_{t=1}^{T}, \varnothing, C, N)$
▷ Attack Generation
$z_{adv}^{(0)} \leftarrow \mathbf{z}$
$\mathbf{I}_{adv}^{(0)} = \text{VAE}_{dec}(z_{adv}^{(0)})$
Initialize Adam parameters: $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
$i \leftarrow 0$
**while** $i < N$ and $\text{Validator}(\mathbf{I}, \mathbf{I}_{adv}^{(i)}) < \theta$ **do**
  $z_{adv}^{(i)}, \mathbf{I}_{mask} = \text{DDIM}_{latent}(\{\varnothing_t\}_{t=1}^{T}, z_{adv}^{(i)})$
  $\mathbf{I}_{adv}^{(i)} = \text{VAE}_{dec}(z_{adv}^{(i)}) \oplus \mathbf{I}_{mask}$
  $\mathcal{L}(\mathbf{I}_{adv}^{(i)}) = \lambda_{ID} \cdot \mathcal{L}_{iden}(\mathbf{I}_{adv}^{(i)}) + \lambda_{self} \cdot \mathcal{L}_{self}(\mathbf{I}_{adv}^{(i)})$
  $t \leftarrow t + 1$
  $g \leftarrow \nabla_{z_{adv}^{(i)}} \mathcal{L}(\mathbf{I}_{adv}^{(i)})$
  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g$
  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g^2$
  $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
  $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
  $z_{adv}^{(i+1)} \leftarrow z_{adv}^{(i)} - \eta \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
  $i \leftarrow i + 1$
**end while**
$z_{adv} \leftarrow z_{adv}^{(i)}, \mathbf{I}_{adv} = \text{VAE}_{dec}(z_{adv})$

the original image's prompt description. This preservation is crucial for our technique, allowing us to extract a cross-attention map that focuses on facial regions requiring protection.

Formally, null-text inversion optimizes a unique unconditional embedding $\varnothing$ initiated with the null-text embedding while keeping the model and the conditional textual embedding unchanged. This process can be either *global*, using a single embedding, or timestamp-specific, with distinct embeddings $\varnothing_t$ optimized for each timestamp $t$, each initialized from the embedding of the previous step $\varnothing_{t+1}$.

The full algorithm utilizes DDIM inversion to produce a sequence of noisy latent codes $z_T^*, \ldots, z_0^*$. For each timestamp $t = T, \ldots, 1$, an optimization is performed for $N$ iterations to minimize $\left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \varnothing_t, C) \right\|^2$. Here, $z_{t-1}(\bar{z}_t, \varnothing_t, C)$ denotes applying the DDIM sampling step using the respective embeddings and conditional embedding. This process, while less expressive than full model fine-tuning, is efficient and well-suited for pivotal inversion, leading to the final edited image using the optimized unconditional embeddings $\{\varnothing_t\}_{t=1}^{T}$, enabling efficient editing operations on the input image. In brief, we can summarize the process as follows:

$$z_T, \{\varnothing_t\}_{t=1}^{T} = \text{Invert}^{Null}(\{\mathbf{z}_t\}_{t=1}^{T}, \varnothing, C, N). \tag{14}$$

The primary aim of this approach is to craft an adversarial image $\mathbf{I}_{adv}$ capable of eluding FR technologies. This is accomplished by modifying the latent code $\mathbf{z}_{adv}$ of an input image $\mathbf{I}$, thus producing a perturbed latent code $\mathbf{z}_{adv}$. Subsequently, this altered code, in conjunction with $\{\varnothing_t\}_{t=1}^{T}$, is subjected to the DDIM decoding process to create $z_{adv}$:

$$z_{adv} = \text{DDIM}_{latent}(\text{prompt}, \{\varnothing_t\}_{t=1}^{T}, \mathbf{z}_{adv}). \tag{15}$$

Like equation (12), particularly when aiming at targeted privacy preservation, the goal is defined by the following optimization:

$$\min_{z_{adv}} \mathcal{L}_{ID}(\mathbf{I}_{adv}) = \mathcal{D}(\mathcal{F}(\mathbf{I}_{adv}), \mathcal{F}(\mathbf{I}_{tgt})), \tag{16}$$

where $\mathbf{I}_{adv} = \text{VAE}_{dec}(z_{adv})$, $\mathbf{I}_{tgt}$ is a target facial image, $\mathcal{D}$ denotes the cosine distance, and $\mathcal{F}$ represents the FR framework.

A cross-attention map is employed to enhance the retention of original image features, such as facial landmarks. This map is a binary mask $\mathbf{I}_{mask}$ highlighting the crucial regions of both the input face $\mathbf{I}$ and the adversarial face $\mathbf{I}_{adv}$. This mask is then utilized on the adversarial image, ensuring that in subsequent iterations, only pivotal facial regions undergo modification: $\mathbf{I}_{adv} = \mathbf{I}_{adv} \oplus \mathbf{I}_{mask}$. Additionally, to ensure the structural integrity of the original image, a self-attention loss $\mathcal{L}_{self}$, derived from the self-attention layers of the U-Net in Stable Diffusion, is integrated into the process.

Thus, the optimization challenge is reformulated as:

$$\min_{z_{adv}} \mathcal{L}_{ID}(\mathbf{I}_{adv}) = \lambda_{ID}\mathcal{L}_{ID} + \lambda_{self}\mathcal{L}_{self}, \tag{17}$$

where $\lambda_{ID}$ and $\lambda_{self}$ are the hyperparameters used to balance the respective loss terms.

Algorithm 2 presents our perturbation method. The inputs consist of the original image $\mathbf{I}$, iteration count $N$, learning rate $\eta$, identity retention weight $\lambda_{ID}$, self-consistency weight $\lambda_{self}$, the null-text inversion function $\text{Invert}^{Null}$, and a threshold $\theta$ for termination. The output is the adversarial image $\mathbf{I}_{adv}$.

The method begins with a null-text inversion step, using the $\text{Invert}^{Null}$ function to generate a latent representation $z_T$ and a null latent path $\varnothing_t t = 1^T$. From the original latent code $\mathbf{z}$, an approximate image $\mathbf{I}adv^{(0)}$ is then reconstructed via a VAE decoder.

At the core of the algorithm is an optimization loop using the Adam optimizer to refine the latent code $z_{adv}^{(i)}$, ensuring it diverges enough to bypass FRS while preserving perceptual similarity to the original image $\mathbf{I}$. The adversarial updates incorporate a dynamically masked version of the image, enhancing the robustness of the resultant adversarial traits. Two loss functions, $\mathcal{L}_{iden}$ and $\mathcal{L}_{self}$, weighted by $\lambda_{ID}$ and $\lambda_{self}$, respectively, guide the training to balance between disguising the identity and maintaining fidelity to the original image. The loop terminates when the adversarial image's validator score stays below the threshold $\theta$, after which the final adversarial image $\mathbf{I}_{adv}$ is reconstructed from the last updated latent code $z_{adv}$ using the VAE decoder.

# 5 Experimental Setup

## 5.1 Datasets

We leverage three significant datasets to evaluate the performance of FR and image generation/editing algorithms: Labeled Faces in the Wild (LFW) [22], Celebrity Faces Attributes (CelebA) [34] and its high-quality variant CelebA-HQ [27], as well as the Flickr-Faces-HQ (FFHQ) [28] dataset. These datasets are chosen for their benchmark status in various domains related to facial image analysis, including recognition, generation, and editing. A detailed description of each dataset is provided in Appendix B.

By combining LFW, CelebA/CelebA-HQ, and FFHQ datasets, our evaluation framework enables a comprehensive assessment of FR and image generation/editing algorithms across different datasets. LFW provides a benchmark for FR accuracy in uncontrolled environments, while CelebA and CelebA-HQ offer a rich source of annotated facial images for attribute editing. FFHQ, with its high-resolution and diverse images, provides cross validation dataset to compensate CelebA/CelebA-HQ. Together, these datasets ensure a robust and diversified testing ground for our algorithms, facilitating meaningful comparisons with related works.

