StyleAdv: A Usable Privacy Framework Against Facial Recognition with Adversarial Image Editing

Minh-Ha Le
Linköping University
Linköping, Sweden

Niklas Carlsson
Linköping University
Linköping, Sweden

ABSTRACT

In this era of ubiquitous surveillance and online presence, protecting facial privacy has become a critical concern for individuals and society as a whole. Adversarial attacks have emerged as a promising solution to this problem, but current methods are limited in quality or are impractical for sensitive domains such as facial editing.

This paper presents a novel adversarial image editing framework called StyleAdv, which leverages StyleGAN's latent spaces to generate powerful adversarial images, providing an effective tool against facial recognition systems. StyleAdv achieves high success rates by employing meaningful facial editing with StyleGAN while maintaining image quality, addressing a challenge faced by existing methods. To do so, the comprehensive framework integrates semantic editing, adversarial attacks, and face recognition systems, providing a cohesive and robust tool for privacy protection. We also introduce the "residual attack" strategy, using residual information to enhance attack success rates. Our evaluation offers insights into effective editing, discussing tradeoffs in latent spaces, optimal edits for our optimizer, and the impact of utilizing residual information.

Our approach is transferable to state-of-the-art facial recognition systems, making it a versatile tool for privacy protection. In addition, we provide a user-friendly interface with multiple editing options to help users create effective adversarial images. Extensive experiments are used to provide insights and demonstrate that StyleAdv outperforms state-of-the-art methods in terms of both attack success rate and image quality. By providing a versatile tool for generating high-quality adversarial samples, StyleAdv can be used both to enhance individual users' privacy and to stimulate advances in adversarial attack and defense research.

KEYWORDS

Adversarial samples, Privacy filter, Facial anonymization

1 INTRODUCTION

Facial recognition systems (FRS:s) have become a prevalent technology extensively used for a wide range of applications, including security, marketing, and social media [53, 58]. However, their diverse use in combination with increased collection of personal images has raised significant concerns about privacy and civil liberties [2].

One of the most significant concerns is how images collected from social media and other online sources may be used without user consent and/or for other purposes than the uploader intended. With the proliferation of digital cameras and social platforms, personal images are widely accessible online. Inappropriate use of such datasets can result in severe privacy breaches and data misuse.

The growing use of FRS:s, coupled with the frequent scanning and analysis of individuals' faces without their consent, has amplified privacy concerns [47]. As facial recognition technology becomes more widespread and personal image collections increase, there is an urgent demand for effective solutions to protect facial privacy on social media platforms and the internet.

Adversarial attacks have emerged as a promising approach to protect facial identity and enhance privacy [18]. These attacks involve subtle image perturbations designed to deceive FRS:s. Ideally, these alterations remain imperceptible to humans while significantly affecting the accuracy of recognition systems.

Despite adversarial attacks being a promising solution to protect the facial identity and enhance privacy, existing adversarial sample methods have been limited in quality and have proven unusable for sensitive domains such as facial editing. More generally, current methods have at least one of the following limitations: requiring white-box access to the models under attack [6, 18], not being practical in real-life settings [30], having low attack success rates or being effectively bypassed or detected by state-of-the-art defense methods [44], or being visible to the "trained" eye [10]. These limitations have hindered the development of effective and practical adversarial attacks that can operate in real-world scenarios.

In this paper, we present StyleAdv, a robust and comprehensive framework for adversarial image editing. StyleAdv exploits the capabilities of deep generative models like StyleGAN [27] and introduces a novel approach for generating high-quality adversarial samples. By obfuscating facial features in a way that is imperceptible to humans but tricks FRS:s, StyleAdv allows users to easily create slightly altered images that they can share on the internet (e.g., social media) without automated FRS:s easily identifying them.

The design of StyleAdv is motivated by the observations that many users perform smaller edits before uploading on many social media websites such as Facebook and Instagram [11, 15, 35, 51], while others may desire to protect their identity with minimal manipulation. To achieve our objectives, StyleAdv incorporates both targeted and untargeted manipulation in the latent spaces of StyleGAN, as well as a unique residual attack strategy. First, StyleAdv seamlessly integrates semantic-aware editing, leveraging leading facial editing methodologies, with a state-of-the-art adversarial attack module and face recognition systems. Second, it incorporates an innovative optimizer, adept at exploiting the non-visible noise introduced by minor face edits, to perform guided attacks within StyleGAN's latent spaces. Finally, it innovatively utilizes a residual attack strategy, which leverages residual information in the generated adversarial images to increase attack success rates. Through

Proceedings on Privacy Enhancing Technologies 2024(2), 106–123

https://doi.org/10.56553/popets-2024-0043

this coordinated and nuanced manipulation of the latent spaces of StyleGAN, our framework can strategically modify the identity of the generated face, aiming for a target identity or an untargeted alteration, while preserving high image quality. As demonstrated by our results, StyleAdv can maintain impressive realism and versatility in the produced adversarial images, surpassing previous methods, making it an efficient, practical, and versatile tool for protecting facial privacy across various digital platforms.

In summary, the contributions of StyleAdv are as follows:

- **Novel Adversarial Editing Framework:** We introduce an innovative adversarial image editing framework that operates within StyleGAN's latent spaces and that offers a unique and powerful tool against FRS:s. Our solution is demonstrated to achieve high success rate, performing meaningful facial editing while preserving image quality, an achievement that remains a challenge for existing methods.
- **Innovative Attack Strategy and Mechanistic Insights:** We introduce the "residual attack" strategy, leveraging residual information within adversarial images to enhance attack success rates significantly. Furthermore, we provide insights into the mechanics of effective editing, detailing the tradeoffs of different latent spaces, the edits that best augment the optimizer, and the benefits of utilizing residual information.
- **Comprehensive Robust Framework and User-friendly Interface:** Our framework integrates various modules, including semantic editing, adversarial attacks, and face recognition systems, to deliver a cohesive and robust tool for privacy protection. Code (with easy-to-run demos) and an easy-to-use interface can be found on github (https://github.com/minha12/StyleAdv). Here, users can upload photos and, in return, receive safeguarded versions.
- **Superior Performance:** Comparison with existing adversarial attack methods show that StyleAdv outperforms prior works both in terms of attack success rate and image quality.

By combining a user-friendly web interface and superior performance compared to prior works, StyleAdv's novel solutions mark a significant advancement in adversarial attacks, empowering users to safeguard their privacy in today's increasingly digital world.

Outline: After presenting background and related work (§2), we present the StyleAdv framework (§3), starting from a problem definition and then stepwise developing the solution approach and the attack variations considered. Experimental results and insights are presented next (§4), before we conclude the paper (§5).

2 BACKGROUND AND RELATED WORKS

This section presents an overview of related works on adversarial attacks against FRS:s face editing, as well as an in-depth discussion on face verification and face identification.

Face Verification: In face verification, the goal is to compare two face images, $\vec{X}_1$ and $\vec{X}_2$, and determine whether they belong to the same individual. To facilitate this, each image is transformed into a low-dimensional embedding ($\vec{e}_1 = \mathcal{F}(\vec{X}_1), \vec{e}_2 = \mathcal{F}(\vec{X}_2)$, respectively) using a trained deep learning facial embedding model $\mathcal{F}(\vec{X})$. A face verification system can then be defined by a distance function that takes these two embeddings as input and outputs a distance metric $D = d(\vec{e}_1, \vec{e}_2)$. To make the final decision, this distance is then compared to a threshold τ:

$$\text{decision} = \begin{cases} 1, & \text{if } D \leq \tau \text{ (same person)}, \\ 0, & \text{otherwise (different persons).} \end{cases} \quad (1)$$

In practice, the chosen threshold τ should carefully balance the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), tailored to the specific requirements of the application at hand.

Face Identification: In face identification, a probe face image X is compared with a gallery set $\mathbb{G}$ of known identities $\mathbb{G} = \{I_1, I_2, ..., I_m\}$, where each identity I_i comprises one or more embeddings. First, the probe image $\vec{X}$ is transformed into an embedding $e = \mathcal{F}(\vec{X})$ using the same model $\mathcal{F}$ that was used to create the embeddings in the gallery set $\mathbb{G}$. Second, a distance function d is used to computes the distance $D_{ij} = d(e, e_{ij})$ between the probe image's embedding e and each embedding e_{ij} $(1 \leq j \leq n_i)$ of every identity I_i $(1 \leq i \leq m)$ in the gallery set $\mathbb{G}$. Depending on the specific implementation, the probe image is then identified as the identity that has either the lowest mean distance (i.e., $i^* = \arg\min_i(\frac{1}{n_i}\sum_{j=1}^{n_i} D_{ij}))$ or the smallest individual distance among its embeddings (i.e., $i^* = \arg\min_i(\min_j D_{ij})$). In addition, the system may incorporate a rank threshold r as part of its decision criteria. If the rank r^* of the best matching gallery identity i^* is less than or equal to a threshold r, the identification is deemed successful; otherwise, the probe identity is classified as unknown.

The rank threshold r plays a pivotal role in determining the system's performance and accuracy. To see this we note that the threshold is intricately tied to the system's FAR and FRR, where the FAR in a face identification scenario corresponds to the likelihood of the system mistakenly associating a probe face with an incorrect identity from the gallery (i.e., incorrectly ranking this identity within the top r matches) and the FRR is the probability of the system failing to correctly identify a face present in the gallery (i.e., not ranking the correct identity within the top r matches).

Adversarial Attacks: Several ways to trick machine learning models to misclassify or produce other incorrect outputs have been proposed. These so-called *adversarial attacks* typically exploit vulnerabilities of the models by injecting malicious inputs. For example, several works have added carefully crafted perturbations to the input data [6, 18, 34] that are small or imperceptible to the human eye but that can cause the machine learning model under attack to produce incorrect outputs. Others have used generative adversarial network (GAN) to create adversarial examples [23, 56] that can fool a machine learning model. This approach has been used in both semi-whitebox and black-box attack settings, as well as a defense method [41, 42]. Lately, there have also been a growing interest in developing physical adversarial attacks [8, 9, 14, 30] in which the input data is manipulated in the real world, with several such attacks having serious consequences in critical applications such as autonomous driving [5, 13, 48] and medical diagnosis [16, 33].

Attacks Against Facial Recognition Systems: Adversarial attacks against FRS:s have been designed both for the physical world [29, 45, 49] and the online world [10, 22, 39, 44]. For the physical-world context, researchers have proposed attacks where eyeglasses are used to evade recognition or impersonate other individuals [45], black-box attacks involving placing a paper sticker on a hat to avoid facial detection [29], and generated patches that can be used to hide a person [49] or object [54] from detection.

Table 1: Comparison to the related works

Method	Image	Protec-	Editing	General-	Shifts
Method	quality	tion	options	izable	to target
Fawkes [44]	low	low	no	no	yes
Lowkey [10]	low	low	no	no	yes
AMT-GAN [22]	high	high	1 (only makeup)	yes	no
SemanticAdv [39]	low	high	N	yes	yes
StyleAdv (ours)	high	high	not limited	yes	yes

Most closely related are works adding perturbations to images to protect user privacy [10, 22, 39, 44]. Here, we briefly describe the four most related state-of-the-art solutions proposed recently, and in Secs. 4.3 and 4.9 we compare our performance against them. (1) Fawkes, proposed by Shan et al. [44], provides users with a method to inoculate their images against unauthorized facial recognition models by adding imperceptible pixel-level changes (called "cloaks") to their photos before releasing them. (2) Cherepanova at al. [10] develop an adversarial filter called Lowkey, shown to significantly degrade the accuracy of Amazon Rekognition and the Microsoft Azure Face Recognition API. Both the above works add noise to the images, compromising image quality, and offer comparatively less protection than the following two solutions as well as our proposed method. (3) Hu et al. [22] propose what they call an adversarial makeup transfer AMT-GAN that uses a GAN to synthesize adversarial face images with makeup transferred from reference images to achieve a desirable balance between the attack strength and the amount of visual changes. (4) Qui et al. [39] propose SemanticAdv, an algorithm that leverages disentangled semantic factors to generate adversarial perturbation by altering controlled semantic attributes to fool the learner towards various "adversarial" targets. The authors demonstrate that the semantic-based adversarial examples can fool different learning tasks and achieve high targeted attack success rate against real-world black-box services such as Azure face verification service based on transferability.

While the latter two solutions offer enhanced protection, as shown in Table 1, all four approaches have their respective shortcomings when compared to our adversarial editing solution. We refer to Secs. 4.3 and 4.9 for performance comparisons and numeric support for the classifications in the table. However, as shown, our approach is the only one that achieves all of the desirable properties.

