GAN-Invert: Unveiling Vulnerabilities in Privacy-Preserving Facial Transformations

Umesh Kashyap
Indian Institute of Technology Bhilai
umeshk@iitbhilai.ac.in

Sk. Subidh Ali
Indian Institute of Technology Bhilai
subidh@iitbhilai.ac.in

Abstract

Face recognition is now widely used in authentication, surveillance, and social media, but it also raises serious privacy risks. Face recognition models enable unauthorized identification of individuals from publicly shared images, support mass surveillance and tracking without consent, and allow inference of sensitive personal attributes such as age, gender, or health conditions. Since biometric data cannot be revoked like a password, once facial embeddings are leaked, they can be exploited for identity theft and cross-platform re-identification. To address these challenges, many deep learning based methods have been proposed to alter facial images so that identity is concealed while the images remain useful for deep learning tasks such as age estimation, attribute recognition, expression analysis, and face recognition for an authorized system. These methods include pixel-level manipulation, generative adversarial makeup, feature disentanglement, and key-based reversible encryptions. However, most of them follow the idea of bounded distortion, where the image is slightly altered for privacy preservation while keeping the image quality and the corresponding deep learning task accuracy intact. In this paper, we perform a detailed security analysis of these deep learning based privacy-preserving methods and show that these defense mechanisms are fundamentally insecure. Using theoretical as well as extensive experimental analysis, we demonstrate that a conditional GAN model can be trained to reconstruct the original image from the privacy-preserving protected image. Our attack analysis on the ten best-known privacy-preserving methods recovers the original from the protected image with high accuracy. Our results expose the key limitations of existing deep learning based privacy preserving methods and stress the need for privacy-preserving solutions based on stronger principles, such as information theory or cryptography, while still ensuring functionality for deep learning tasks.

Keywords

Privacy-Preserving, GAN-Invert, Face-Recognition, Facial Attribution

1 Introduction

Face recognition has witnessed tremendous advancements over the past decade, largely driven by deep learning [31, 37, 46]. Modern convolutional and transformer-based architectures have achieved remarkable performance in deep learning tasks such as face recognition [37, 46], expression recognition, age estimation, and demographic analysis [26, 41], etc. These systems now support a variety of real-world applications, ranging from smartphone authentication, border security to social media content moderation, and personalized recommendation services [1, 2]. The success of these models relies on the ability to automatically extract robust, discriminative features from images, enabling high recognition accuracy even under variations in pose, lighting, and expression [26, 37, 41]. As a result, facial recognition technology has become both highly practical and deeply integrated into everyday digital experiences [1, 31].

Despite these advancements, the widespread deployment of facial recognition has raised significant privacy concerns. Unlike passwords or other revocable credentials, biometric identifiers such as facial features are permanent and inherently sensitive to misuse [27, 44]. Unauthorized access to these data can lead to identity theft, profiling, or surveillance by malicious actors, service providers, or third parties. Users increasingly demand services that rely on facial analysis, such as age-appropriate content filtering, automatic photo tagging, or emotion-aware interfaces, without compromising the confidentiality of their biometric information. This creates a fundamental trade-off between the accuracy of privacy and deep learning task accuracy on images [18, 27, 44]. Increasing privacy will also degrade the image quality, which in turn lowers the performance of the corresponding deep learning task.

To address these issues, a diverse set of Deep Learning-based Privacy-Preserving (*DLPP*) transformation methods has been proposed. Pixel-level obfuscation methods, such as $VisualMixer$ [21] leverages Visual Feature Entropy (VFE) to guide adaptive pixel shuffling in the input image, where identity-revealing regions are shuffled more aggressively, while preserving the textured areas important for deep learning tasks. This region-adaptive method effectively protects identity information without compromising features, which is important for deep learning tasks. Adversarial and makeup-based methods, including *CLIP2Protect* [32] and $Diff\text{-}Privacy$ [10], introduce subtle structured perturbations that are visually realistic while effectively misleading face recognition systems. *CLIP2Protect* generates protected faces by adversarially exploring the latent space of a pretrained generative model, restricting modifications to identity-preserving codes and applying natural, makeup-like perturbations guided by text prompts. $Diff\text{-}Privacy$ uses a diffusion-based framework, where conditional embeddings are learned to guide the generation, and an identity-guidance module ensures that the protected image deviates from the original identity while maintaining image quality. The original image can later be recovered only by an authorized user using the correct keys.