## 5.2 Evaluation Settings

In our evaluation, we explore different options, including targeted/untargeted whitebox/blackbox attacks against FRS. Our approach uses five well-known FR models: Facenet [48], IR152, IRSE50 (ArcFace) [9], MobileFaceNet (termed as MobileNet in this paper) [5], and IR101 (CurricularFace) [23]. To enhance the transferability of the attack across different systems and inspired by StyleAdv [31], we designate one of these models as a primary model (a whitebox model) and one a second model as validator model (semi-blackbox). Like StyleAdv, we use the first model for identity loss in our optimization process and the validator model (e.g., a third-party service or API), which we query up to three times in our experiments. This limitation simulates realistic attack scenarios where an attacker might have restricted opportunities to test their approach without being detected. Finally, we assume no access to the other three models. This means that we assume no access to their internal details or parameters, which is a common scenario in real-world attacks.

When reporting our results, we call attacks against the primary models as *whitebox* attacks (as we have access to the primary model), attacks against the validator model as *semi-blackbox* [31] attack (as we only can query its API a very limited number of times but do not assume any access to the model itself, only the primary model), and we consider attacks against other models as *blackbox* attacks (assuming no access to them).

## 5.3 Evaluation Metrics

To assess the effectiveness of the privacy protection mechanism, we introduce a metric called the Privacy Protection Rate (PPR). The PPR is designed to directly measure the ability of the altered images to prevent correct recognition by an FRS. By evaluating the likelihood of mismatching the altered image with the original identity, it offers a practical assessment of our approach's ability to prevent identity leaks. PPR's alignment with standard benchmark practices, such as the LFW benchmark (which typically assess whether a pair of images belong to the same identity based on a threshold), makes it more relevant than traditional metrics like recognition accuracy or true positive rates, which may not fully reflect privacy protection in real-world scenarios. The PPR metric is based on the cosine similarity distance, $d_{\text{cosine}}$, and is defined as follows:

$$\text{PPR} = \frac{1}{N} \sum_{\mathbf{I}} \mathbb{1}\left( d_{\text{cosine}}(\mathcal{F}(\mathbf{I}_{org}), \mathcal{F}(\mathbf{I}_{ptd})) < \tau \right) \times 100\%, \quad (18)$$

where $\mathbb{1}$ is the indicator function, $N$ the total number of face images $\mathbf{I}$, $\tau$ a predefined threshold value, and $\mathbf{I}_{org}$ and $\mathbf{I}_{ptd}$ are the original and protected face images, respectively. In this context, $\mathcal{F}$ is a face embedding model (feature extractor), referred to as FRS throughout our paper, which generates a vector representation of facial features for each image. Here, $d_{\text{cosine}}$ is defined as:

$$d_{\text{cosine}}(\vec{u}, \vec{v}) = \frac{1}{\pi} \arccos\left( \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|} \right), \quad (19)$$

reflecting the distance based on cosine similarity, as described previously. To align with privacy protection standards, we set the threshold $\tau$ for each victim model to where they achieve $1 \times 10^{-3}$ False Acceptance Rate (FAR). Deciding on this threshold for a particular system is critical, as it directly impacts the protection rate. Furthermore, we adhere to a rigorous LFW benchmark to establish this threshold, serving as a standard for evaluating all related works when compared with ours. More information regarding the determination of this threshold can be found in Appendix D.

While users employing our approach do not need to know the threshold of the target model they are protecting against or the structure of blackbox models, the threshold of the primary model is important. Depending on what model is used, these thresholds can be obtained either by extracting them from the evaluation section of the corresponding papers, by following the process described in Appendix D, or by using the results we provide in that section. In general, there is a tradeoff between ease of crafting adversarial samples (easier when threshold selected for a small FAR target) and transferability across models (worse with small FAR target).

In addition to PPR, we employ the Learned Perceptual Image Patch Similarity (LPIPS) [68] to evaluate the naturalness and quality of the protected face images. LPIPS has been chosen specifically for its ability to mirror human visual perception more closely compared to traditional metrics like SSIM [59] or MS-SSIM [61]. LPIPS is designed to assess the perceptual difference between images based on high-level features extracted by deep neural networks, which tend to align better with human judgment of image quality and similarity. This is particularly important in our context where subtle visual changes, which might significantly alter machine recognition performance, might still be perceived as minor by human observers.

While SSIM and MS-SSIM effectively measure structural similarity, they often do not adequately account for perceptual aspects like texture and color dynamics, which are important for assessing natural appearance on social media. Since our goal is to create images that appear natural and consistent with human perception while deceiving FRS, LPIPS serves as a more appropriate metric, ensuring that the modifications made preserve the overall aesthetic and perceptual quality perceived by human viewers.

Combined, PPR and LPIPS provide a comprehensive assessment of the efficacy of our privacy protection mechanism, covering aspects of privacy, visual fidelity, and perceptual quality.

**Table 1: Comparison with state-of-the-art methods on CelebA-HQ and FFHQ datasets for targeted blackbox attacks.**

| Methods | CelebA-HQ | | | | FFHQ | | | |
|---|---|---|---|---|---|---|---|---|
| | PPR (%) ↑ | | | LPIPS ↓ | PPR (%) ↑ | | | LPIPS ↓ |
| | IRSE50 | IR101 | IR152 | | IRSE50 | IR101 | IR152 | |
| PGD [35] | 0.00 | 0.53 | 0.00 | 0.06 | 0.00 | 0.00 | 0.53 | 0.07 |
| CW [3] | 13.02 | 1.03 | 3.54 | 0.18 | 7.04 | 2.01 | 6.05 | 0.18 |
| Fawkes [50] | 1.52 | 1.53 | 14.03 | 0.04 | 1.02 | 1.04 | 10.52 | 0.04 |
| AMT-GAN [21] | 2.04 | 0.52 | 5.09 | 0.09 | 1.53 | 1.52 | 4.52 | 0.12 |
| StyleAdv [31] | 41.53 | 45.54 | 75.53 | 0.14 | 20.52 | 25.03 | 60.04 | 0.16 |
| DiffPrivate | 88.02 | 86.54 | 98.03 | 0.05 | 92.03 | 95.02 | 98.53 | 0.06 |

## 6 Qualitative and Quantitative Comparisons

We first present a direct comparison against some of the most related works (see Sec. 2): CW [3], PGD [35], Fawkes [50], AMT-GAN [21], and StyleAdv [31]. Here, our selection of specific models for the blackbox context, shown in Table 1, is based on rigorous prior investigations. For instance, Fawkes primarily uses FaceNet in its optimization, making it unsuitable as a blackbox model for our evaluations. Instead, we need at least one model, such as MobileFace, as a validator due to its characteristics as the shallowest model, which, as we discuss in Sec. 7, presents the greatest challenge in terms of attack resistance.

**Qualitative Comparison:** Fig. 5 offers a visual example comparison between our work and related works, in which we include direct “face-to-face" comparisons for representative set of example faces. We observed that: (1) perturbation-based methods in pixel space, such as PGD, CW, and Fawkes, exhibit a common drawback wherein the noisy effect is visibly apparent. This can hinder usability for applications like posting protected images on social media, given the human sensitivity to artifacts on faces. (2) GAN-based methods, including AMT-GAN and StyleAdv, suffer from some degree of visible distortions, an inherent limitation of GAN models. (3) Our method, which employs semantic editing to make images more attractive, yields the most visually pleasing results. This approach has potential applications as a beautifying face filter, akin to those frequently used in popular apps such as Instagram or Camera360. (4) Our perturbation approach in the latent space demonstrates an advantage in preserving small details in photos, including background elements, hair, and more obscure objects like hands, phones, hats, etc. The comparisons demonstrate that our diffusion-based method for creating protected images produces higher image quality relative to the compared methods. It surpasses GAN-based alternatives in achieving consistent results, a critical factor for applications within the facial domain.

**Image Quality:** The evaluation of image quality, as quantified by the LPIPS metric, reveals significant insights into the efficacy of various adversarial methods when using DiffPrivate in Perturb-based mode for this evaluation.