StyleGAN and Semantic Editing in Latent Space: Generative Adversarial Networks (GANs) [17] have proven very successful for generating realistic images [3, 26, 27]. One of the most influential such models is StyleGAN [27], which introduced a new generator architecture for GANs that enables intuitive, scale-specific control of image synthesis, and that outperformed the state-of-the-art.

Inspired by its high-quality results, several follow-up works have explored and analyzed StyleGAN's disentangled latent spaces and propose semantic image editing solutions. For example, GANSpace [20] performs a comprehensive PCA analysis and shows that a set of semantic attributes can be mapped to particular principal components. InterFaceGAN [46] provides methods to find attribute boundaries on particular attributes with the help of an SVM model.

In our study, we utilize two recent advancements, StyleSpace [55] and StyleFlow [1], for face editing purposes. StyleSpace [55] discovered a higher disentanglement space within StyleGAN's architecture called S-Space, and performs editing within this space. In contrast, StyleFlow [1] uses normalizing flows to improve image editing in StyleGAN's latent space. In other related work, Patashnik et al. [38], propose StyleCLIP, that effectively combines a CLIP model with StyleGAN latent space editing methods to provide an interesting text-to-image editing interface. No prior work has used any of the above semantic editing tools for the purpose of creating adversarial samples. Here, we incorporate them into our solution, and show how the meaningful edits that they allow can be used as meaningful noise into our optimizer that pushes the identity away from the original source identity and towards a target identity. The resulting tool provides meaningful adversarial image editing.

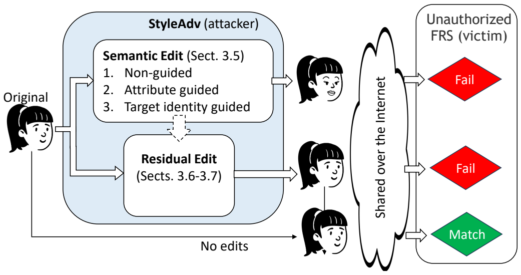

StyleAdv (attacker)
Semantic Edit (Sect. 3.5)
1. Non-guided
2. Attribute guided
3. Target identity guided
Original
Residual Edit (Sects. 3.6-3.7)
No edits
Shared over the Internet
Unauthorized FRS (victim)
Fail
Fail
Match

Figure 1: Use case example and workflow overview

Use Case Comparison: We present novel privacy filter tailored to combat state-of-the-art FRS:s, which typically deploy embedding-based models. This is in contrast to most prior adversarial attack studies, which have target classifier models. StyleAdv also stands distinct from Fawkes and the broader category of data poisoning techniques. While data poisoning, exemplified by Fawkes, focuses on altering the training data to mislead or degrade the performance of machine learning models, StyleAdv aims to protect against already trained FRS:s. A key advantage here is StyleAdv's ability to seamlessly integrate semantic-aware editing and residual editing with state-of-the-art adversarial attack modules, resulting in images that are visually appealing to humans but confounding to FRS:s. Another advantage of StyleAdv is its flexibility and ability to work on individual images, allowing users to choose which images to safeguard. While Fawkes does not require every image of the user to be cloaked, for maximum effectiveness, users must apply the cloaking to the majority of their images before posting online.

3 ADVERSARIAL EDITING WITH StyleAdv

We first describe the general facial recognition model under attack and the general approach (and its variations) that we use to create our adversarial samples. Later in the paper we show how our novel approach and its variations (including a residual-based extension) are applicable to any FRS and how they can be used together with a broad range of face editing approaches operating in the different latent spaces of StyleGAN.

3.1 Use Case Scenario and Workflow Overview

Fig. 1 presents an overview of the workflow of our privacy-preserving filter. Here, a user (1) uploads a facial image to StyleAdv, (2) selects which editing approach should be used to protect their identity, and then (3) uses the edited image generated by StyleAdv to upload to websites and social media accounts, for example. The goal of StyleAdv is to generate an edited images that can bypass unauthorized FRS:s, while ensuring that the edits are visually subtle and retain resemblance with the original image.[1] To achieve its goal, StyleAdv implements two high-level editing approaches.

- **Semantic-aware Adversarial Editing:** With this approach, semantic editing is used to modify the face in one of three ways: non-guided, attribute guided, or target identity guided.
- **Residual Editing:** With this approach, invisible noise is added to the image so as to hide the identity of the individual while otherwise trying to preserve the image.

In the figure, we illustrate that the identity is successfully protected when the FRS used by the unauthorized system fails to match the identity of the edited image (red diamonds). In contrast, an image that does not undergo any such edits typically allows the FRS to successfully identify the individual (green diamond). To try out the solution, we refer to our user-friendly demo.

3.2 Threat Model

In the above scenario, StyleAdv is considered the "attacker" and the system performing unauthorized facial recognition is considered the "victim" of our attack. To facilitate our unique design we assume that the attacker (i.e., StyleAdv) has (1) white-box access to one model (referred to as the "adversarial" model $\mathcal{M}^A$) and (2) black-box access to a second model (referred to as the "victim" model $\mathcal{M}^V$).

At the core of our attack is an optimization using the adversarial model $\mathcal{M}^A$, which we use to improve the adversarial samples in a semantically meaningful way. However, motivated by the observation that many FRS:s today are public (even when their internals may not be public), we incorporate a "black-box" victim model $\mathcal{M}^V$ that we use for a limited number of embedding queries (e.g., 1 to 3 attempts) as part of the stopping criteria for the editing optimizations. When we know the victim model, such queries offer valuable feedback for adjusting edit levels to maximize success rates. By limiting the number of queries, the attack can remain stealthy even when the victim model can only be queried online.

Knowledge of Victim Model: We note that our attack is applicable in three cases. First, in the case that the victim model is known, and the attacker has full access to it, then this model can be used as also advertorial model (this would give the best results). Second, in the case that the victim model is known but the attacker can only query it (in online or offline mode), then the victim model can be used for the stopping criteria in StyleAdv (resulting in small number of queries to the model). Third, in the case that the victim model is not known, the attacker can pick the adversarial and the victim models so as to optimize StyleAdv's chance based on experiments with different combinations of models. We refer to Sec. 4.8 for transferability results and a discussion regarding which models may be best for both general and targeted attacks.

3.3 Problem Definition & Solution Approach

Unless stated otherwise, in the following, we describe the attack done using the adversarial model $\mathcal{M}^A$ and drop the superscript A.

Facial Recognition Model under Attack: We consider a machine learning model $\mathcal{M}$ that is trained on a dataset $\mathcal{D} = \{(\vec{X}_i, \vec{y}_i)\}$ consisting of image-label pairs. Here, $\vec{X}_i \in \mathbb{R}^{H\times W\times C}$ represents an image with height H, width W, and C color channels, and $\vec{y}_i \in \mathbb{R}^K$ denotes the corresponding ground-truth label with K dimensions. Given an image $\vec{X}_i$, the goal of the model $\mathcal{M}$ is to produce a prediction $\hat{\vec{y}}_i = \mathcal{M}(\vec{X}_i) \in \mathbb{R}^K$, which can be used to classify the image.

Our Problem: As an attacker of the above model, our aim is to synthesize adversarial samples $\vec{X}^{\text{adv}}$ that the model under attack would label as the target $\vec{y}^{\text{tgt}}$ or some other label than the true label $\vec{y}_i$. In our evaluation, we also consider how "far away" the sample is from being labelled correctly (using distance- and rank-metrics).

A Naive Approach: A traditional method for this is to generate an adversarial example $\vec{X}^{\text{adv}}$ such that $\mathcal{M}(\vec{X}^{\text{adv}}) = \vec{y}^{\text{tgt}}$ by either adding pixel-wise perturbations or by spatially transform the original image $\vec{X}_i$. As noted above, this approach is taken by some prior works (e.g., Fawke [44] and Lowkey [10]) at the expense of (among other things) image quality. In this paper, we take a different approach, which we refer to as a semantic attacker. We next describe our high-level approach and a residual-based extension.

Our Semantic-aware Adversarial Editing Approach: This approach uses face edits in latent space to guide the attack in a semantically meaningful direction. To achieve greater control in the generation of adversarial samples, we edit a single semantic aspect of an image using attribute-conditioned image editing with a conditional generative model $\mathcal{G}$. More specifically, given a target image-label pair $(\vec{X}^{\text{tgt}}, \vec{y}^{\text{tgt}})$ and $\vec{y}_i \neq \vec{y}^{\text{tgt}}$, our semantic attack generates adversarial samples by editing a single semantic aspect of the original image $\vec{X}_i$ such that $\mathcal{M}(\vec{X}^{\text{adv}}) = \vec{y}^{\text{tgt}}$, while preserving the other semantic aspects of the original. This is achieved by conditioning the generative model $\mathcal{G}$ on the attribute corresponding to the semantic aspect that we want to edit, and by keeping the other attributes constant or close to the original values. The resulting adversarial example $\vec{X}^{\text{adv}}$ is then passed to the model $\mathcal{M}$ to verify whether it produces the desired label $\vec{y}^{\text{tgt}}$. Compared to adding noise at the pixel level, we have found that this approach provides more interpretable and semantically meaningful perturbations.

Our Residual-based Adversarial Attack: Let $\mathcal{E}$ denote the encoder model that projects the latent codes from an image. Given a target image $\vec{X}_i$, we can reconstruct it in the latent space as $\vec{X}'_i = \mathcal{G}(\mathcal{E}(\vec{X}_i))$. In this method, we define the residual $\vec{R}_i$ between the original and reconstructed image as $\vec{R}_i = \vec{X}_i - \vec{X}'_i$. This residual captures the differences that the generative model $\mathcal{G}$ could not reconstruct. We then use an encoder $\mathcal{E}_R$ to encode this residual and add it back to the reconstructed image in the generative model to obtain an enriched image $\vec{X}^*_i$ as $\vec{X}^*_i = \mathcal{G}(\mathcal{E}(\vec{X}_i), \mathcal{E}_R(\vec{R}_i))$.

The adversarial attack is then performed by trying to perturb this residual $\vec{R}_i$ by $\Delta\vec{R}_i$ so as to obtain the adversarial residual $\vec{R}^{\text{adv}}_i$ that produces an adversarial image $\vec{X}^{\text{adv}}_i = \mathcal{G}(\mathcal{E}(\vec{X}_i), \mathcal{E}(\vec{R}^{\text{adv}}_i))$ that results in the model $\mathcal{M}$ outputting the target label $\vec{y}^{\text{tgt}}$.

This attack thus seeks to find the minimal perturbation $\Delta\vec{R}_i = \vec{R}^{\text{adv}}_i - \vec{R}_i$ that fulfills $\mathcal{M}(\vec{X}^{\text{adv}}_i) = \vec{y}^{\text{tgt}}$, where $\vec{y}_i \neq \vec{y}^{\text{tgt}}$. Since $\mathcal{E}_R(\vec{R}^{\text{adv}}_i) = \mathcal{E}_R(\vec{R}_i + \Delta\vec{R}i)$, we can formulate this as an optimization problem:

$$\min_{\Delta\vec{R}_i} |\Delta\vec{R}_i|_2^2, \tag{2}$$

subject to

$$\mathcal{M}(\mathcal{G}(\mathcal{E}(\vec{X}_i), \mathcal{E}_R(\vec{R}_i + \Delta\vec{R}_i)) = \vec{y}^{\text{tgt}}. \tag{3}$$

[1] Note that our focus is to bypass automated FRS:s, not preventing manual/human recognition (Appendix D). These defense targets (i.e., automated FRS:s) are far more prevalent and scalable than manual reviews and allow us to provide useful images.

The above optimization problem seeks to find the smallest perturbation in the encoded residual space that would cause the facial recognition model to misclassify the image into the desired target label. This residual-based adversarial attack offers a nuanced strategy for creating adversarial examples, optimizing the balance between imperceptibility and misclassification effectiveness.

3.4 Loss Terms and Optimization Loop

To form the adversarial attacks as an optimization problem, we introduce three types of loss functions: (1) the identity loss $\mathcal{L}_{\text{ID}}$, (2) the perceptual loss $\mathcal{L}_{\text{P}}$, and (3) the square error loss $\mathcal{L}_{\text{SE}}$. Before going into the subtle differences in how these losses are combined and used, we first provide a high-level description of each term.

Identity Loss: For identity loss, we employ state-of-the-art facial embedding models (e.g., FaceNet [43], ArcFace [12], CurricularFace [25], and MobileFaceNet [7]), denoted as $\mathcal{F}$. These models encode facial images into low-dimensional vectors, capturing the face's identity while disregarding variations like pose and lighting.