Disentangled De-ID [43] separates identity and attribute features, keeping non-identity attributes such as expression and pose

Proceedings on Privacy Enhancing Technologies 2026(2), 76–91

https://doi.org/10.56553/popets-2026-0037

unchanged while modifying the identity based on a password and privacy level. The modified identity is then combined with the preserved attributes to reconstruct de-identified images that retain deep learning task-relevant features while protecting identity. Key-based or reversible masking methods, including *RiDDLE* [20] and *Identity Transformers* [9], rely on invertible mappings and secret keys to hide identity. Despite their methodological diversity, nearly all these methods operate under the principle of bounded distortion, *i.e.*, the transformed images remain visually and structurally similar to the original in order to preserve deep learning task performance and realism [4, 29].

Several recent studies have attempted to analyze the security of *DLPP* transformations to understand whether maintaining deep learning task performance inevitably compromises privacy. In other words, whether the existing *DLPP* transformations are vulnerable to a learnable reconstruction attack, where the reconstruction method is trained on protected and original image pairs to reconstruct the original image from the protected image. *Fantomas* [38] introduced a unified framework for testing the reversibility of transformation methods, showing that many visual transformations remain partially invertible due to the preservation of structural cues. Master-Key GAN [17] extended this analysis to deep learning–based protection methods, namely, perceptual encryption methods such as Learnable Encryption (*LE*) [33], Encryption-then-Compression (*EtC*) [34], and *AVIH* [36]. The analysis showed that these perceptual encryption methods also retain recoverable visual information. Theoretical studies, such as the Least-Privilege Principle (*LPP*) [35], and Conditional Entropy Bottleneck (*CEB*) [8], further formalize this trade-off between privacy and deep learning task performance, proving that any representation-preserving task inevitably reveals unintended information. Similarly, *Cloak* [24] also highlighted this trade-off by empirically showing that suppressing identity features enhances privacy but degrades task accuracy.

These studies collectively suggest that the principle of bounded distortion, maintaining deep learning task performance by preserving visual or structural similarity between original and protected images, remains the core source of vulnerability in existing methods. By preserving visual and structural features of the original image, much of the underlying identity information remains encoded in the obfuscated image [8, 17, 24, 35]. The key question, therefore, is whether pixel shuffling, subtle adversarial perturbations, or latent-space manipulations can fully remove identity cues while retaining task-relevant features. If not, can an adversary reconstruct the original image from these residual cues left by *DLPP* transformations? Despite these critical concerns, no systematic evaluation has yet been conducted to assess the true security effectiveness of these defense methods.

In this work, we perform a detailed security analysis of existing deep learning based privacy-preservation methods. We propose *GAN-Invert*, a method based on a conditional *GAN* trained on pairs of protected images and original images. The network learns to approximate the inverse mapping of a bounded-distortion transformation, effectively reconstructing the original image. *GAN-Invert* applicable across multiple *DLPP* transformation categories, such as pixel-level manipulation, generative adversarial makeup, feature disentanglement, and key-based reversible encryption. Through extensive experiments on the ten best-known *DLPP* transformation methods, we demonstrate that high-fidelity reconstruction is possible in all cases, highlighting the universal vulnerability of bounded-distortion defenses.

The contributions of this paper are summarized as follows:

- We formally establish the limits of achievable privacy under bounded-distortion constraints, showing that transformations constrained by deep learning tasks are inherently invertible.
- We propose *GAN-Invert*, a framework that can be trained to reconstruct the original images from their protected images generated by *DLPP* transformation methods.
- We evaluate our attack across ten well-known representative defenses, demonstrating consistent high-fidelity reconstruction.