As shown in Table 1, pixel-perturbation based methods, including PGD, CW, and Fawkes, exhibit relatively low LPIPS scores. Notably, Fawkes achieves the lowest score among them, which can be attributed to Fawkes incorporating LPIPS into its loss function during optimization to minimize perceptual differences. In contrast, GAN-based methods such as AMT-GAN and StyleAdv demonstrate higher LPIPS scores, suggesting more substantial alterations to the image that potentially compromise its natural appearance. Our method, DiffPrivate, effectively maintains a balance between privacy enhancement and utility preservation. It achieves a significantly lower LPIPS score compared to GAN-based approaches, aligning more closely with the scores of pixel-perturbation methods. This outcome underscores DiffPrivate’s ability to maintain the visual integrity of images while providing robust privacy protection, highlighting its advantageous position in reconciling the tradeoffs between adversarial effectiveness and image quality.

**Privacy Protection Rate:** In our comprehensive comparison with state-of-the-art methods on the CelebA-HQ and FFHQ datasets for targeted blackbox attacks, as presented in Table 1, we rigorously use the LFW benchmark to set the threshold at a specified False Acceptance Rate (FAR), as detailed in Appendix D. This rigorous approach, coupled with the employment of advanced FRS models, significantly influences the evaluation of protection rates, providing a stark contrast to the methods used in prior works such as Fawkes (and PGD, CW-based methods such as Face-off [4], and FoggySight [12]). These earlier studies either relied on classifier models or focused solely on the FaceNet model, which may not reflect the efficacy against more current and accurate FRS models.

A notable observation is that the PPR for CelebA-HQ are generally higher than those for FFHQ. This discrepancy can be attributed to the intrinsic differences between the two datasets. CelebA-HQ, with its more uniform and curated collection of images, may inherently facilitate the generation of adversarial examples that are more effective across a variety of models. In contrast, FFHQ, known for its diversity in age, ethnicity, and image quality, presents a more challenging scenario for adversarial attacks, likely due to the increased variability and complexity of the dataset. This further underscores the necessity of a robust evaluation framework that encompasses a wide range of real-world conditions.

Furthermore, our evaluation strictly observes the blackbox criteria, setting us apart from semi-blackbox settings reported in studies like StyleAdv [31], which show near-perfect protection. Under our stringent blackbox evaluation framework, where we assume no knowledge of or query access to the model (except for a limited use of a validator model not included in the blackbox models list), we find that pixel-perturbation based methods such as PGD, CW, and Fawkes exhibit negligible transferability to blackbox models. This is evidenced by their low or zero PPR scores across all evaluated FRS models. Similarly, AMT-GAN shows limited effectiveness, attributed to a training process that may not adhere to the stricter threshold levels we implement. StyleAdv, while achieving moderate results, falls short of its previously reported outcomes under a less stringent semi-blackbox setup.

The standout performance of our method, *DiffPrivate*, demonstrates its advantages, achieving the highest PPR scores across both datasets and all models tested. This underscores the importance of both a rigorous evaluation protocol and the selection of evaluation models that closely mirror real-world application scenarios. The results, as shown in Table 1, highlight the pivotal role of choosing appropriate thresholds and models in accurately assessing the protection rate, affirming the effectiveness of our approach in safeguarding against unauthorized FR attempts. Finally, we note that the high achieved success in the blackbox model scenario, the most challenging of the three threat models, suggests effectiveness in the less restrictive whitebox and semi-blackbox contexts as well.

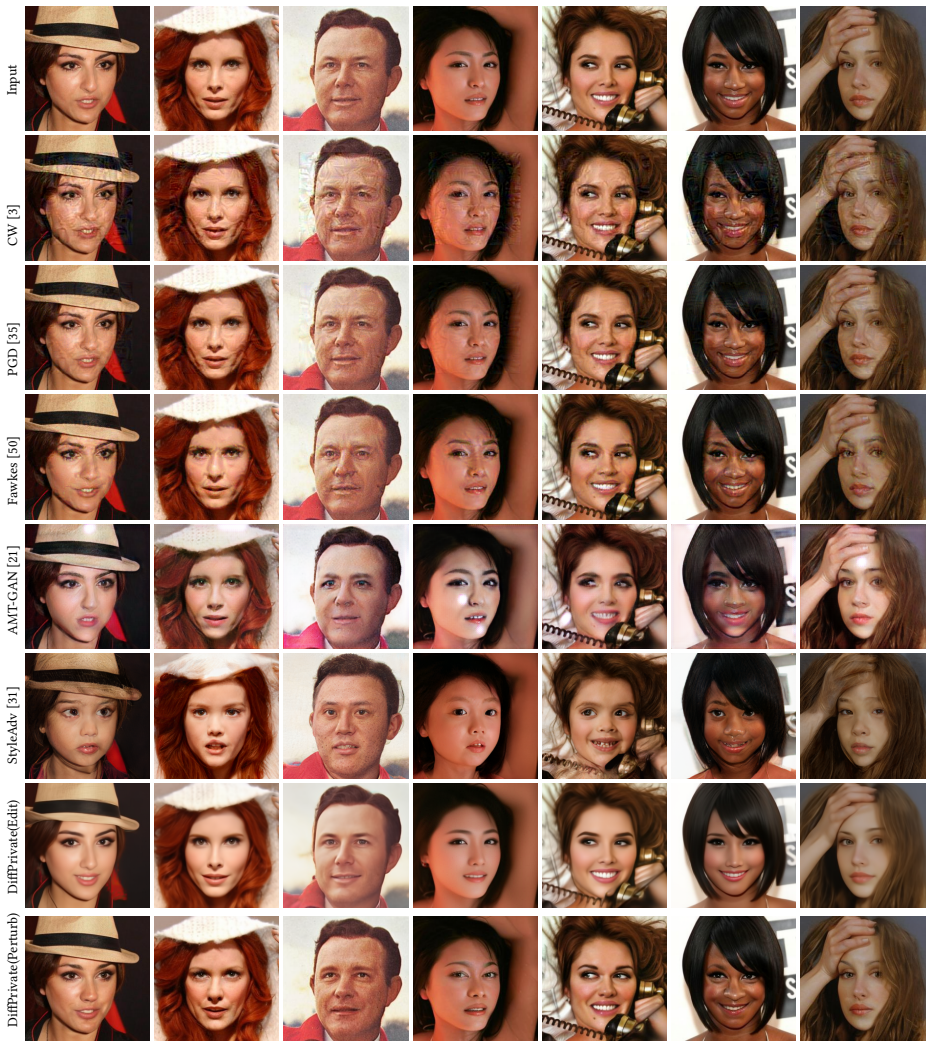



**Figure 5: Visualizations of protected face images generated by different face protection methods on CelebA-HQ.**

## 7 Transferability of Protection

**Semi-blackbox Attack:** In the *semi-blackbox* evaluation, detailed in Fig. 6, we explore the effectiveness of DiffPrivate under the assumption that users have whitebox access to a primary model and limited query access to the validator model that (in this case) represents the unauthorized FRS they aim to protect against. This evaluation, deemed resource-intensive, was conducted as a targeted adversarial attack on the FFHQ dataset due to its greater challenge and closer representation of real-world scenarios compared to CelebA-HQ (as discussed in Sec. 6).

We utilize Stable Diffusion as the base generative model with diffusion steps set to 20, a guidance scale of 2.5, and a default prompt of "a person", our approach employs an Adam optimizer with a learning rate of $5 \times 10^{-3}$. Crucially, we limit the optimization loop to a maximum of 250 iterations, conducting only three occasional checks with the validator at steps 50, 100, and 250 to determine if an adversarial image meets the validator's requirements. If the adversarial image does not pass the validator's criteria by the final check at iteration 250, the process is terminated.

The resultant heatmap illustrates the PPR achieved across various combinations of primary and validator models. For instance, with MobileFace as the primary model and IR101 as the validator, a PPR of 65% indicates that, within the 250-iteration limit, 65% of the test samples succeeded in generating an adversarial image that could evade detection.

The heatmap reveals significant variability in protection efficacy, highlighting the impact of model combinations on adversarial success. Notably, when models serve as their own validators (known as whitebox attacks), the PPR approaches or reaches 100%. Conversely, lower PPRs in cross-model evaluations, such as the 22.5% observed when IR152 and IRSE50 serve as validators for each other, reflect the increased difficulty of evading unfamiliar FRSs. These observations underline the critical role of model familiarity and the adversarial model's adaptability in designing effective privacy protection strategies.