Formally, given an original image $\vec{X}_i$ and its adversarial image $\vec{X}_i'$, the identity loss $\mathcal{L}_{\text{ID}}$ is defined as the pairwise distance

$$\mathcal{L}_{\text{ID}} = d(\vec{e}_i, \vec{e}_i') \tag{4}$$

between the two embeddings $\vec{e}_i = \mathcal{F}(\vec{X}_i)$ and $\vec{e}_i' = \mathcal{F}(\vec{X}_i')$. This distance measures how close the identity of the adversarial image is to the identity of the original image. (In the case of our target-identity guided attack, we instead use an identity loss defined relative to the embedding of a target identity.)

Perceptual Loss: Perceptual loss, denoted by $\mathcal{L}_{\text{P}}$, measures the perceptual similarity between the original and adversarial images. It is defined based on the Learned Perceptual Image Patch Similarity (LPIPS) [57], a perceptual metric that uses a deep neural network, denoted by $\mathcal{N}$, to learn a distance function consistent with human perception. It compares the feature representations of two images at multiple layers in the network, taking into account the statistics of natural images and the hierarchical structure of visual perception. Formally, we can express the perceptual loss as follows:

$$\mathcal{L}_{\text{P}} = \sum l \in \mathcal{L} \left| \mathcal{N}_l \left(\mathcal{G} \left(\vec{W}_{\text{new}} \right) \right) - \mathcal{N}_l \left(\vec{X}_i \right) \right|_2^2, \tag{5}$$

where $\mathcal{L}$ is the set of layers used for the perceptual loss, $\mathcal{N}_l$ represents the feature extraction of network $\mathcal{N}$ at layer l, $\mathcal{G}$ is the generator that produces the adversarial image from the manipulated latent code $\vec{W}_{\text{new}}$, and $\vec{X}_i$ is the original image.

Square Error Loss: The square error loss, denoted by $\mathcal{L}_{\text{SE}}$, is a regularization term designed to penalize deviations between the original and the manipulated versions of a given input. This loss is versatile and can be applied to various types of inputs, such as latent codes $\vec{W}$, images $\vec{X}$, or residual vectors $\vec{R}$. In a generalized context, let $\vec{V}_{\text{org}}$ represent the original input (which could be either $\vec{W}$, $\vec{X}$, or $\vec{R}$) and $\vec{V}_{\text{new}}$ denote its manipulated counterpart. The generalized Square Error Loss can then be formulated as:

$$\mathcal{L}_{\text{SE}}(\vec{V}_{\text{org}}, \vec{V}_{\text{new}}) = \left| \vec{V}_{\text{new}} - \vec{V}_{\text{org}} \right|_2^2. \tag{6}$$

Combined Loss and Optimization Formulation: Finally, the overall loss function is formulated as a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{P}} \cdot \mathcal{L}_{\text{P}} + \lambda_{\text{SE}} \cdot \mathcal{L}_{\text{SE}} - \lambda_{\text{ID}} \cdot \mathcal{L}_{\text{ID}}, \tag{7}$$

Algorithm 1 StyleAdv's optimization loop

```
**Access to adversarial model** $\mathcal{M}^A$**:** Ability to quickly calculate
    $\mathcal{F}^A$ (for $\mathcal{L}_{ID}^A$ and $\mathcal{L}_{total}^A$)
**Access to victim model** $\mathcal{M}^V$ **(optional):** Limited $(q+1)$
    queries to $\mathcal{F}^V$ (If not, set $q = 0$ or use $\mathcal{F}^A$ as proxy for $\mathcal{F}^V$)
**Input:** Image $\mathbf{X}_i$ and parameters $N$, $q$, $\eta$, $\lambda_P$, $\lambda_{SE}$, $\lambda_{ID}$, and $\theta$
**Output:** Protected image $\mathbf{X}_i^{adv}$
$\mathbf{W} \leftarrow \mathcal{E}(\mathbf{X}_i)$
$w^{(0)} \leftarrow \mathbf{W}$
$\mathbf{X}_i^{rec} \leftarrow \mathcal{G}(w^{(0)})$
$R^{(0)} \leftarrow (\mathbf{X}_i - \mathbf{X}_i^{rec})$
**for** $j = 0$ to $N - 1$ **do**
    **Case:** Editing attack
        $\mathbf{X}_i^{adv} \leftarrow \mathcal{G}(w^{(j)})$
        Update $w^{(j+1)}$ minimizing $\mathcal{L}_{total}^A(\mathbf{X}_i^{adv})$ (per Sec. 3.5)
    **Case:** Residual attack
        $\mathbf{X}_i^{adv} \leftarrow \mathcal{G}(w^{(0)}, \mathcal{E}_R(R^{(j)}))$
        Update $R^{(j+1)}$ minimizing $\mathcal{L}_{total}^A(\mathbf{X}_i^{adv})$ (per Sec. 3.6)
    **if** $j \mod \lfloor N/q \rfloor == 0$ or $j == N - 1$ **then**
        **if** $d(\mathcal{F}^V(\mathbf{X}_i^{adv}), \mathcal{F}^V(\mathbf{X}_i)) > \theta$ **then**
            **break**
        **end if**
    **end if**
**end for**
$\mathbf{X}_i^{adv} \leftarrow \begin{cases} \mathcal{G}(w^{(j+1)}), & \text{if editing attack} \\ \mathcal{G}(w^{(0)}, \mathcal{E}_R(R^{(j+1)})), & \text{if residual attack} \end{cases}$
```

where λ_{ID}, λ_{P}, and λ_{SE} are hyperparameters to balance the three terms. Now, our optimization problem can be formulated as:

$$\min_{\vec{W}'} \mathcal{L}_{\text{total}}, \tag{8}$$

subject to $\vec{W}' = \mathcal{E}(\vec{X}_i')$, where $\vec{X}_i'$ is the adversarial image. The solution of this optimization problem yields the adversarial image $\vec{X}_i'$ that minimizes the total loss; thus, ensuring that it retains the identity of the original image, appears perceptually similar to it, and yet is classified as the target label by the attacked model.

Optimization Loop: To generate an adversarial image $\mathbf{X}_i^{adv}$ from an input image $\mathbf{X}_i$, we first encode the image into a latent space representation $\mathbf{W}$, using a GAN-based encoder function $\mathcal{E}$. This encoded representation is then used as the initial adversarial weight $w^{(0)}$ that will be iteratively improved upon during the optimization loop. For the residual attack, this encoding is also used to create a low fidelity reconstruction $\mathbf{X}_i^{rec}$ and the initial residual vector $R^{(0)} = (\mathbf{X}_i - \mathbf{X}_i^{rec})$. For both attacks, the algorithm iteratively employs a generator function $\mathcal{G}$ to produce an improved adversarial sample $\mathbf{X}_i^{adv}$ based on the previous weight $w^{(j)}$ and residual vector $R^{(j)}$ and then updates these weights (using gradient decent) so as to minimize the attack specific loss functions $\mathcal{L}_{total}^A$. This loop continues until either N optimization steps have been done or the identity distance d_{ID}^V using the victim model $\mathcal{M}^V$ (when available) is found bigger than θ during one of up-to q distance checks. Finally, the protected image $\mathbf{X}_i^{adv}$ is calculated using the most recent adversarial weight $w^{(j)}$ and residual vector $R^{(j)}$, respectively.

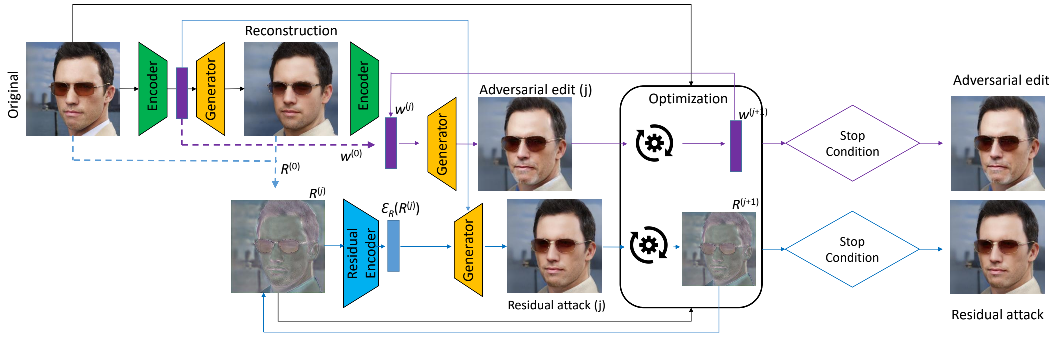


Figure 2: Overview of the attack strategies: Semantic editing (purple) and residual editing (blue). Initialization of the adversarial weight $w^{(0)}$ and residual $R^{(0)}$ are shown using dotted lines, while the (iterative) optimization loop is shown using solid lines. In each step, the optimizer finds the $w^{(j+1)}$ or $R^{(j+1)}$ that minimizes the total loss $\mathcal{L}^{A}_{total}$ (see Secs. 3.5 and 3.6-3.7, respectively).

Algorithm 1 and Fig. 2 provide an overview of the optimization loop, while the next two subsections provide a detailed description of the optimization for each type of attack. Although our framework easily allows hybrid attacks in which the residual attack use $w^{(j+1)}$ of the editing attack as starting point, for simplicity, we keep the presentation of the attack types separately here.

3.5 Semantic-aware Adversarial Attacks

We next describe three attack variations based on our semantic-aware editing approach (purple in Fig. 2): our non-attribute-guided attack, our attribute-guided attack, and our target identity attack.

Non-Attribute-Guided Attack: Starting with the original latent code $\vec{W}$, or a random one, we employ gradient descent to discover a new latent code $\vec{W}'$ that minimizes the total loss $\mathcal{L}^{NAGA}_{\text{total}}$:

$$\mathcal{L}^{NAGA}_{\text{total}} = \lambda_{\text{P}} \cdot \mathcal{L}_{\text{P}} + \lambda_{\text{SE}} \cdot \mathcal{L}_{\text{SE}} - \lambda_{\text{ID}} \cdot \mathcal{L}_{\text{ID}}, \tag{9}$$

where $\mathcal{L}_{\text{P}}$, $\mathcal{L}_{\text{SE}}$, and $\mathcal{L}_{\text{ID}}$ denote the perceptual loss, the square error loss, and the identity loss, respectively. The constants λ_{P}, λ_{SE}, and λ_{ID} are the weights for each loss. The goal is to find a latent code $\vec{W}'$ that maximizes the identity loss while preserving perceptual similarity and keeping changes in the latent code minimal.

Attribute-Guided Attack: In this attack, we first identify a target latent code $\vec{W}_{\text{tgt}}$. Then, we compute the new latent code $\vec{W}'$ by blending the original latent code $\vec{W}$ and the target latent code:

$$\vec{W}' = \vec{W} \odot \vec{\alpha} + \vec{W}_{\text{tgt}} \odot (1 - \vec{\alpha}), \tag{10}$$

where $\vec{\alpha}$ is a vector of the same size as W vector that determines the degree of change towards the target attribute. The vector $\vec{\alpha}$ is optimized to minimize the total loss $\mathcal{L}_{\text{total}}$ in the equation above.

Target Identity Attack: In the context of our semantic-aware adversarial editing approach, we introduce a variant that aims to alter an image such that it resembles a target identity while still being adversarial. We call this approach “Target Identity Attack".

In this attack, we start by encoding an image into the latent space, similar to the attribute guided and non-guided attacks. However, in this case, we introduce a modification to the identity loss component. Instead of maintaining the identity of the original image, we strive to minimize the distance to a target identity, thus creating an adversarial image that resembles the target.

Let us denote the target identity's image as $\vec{X}_{\text{tgt}}$. The target identity's embedding can then be calculated as $\vec{e}_{\text{tgt}} = \mathcal{F}(\vec{X}_{\text{tgt}})$, where $\mathcal{F}$ is the facial embedding model.

The revised identity loss for the target identity attack, denoted as $\mathcal{L}_{\text{ID}^{tgt}}$, is formulated as the distance between the embeddings of the adversarial image and the target identity:

$$\mathcal{L}^{tgt}_{\text{ID}} = d(\mathcal{F}(\mathcal{G}(\vec{W}_{\text{new}})), \vec{e}_{\text{tgt}}). \tag{11}$$

In the context of the optimization problem for the adversarial attack, the total loss for the target identity attack is as follows:

$$\mathcal{L}^{TIA}_{\text{total}} = \lambda_{\text{P}} \cdot \mathcal{L}_{\text{P}} + \lambda_{\text{SE}} \cdot \mathcal{L}_{\text{SE}} + \lambda_{\text{ID}} \cdot \mathcal{L}^{tgt}_{\text{ID}}, \tag{12}$$

where λ_{P}, λ_{SE}, and λ_{ID} are the balancing factors.

Two Versions of TIA: Note that this target identity attack can be applied in both attribute guidance and non-guidance settings. In the attribute guidance setting, the optimization process is steered by a predefined attribute target. The guidance is used to constrain the direction of the adversarial perturbations in the latent space, thus ensuring that the perturbed image will not deviate too significantly in terms of identity from the original image. This method balances the task of preserving identity resemblance and simultaneously achieving the desired attribute modification.