2 Related Works

Privacy-preserving facial analysis has become an active research area due to the widespread adoption of face recognition in authentication, surveillance, and social applications. These methods aim to protect sensitive identity information while maintaining usability for deep learning tasks such as demographic estimation, emotion recognition, and face verification. Existing methods can be broadly categorized based on the underlying mechanism and design philosophy, as described below.

2.1 Pixel-Level Manipulation

Pixel-Level Manipulation methods focus on altering pixel-level structures to disrupt facial identity recognition while preserving sufficient information for detection and attribute inference. *VisualMixer* [21] introduces a *DLPP* transformation that uses Visual Feature Entropy (*VFE*) to guide adaptive pixel shuffling. *VFE* quantifies local pixel variation, where low *VFE* areas (smooth regions) often reveal facial identity cues, while high *VFE* areas (textured regions) contain important features for deep learning tasks. Based on this, *VisualMixer* applies region-based shuffling, assigning larger windows to low *VFE* areas for stronger obfuscation and smaller windows to high *VFE* areas to retain deep learning task-relevant information. Overall, *VisualMixer* demonstrates that entropy-guided, region-adaptive transformations can effectively protect identity information while preserving features critical for deep learning tasks. Extending this idea, *AVIH* [36] combines pixel-level disruption with carefully crafted adversarial perturbations that retain important features for target deep learning tasks. *AVIH* generates protected images that are difficult for unauthorized models to reconstruct, while allowing selective recovery by a designated key model.

2.2 Generative Adversarial Makeup

Generative Adversarial Makeup methods leverage deep generative models, such as *GAN*s or diffusion models, to produce visually realistic but privacy-preserving images. *CLIP2Protect* [32] protects facial privacy by slightly changing the face in the latent space of a pretrained generative model. It modifies only the parts related to identity, keeping the face close to the original to avoid visible

artifacts. The method uses text-based makeup prompts (*e.g.*, red lipstick and purple eyeshadow) to guide these changes, making them look natural. This allows the protected faces to fool commercial face recognition systems while still looking realistic in human eyes. *Diff-Privacy* [10] adopts a diffusion-based generative framework to obfuscate identity while retaining deep learning task-relevant information. It adds controlled noise to an image so the generated protected image looks natural, but hides the original identity. The identity features are changed while keeping pose, expression, and background intact. *Diff-Privacy* uses an identity guidance module to control denoising to ensure the protected image does not match the original image in a face recognition system. A noise is saved as a key, which also allows an authorized user to recover the original image if needed. *AMT-GAN* [11] uses synthetic makeup as a medium for embedding adversarial features in natural-looking images through *GAN*s to enhance privacy without affecting visual realism.

2.3 Feature Disentanglement

Feature Disentanglement methods operate in the latent representation space of deep neural networks, aiming to separate identity-related features from deep learning task-relevant features. *Feature Subtraction* [23] presents a *DLPP* framework that generates protected images through feature subtraction. The process begins by extracting latent features using an encoder, followed by training a generative model to accurately reconstruct these latent features. The residual latent features, obtained by subtracting the reconstructed latent feature from the original latent features, serve as the protective latent feature. These residual latent features suppress fine-grained visual details while retaining essential identity-related information. To further enhance privacy, random channel shuffling is applied to the residual latent features before decoding them back into a protected image. This operation obscures the visual appearance of a person in the protected image while preserving sufficient identity information for deep learning tasks.

Disentangled De-ID [43] protects facial images by separating identity and other facial features into different latent embeddings. The identity embedding is modified in a controlled, password-guided way to create a new, unreal identity, while features like pose, expression, and background are preserved. A generator reconstructs a protected image that does not match the original identity but retains the original image attributes for deep learning tasks. *PRO-Face C* [42] protects facial privacy by concealing visual identity cues, while preserving task-relevant information for deep learning tasks. It uses a two-stage process: first stage, use obfuscation methods like blurring, pixelation, or face morphing to hide the individual's identity. In the second stage, a feature compensation network restores non-identity attributes such as pose and expression. These features are combined with the obfuscated image to enable deep learning tasks like emotion recognition, age estimation, or identity verification without revealing the person's identity. Despite operating in high-level feature space, these methods are limited by the near-bijective nature of latent transformations [3, 19].