However, there are also several semi-blackbox attacks (where the primary model differ from the validator model) that are highly

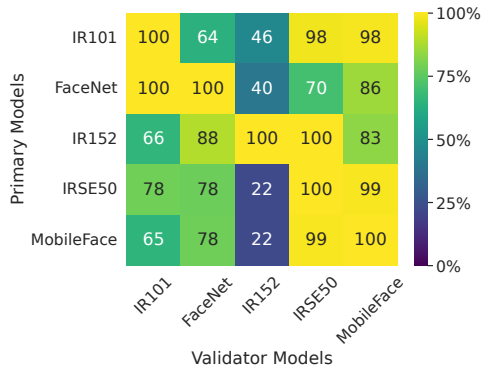



**Figure 6: The success in the PPR of crossing attack between primary and validator models**

successful (e.g., IR101 against IRSE50 and MobileFace). The effectiveness in these cases suggests that certain features learned by IR101 are robust enough to generalize across different models, including IRSE50 and MobileFace. This indicates a transferability of the adversarial modifications that is not model-specific but potentially applicable across a spectrum of models with varying architectures. Here, IRSE50 is most sensitive to such semi-blackbox attacks in 3 out of 4 cases (i.e., with IR101, IR152, MobileFace as primary models). When FaceNet is the primary model, IR101 is most vulnerable and when IRSE50 is the primary model, MobileFace is most vulnerable.

**Blackbox Attacks:** Fig. 7 presents the results of our evaluation on the transferability of adversarial protection across different blackbox models, utilizing a combination of primary models and validators. This setup aims to assess the effectiveness of adversarial edits against models to which the user is assumed to have no knowledge or access. The evaluation highlights a pattern of transferability that bears resemblance to the findings from the heatmap presented in Fig. 6, underlining the consistency of our adversarial protection strategy across varied settings.

Two distinct cases emerge from our analysis: (1) Utilizing IR152 as a validator, which is a deep CNN model renowned for its high accuracy on the LFW benchmark, results in the lowest PPR, particularly when paired with FaceNet or IRSE50 as primary models. This observation suggests that despite IR152's high accuracy, it may be more susceptible to adversarial attacks, corroborating the tradeoff between accuracy and robustness documented in prior research [15, 35, 57]. (2) Conversely, selecting MobileFace as the validator consistently yields the highest PPR across all primary models, especially when used in conjunction with FaceNet or IR101. This indicates that MobileFace, despite being the model with the shallowest architecture and the lowest accuracy among those evaluated, offers superior robustness against adversarial attacks. These results underscore the nuanced relationship between model complexity, accuracy, and vulnerability to adversarial manipulation, reinforcing the importance of considering these factors in the development and evaluation of privacy-enhancing technologies.

## 8 Robustness of Protection

In our analysis, represented in Fig. 8, we specifically assess the efficacy of various adversarial protection methods on the CelebA-HQ and FFHQ datasets, with a particular focus on the CelebA-HQ dataset for evaluating the DiffPure method. This targeted evaluation stems from the prerequisite that DiffPure requires a DDPM model trained exclusively on CelebA-HQ for faces, thereby limiting its purification effects to this dataset alone. Given the ineffectiveness of pixel-based perturbation methods in blackbox settings—often resulting in negligible protection rates—we opt for a whitebox evaluation framework to discern robustness.

Across the datasets, our method consistently exhibits high robustness against purifying methods when compared to alternatives. This is particularly evident on the CelebA-HQ dataset, where the robustness of our protection method outperforms that on the FFHQ dataset. Notably, pixel-perturbation methods such as PGD, CW, and Fawkes are highly susceptible to noise-canceling techniques like Gaussian Blurring and Total Variance Minimization, which can significantly diminish their protection rates. Other purifying methods generally reduce the protection efficacy of these pixel-perturbation methods by approximately half. Remarkably, DiffPure effectively neutralizes the protective capability of all methods except for Ours and StyleAdv. AMT-GAN, while consistently holding a moderate protection rate against other purifying methods, is also vulnerable to DiffPure.

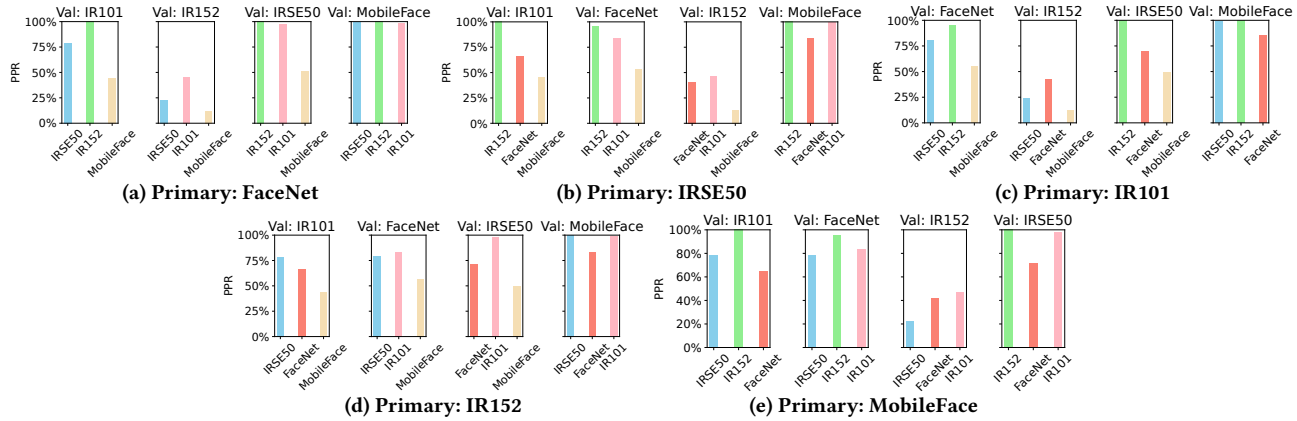
The observed patterns underscore the nuanced interaction between adversarial protection methods and purifying techniques, highlighting the critical importance of designing protection strategies that are resilient not only to direct adversarial attacks but also to subsequent purification attempts. Our method's strong performance against various purifying methods on CelebA-HQ validates its effectiveness and highlights potential limitations of pixel-perturbation approaches for robust adversarial protection.

## 9 Protection with Attributes Editing

We next analyze the relative protection offered by editing different attributes. For each scenario, we show results from representative models, noting that our model selections and comparisons are consistent with the model selection in Sec. 6 (i.e., the primary model is FaceNet and the validator model is MobileFace) and are generally in line with the tradeoff between model accuracy and robustness captured in Sec. 7 (e.g., blackbox results for FaceNet with MobileFace as validator (Fig. 7a) are outstanding for the LFW datasets, although there will be deviations when we test on other datasets).

**Semi-Blackbox Setting:** Fig. 9 illustrates the PPR across various attributes for the DiffPrivate Edit-based method in a semi-blackbox setting, where MobileFace serves as both the validator and the target unauthorized FRS. Notably, the performance across attributes on the CelebA-HQ dataset is predominantly high, a result that aligns with expectations considering CelebA-HQ's use in training the classifier for DiffAE, which guides the attribute editing process. This training alignment ensures a high degree of attribute manipulation success.