In contrast, for the non-guidance setting, the perturbations primarily aim to make the image resemble the target identity without the constraint of an attribute direction. Here, the perturbation could potentially cause more pronounced changes in the identity of the resulting adversarial image. In both settings, the aim of the target identity attack is to manipulate the image in such a way that it is misclassified as the target identity, while maintaining a balance between perceptual similarity and effectiveness of the attack.

3.6 Residual-based Adversarial Attack

Residual-based adversarial attacks differ from conventional semantic-aware adversarial editing, which typically perturbs the latent codes directly. Instead, this methodology manipulates the residuals between original and reconstructed images. The primary goal is to calculate a perturbation that results in the smallest perceptual deviation while still maximizing the identity loss.

A high-level description of our residual-based adversarial attack (blue in Fig. 2) is described next, followed by a more detailed description of our residual encoder (used in step 2):

(1) **Residual Computation:** Starting from the original image $\vec{X}_i$, we first derive its reconstructed counterpart $\vec{X}_i' = \mathcal{G}(\mathcal{E}(\vec{X}_i))$. The residual $\vec{R}_i = \vec{X}_i - \vec{X}_i'$ is then computed, capturing the differences that the generative model $\mathcal{G}$ could not reconstruct.
(2) **Residual Encoding:** We then use an encoder $\mathcal{E}_R$ to encode the computed residual into a latent residual space, offering an enriched representation of the image. Next, this encoded residual is combined with the original latent code to generate an enriched image $\vec{X}_i^* = \mathcal{G}(\mathcal{E}(\vec{X}_i), \mathcal{E}_R(\vec{R}_i))$.
(3) **Optimization Process:** Finally, with the enriched image in hand, we perform an optimization process that seeks a perturbed residual that minimizes the perceptual loss and the square error loss while maximizing the identity loss. In this context, the square error (SE) loss $\mathcal{L}_{\text{SE}}$ is calculated as the squared Euclidean distance between the original and perturbed residuals in latent residual space; i.e., $\mathcal{L}_{\text{SE}}{=}\|\Delta\vec{R}_i\|_2^2$.

Let us now look closer at the optimization problem. In basic terms, the optimization problem is formalized as:

$$\min_{\vec{R}_i} \left\{ \lambda_{\text{P}} \mathcal{L}_{\text{P}}(\vec{X}_i', \vec{X}_i^*) + \lambda_{\text{SE}} \mathcal{L}_{\text{SE}}(\vec{R}_i, \vec{R}_i^*) - \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(\vec{X}_i, \vec{X}_i^*) \right\}, \quad (13)$$

where $\mathcal{L}_{\text{P}}$, $\mathcal{L}_{\text{SE}}$, and $\mathcal{L}_{\text{ID}}$ denote the perceptual loss, square error loss, and identity loss, respectively, and λ_{P}, λ_{SE}, and λ_{ID} represent the balancing factors for each loss component.

Residual-based adversarial attacks thus provide a distinct way to craft misclassification-triggering examples that closely resemble the original images. By leveraging the concept of residuals, these attacks subtly, yet effectively perturb image information, thereby contributing a novel approach to adversarial image generation.

3.7 Residual Encoder

The `ResidualEncoder` used in our approach includes convolutional layers, residual blocks, and feature scaling and shifting operations.

The encoder takes an input tensor $\vec{X} \in \mathbb{R}^{C\times H\times W}$ and processes it through convolutional layers with a Parametric ReLU (PReLU) activation function [21]. We denote the output as:

$$\boldsymbol{feat}1 = \text{PReLU}(\text{BN}(\text{Conv}(\vec{X}))), \quad (14)$$

The following two layers, each consisting of three residual blocks, involve a bottleneck operation (`bottleneck_IR`) inspired by ResNet-IR [4]. These layers take $\boldsymbol{feat}1$ or $\boldsymbol{feat}2$ as input, and output:

$$\boldsymbol{feat}2 = \text{bottleneck\_IR}(\boldsymbol{feat}1), \boldsymbol{feat}3 = \text{bottleneck\_IR}(\boldsymbol{feat}2). \quad (15)$$

Scale and shift conditions are then generated using equalized convolutional layers and scaled leaky ReLU activation functions [27]:

$$\boldsymbol{s}, \boldsymbol{t} = \text{ScaledLeakyReLU}(\text{EqualConv2d}(\boldsymbol{feat}3)) \circ \vec{I}_{64\times64}. \quad (16)$$

Given an image $\vec{X}'$ generated from the latent code $\boldsymbol{W}$ and the real image $\vec{X}$, the residual $\vec{R}$ is calculated as $\vec{R}{=}\vec{X}{-}\vec{X}'$ and passed through the Residual Encoder, yielding the scale and shift conditions:

$$\boldsymbol{s}, \boldsymbol{t} = \texttt{ResidualEncoder}(\vec{R}). \quad (17)$$

The generator $\mathcal{G}$, a StyleGAN2 based model, takes the latent code $\vec{W}$ and the conditions to generate the image:

$$\vec{I} = \mathcal{G}(\vec{W}, \boldsymbol{s}, \boldsymbol{t}). \quad (18)$$

This image is designed to resemble the original while the FRS should recognize a different individual. See details in Appendix A.

4 EVALUATION RESULTS

4.1 Experiment Setups

4.1.1 Datasets: To thoroughly evaluate our proposed method in diverse environments, we employ four datasets. First, we use the Labelled Faces in the Wild (LFW) dataset [24], encompassing over 13K images of 5,749 identities, to evaluate the effectiveness against facial verification systems. LFW's diverse range of images, taken in uncontrolled environments, presents a challenging test for evaluating our approach's resilience in realistic conditions.

Second, we use the FaceScrub dataset [36], comprising 107K face images of 530 celebrities (about 200 images/person). Developed through a combination of automated face detection and subsequent manual cleaning, this dataset contains images of public figures collected from the Internet under real-world, uncontrolled conditions. Its diversity and scale allow us to validate our method's versatility and accuracy across a wide range of scenarios and identities.

Third, we use the MS-Celeb-1M dataset. This dataset includes 10M image samples from 100K individuals and is one of the most challenging datasets for facial recognition due to its sheer volume and diversity. It offers a comprehensive collection of celebrity images, capturing a myriad of poses, lighting conditions, and occlusions. Like LFW, MS-Celeb-1M is apt for evaluating face identification systems as it encompasses a vast number of identities; thus, offering a more expansive evaluation compared to FaceScrub.

Finally, we use the CelebA [31] and CelebA-HQ datasets [26], renowned for their large-scale collection of celebrity images. These datasets, totaling over 200K images and covering more than 10K unique identities with 40 attribute labels per image, enable us to evaluate the precision and realism of attribute transformations introduced by our method. CelebA-HQ, a high-quality subset of CelebA, offers 30,000 uniformly high-resolution images (1024x1024 pixels) and allows for more detailed attribute editing analysis.

Collectively, these datasets, with their substantial diversity in scale, resolution, and annotated attributes provide a solid base for our evaluation, ensuring our results are both credible and widely applicable across different facial recognition and editing scenarios.

4.1.2 Pre-trained Models. In our experimental setup, we employ several state-of-the-art pre-trained models to ensure an effective mechanism for image manipulation and evaluation.

StyleGAN2: StyleGAN2 [28] serves as our primary image generator. This generative adversarial network model is known for generating high-quality, diverse, and photorealistic human face images from points in the model's latent space. StyleGAN2 resolves certain limitations observed in StyleGAN, leading to improvements in the quality of generated images and better disentanglement in the latent space. This model was trained on the FFHQ dataset, which comprises 70K high-quality face images at 1024×1024 resolution.

e4e Encoder: To map real-world images into the latent space of StyleGAN2, we employ the e4e (encoder for editing) model. The e4e encoder [50] works by minimizing the perceptual distance between the original image and the one generated from the projected latent code, thereby ensuring accurate and visually consistent projection. This model has been pre-trained on the CelebA-HQ dataset.

Facial Embedding Models: Five facial embedding models are incorporated into StyleAdv (both for attacks and evaluation). ArcFace

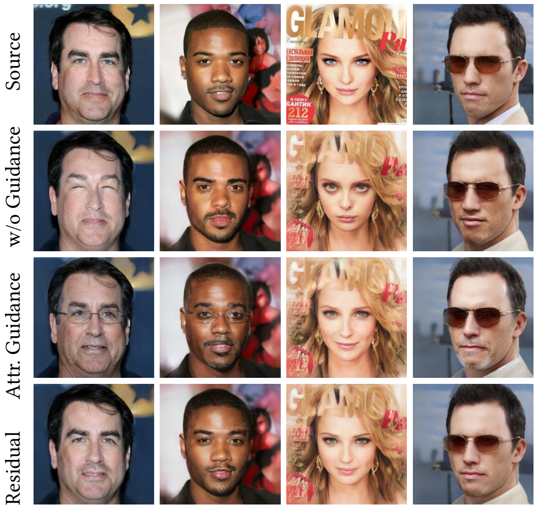


Figure 3: Visual example results illustrating the primary differences between the proposed attack approaches.

[12], trained on the MS1MV2 dataset [19], is known for its powerful face recognition performance, with accuracy reaching 99.82% on the LFW dataset. We use two versions of ArcFace: irse50 and ir152. FaceNet [43], an inception-resnet-based model trained on the VGGFace2 dataset [4], demonstrating remarkable performance with 99.65% accuracy on the LFW dataset [24]. MobileFaceNet [7], optimized for mobile and embedded vision applications, is also included, offering a good tradeoff between computational efficiency and performance. Additionally, CurricularFace [25] is integrated, known for its dynamic adjustment of the learning objective based on the training status, yielding robust facial recognition.

We refer to Appendix B for further implementation details.

4.2 Visual Example Results

We first provide a visual comparison of the main variations of the attack approaches proposed and evaluated. Fig. 3 shows representative example results for three distinct approaches, each offering insights into their unique characteristics and relative tradeoffs. For ease of comparison, we include the original unaltered source images (top row) as a baseline against which each attack can be compared.

Latent Space Attack without Attribute Guidance (row two): This approach generally converges quicker compared to the other attacks. This makes it the most advantages from a computation time perspective, but due to its lack of guidance, it offers less control of the properties of the final results. In the example results, this is evident in the instances where the facial features have been significantly altered, leading to a noticeable shift in identity resemblance.

Latent Space Attack with Attribute Guidance (third row): This method provides a better balance between identity modification and resemblance preservation. By incorporating attribute guidance, we are able to diversify the representation of identity in the feature space while effectively controlling the extent of change in pixel space. The outcome, seen on row three, shows that even though the faces have been altered modestly to mislead FRS:s, the identities still maintain their visual resemblance to the original, making it more suitable for practical applications.

Table 2: Image quality results. (Best values in bold.)

Method	LPIPS ↑	MS-SSIM ↑	MSE ↑
Residual Attacks (ours)	**0.1864**	**0.1368**	**0.0312**
Latent Edit Attack (ours)	0.2665	0.2284	0.0548
Fawkes	0.4681	0.6256	0.4456
SemanticAdv	0.5331	0.7894	0.4745

Residual Attack (fourth row): The distinct advantage of this method is its high-quality results that require less time to converge, making it computationally efficient. However, it is also worth noting that this approach may sometimes introduce minor artifacts that become noticeable upon closer inspection. Despite these, the residual attack method presents another feasible approach to achieving our privacy protection objective.

The choice of attack method depends on the specific requirements of the use-case, with considerations for factors like computational efficiency, controllability of identity changes, and preservation of visual resemblance. The evaluation in the following sections compare our techniques with related works and provide deeper insights into the relative tradeoffs between our proposed approaches.

4.3 Image Quality

One important utility aspect is the image quality. In Table 2 we use three different image quality metrics to illustrate the relative image quality achieved using our adversarial techniques (i.e., "Residual Attack" and "Latent Edit Attacks") as well as the closest relate works: Fawkes [44] and SemanticAdv [39]. Here, we group the image quality results for the "guided" and "non-guided" latent edit attack variations, as they are statistically similar.

To effectively quantify the degree of distortion or alteration in image quality that each method produces, we calculate and report three image quality metrics: the Learned Perceptual Image Patch Similarity (LPIPS) [57], the Multi-Scale Structural Similarity Index (MS-SSIM) [52], and the Mean Squared Error (MSE). In each case, the metrics were computed by contrasting the image before and after the attack. Furthermore, in each case, a lower score suggests a lower degree of image distortion, hence higher image quality, as the adversarial image stays perceptually closer to the original one.