2.4 Key-Based Reversible Encryption

Key-Based Reversible Encryption methods provide controlled, password protected transformations of facial identities that can be fully reversed. Instead of permanently altering a face, these methods temporarily replace the original identity with an anonymous person's face while preserving all other facial features and image quality of the original images. *RiDDLE* [20] demonstrates this idea by working in the latent space of *StyleGAN2* [16], where facial identity is encrypted with password-based keys. This allows the same face to be protected in multiple ways using different passwords. The original identity can only be restored with the correct password, while incorrect ones generate alternate protected faces, ensuring privacy even against unauthorized access. Extending this idea, *Identity Transformers* [9] advance this concept by providing a unified architectural framework that seamlessly integrates both protection and recovery processes within a single deep learning model. Conditioned on discrete password tokens, the model performs automated, photo-realistic identity transformations while ensuring that authorized recovery requires exact password matching. This unified method eliminates the need for separate protection and reconstruction models, streamlining the bidirectional identity transformation process while maintaining security through password conditioning.

A key question arises: Does preserving visual realism and functionality for deep learning tasks inherently retain significant identity information? Are *DLPP* methods across categories, such as pixel-level manipulation, generative adversarial makeup, feature disentanglement, and key-based reversible encryption, vulnerable to learning-based reconstruction attacks due to bounded-distortion constraints? This motivates the need to systematically evaluate the effectiveness of *DLPP* transformation methods in protecting identities against reconstruction attacks.

2.5 Existing Learnable Reconstruction Attack

Recent studies have attempted to analyze the security of *DLPP* transformations, particularly to examine whether transformations preserving deep learning tasks retain recoverable identity features. *Fantomas* [38] proposed a reconstruction attack to assess reversibility of *DLPP* transformations using a reconstruction model with convolutional and linear layers trained on paired original and protected images (generated by blurring, pixelation, color transformation, and synthetic replacement). The attack was able to reconstruct the original image from their protected image due to residual spatial correlations between them. However, *DLPP* transformation methods [13, 22] are shown to be secure against the attack proposed in *Fantomas*. Subsequently, the Master-Key GAN attack [17] was proposed to evaluate deep learning-based perceptual encryptions, which is a type of *DLPP* transformation method for protecting the visual content of the original image. A generator–discriminator pair was trained by introducing adversarial, reconstruction, and perceptual consistency losses to reconstruct protected images, exposing that *DLPP* transformation methods such as *LE* [33], *EtC*[34], and *AVIH* [36] inherently preserve identity recoverable features. In a similar line, *LPP* [35] theoretically established that preserving features in the protected image for the deep learning task inevitably

reveals unintended information, thus leading to a trade-off between privacy and deep learning task performance. *CEB* [8] extended this idea by showing that the original image can be recovered up to a certain extent with a minimal deep learning task features preserved in the protected image. Complementarily, *Cloak* [24] highlighted the inherent trade-off between privacy and the performance of deep learning tasks. Collectively, these works confirm that existing *DLPP* methods achieve privacy by maintaining bounded distortion, which inevitably preserves identity features.