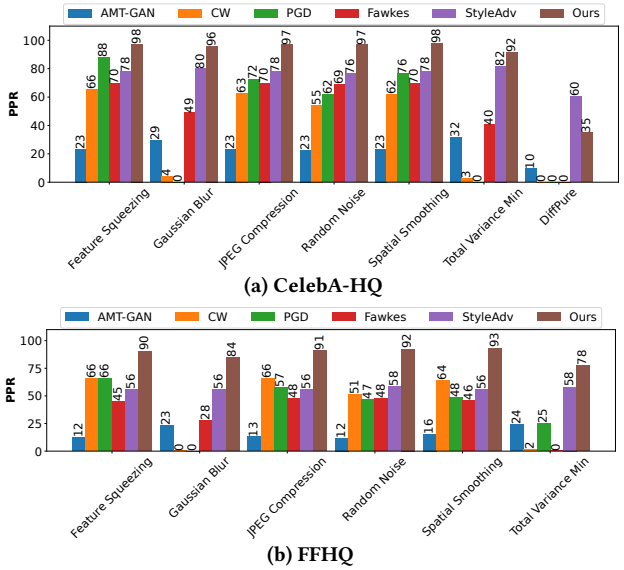
Conversely, the FFHQ dataset presents a more varied set of outcomes, with some attributes experiencing a protection rate decline to as low as 70-80%. Attributes such as *Blurry* and *Pale Skin* demonstrate lower protection rates, which could be attributed to the inherent challenges in editing these features in a manner that significantly alters the FRS's perception without compromising natural appearance. This variability in success rates underscores the complexity of attribute-based adversarial editing in diverse




(a) Primary: FaceNet (b) Primary: IRSE50 (c) Primary: IR101

(d) Primary: IR152 (e) Primary: MobileFace

**Figure 7: Attacker transferable between cross blackbox models**




(a) CelebA-HQ

(b) FFHQ

**Figure 8: Protection percentage on CelebA-HQ and FFHQ. Note that DiffPure is not applicable in the case of FFHQ as this method requires a DDPM model trained on CelebA-HQ and the purification only works effectively on this dataset**

datasets. The high protection rates for attributes like *Smiling* or *Wearing Lipstick* in CelebA-HQ, compared to FFHQ, suggest a direct correlation between the training data of the guiding classifier and the effectiveness of attribute manipulation. This observation points to the potential need for selecting the right attributes in optimizing the privacy protection efficacy of edit-based methods, particularly in semi-blackbox settings with only limited access to target FRS.

**Blackbox Setting:** Here, we summarize key observations from a comprehensive evaluation of the protection rates across various attributes in a blackbox setting (primary: FaceNet, validator: MobileFace) comparing performance on the CelebA and FFHQ datasets across multiple FRS models: IRSE50, IR152, IR101, and MobileFace (numeric results are provided in Table 2 of Appendix C.1).

Notably, attributes such as *High Cheekbones* and *Smiling* demonstrate high protection rates across both datasets, indicative of the effectiveness of our protection methods in preserving key facial features while ensuring privacy. In contrast, attributes like *Blurry* and *Wearing Hat* exhibit variable protection rates, reflecting the challenges in consistently obfuscating certain features across different FRS models. A particularly intriguing observation is the generally higher protection rates on CelebA compared to FFHQ, which may be attributed to the former's training alignment with the classifier used in DiffAE for guiding attribute editing. The varied success rates across attributes underscore the nuanced complexity of editing facial features for privacy protection, with some attributes (e.g., *5 o Clock Shadow*, *Big Nose*) achieving lower protection rates on FFHQ, potentially due to the dataset's greater diversity and complexity. These findings highlight the importance of adapted approaches in adversarial editing to maximize protection efficacy across diverse facial attributes and recognition systems.

## 10 Ablation Study

In our ablation study, we explore the efficacy of DiffPrivate on Edit-based and Perturb-based methods: for Edit-based we selectively choose the target attribute as "Attractive". This study utilizes IRSE50 as the primary model and FaceNet as the validator, with FaceNet's detection threshold $\theta$ varying from 0.235 to 0.425, corresponding to a False Acceptance Rate (FAR) ranging from $1\times10^{-4}$ to $1\times10^{-1}$. This range is indicative of the privacy budget, essentially representing the degree of image modification permissible for crafting adversarial samples. It is critical to note that a lower FAR equates to a less stringent detection threshold, thereby facilitating the generation of adversarial samples with fewer alterations (details in Appendix D).

Dividing the threshold spectrum into 10 discrete steps, we present in Fig. 10 a visualization depicting how variations in the privacy budget influence image quality. Our observations reveal that the Edit-based method, which allows for targeted modifications on selected attributes, tends to yield visually appealing outcomes. Nonetheless, it is observed that the extent of identity alteration becomes more pronounced as the budget increases. Conversely, the DiffPrivate approach employing a Perturb-based methodology, typically results in less visual modifications when compared to the Edit-based strategy. This distinction underscores the inherent tradeoff between maintaining visual fidelity and achieving desired privacy levels.

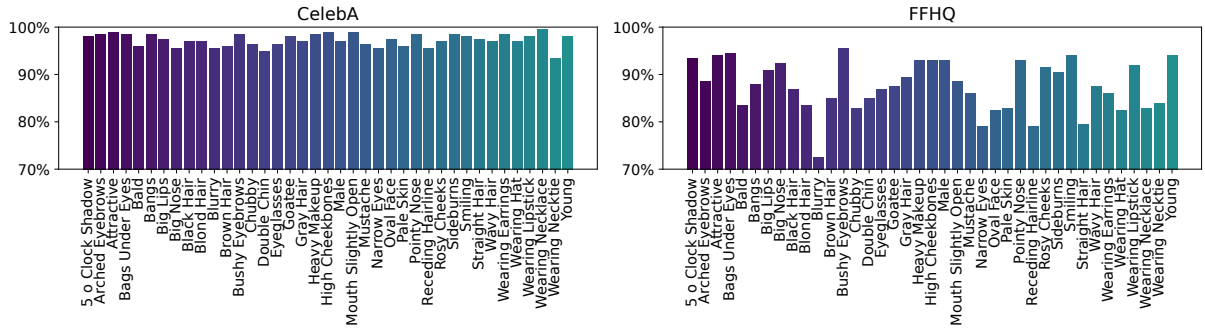



**Figure 9: The PPR crossing attributes for Edit-based approach**

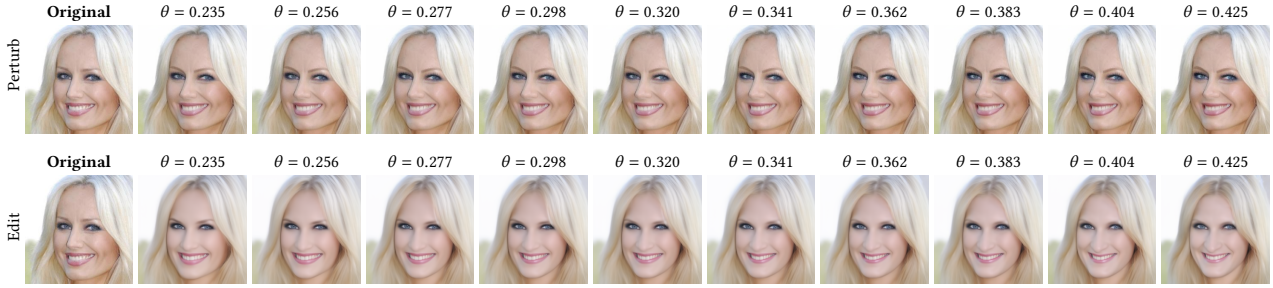



**Figure 10: Visualizations of protected face images using DiffPrivate (Perturb-based method first row and Edit-based method second row) at different protection budgets denoted by $\theta$.**

## 11 Limitations and Future Work

**Target Scope and Vulnerability to Advanced Attacks:** Referring to the SoK paper by Wenger et al. [62], our paper falls within the category of "Attacking to Evade Identification". Furthermore, our method, similar to approaches like Fawkes [50], focuses on fooling FRS rather than deceiving human perception. However, as noted by Todt et al. [58], adversarial attack methods, including ours, are relatively vulnerable to advanced attackers who can build/train reverse models with knowledge of the method. This is an inherent drawback of adversarial approaches (see [58] for detailed analysis of robustness against reversion and image quality based on user study and human evaluation). For enhanced robustness against such reversibility, we recommend anonymization methods like DeepPrivate [24], StyleID [30], and AnonFACES [32], which completely alter facial resemblance, and at the cost of lower usability.

**Dependency on Generative Model Quality:** The effectiveness of DiffPrivate depends on the performance of the underlying generative model, Stable Diffusion in our case. While Stable Diffusion excels in generativity, flexibility, and reduced bias compared to GANs [55, 69], the quality of reconstructed images is directly linked to its capabilities. Although it generally performs well, it occasionally still shows biases towards certain demographic groups that could impact the uniformity and fairness of our privacy protection.

**Computational Efficiency:** The reliance on deep diffusion models, which require significant computational resources for training and inference, might limit the accessibility of our methods to users with lower computational capabilities. Optimizing efficiency without compromising effectiveness remains future work.

**Adaptability to Advances in FRS:** Our method is tested against state-of-the-art FRS, but as adversarial defenses evolve, its future effectiveness is not guaranteed. The fast-changing landscape demands ongoing updates to ensure our approach remains effective.