We note that our two proposed methods display outstanding performance in maintaining image quality across all tested metrics and significantly outperform the related works. Notably, the residual attack method achieves an LPIPS score of 0.1864, an MS-SSIM score of 0.1368, and an MSE score of 0.0312, implying low distortion. The results for the latent edit attack methods are similarly impressive, yielding scores of 0.2665, 0.2284, and 0.0548, respectively.

The two related works (Fawkes and SemanticAdv) have significantly higher distortion, indicating that they provide inferior image quality compared to our methods. For example, Fawkes obtains LPIPS, MS-SSIM, and MSE scores of 0.4681, 0.6256, and 0.4456 respectively, and SemanticAdv even worse (0.5331, 0.7894, and 0.4745).

These findings highlight the attack's superior ability to preserve the image quality. The high visual quality of the adversarial images has high practical utility, a critical factor for privacy filters.

4.4 Targeted and Untargeted Attacks

Targeted Example Attack: We next demonstrate the robustness of our "Latent Space Attack with Attribute Guidance" method.

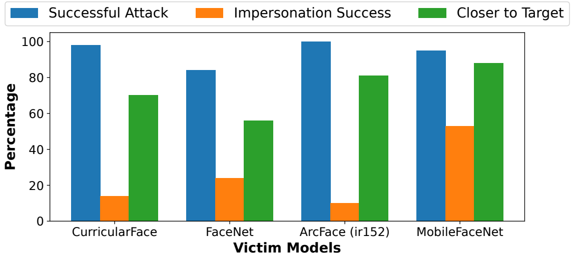


Figure 4: Success rates of targeted semantic editing attack with AcrFace(irse50) as attacker model.

Fig. 4 shows the success rate when performing our attack using ArcFace(irse50) [12] on four (different) victim models: CurricularFace [25], FaceNet [43], ArcFace(ir152) [12], and MobileFaceNet [7].

Here, for each attacker-victim pair, we measure and report three metrics. (1) How frequently does the attack successfully "hide" the source identity from the victim model (i.e., the identity distance d^{src} is no longer within the victim model's identification threshold $\hat{d}$). (2) How frequently does the victim model consider the new identity to be closer to the target identity than the source identity (i.e., $d^{tgt} < d^{src}$). (3) How frequently is an impersonation attack performed in which the victim model considers the identity distance to the target d^{tgt} to be within the model's identity threshold $\hat{d}$.

Across the different attacker-victim pairs, our attack is highly successful hiding the source identity, demonstrating the robustness of our method. The results are most successful when using the irse50 model for the attack, providing a success rate above 84% across all victim models: CurricularFace 98%, FaceNet 84%, ir152 100%, and MobileFaceNet 95%. While we observe noticeable differences between models, the overall good performance demonstrates the capability of our attack to sufficiently increase the pairwise distance relative to the source identity so as to simultaneously surpass the respective detection thresholds of several models in parallel.

While our goal here is to alter the original identity sufficiently to hide the original identity, several observations are possible from when looking at how close to the target that the attacker model is able to (simultaneously) move the identity. First, we note that the success rate for (pure) impersonation is substantially smaller than simply hiding the identity, underscoring an inherent limitation of the method. Here, it should be noted that cross-model identity impersonation is a non-trivial challenge as it involves impersonating another identity while maintaining the pairwise distance within the detection threshold of the respective models. Here, we emphasize that our goal simply is to move the identity far enough from the original identity (which we do successfully) and instead note that the impersonation rates differ somewhat between the victim models, suggesting that there are some differences in how far an identity must be moved to achieve successful impersonation. Yet, also for these models, we are successful in hiding the original identities.

Second, we note that the relative fraction of cases where the identity was moved closer to the target identity than the source identity (i.e, $d^{tgt} < d^{src}$) was substantial in many of the cases, confirming that the target identity indeed helped guide the attack. To provide some visual intuition how far the identities are moved away from the original image (from the perspective of a subjective human) we refer to the example images in Figs. 3 and 9.

Untargeted Attack Against Facial Identification: Two distinct databases were used for this evaluation: one constructed from the LFW dataset and one from the MS-Celeb-1M dataset (MS1M). In the LFW-based evaluation, we selected 1,674 unique identities, with each identity represented by more than one image sample. For the MS-Celeb-1M dataset, we used 10 batches, each with 10K unique identities. The objective of these evaluations was to assess the systems' resilience against two adversarial attacks: the "Residual Attack" and the "Latent Attack".

The results (Table 3) reveal some stark differences between the datasets. With LFW, the system's top-1 accuracy consistently was 0% for all tested models and attacks. In contrast, MS-Celeb-1M was more challenging, with the FRS:s showing slightly better resilience. Here, the residual attack was consistently more effective, with the top-1 accuracy varying from 1.74% (CurricularFace) to 6.46% (ArcFace(ir152) for the residual attack compared to between 3.28% (MobileFaceNet) and 8.90% (FaceNet) for the latent edit attack. (To provide insights into the very high success rate with LFW, we include top-k accuracy results also for other k in Appendix F.)

The average rank results, which reflect the system's ability to correctly identify true identities, were consistently high for both attacks and across both datasets. For most models, the average rank exceeds 1,000, which in the case of LFW puts the identity closer to the end of the list and in the case of MS-Celeb-1M puts the rank among the top-10% closest matches.

Overall, these results highlight facial identification systems' vulnerability to both residual and latent edit attacks. While MS-Celeb-1M exhibited greater resilience, all models across different datasets showed significant susceptibility to these adversarial techniques.

Untargeted Attack Against Face Verification: Using the same two datasets (LFW and MS-Celeb-1M), we evaluated the percentage of instances in which the adversarial attack effectively misled the facial identification system (Table 4).

The residual attack demonstrated formidable effectiveness on both datasets. On LFW, the attack achieved an impressive success rate of close to 100% for most models and a success rate of 95.3% against the most resilient model (FaceNet). For MS-Celeb-1M, the rates were almost as high, with the most resilient model (FaceNet) only reducing the success rate of our attack to 94.9%.

The latent attack also showcased consistently high success rates across all models. On LFW, the rates ranged from 94.25% to 99.76% and on the MS-Celeb-1M dataset, the rates were equally concerning, ranging between 93.2% and 96.1%.

These findings underline the pronounced vulnerability of facial identification systems, and the threat that both residual and latent edit attacks presents, with all models manifesting high susceptibility to these adversarial interventions, regardless of dataset.

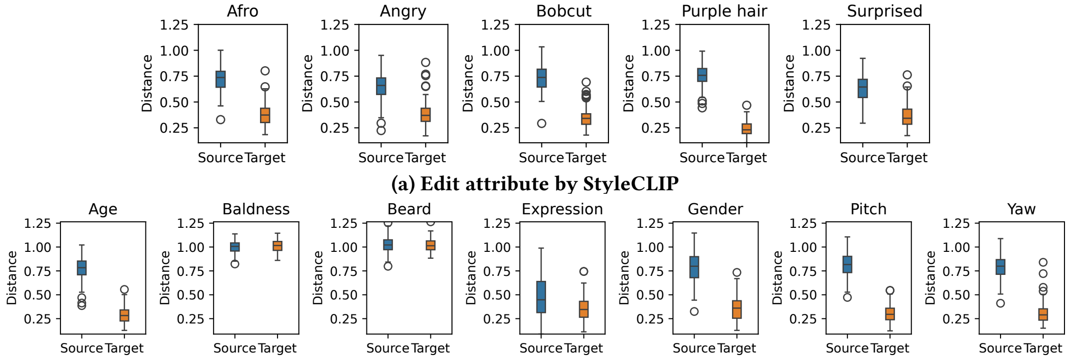
Implications of High Success: The consistently high success rates across models (Tables 3 and 4) demonstrate that the attacks effectively can render many FRS:s ineffective for both face identification and face verification tasks. Note that, similar to our results on LFW, Fawkes [44] also reported perfect protection but for older facial recognition based on VGGFace (2016). In comparison, our results are shown to be more effective on recent models; e.g., see Table 7 for comparison with Fawkes (on FaceScrub dataset). This is a particularly strong result since our attack only uses a limited number of queries to determine when to stop the optimization loop

Table 3: Facial identification results with untargeted attack: Residual vs. latent edit on the two datasets: LFW and MS-Celeb-1M (MS1M)

Model	Residual				Latent Edit			
	Top-1 Acc.		Ave. Rank		Top-1 Acc.		Ave. Rank	
	LFW	MS1M	LFW	MS1M	LFW	MS1M	LFW	MS1M
ArcFace(irse50)	0.00%	5.01%	1,445	1,323	0.00%	6.83%	1,432	1,318
ArcFace(ir152)	0.00%	6.46%	1,440	1,174	0.00%	8.89%	1,435	1,170
FaceNet	0.00%	5.97%	1,461	1,393	0.00%	8.90%	1,463	1,308
CurricularFace	0.00%	1.74%	1,437	1,252	0.00%	3.91%	1,440	1,173
MobileFaceNet	0.00%	3.91%	1,442	1,290	0.00%	3.28%	1,446	1,136

Table 4: Success rate (percentage) of residual and latent attacks (un-targeted) on face verification

Model	Residual		Latent	
	LFW	MS1M	LFW	MS1M
ArcFace(irse50)	99.77%	97.6%	94.25%	93.2%
ArcFace(ir152)	99.63%	98.3%	96.87%	94.8%
FaceNet	95.30%	94.9%	96.51%	94.2%
CurricularFace	100%	98.2%	98.44%	95.6%
MobileFaceNet	100%	98.0%	99.76%	96.1%



(a) Edit attribute by StyleCLIP

(b) Edit attribute by StyleFlow

Figure 5: Attributes evaluations using both attributes edited using StyleCLIP (top row) and StyleFlow (bottom row). Higher distance to source implies more successful attack and shorter distance to target implies greater degree of impersonation.

(not for the optimization itself). Our success with this black-box assumption demonstrates that an attacker can exploit the victim's vulnerabilities without the need for in-depth knowledge of the victim model's architecture or parameters. While queries in online mode may raise some suspicion, the stealthy nature of our attack (i.e., limited number of queries) means that the attacks can go relatively undetected in real-world scenarios. From the perspective of the FRS:s, the results are concerning, as our results show that they easily can be fooled with limited or no access to the victim model.

4.5 Editing Different Attributes

To underline the flexibility and versatility of our approach, we next demonstrate how our proposed methods can be integrated and adapted into two popular frameworks for editing facial attributes: StyleCLIP and StyleFlow. These frameworks were selected for several compelling reasons. Most importantly, their ability to manipulate facial features with high fidelity, preserving the overall quality of the image while making precise changes.

While these two examples help highlight the flexibility and versatility of our approach, our methods are more broadly designed for seamless integration into diverse editing frameworks.

Adversarial Editing with Attribute Guidance using SyleCLIP: The boxplots in Fig.5a demonstrate the efficacy of StyleCLIP's mapper in conducting latent attacks with various attribute guidance. Here, a higher distance towards the source implies a more successful attack and a lower distance towards the target indicates a more successful impersonation attack.

From the results, it is clear that different attributes experience varying degrees of success. For instance, the "afro" and "purple hair" attributes achieve a higher mean distance towards the source (0.75 and 0.76, respectively), indicating a more effective attack when these attributes are used for guidance. (See Appendix B for individual decision thresholds for the different facial recognition models.) On the contrary, "angry" and "surprised" attributes have lower mean distances to the source, suggesting a less successful attack.

With regards to impersonation, "purple hair" seems to be the most successful attribute guidance, with a mean distance of 0.24 (lower mean is better here). In contrast, attributes like "afro" and "angry" show higher mean distances to the target, indicating their limited utility for impersonation attacks.

Fig. 5a illustrates the nuanced impact of various attribute guidance on the success of latent attacks. These results offer insights into both the potentials and limitations of StyleCLIP's mapper, valuable for enhancing its efficacy in robust and efficient latent attacks.

Adversarial Editing with Attribute Guidance using StyleFlow: Fig. 5b shows the corresponding results when using StyleFlow to conduct latent attacks with various attribute guidance. Here, the attack achieves the highest success and mean distances towards the source (above 1.0) when "beard" and "baldness" are used for guidance. In contrast, the "expression" attribute exhibits the lowest average source distance, indicating it is relatively less useful.

In impersonation attacks, attributes like "age", "pitch", and "yaw" exhibit the lowest mean distances to the target, showcasing StyleFlow's high effectiveness in imitating the target with these attributes. Conversely, "baldness" and "beard" attributes show higher mean distances to the target, indicating lower impersonation attack success. These findings emphasize both the strengths and areas for improvement in StyleFlow's latent attacks.

Comparing use of StyleCLIP and StyleFlow for Attribute Guidance: The statistical and graphical analyses presented above offer a compelling comparison of the two frameworks - StyleCLIP's mapper and StyleFlow - in terms of their effectiveness in conducting latent attacks under different attribute guidance.