Algorithm 1: GAN-Invert Training Procedure (Batch-based Optimization)

Input: Training set $(x, \hat{x})$, batch size B, learning rates η_G, η_D, loss weights $\alpha, \beta, \gamma, \lambda$, number of epochs N_{epoch}, validation interval N_{val}, patience N_{patience}

Output: Trained generator G_θ and discriminator D_θ

for *epoch = 1* **to** N_{epoch} **do**
  for *each batch* $\{(x_j, \hat{x}_j)\}_{j=1}^{B}$ *in Training set* **do**
    // Generate reconstructions for batch;
    $\tilde{x} \leftarrow G_\theta(\hat{x})$;
    //Discriminator update
    $\mathcal{L}_D \leftarrow -\frac{1}{B}\sum_{j=1}^{B}\left[\log D_\theta(x_j) + \log(1 - D_\theta(\tilde{x}_j))\right]$;
    $D_\theta \leftarrow D_\theta - \eta_D \cdot \nabla_{D_\theta}\mathcal{L}_D$;
    //Generator update
    $\mathcal{L}_{\text{pixel}} \leftarrow \frac{1}{B}\sum_{j=1}^{B}\|\tilde{x}_j - x_j\|_2^2$;
    $\mathcal{L}_{\text{feat}} \leftarrow \frac{1}{B}\sum_{j=1}^{B}\sum_l \|\phi_l(\tilde{x}_j) - \phi_l(x_j)\|_2^2$;
    $\mathcal{L}_{\text{attr}} \leftarrow \frac{1}{B}\sum_{j=1}^{B}\|\mathcal{A}(\tilde{x}_j) - \mathcal{A}(x_j)\|_2^2$;
    $\mathcal{L}_{\text{adv}} \leftarrow -\frac{1}{B}\sum_{j=1}^{B}\log D_\theta(\tilde{x}_j)$;
    $\mathcal{L}_G \leftarrow \alpha\mathcal{L}_{\text{pixel}} + \beta\mathcal{L}_{\text{feat}} + \gamma\mathcal{L}_{\text{attr}} + \lambda\mathcal{L}_{\text{adv}}$;
    $G_\theta \leftarrow G_\theta - \eta_G \cdot \nabla_{G_\theta}\mathcal{L}_G$;
  end
  // Validation
  if *epoch* % N_{val} == 0 **then**
    $\mathcal{L}_{\text{val}} \leftarrow \text{Validate}(G_\theta, (x, \hat{x}))$;
    if *no improvement for* $N_{patience}$ *epochs* **then**
      break;
    end
  end
end
return G_θ, D_θ

3 Methodology

In this section, we present our method, *GAN-Invert*, which investigates the inherent invertibility of *DLPP* transformations, particularly those constrained by bounded-distortion. Such transformations are designed to obscure personal identity while retaining information for deep learning tasks. However, because they operate under the restriction of introducing only limited distortion, they inevitably leave behind structured cues that can be exploited to reconstruct the original image. We combine theoretical insights with a practical reconstruction framework in *GAN-Invert*, with the aim of reconstructing the original images from their protected images.

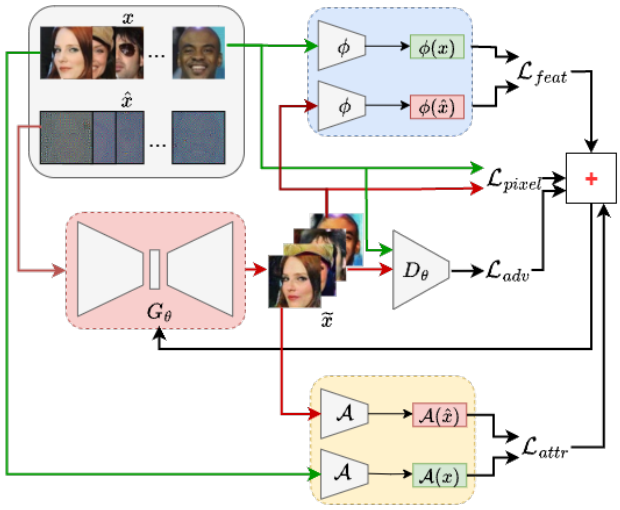


Figure 1: Overview of GAN-Invert training pipeline.