**Extended Use Cases:** Our diffusion-based models, DiffPrivate, are designed not only to protect individual users' online personas on social media but can also be used to address broader privacy concerns, including as a component to help safeguarding bystander privacy in public datasets [17]. When images captured in social or public spaces are shared, bystanders - whose privacy may be compromised - can have their faces anonymized using DiffPrivate. This approach preserves the utility of these datasets for research or public safety while respecting individual privacy rights. Interesting future work include user studies determining people's current interest, expectation, and desirable privacy levels from such systems.

## 12 Conclusion

In conclusion, DiffPrivate leverages the diffusion process to create facial images that protect users' privacy. Our thorough testing highlights the important technical decisions behind DiffPrivate and its ability to work effectively with blackbox models. This work represents one of the initial attempts to use diffusion models for adversarial attacks aimed at safeguarding biometric privacy. Through this approach, we offer a novel solution that not only challenges FRS but also maintains the natural appearance of the images. Our findings underscore DiffPrivate's potential as a robust tool against the evolving landscape of FR technology, pushing forward the capabilities for privacy protection in digital spaces. Our code can be found here: https://github.com/minha12/DiffPrivate.

## Acknowledgments

This work was supported by the Swedish Research Council (VR) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors used ChatGPT to correct typos, grammar, and wordings.

## References

[1] Amazon Web Services. 2023. Amazon Rekognition. https://aws.amazon.com/rekognition/. Accessed: 2024-01-26.
[2] Nicholas Carlini, J Zico Kolter, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, and Mingjie Sun. 2023. (Certified!!) Adversarial Robustness for Free!. In *ICLR*.
[3] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. 39–57.
[4] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. 2020. Face-off: Adversarial face obfuscation. *arXiv preprint arXiv:2003.08861* (2020).
[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference Biometric Recognition (CCBR)*. Springer, 428–438.
[6] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. 2021. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *ICLR*.
[7] Thomas Cilloni, Wei Wang, Charles Walter, and Charles Fleming. 2020. Ulixes: Facial recognition privacy with adversarial machine learning. *arXiv preprint arXiv:2010.10242* (2020).
[8] Jim Cross. 2019. Valley attorney: Facebook facial recognition carries identity theft risk. *KTAR News* (September 2019). https://ktar.com/story/2735815/valley-attorney-facebook-facial-recognition-carries-identity-theft-risk/
[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.
[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *CVPR*.
[11] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. 2019. Efficient decision-based black-box adversarial attacks on face recognition. In *Proc. IEEE/CVF CVPR*. 7714–7722.
[12] Ivan Evtimov, Pascal Sturmfels, and Tadayoshi Kohno. 2020. Foggysight: A scheme for facial lookup privacy. *arXiv preprint arXiv:2012.08588* (2020).
[13] Face++. 2023. Face++ Cognitive Services. https://www.faceplusplus.com/. Accessed: 2024-01-26.
[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*.
[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
[16] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan A. Calian, and Timothy Mann. 2021. Improving Robustness using Generated Data. *arXiv preprint arXiv:2110.09468* (2021). https://arxiv.org/pdf/2110.09468
[17] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 318–335.
[18] Yonghao He, Dezhong Xu, Lifang Wu, Meng Jian, Shiming Xiang, and Chunhong Pan. 2019. LFFD: A light and fast face detector for edge devices. *arXiv preprint arXiv:1904.10633* (2019).
[19] Kashmir Hill. 2022. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*. Auerbach Publications, 170–177.
[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).
[21] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. 2022. Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-robust Makeup Transfer. In *CVPR*.
[22] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.
[23] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *IEEE/CVF CVPR*. 5901–5910.
[24] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*. Springer, 565–578.
[25] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. 2022. Adv-Attribute: Inconspicuous and Transferable Adversarial Attack on Face Recognition. In *NeurIPS*.
[26] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. 2019. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proc. IEEE/CVF ICCV*. 4773–4783.
[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.
[28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.
[29] Stepan Komkov and Aleksandr Petiushko. 2021. AdvHat: Real-world adversarial attack on ArcFace face ID system. In *ICPR*.
[30] Minh-Ha Le and Niklas Carlsson. 2022. Styleid: Identity disentanglement for anonymizing faces. *arXiv preprint arXiv:2212.13791* (2022).
[31] Minh Ha Le and Niklas Carlsson. 2024. StyleAdv: A Usable Privacy Framework Against Facial Recognition with Adversarial Image Editing. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*.
[32] Minh-Ha Le, Md Sakib Nizam Khan, Georgia Tsaloli, Niklas Carlsson, and Sonja Buchegger. 2020. Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security. In *Proc. WPES*. 87–100.
[33] Tao Li and Lei Lin. 2019. AnonymousNet: Natural face de-identification with measurable privacy. In *Proc. of CVPR*.
[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *International Conference on Computer Vision (ICCV)*.
[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
[36] Angelica Mari. 2019. Brazilian retailer quizzed over facial recognition tech. *ZDNet* (March 2019). https://www.zdnet.com/article/brazilian-retailer-quizzed-over-facial-recognition-tech/
[37] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proc. IEEE/CVF CVPR*. 6038–6047.
[38] Paul Mozur and Aaron Krolik. 2019. A surveillance net blankets China's cities, giving police vast powers. *The New York Times* 17 (2019).
[39] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*.
[40] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion Models for Adversarial Purification. In *ICML*.
[41] Kate O'Flaherty. 2019. Facial Recognition At U.S. Airports. Should You Be Concerned? *Forbes* (March 2019). https://www.forbes.com/sites/kateoflahertyuk/2019/03/11/facial-recognition-to-be-deployed-at-top-20-us-airports-should-you-be-concerned/
[42] Charles C-F Or, Kester YJ Ng, Yiik Chia, Jing Han Koh, Denise Y Lim, and Alan LF Lee. 2023. Face masks are less effective than sunglasses in masking face identity. *Scientific reports* 13, 1 (2023), 4284.
[43] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *CVPR*.
[44] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. 2020. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *ECCV*.
[45] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Fixing Data Augmentation to Improve Adversarial Robustness. *arXiv preprint arXiv:2103.01946* (2021). https://arxiv.org/pdf/2103.01946
[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF CVPR*. 10684–10695.
[47] Adam Satariano. 2019. Police use of facial recognition is accepted by British court. *The New York Times* 4 (2019).
[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*.
[49] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Proc. of NeurIPS*.
[50] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*. 1589–1604.
[51] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2019. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)* 22, 3 (2019), 1–30.
[52] Maya Shwayder. 2020. Clearview AI's facial-recognition app is a nightmare for stalking victims. *Digital Trends* (January 2020). https://www.digitaltrends.com/news/clearview-ai-facial-recognition-domestic-violence-stalking/
[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.
[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *ICLR*.
[55] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. 2024. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proc. IEEE/CVF WACV*. 5091–5100.

[56] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. 2018. A hybrid model for identity obfuscation by face replacement. In *Proc. of ECCV*.
[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv:1312.6199* (2013).
[58] Julian Todt, Simon Hanisch, and Thorsten Strufe. 2022. Fant\^ omas: Understanding Face Anonymization Reversibility. *arXiv preprint arXiv:2210.10651* (2022).
[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
[60] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. 2023. Better Diffusion Models Further Improve Adversarial Training. *arXiv preprint arXiv:2302.04638* (2023).
[61] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.
[62] Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y Zhao. 2023. Sok: Anti-facial recognition technology. In *IEEE S&P*. 864–881.
[63] Eric Wong and J Zico Kolter. 2021. Learning perturbation sets for robust machine learning. In *ICLR*. https://openreview.net/forum?id=MIDckA56aD
[64] Yifan Wu, Fan Yang, and Haibin Ling. 2018. Privacy-Protective-GAN for face de-identification. *arXiv:1806.08906* (2018).
[65] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. In *IJCAI*.
[66] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. 2021. Towards face encryption by generating adversarial identity masks. In *CVPR*.
[67] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. 2021. Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition. *arXiv preprint arXiv:2105.03162* (2021).
[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc IEEE CVPR*. 586–595.
[69] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems* 32 (2019).
[70] Chen Zhu, W Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2019. Transferable clean-label poisoning attacks on deep neural nets.