In comparison to StyleCLIP, StyleFlow demonstrates a higher mean distance to the source, particularly for the “beard" and “baldness" attributes. This suggests a stronger efficacy in conducting attacks when using these attributes with StyleFlow. For impersonation, the results are more mixed, although we observe a clear trend with StyleFlow indicating that attributes that are good for obfuscation tend to be bad for imitation. Appendix C provides a more detailed discussion on the duality of obfuscation and imitation.

Interestingly, the “purple hair" attribute stands out in StyleCLIP, with relatively high mean distances towards the source but low towards the target. In contrast, “beard" and “baldness" demonstrate relatively high mean distances to both source and target. This might hint at a higher degree of complexity and challenge involved in manipulating these attributes, which can be an area of further exploration. The choice between StyleCLIP and StyleFlow depends on the specific attribute guidance and the nature of the attack, with both having their own strengths and weaknesses. The results also indicate that there are nuanced differences between the two, providing valuable insights for researchers and developers looking to further improve or adapt these frameworks for different purposes.

4.6 Resistance to Standard Defense Methods

We next demonstrate the resilience of our attacks against various defense methods, including JPEG compression, total variance minimization, feature squeezing, spatial smoothing, Gaussian blur, and random noise. These methods are available as defense's preprocessors in Adversarial Robustness Toolbox [37]. For these experiments, we measure the pairwise distance between the original image and the protected images. Figure 6 summarizes these results for both the "Latent Edit Attack with Attribute Guidance" and the "Residual Attack", as well as when no defense is applied (baseline).

Latent Edit Attack with Attribute Guidance: The boxplot in Fig. 6a clearly depicts that there is no significant reduction in the pairwise distance when comparing the no defense scenario to the use of different defense methods. In other words, the attack samples retained their altered identity even after the application of these defense strategies. This finding implies that our “Latent Edit Attack” is effective against these defensive techniques. For example, the average pairwise distance for “JPEG compression” is approximately 0.30 and 0.32 for “Random noise”. The slight variations in pairwise distances among defenses suggest that our “Latent Edit Attack" effectively preserves the altered identity against these defenses.

Moreover, it is noteworthy that the performance of defense methods like “Total variance minimization”, “Spatial smoothing”, and “Gaussian blur” were particularly close to the “No defense” scenario. These results indicate that the attack has shown remarkable resilience in the face of these specific defenses.

This evaluation therefore affirms the robustness of our “Latent Edit Attack", suggesting that it can be effectively employed for identity alteration, while resisting commonly employed defense techniques. This could be a pivotal finding for advancing privacy-enhancing technologies, especially in the realm of facial images.

Residual Attack: We observe that also the “Residual Attack" is resilient to the various defenses. The results for this attack are shown in Figure 6b. Here, we note that the average pairwise distance for the “No defense" (0.42) is similar to that of most of the defense classes. For example, the average pairwise distance for “JPEG compression" is 0.41 and “Random noise" is 0.43. There is also an evident consistency in these pairwise distances, indicating that our “Residual Attack" maintains its effect against these defenses. The observation that almost all averages are the same again highlight the robustness of our attack approach.

Summary: Both our attack approaches – “Latent Edit" and “Residual" – exhibit significant resilience against a range of defense methods. The minimal differences in the pairwise distances for each method further bolster this claim. Consequently, these approaches prove to be highly efficient for identity alteration in facial images, even in the face of common defense techniques, thus showcasing potential for advancements in privacy-focused technologies.

4.7 Resistance to Adaptive (retrained) Defenses

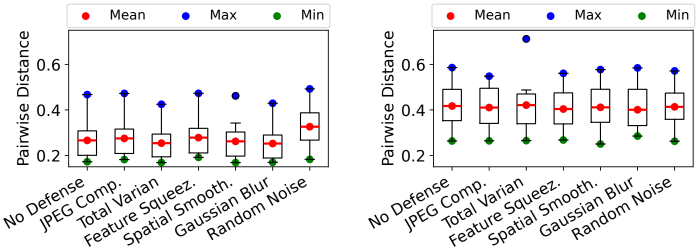
The most potent defense strategy known for countering data poisoning is to retrain the FRS model (of the victim) using perturbed (or protected) images [40]. While this defense mechanism has been found to be exceptionally effective, we note that the computational cost associated with this method is exceptionally high, making it an impractical choice in many real-world scenarios. Yet, it provides a good stress test when comparing our robustness against prior works. We next present such a comparison.

Similar to the original adaptive defense setup in [40], we randomly select one user from a pool of 530 FaceScrub identities. We then perturb 70% of this user's training set images using StyleAdv, Fawkes, or SemanticAdv. These altered images together with the training images of the other 529 FaceScrub users are then used to train the model (training dataset). This retraining process is intended to mirror real-world scenarios where FRS:s come across both conventional and privacy-enhanced images.

For the evaluation, we use ArcFace(irse50) to assess the error rate of the deceptively retrained model on the test images of the subject that had not been altered. We averaged the error rates across 20 separate experiments, each involving a randomly selected user. Fig. 7 shows the errors rates when using each of the three privacy protection methods: Fawkes, SemanticAdv, and StyleAdv. Here, higher error rate signifies better protection (performance), with such attacks presenting the FRS with a bigger challenge to accurately identifying the face and hence also the user with better privacy protection. while all three methods, as expected, see a noticeable drop in the protection they provide when ArcFace(irse50) is adaptively retrained (bars on the right) versus with the original model (bars on the left), we note that StyleAdv with latent edits achieve the highest error rates, followed by SemanticAdv. These results demonstrate another significant advantage of semantic editing techniques such as StyleAdv (with latent edits) and SemanticAdv. While traditional privacy protection methods may become less effective against adaptive models (as seen by Lowkey results), semantic editing techniques continue to provide some defense even against such highly advanced FRS:s.

4.8 Using FRS:s to Attack other FRS:s

We next evaluate the transferability of the attacks. For this analysis we apply pairwise tests in which we use each of the following five face recognition models to attack each other: CurricularFace,



(a) Latent attack w. attribute guidance **(b) Residual attack**

Figure 6: Defense comparisons: Pairwise distance between protected images and original images

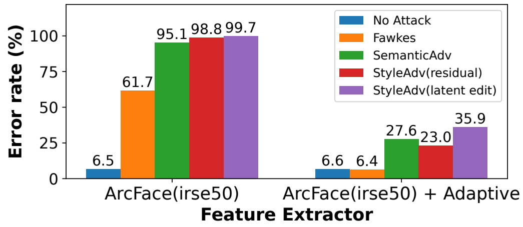
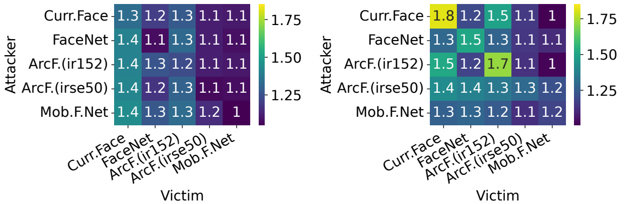


Figure 7: Adaptive defense using model retraining.



(a) Latent attack (w. guidance) **(b) Residual attack (w/o. target)**

Figure 8: Transferability: Pairwise tests using one FRS to attack another FRS.

FaceNet, ArcFace(ir152), ArcFace(irse50), and MobileFaceNet. Figure 8 shows the pairwise results as one heatmap per attack type, where each cell indicates the pairwise distance between the identity before and after the attack, with lighter colors representing higher values and thus a more successful attack.

While the relative success of the latent space attack (Figure 8a) is most dependant on the victim model, the residual attack (Figure 8b) is typically most successful when a model is applied on itself, with the CurricularFace model (applied on itself) resulting in the overall highest normalized distance (1.813).

Latent Space Attack with Attribute Guidance: Referring to Figure 8a, it is clear that the normalized pairwise distances for all attacks significantly exceed the detection thresholds of the respective models. For example, CurricularFace has a normalized detection threshold of 1, yet the normalized pairwise distances for the latent attack on CurricularFace (regardless of the attacker model) range from 1.334 to 1.434 (these numbers are rounded to better fit the figure). This shows that the attacks successfully alter the identity to a point that is well beyond the model's detection boundary.

Similarly, FaceNet, with a normalized detection threshold of 1, experiences attacks resulting in pairwise distances between 1.077 to 1.328 when subjected to attacks from all the models. These distances again fall well outside of the expected boundaries of a successful identification, showing the efficacy of the attack.

These results emphasize the robustness and transferability of the "Latent Space Attack with Attribute Guidance" method. The fact that the normalized distances consistently surpass the detection thresholds indicates that the attack effectively alters the identity across all tested face recognition models. This makes it a promising strategy for privacy-focused applications as it can effectively thwart the face recognition models from correctly identifying the subject.

Residual Attack: Fig. 8b shows the corresponding results for the residual attack, which exhibits greater transferability compared to the latent space attack, as evidenced by the heatmap. This attack consistently pushes the verification distances beyond the normalized detection thresholds of the respective models, effectively altering the identity across all tested models. This robustness signifies the attack's transferability and effectiveness.

The heatmap further unveils intriguing inter-model transferability patterns. For example, the residual attack developed for ArcFace(ir152) transfers very effectively to the CurricularFace model (normalized distance of 1.511) but does not achieve as high protection when applied on MobileFaceNet (1.014). While also this distance surpasses the identity verification threshold, its success is relatively lower than for the other cross-model attacks. The overall positive results (with consistently normalized values above 1), demonstrate that especially the residual attack (right plot) is generally robust and transferable, although its effectiveness can vary depending on the combination of attacker and victim models.

Discussion: For both attacks, we have observed distinct variations in the verification distances across different facial recognition models, with FaceNet and MobileFaceNet being relatively more resilient to the proposed attacks. We expect that these differences in resilience stem from inherent differences in the architectures, training/pre-processing methodologies, and the latent representations of these models. For example, if the latent spaces within such a model is less amenable to certain manipulations this will influence how susceptibility it is to our adversarial attacks. While a deeper exploration into these specific models and their latent spaces are outside the scope of this paper, we expect that even stronger adversarial attacks could be achieved if taking into account their internals. Here, we present a generic approach that provides good transferability across models.

Summary: The preceding analysis underscores the remarkable transferability of both the “Latent Space Attack with Attribute Guidance" and the “Residual Attack" methods across various face recognition models. This characteristic is crucial for designing robust privacy-enhancing applications. These applications, armed with such transferable attacks, can maintain user privacy across different face recognition models, bolstering their resilience against varying recognition technologies.

4.9 Comparisons with Related Works

In this section we provide a quantitative performance comparison to complement and support our qualitative high-level comparison

Table 5: Average verification distance

Method	ArcFace (irse50)	ArcFace (ir152)	FaceNet	Mobile-FaceNet	CurricularFace
Residual (Ours)	**0.9603**	**0.8053**	0.4361	0.5798	**0.7895**
Latent Edit (Ours)	0.4803	0.7064	0.4408	0.4384	0.6717
Fawkes	**0.3631**	0.5703	**0.5455**	**0.3508**	0.4472
SemanticAdv	0.5644	0.5776	0.4444	**0.6202**	0.5881

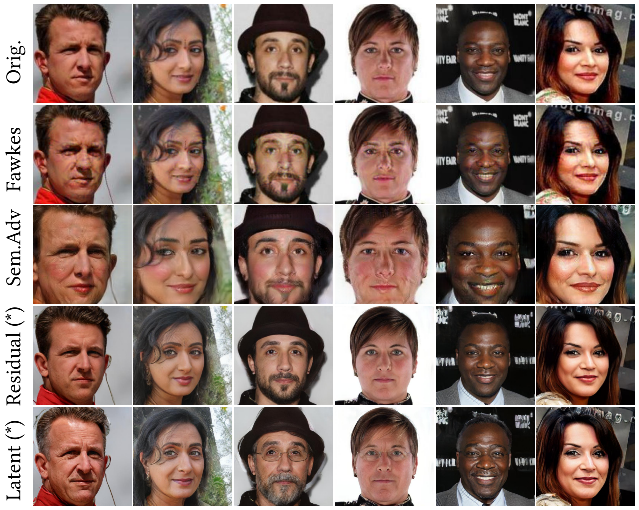


Figure 9: Visual comparison to related works.

presented in Table 1. Here, we only preset results for the most representative related works: Fawkes and SemanticAdv.

Image Quality: First, referring back to Table 2 and the discussion in Sec. 4.3 we note that our methods provide superior image quality to both Fawkes and SemanticAdv.

Verification Distance Comparison: Second, Table 5 presents the average verification distances of various adversarial methods, including when using our attacks (both “Residual Attack" and “Latent Edit Attack") and the most closely related works (Fawkes and SemanticAdv) tested using five different facial recognition models: ArcFace(irse50), ArcFace(ir152), FaceNet, MobileFaceNet, and CurricularFace. Both our proposed methods on average outperform the other models, typically achieving higher verification distances for most of the tested models. This implies that our methods are more effective in obfuscating the original identity, thereby offering stronger defense against FRS:s.

Against four out of five models, Fawkes has the smallest distances. In fact, against two models (ArcFace(irse50) and MobileFaceNet) the average distance with Lowkey is not even below or close the detection thresholds of 0.36 and 0.425 (i.e., 0.3631 and 0.3508, respectively). We therefore marked these instances in red, indicating less successful adversarial attacks. Thus, despite its ability to evade detection by some models (i.e., FaceNet), Fawkes most often achieved the least protection of the three attacks.

While SemanticAdv, similar to our latent attack, can achieve effective adversarial attacks by editing in the latent space, it is noteworthy that our “Residual Attack” method outperforms SemanticAdv for 3 out of 5 models.

Visual comparison: We next turn our attention to the visual results. Figure 9 illustrates a comparative analysis of the quality and effectiveness of our methods against the same related works.

We have found that both of our approaches are able to yield high-fidelity image quality and high attack success rate, although our “Residual Attack" method sometimes produces minor artifacts. In this regard, the “Latent Space Attack with Attribute Guidance" (i.e., “Latent Edit") is more reliable in that it consistently maintains superior image quality. Interestingly, this editing requirement can be viewed as advantageous rather than limitation, considering that users often tend to edit their images prior to sharing [11, 15, 35, 51].

In comparison to our methods, we see a significant degradation in image quality with Fawkes (which only achieve high protection when the method is used in “high" mode).

While SemanticAdv is effective in some instances, we have found that it generally suffers from low image fidelity and is prone to significant artifacts. This method also relies on the attributes of the CelebA dataset, which considerably limits its scope of application. For example, SemanticAdv falls short when tested on images that do not belong to the CelebA dataset, or specifically those on which the StarGAN model has not been trained. This is a significant limitation that impedes its usefulness.

Summary: Our proposed “Latent Space Attack with Attribute Guidance" and “Residual Attack" methods demonstrate a compelling balance between maintaining high image quality and offering effective protection against facial recognition technologies. They evidently outperform the other methods considered here, making them robust tools for privacy-preserving image sharing.

5 CONCLUSIONS

In this paper we have presented StyleAdv, a robust and comprehensive framework for adversarial image editing, which we have demonstrated is a promising solution to protect facial identity and enhance privacy. By leveraging the latent spaces of StyleGAN and incorporating a novel residual attack strategy, StyleAdv generates high-quality adversarial samples that surpass prior works in image quality, realism, and attacks success rate. The framework seamlessly integrates semantic-aware editing, adversarial attack modules, and face recognition systems, offering a cohesive and practical tool for privacy protection. The paper demonstrates the effectiveness of StyleAdv through high attack success rates achieved while preserving image quality, which has remained a challenge for existing methods. Additionally, we provide insights into effective editing techniques, discuss tradeoffs in latent spaces, and highlight the impact of utilizing residual information. With an easy-to-use web interface and a comparison against existing adversarial attack methods, StyleAdv represents a significant advancement in the field, empowering users to safeguard their privacy in the digital age.

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.

REFERENCES

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)* 40, 3 (2021), 1–21.

[2] Kevin W Bowyer. 2004. Face recognition technology: security versus privacy. *IEEE Technology and society magazine* 23, 1 (2004), 9–19.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.

[5] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2267–2281.

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.

[7] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*. Springer, 428–438.

[8] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. 2019. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 52–68.

[9] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. 2020. Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices.. In *USENIX Security Symposium*. 2667–2684.

[10] 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. *arXiv preprint arXiv:2101.07922* (2021).

[11] Trudy Hui Hui Chua and Leanne Chang. 2016. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in human behavior* 55 (2016), 190–197.

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[13] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. 2020. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 1–10.

[14] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945* 2, 3 (2017), 4.

[15] Jasmine Fardouly, Phillippa C Diedrichs, Lenny R Vartanian, and Emma Halliwell. 2015. Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. *Body image* 13 (2015), 38–45.

[16] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (2019), 1287–1289.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 87–102.

[20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems* 33 (2020), 9841–9850.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15014–15023.

[23] Weiwei Hu and Ying Tan. 2023. Generating adversarial malware examples for black-box attacks based on GAN. In *Data Mining and Big Data: 7th International Conference, DMBD 2022, Beijing, China, November 21–24, 2022, Proceedings, Part II*. Springer, 409–423.

[24] 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*.

[25] 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 *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5901–5910.

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[27] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[29] Stepan Komkov and Aleksandr Petiushko. 2021. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 819–826.

[30] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 99–112.

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August* 15, 2018 (2018), 11.

[32] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[33] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* 110 (2021), 107332.

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[35] Siân A McLean, Susan J Paxton, Eleanor H Wertheim, and Jennifer Masters. 2015. Photoshopping the selfie: Self photo editing and photo investment are associated with body dissatisfaction in adolescent girls. *International Journal of Eating Disorders* 48, 8 (2015), 1132–1140.

[36] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*. IEEE, 343–347.

[37] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. 2018. Adversarial Robustness Toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069* (2018).

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[39] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. 2020. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 19–37.

[40] Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramèr. 2021. Data poisoning won't save you from facial recognition. *arXiv preprint arXiv:2106.14851* (2021).

[41] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605* (2018).

[42] Gokula Krishnan Santhanam and Paulina Grnarova. 2018. Defending against adversarial attacks by leveraging an entire GAN. *arXiv preprint arXiv:1805.10652* (2018).

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[44] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Security Symposium*.

[45] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 1528–1540.

[46] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence* 44, 4 (2020), 2004–2018.

[47] Scott Skinner-Thompson. 2020. *Privacy at the Margins*. Cambridge University Press.

[48] Jiachen Sun Sun, Yulong Cao Cao, Qi Alfred Chen, and Z Morley Mao. 2020. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security Symposium (Usenix Security'20)*.

[49] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[50] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.

[51] Jill Walker Rettberg. 2014. *Seeing ourselves through technology: How we use selfies, blogs and wearable devices to see and shape ourselves.* Springer Nature.

[52] 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.

[53] Harry Wechsler, Jonathon P Phillips, Vicki Bruce, Francoise Fogelman Soulie, and Thomas S Huang. 2012. *Face recognition: From theory to applications.* Vol. 163. Springer Science & Business Media.

[54] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16.* Springer, 1–17.

[55] Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12863–12872.

[56] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* (2018).

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[58] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35, 4 (2003), 399–458.

APPENDIX

A RESIDUAL ENCODER DETAILS

A key component of the "Residual Attack" is the ResidualEncoder. In short, this is a model consisting of a sequence of convolutional layers and residual blocks, along with feature scaling and shifting operations. Table 6 provides a summary of the important layers in our residual encoder model.

Step-by-Step Processing Sequence: The encoder operates on an input tensor $\vec{X} \in \mathbb{R}^{C \times H \times W}$, where C is the number of channels, and H and W are the height and width of the input image.

The initial convolutional layer, conv_layer1, applies a convolution operation with a 3×3 kernel and stride 1, followed by batch normalization and a Parametric ReLU (PReLU) activation function [21]. Formally, we can denote the output of this layer as:

$$\boldsymbol{feat}1 = \text{PReLU}(\text{BN}(\text{Conv}(\boldsymbol{x}))), \tag{19}$$

where Conv denotes the convolution operation, BN denotes the batch normalization operation, and PReLU denotes the PReLU activation function.

The following two layers, conv_layer2 and conv_layer3, each consist of a sequence of three residual blocks. Each residual block involves a bottleneck operation, denoted as bottleneck_IR, inspired by the ResNet-IR architecture [4]. Thus, the output of these layers can be represented as:

$$\boldsymbol{feat}2 = \text{bottleneck\_IR}(\boldsymbol{feat}1), \tag{20}$$

$$\boldsymbol{feat}3 = \text{bottleneck\_IR}(\boldsymbol{feat}2). \tag{21}$$

The condition_scale3 and condition_shift3 operations each consist of a sequence of equalized convolutional layers (denoted as EqualConv2d) and scaled leaky ReLU (denoted as ScaledLeakyReLU) activation functions, as introduced in the StyleGAN work [27]. We perform these operations on the output of the third convolutional layer, interpolate the result to a resolution of 64×64, and then clone it to generate the scale and shift conditions.

Table 6: Summary of the most important layers in the residual encoder model.

Layer	Description	Output Shape
conv_layer1	Convolution layer	[32, 256, 256]
conv_layer2	Convolution layer applied to feat1	[48, 128, 128]
conv_layer3	Convolution layer applied to feat2	[64, 64, 64]
condition_scale3	Scale condition applied to feat3	[512, 64, 64]
interpolation (scale)	Bilinear interpolation applied to scale	[512, 64, 64]
condition_shift3	Shift condition applied to feat3	[512, 64, 64]
interpolation (shift)	Bilinear interpolation applied to shift	[512, 64, 64]

Formally, the scale and shift conditions are represented as:

$$\boldsymbol{s} = \text{ScaledLeakyReLU}(\text{EqualConv2d}(\boldsymbol{f}_3)) \circ \boldsymbol{I}_{64\times64}, \tag{22}$$

$$\boldsymbol{t} = \text{ScaledLeakyReLU}(\text{EqualConv2d}(\boldsymbol{f}_3)) \circ \boldsymbol{I}_{64\times64}. \tag{23}$$

The model then returns these conditions as its output.

Using the Residual Encoder to Generate Images: Given an image $\vec{X}'$ generated from the latent code $\vec{W}$ and the corresponding real image $\vec{X}$, the residual $\vec{R}$ is calculated as $\vec{R} = \vec{X} - \vec{X}'$. This residual $\vec{R_i}$ is then processed through the "Residual Encoder" to yield the scale and shift conditions, denoted as s and t:

$$s, t = \texttt{ResidualEncoder}(\vec{R}). \tag{24}$$

These conditions, along with the latent code $\vec{W}$, are passed into the generator $\mathcal{G}$, where $\mathcal{G}$ is a StyleGAN2 based model designed to accept and manipulate these latent codes W of size [18, 512].

At a specific layer in $\mathcal{G}$ (e.g., the 7th layer for a 64×64 output resolution), the scale and shift conditions $\boldsymbol{s}$ and $\boldsymbol{t}$ are applied to modulate the output features $\boldsymbol{f}$:

$$\boldsymbol{f} = \boldsymbol{f} \cdot (1 + \boldsymbol{s}) + \boldsymbol{t}. \tag{25}$$

This effectively performs an element-wise scaling and shifting of the features at this layer, enabling the modification of the generated output in response to the residual.

Finally, $\mathcal{G}$ generates an image $\vec{I}$ based on the conditioned features and latent code:

$$\vec{I} = \mathcal{G}(\vec{W}, \boldsymbol{s}, \boldsymbol{t}). \tag{26}$$

This transformed version of the original input image serves as the new representation of the user. It is worth noting that this image is designed to maintain visual similarity to the original one, while still being identified as a different individual by the face recognition system. Thus, it plays a crucial role in the success of our privacy-focused attacks, and is the tangible outcome of these attacks.

B IMPLEMENTATION DETAILS

In our implementation, we have considered several specific aspects for both the latent and residual attacks.

- **Identity Threshold**: The identity threshold of a victim model is used to decide when to end the optimization loop. This threshold, denoted as ‘ID_threshold‘, varies depending on the facial recognition model in use. The models include CurricularFace, FaceNet, ArcFace(ir152), ArcFace(irse50), and

MobileFaceNet, with corresponding ‘ID_threshold‘ values of ‘[0.43, 0.36,0.42, 0.412, 0.425]‘.

- **Optimizer for Latent Attack**: The latent attack uses the Adam optimizer. This optimizer employs a learning rate schedule named 'Cosine Annealing with Warm Restarts' [32] for the optimization process. The initial learning rate for this scheduler is set to 0.01 and the ramp-up factor is 0.05.
- **Optimizer for Residual Attack**: The RMSprop optimizer is used for the residual attack with a fixed learning rate of 0.004. We do not use a learning rate scheduler for this optimizer.
- **Regularization Parameters (λ)**: For the latent attack, λ_{ID} is set in a range from 0.2 to 1, depending on the dataset. λ_{LPIPS} is set to 1 and λ_{SE} is set to 0.008. For the residual attack, λ_{ID} is set in a range between 1 and 10, depending on the dataset, and λ_{LPIPS} and λ_{SE} are set to 1 and 0.001, respectively.
- **Training the Residual Encoder**: For the training of the Residual Encoder, we make use of the Adam optimizer. A specific learning rate scheduler, known as 'Cosine Annealing with Warm Restarts' [32], is employed during the optimization process. The learning rate is set to 1×10^{-4} initially and the ramp-up factor is 0.05. We train the Residual Encoder for 80,000 steps with a batch size of 4. The training is conducted on an RTX 3090 graphics card with 24GB of VRAM. This approach ensures a comprehensive and efficient training process for the Residual Encoder, ensuring that it can accurately and effectively encode the residuals for use in the attack process.