# Supplementary Material

## A Facial Recognition System

FRS typically have two primary functions: face verification and face identification.

**Face Verification:** In face verification, we aim to determine whether two face images, $\vec{X}_1$ and $\vec{X}_2$, belong to the same individual. We achieve this by:

(1) *Embedding*: Each image is transformed into a low-dimensional space using a deep learning model denoted by $\mathcal{F}$. This results in embedding vectors $\vec{e}_1$ and $\vec{e}_2$ for the respective images.
(2) *Distance Calculation*: We compute the distance between these vectors using a distance function $D = d(\vec{e}_1, \vec{e}_2)$.
(3) *Thresholding*: This distance is compared to a threshold $\tau$. If $D \leq \tau$, we conclude the images belong to the same person; otherwise, they differ.

The threshold $\tau$ balances the False Acceptance Rate (FAR) and False Rejection Rate (FRR) of the system. A lower $\tau$ reduces FAR but increases FRR, and vice versa.

**Face Identification:** Given a probe image $X$ and a gallery set $G$ of known identities, face identification seeks to identify the corresponding identity in $G$ for $X$. The process follows:

(1) *Embedding*: Similar to face verification, $X$ is transformed into a low-dimensional embedding $e$ using the same model $\mathcal{F}$.
(2) *Distance Calculation*: We calculate the distance $D_{ij}$ between $e$ and each embedding $e_{ij}$ in $G$ for every identity $I_i$ (containing $n_i$ embeddings).
(3) *Thresholding*: Depending on the implementation, the probe is identified with the identity having either the: (a) minimum mean distance $i^* = \arg\min_i \left(\frac{1}{n_i}\sum_{j=1}^{n_i} D_{ij}\right)$ or the (b) smallest individual distance $i^* = \arg\min_i \left(\min_j D_{ij}\right)$
(4) *Optional Thresholding*: A rank threshold $r$ can be included. In this case, if the rank $r^*$ of the best-matching identity $i^*$ is less than or equal to $r$, the identification is successful; otherwise, $X$ is classified as unknown.

Similar to face verification, a lower rank threshold $r$ decreases FAR but increases FRR, while a higher $r$ increases FAR but lowers FRR. Therefore, like threshold $\tau$, $r$ must be carefully chosen to optimize system performance and balance FAR and FRR according to the specific application.

## B Detailed Dataset Description

**Labeled Faces in the Wild (LFW):** The LFW dataset consists of over 13,000 facial images of more than 5,700 individuals. It has been designed to assess the performance of FR algorithms under uncontrolled, real-world conditions. The standard LFW benchmark, focusing on face verification, involves comparing 6,000 pairs of faces to determine whether they depict the same person. This benchmark is critical for evaluating the effectiveness of FRS in diverse and challenging scenarios. We utilize LFW in our evaluation framework to benchmark the FR capabilities of the systems.

**CelebA and CelebA-HQ:** CelebA is a large-scale facial attributes dataset designed to support research in facial attribute recognition, face detection, and landmark localization. It contains over 200,000 images of more than 10,000 celebrities, each annotated with 40 attribute labels and five landmark locations. CelebA-HQ, a high-quality subset of CelebA, comprises 30,000 images at a resolution of 1024×1024 pixels. These datasets are benchmarks for general tasks in image generation and editing, providing a diverse and comprehensive resource for evaluating algorithm performance in these domains. In our work, CelebA serves as the training dataset for the classifier of the DiffAE model, while CelebA-HQ is used as training data for the DiffPure model on faces.

**Flickr-Faces-HQ (FFHQ):** The FFHQ dataset consists of 70,000 high-quality images at 1024×1024 resolution, showcasing a wide diversity in age, ethnicity, and background among the subjects. It is particularly useful for training and testing style-based generative models, such as StyleGAN, for tasks like image generation, manipulation, and style transfer. The FFHQ dataset's emphasis on image quality and diversity makes it an ideal training dataset (in-domain) for the DiffAE model, setting a high standard for generative image quality and style transfer evaluation.

## C Additional Results

### C.1 Blackbox Evaluation with Attribute Editing

Table 2 provides a comprehensive evaluation of the protection rate across various attributes in a blackbox setting, comparing performance on the CelebA and FFHQ datasets across multiple FRS, including IRSE50, IR152, IR101, and MobileFace.

**Table 2: Comparison of different blackbox models across attributes on CelebA and FFHQ datasets.**

| Attribute | CelebA | | | | FFHQ | | | |
|---|---|---|---|---|---|---|---|---|
| | IRSE50 | IR152 | IR101 | MobileFace | IRSE50 | IR152 | IR101 | MobileFace |
| 5 o Clock Shadow | 26.0 | 55.5 | 30.5 | 24.5 | 41.5 | 62.5 | 31.5 | 36.5 |
| Arched Eyebrows | 43.0 | 64.0 | 34.0 | 35.5 | 42.5 | 63.5 | 36.0 | 34.5 |
| Attractive | 41.0 | 68.5 | 40.5 | 34.5 | 46.0 | 69.5 | 45.0 | 39.0 |
| Bags Under Eyes | 46.0 | 65.5 | 44.0 | 42.5 | 57.5 | 70.0 | 38.5 | 45.0 |
| Bald | 41.0 | 64.0 | 32.0 | 30.0 | 47.0 | 66.0 | 27.5 | 35.5 |
| Bangs | 48.0 | 65.0 | 33.0 | 36.0 | 49.5 | 64.0 | 38.0 | 42.0 |
| Big Lips | 34.0 | 63.0 | 33.0 | 25.0 | 45.5 | 63.0 | 37.5 | 34.0 |
| Big Nose | 32.5 | 64.0 | 36.0 | 30.5 | 50.5 | 66.5 | 39.0 | 36.5 |
| Black Hair | 32.0 | 60.0 | 31.0 | 26.0 | 40.0 | 66.5 | 32.0 | 33.5 |
| Blond Hair | 37.0 | 59.0 | 36.0 | 23.0 | 42.0 | 57.0 | 31.5 | 30.5 |
| Blurry | 48.0 | 62.0 | 33.0 | 29.5 | 38.0 | 50.0 | 19.0 | 23.5 |
| Brown Hair | 42.0 | 61.0 | 35.0 | 31.0 | 45.0 | 62.0 | 39.0 | 34.5 |
| Bushy Eyebrows | 29.5 | 57.5 | 28.0 | 24.0 | 32.0 | 55.0 | 24.5 | 25.5 |
| Chubby | 44.0 | 66.5 | 37.0 | 31.5 | 43.5 | 60.0 | 30.5 | 35.5 |
| Double Chin | 44.5 | 64.0 | 36.5 | 34.5 | 49.0 | 67.5 | 40.5 | 37.0 |
| Eyeglasses | 47.5 | 54.0 | 31.0 | 42.0 | 51.5 | 57.5 | 28.0 | 42.0 |
| Goatee | 32.5 | 53.0 | 32.0 | 27.0 | 48.0 | 58.5 | 27.0 | 40.5 |
| Gray Hair | 32.5 | 53.0 | 27.0 | 27.5 | 43.5 | 62.0 | 30.5 | 35.0 |
| Heavy Makeup | 35.5 | 68.5 | 39.0 | 34.0 | 42.0 | 68.0 | 41.5 | 41.5 |
| High Cheekbones | 51.0 | 74.5 | 47.5 | 40.0 | 64.0 | 78.0 | 53.5 | 53.0 |
| Male | 34.0 | 54.0 | 25.5 | 28.0 | 52.5 | 62.0 | 32.5 | 40.0 |
| Mouth Slightly Open | 62.0 | 69.0 | 53.5 | 46.5 | 65.5 | 71.0 | 51.0 | 46.5 |
| Mustache | 31.0 | 55.5 | 27.5 | 30.5 | 40.0 | 52.0 | 24.0 | 36.5 |
| Narrow Eyes | 63.5 | 72.0 | 51.5 | 44.5 | 61.0 | 69.0 | 45.5 | 48.0 |
| Oval Face | 51.0 | 70.5 | 40.0 | 35.5 | 44.0 | 62.5 | 37.5 | 32.5 |
| Pale Skin | 42.0 | 59.5 | 36.5 | 32.5 | 41.5 | 59.5 | 33.5 | 30.5 |
| Pointy Nose | 42.0 | 62.0 | 37.0 | 33.5 | 46.0 | 67.5 | 34.5 | 41.0 |
| Receding Hairline | 51.0 | 64.0 | 38.5 | 36.0 | 45.0 | 61.5 | 26.0 | 39.0 |
| Rosy Cheeks | 41.0 | 67.0 | 35.5 | 27.0 | 46.0 | 63.0 | 37.5 | 36.0 |
| Sideburns | 27.5 | 50.5 | 25.5 | 28.0 | 51.5 | 65.5 | 32.0 | 45.0 |
| Smiling | 57.0 | 67.0 | 49.5 | 44.0 | 66.5 | 73.0 | 58.0 | 52.5 |
| Straight Hair | 46.5 | 61.5 | 35.5 | 34.5 | 52.0 | 65.0 | 34.5 | 37.5 |
| Wavy Hair | 51.5 | 64.5 | 38.0 | 36.0 | 46.5 | 62.5 | 34.5 | 36.5 |
| Wearing Earrings | 39.0 | 63.5 | 33.0 | 29.5 | 38.5 | 61.5 | 34.0 | 34.0 |
| Wearing Hat | 52.5 | 69.0 | 41.5 | 37.5 | 55.5 | 65.0 | 35.5 | 39.0 |
| Wearing Lipstick | 33.5 | 71.0 | 36.0 | 31.0 | 44.0 | 69.0 | 44.0 | 40.5 |
| Wearing Necklace | 43.5 | 62.5 | 35.0 | 35.0 | 45.0 | 61.5 | 32.0 | 38.0 |
| Wearing Necktie | 34.5 | 54.0 | 27.0 | 28.0 | 48.0 | 62.0 | 30.5 | 37.0 |
| Young | 36.0 | 64.0 | 36.5 | 28.5 | 45.0 | 70.5 | 40.5 | 38.5 |