These specific choices play crucial roles in balancing the objectives of our attacks and ensuring their effective performance.

C DISCUSSION ON THE DUALITY OF GUIDED ATTRIBUTES

Our results using attribute guidance (Fig. 5) show that attributes that are good for obfuscation often are bad for imitation. We have found that this contrast between obfuscation and imitation stems from their distinct objectives. While imitation emphasizes capturing and replicating unique attributes, obfuscation aims to generalize and make features less distinct. This duality provides valuable insights into facial recognition datasets and the associated privacy implications of different facial attributes.

First, consider the contrasting objectives of obfuscation and imitation. Obfuscation is often achieved by generalizing facial features, making the individuals shown in an image less distinctive and thereby harder to identify. By creating faces that appear more “generic", the faces therefore become less unique and more easily “blending in" within a set of faces with other identities. In contrast, imitation is typically best achieved by focusing on specific (often unusual) characteristics that help mimic a particular identity. By striving to capture the uniqueness of the target identity, features that are more distinctive are therefore often leveraged to impersonate a specific individual. This inherent contrast between the two approaches (i.e., one benefiting from generalization and the other specificity) suggests that there may be inherent differences in which features are better or worse for each of the two objectives.

Second, we note that the two contrasting objectives are greatly impacted by the way facial recognition datasets are constructed and utilized. For example, datasets are often populated with a myriad of facial features, with a bias towards more common or generic features that may be beneficial for obfuscation. Focusing on these generic features, faces can be effectively concealed within a large pool of similar data points, rendering recognition more difficult. In contrast, imitation may be hampered by the more unusual attributes (capturing distinctive features) being less represented in such datasets. This dynamic underscores the broader privacy implications of facial features. In general, features that are common or generic offer a natural cloak of anonymity, while unique or unusual features can be both a strength (for personal recognition) and a vulnerability (for impersonation). Reflecting on this interplay can guide future endeavors in facial recognition, emphasizing the need for balanced datasets and more nuanced recognition algorithms that can discern between genuine uniqueness and adversarial manipulation.

Third, it is worth noting the delicate balance between the extent of editing applied to the original image and the consequent level of privacy protection achieved. Our results indicate that while more pronounced edits can indeed enhance privacy, they might deviate significantly from the original image, potentially compromising its authenticity or the user's intent. Recognizing this challenge, our approach with the residual attack strategy is particularly noteworthy. By leveraging residual information, we manage to strike a balance, ensuring minimal deviations from the original image while still bolstering its resistance to FRS:s. This nuanced approach not only preserves the integrity of the image but also respects the user's intent, offering a tailored solution that does not force users to choose between authenticity and privacy.

Finally, the interplay between imitation and obfuscation, as evidenced from our results, underscores the complexities involved in adversarial image editing. As we continue to refine our techniques, a deeper exploration of this duality will be valuable in shaping more effective and user-centric privacy solutions.

D ADVERSARIAL EDITING: AUTOMATED FRS:S VS. MANUAL EXAMINATION

When designing a privacy filter, it is crucial to strike a balance between adversarial perturbation and image fidelity. While StyleAdv successfully produces high-quality images that can bypass unauthorized automated FRS:s, it is worth noting that these edited images may still allow manual recognition in some cases. This tradeoff stems from our focus on countering automated FRS:s, which are more prevalent and scalable than manual reviews in today's digital landscape. With algorithms conducting the majority of facial scans at a pace far beyond manual inspection, privacy filters that target automated FRS:s become increasingly important.

However, we also acknowledge that some users may desire even greater protection and/or require a different threat model. Future iterations of StyleAdv may therefore explore different thresholds to provide a range of options. This will allow users to choose the level of privacy protection and image authenticity that best suits their specific needs and threat scenarios. While the current version of StyleAdv may not deter a determined attacker from conducting manual checks, it offers robust defense against the more common threat of large-scale, automated facial recognition scans.

E DISCUSSION REGARDING DATA POISONING VS. FACIAL RECOGNITION

In the ever-evolving landscape of privacy and facial recognition technology, several critical considerations come to light. First, it should be noted that data poisoning, where users modify online images to fool future facial recognition models, has limitations. For example, as argued in [40], data poisoning provides a false sense of privacy security as this strategy overlooks a key imbalance: once a user uploads a modified image that is then scraped, the perturbation becomes permanent. Later, model trainers, aware of these perturbation methods, can then adapt their facial recognition models, nullifying the perturbations' effectiveness. Two systems for poisoning attacks, Fawkes and LowKey, were evaluated in [40], with findings suggesting that "oblivious" model trainers can bypass these poisoning protections by merely waiting for advancements in computer vision. In addition, it was shown that adversaries with black-box access can both resist perturbed image effects and detect altered images online. A crucial point is that once a picture is altered and scraped, the changes are irreversible, making it vulnerable to any new recognition technology developed later. Recent evidence shows that some state-of-the-art poisoning strategies are already compromised by newer training models, suggesting that poisoning techniques may not provide a sustainable "arms race" between attack and defense methods.

However, in the context of today's fast-paced technological advancements, it is also crucial to recognize the ever-evolving nature of adversarial techniques such as those developed and employed in StyleAdv. While the discussed challenges point towards an inherent disadvantage for users, the technological landscape is constantly shifting, bringing forth advancements that can effectively combat these challenges. Just as facial recognition technologies advance, so do the countermeasures, as exemplified by the diverse attack methods of StyleAdv. The mere existence of diverse approaches highlights the potential to discover strategies that are more robust in the future, ensuring that we do not remain stagnant in the face of surveillance concerns. Furthermore, adversaries aiming to counter such advanced adversarial methods will invariably face increasing model training costs. Continual adaptation and retraining of models to neutralize adversarial attacks not only require significant computational resources but also time and expertise. This escalation in costs serves as a deterrent, making it less economical for entities to persistently undermine user-driven privacy measures. Thus, while challenges persist, tools like StyleAdv exemplify the potential of ongoing technological innovation to safeguard individual privacy in an era of ubiquitous surveillance.

We also note that the arguments presented in [40] might elicit a form of categorical rejection of adversarial attacks as viable defenses. Such a rejection, based solely on the limitations of a few methods, could be premature and overly deterministic. Instead of viewing adversarial techniques in isolation, it is beneficial to consider them as components of a broader, multi-pronged strategy. A layered defense, for instance, merges the capabilities of adversarial attacks like StyleAdv with other privacy-enhancing measures, offering a more robust barrier against invasive recognition tools. When multiple protective layers are combined, each covering the potential shortcomings of the others, the defense's resilience is naturally heightened.

Widespread deployment of techniques such as StyleAdv would also instigate a paradigm shift in data collection practices. As the reliability of scraped data becomes questionable due to prevalent adversarial interventions, companies might rethink the feasibility of indiscriminate data harvesting. There is potential for a transition towards more consensual and transparent data-gathering practices, driven by the reduced reliability of unconsented data sources tainted by adversarial modifications. In essence, tools such as StyleAdv can catalyze shifts in broader industry practices, emphasizing the necessity of ethical and effective data collection in a world increasingly wary of privacy infringements.

In conclusion, the limitations of data poisoning highlight the need for dynamic and evolving strategies to safeguard individual privacy in the face of advancing surveillance tools. Adversarial techniques, such as those developed and employed in StyleAdv, represent one promising strategy. However, we also note that the complexities of the current privacy-safeguarding methods and their interplay with surveillance tools emphasize the need for continual innovation and a shift towards a more holistic, layered defense approach. When combined with other protective measures, the efficacy of these defenses can be significantly amplified, helping drive a transformation in data collection standards and pushing for more ethical, transparent, and consensual practices.

F ADDITIONAL FRS RESILIENCE ANALYSIS ON THE LFW DATASET

To provide some further insights into the high success rates reported for the residual attack when evaluated on the LFW dataset (Section 4.4), this appendix presents some additional analysis for this attack and dataset.

F.1 Face Identification: Top-k Accuracy Analysis

We built a face identification system based on the 1,674 individuals in the LFW dataset that have two or more images. Focusing on this subset of the dataset allows for both intra-personal (same person, different images) and inter-personal (different people) comparisons. Originally, the LFW dataset contains 1,680 individuals with 2 or more images, summing up to 7,701 images. However, six individuals were removed due to error matching in the dataset.

Table 7: Top-k accuracy for five FRS:s under attack.

FRS	Top-1 Acc.	Top-5 Acc.	Top-50 Acc.	Top-100 Acc.
ArcFace (irse50)	0%	0.5%	3.5%	7.0%
ArcFace (ir152)	0%	0.6%	3.6%	7.1%
FaceNet	0%	0.4%	3.2%	6.8%
CurricularFace	0%	0.6%	3.7%	7.2%
MobileFaceNet	0%	0.5%	3.5%	7.0%

Table 7 shows the top-k accuracy for the five FRS:s under adversarial attacks. Despite all systems experiencing a significant drop in their top-1 accuracy to 0%, their resilience improve if allowing higher ranks. ArcFace(irse50) has a top-5 accuracy of 0.5% and improves to 7.0% for top-100. Its counterpart, ArcFace(ir152)

Table 8: Confusion matrices for face verification experiments on LFW dataset

(a) ArcFace(irse50)

	Pred +	**Pred -**
Act +	14	2986
Act -	0	3000

(b) ArcFace(ir152)

	Pred +	**Pred -**
Act +	22	2978
Act -	0	3000

(c) FaceNet

	Pred +	**Pred -**
Act +	282	2718
Act -	0	3000

(d) CurricularFace

	Pred +	**Pred -**
Act +	0	3000
Act -	0	3000

(e) MobileFaceNet

	Pred +	**Pred -**
Act +	0	3000
Act -	0	3000

(f) Notions

	Pred +	**Pred -**
Act +	TP	FN
Act -	FP	TN

slightly outperforms it with a top-5 accuracy of 0.6% and 7.1% for top-100. FaceNet, with an average rank of 1,461, has the lowest top-5 accuracy at 0.4%, and its top-100 accuracy stands at 6.8%. CurricularFace demonstrates a robust performance with a top-5 accuracy of 0.6%, increasing to 7.2% for top-100. Finally, MobileFaceNet's performance closely mirrors that of ArcFace(irse50), with a top-5 accuracy of 0.5% and a top-100 accuracy of 7.0%.

F.2 Face Verification: Confusion Matrices

The Labeled Faces in the Wild (LFW) verification benchmark serves as a standard evaluation protocol for assessing the performance of facial verification algorithms. The benchmark comprises a total of 6,000 pairings, evenly divided into 10 distinct “folds" or subsets. Each of these subsets contains 600 image pairs, half of which (300 pairs) are of the same individual (matched), while the other half represent two different individuals (non-matched). To ensure comprehensive assessment, evaluations iterate for all ten folds. The resulting performance metrics, derived from these ten-fold cross-validations, offer insights into an algorithm's true positive rate, false positive rate, and overall accuracy in distinguishing between matched and non-matched face pairs.

Table 8 shows the confusion matrices for five distinct facial recognition models under adversarial attack conditions. Each table represents the outcomes from testing on the LFW benchmark. The matrices provide a detailed breakdown of the True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN) for each system (Table 8f). Notably, ArcFace(irse50) and ArcFace(ir152) displayed some recognition capability for matched pairs, albeit limited. In contrast, both CurricularFace and MobileFaceNet failed to correctly identify any matched pairs. Despite these discrepancies in matched pair recognition, all models consistently identified non-matched pairs with perfect accuracy, as indicated by the TN values. These results underscore the resilience of the tested models to false positives under the given attack but reveal vulnerabilities in their true positive recognition rates.

The adversarial attack's primary effect on the facial recognition systems was a significant inhibition in their ability to accurately detect matched pairs. This is most evident from the elevated False Negative (FN) counts across the models, indicating the systems' failures to recognize legitimate matches. Conversely, the True Negative (TN) values remained consistently high for all systems, showing that the attack did not impair their ability to correctly identify non-matched pairs. This skewed impact of the attack, predominantly affecting the True Positives (TP) while leaving the False Positives (FP) largely unchanged, suggests a targeted vulnerability in these models' match-recognition mechanisms. Such a specific degradation in performance, while other aspects remain unaffected, underscores the need for refined defenses against adversarial perturbations targeting the match-detection capability of facial recognition systems.