# D Detection Threshold at FAR Target

The effectiveness of facial recognition systems is essential in applications where high security and accurate identity verification are necessary. A crucial part of evaluating these systems involves setting an appropriate detection threshold $\theta$ that determines whether two facial images represent the same individual. This threshold plays a vital role in balancing the tradeoff between system usability and security, specifically in reducing false acceptances without significantly increasing false rejections. This section presents the mathematical formulation and approach for setting the detection threshold $\theta$ to achieve a desired FAR using the LFW dataset.

## D.1 Operational Definitions

Before explaining the methodology, it is important to define the key metrics:

- **True Positive (TP):** Correctly identifying a match between different images of the same individual.
- **False Positive (FP):** Incorrectly identifying a match between images of different individuals.
- **True Negative (TN):** Correctly identifying no match between images of different individuals.
- **False Negative (FN):** Incorrectly identifying no match between different images of the same individual.
- **False Accept Rate (FAR):** The probability of incorrectly granting access to an unauthorized individual, calculated as $\text{FAR} = \frac{\text{FP}}{\text{FP+TN}}$.

## D.2 Methodology

The process of determining the detection threshold $\theta$ for a specified FAR involves the following steps:

**Table 3: Detection thresholds and performance metrics for various FRS models at specified FAR targets on the LFW dataset**

| Model | FAR Target | Threshold | FAR Achieved | AUC | Accuracy |
|---|---|---|---|---|---|
| MobileFace | $1 \times 10^{-5}$ | 0.37153 | 0.00067 | 0.95012 | 0.9285 |
| | $1 \times 10^{-4}$ | 0.37177 | 0.00067 | | |
| | $1 \times 10^{-3}$ | 0.3875 | 0.00133 | | |
| | $1 \times 10^{-2}$ | 0.42018 | 0.00967 | | |
| | $1 \times 10^{-1}$ | 0.45514 | 0.10033 | | |
| IRSE50 (ArcFace) | $1 \times 10^{-5}$ | 0.39523 | 0.00033 | 0.95068 | 0.93667 |
| | $1 \times 10^{-4}$ | 0.39547 | 0.00033 | | |
| | $1 \times 10^{-3}$ | 0.4 | 0.001 | | |
| | $1 \times 10^{-2}$ | 0.4231 | 0.01 | | |
| | $1 \times 10^{-1}$ | 0.4571 | 0.09967 | | |
| IR152 (ArcFace) | $1 \times 10^{-5}$ | 0.38873 | 0.00033 | 0.94749 | 0.94233 |
| | $1 \times 10^{-4}$ | 0.38897 | 0.00033 | | |
| | $1 \times 10^{-3}$ | 0.4277 | 0.00133 | | |
| | $1 \times 10^{-2}$ | 0.44566 | 0.01033 | | |
| | $1 \times 10^{-1}$ | 0.47085 | 0.10067 | | |
| FaceNet | $1 \times 10^{-5}$ | 0.23503 | 0.00067 | 0.95338 | 0.9405 |
| | $1 \times 10^{-4}$ | 0.23527 | 0.00067 | | |
| | $1 \times 10^{-3}$ | 0.3352 | 0.001 | | |
| | $1 \times 10^{-2}$ | 0.37393 | 0.01 | | |
| | $1 \times 10^{-1}$ | 0.42536 | 0.1 | | |
| IR101 (CurricularFace) | $1 \times 10^{-5}$ | 0.41383 | 0.00033 | 0.94774 | 0.94183 |
| | $1 \times 10^{-4}$ | 0.41407 | 0.00033 | | |
| | $1 \times 10^{-3}$ | 0.4332 | 0.00133 | | |
| | $1 \times 10^{-2}$ | 0.45104 | 0.01033 | | |
| | $1 \times 10^{-1}$ | 0.47432 | 0.10067 | | |

(1) **Similarity Score Calculation:** Calculate similarity scores for pairs of images within the LFW dataset. These scores represent the degree of similarity between images, as determined by the facial recognition model.
(2) **FAR Calculation for Various Thresholds:** For a range of thresholds, calculate the FAR as $\text{FAR}(\theta) = \frac{\text{FP}(\theta)}{\text{FP}(\theta)+\text{TN}(\theta)}$, where $\text{FP}(\theta)$ and $\text{TN}(\theta)$ are the counts of false positives and true negatives, respectively, for the threshold $\theta$.
(3) **Target FAR Achievement:** Find the threshold $\theta$ that yields an FAR closest to the desired target $\text{FAR}_{\text{target}}$, thus optimizing the balance between security and usability.

## D.3 Optimization Problem

The optimization problem to identify the ideal detection threshold can be expressed as:

$$\min_{\theta} |\text{FAR}(\theta) - \text{FAR}_{\text{target}}| \tag{20}$$

This problem aims to minimize the absolute difference between the FAR at a given threshold and the target FAR, ensuring the facial recognition system operates within the desired security levels.

## D.4 Implementation Considerations

In practice, applying this method requires careful analysis of the similarity score distribution and may involve calibration techniques to adjust scores before applying the detection threshold. Additionally, the choice of optimization technique, such as binary search, depends on whether the FAR changes monotonically with respect to $\theta$.

## D.5 Experimental Results

We evaluated several facial recognition models on the LFW dataset to determine the detection threshold ($\theta$) required to achieve specific FAR. The models tested include MobileFace, IRSE50 (ArcFace), IR152 (ArcFace), FaceNet, and IR101 (CurricularFace). Each model was assessed at FAR targets of $1 \times 10^{-5}$, $1 \times 10^{-4}$, $1 \times 10^{-3}$, $1 \times 10^{-2}$, and $1 \times 10^{-1}$ to determine the corresponding threshold ($\theta$) and actual FAR achieved. Additionally, the Area Under the Curve (AUC) and accuracy of each model were recorded to provide a comprehensive evaluation of performance.

As indicated in Table 3, with a lower FAR, the threshold requirement is lower, making it easier to craft adversarial samples. However, this also leads to less transferability across models. For instance, lower thresholds provide effective protection but may not generalize well across different facial recognition systems.

These results highlight the variability in performance across different models and FAR targets. Notably, while some models perform exceptionally well at lower FAR targets, the accuracy and AUC provide additional layers of insight into their overall effectiveness in facial recognition tasks.