3.1 Problem Formulation

Recent *DLPP* methods aim to hide personal identity information while keeping only the features needed for tasks such as age, gender, or emotion classification. These methods usually assume that once an image is transformed using a privacy-preserving method, it is impossible to recover the original image. Many frameworks also allow users to apply these transformations locally with a private key or by adjusting the distortion level before sharing the protected image online or storing it on cloud platforms [9–11, 20, 21, 23, 32, 36, 42, 43]. However, this assumption is not always true in real-world situations. Having access to protected images, an attacker can perform a learnable reconstruction attack to recover the original image [8, 17, 24, 35, 38]. For example, the attacker can be an administrator of the cloud storage where the protected images are stored, or he/she may be a common friend of the owner of the protected image on social media. Therefore, the security of *DLPP* methods must be evaluated before being deployed in real-world applications.

Formally, let $\mathbf{x}$ be an original image and let f be a *DLPP* transformation that creates a protected image $\hat{\mathbf{x}} = f(\mathbf{x})$. The attacker's goal is to learn an inverse mapping G_θ such that $G_\theta(\hat{\mathbf{x}}) \approx \mathbf{x}$. This reconstruction task is treated as a conditional image generation problem, where G_θ tries to produce a realistic and identity-consistent reconstruction $\tilde{\mathbf{x}} = G_\theta(\hat{\mathbf{x}})$. The effectiveness of our attack is evaluated by its ability to reveal the correct identity from the reconstructed image. However, the high visual quality reconstructions of our attack shows that the attack is not limited to identity recovery but also overcomes facial privacy. For example, methods such as *AVIH*, *VisualMixer*, *Feature Subtraction*, and *Pro-Face C* aim to prevent high-quality face recovery, yet our results demonstrate that the original face can still be reconstructed.

Threat Model: We consider adversary operating under realistic yet powerful assumptions. The adversary knows the transformation

function f and has access to a set of protected images $\hat{\mathbf{x}}$, however, he/she does not have access to the corresponding original images $\mathbf{x}$. Furthermore, the adversary does not have access to user-specific parameters, such as keys, identity labels, random seeds, or distortion levels, used during the transformation f. The attacker also has access to a set of original images from a public dataset, which does not include any individuals from the test set. The adversary constructs a surrogate dataset using a publicly available dataset (e.g., *CelebA-HQ*) to train the reconstruction model G_θ. As the transformation method f is known to the adversary, therefore, it applies the transformation method with its own key and distortion level to collect a surrogate protected and original image pairs. These pairs are then used to train G_θ that learns to approximate the inverse mapping of f.

Once trained, G_θ is applied to unseen protected images, generated by the same transformation methods f with different parameters or user keys, to reconstruct the original images. The reconstructed outputs $\tilde{\mathbf{x}} = G_\theta(\hat{\mathbf{x}})$ are evaluated in terms of identity recognition (The adversary can use any public or fine-tune face recognition model for testing the result). The adversary matches the reconstructed images against a gallery of known identities collected from public sources, representing individuals who are likely to appear in the protected dataset. If there is a match, then the attack is successful.

3.2 GAN-Invert Architecture

GAN-Invert is designed as a conditional generative framework to reconstruct original images x from protected images $\hat{x}$ as shown in Algorithm 1 and Figure 1. The training pipeline consists of a generator G_θ, a discriminator D_θ, a feature extractor ϕ, and an attribute predictor $\mathcal{A}$. The generator G_θ produces a reconstructed image $\tilde{x} = G_\theta(\hat{x})$ from a protected image $\hat{x}$, while the discriminator D_θ simultaneously learns to distinguish between original images x and reconstructed images $\tilde{x}$. The feature extractor ϕ is a pretrained face recognition network that guides the generator to preserve identity-related information by aligning high-level feature representations of $\tilde{x}$ and x. The attribute predictor $\mathcal{A}$ ensures that reconstructed images maintain facial attributes such as age, gender, and expression, while reconstructing the original image from a protected image.

To effectively perform the reconstruction, the generator adopts a *U-Net* [28] style encoder-decoder architecture with skip connections, designed to perform conditional image-to-image translation from protected images to reconstructions. The encoder consists of convolutional layers with stride-2 downsampling, each followed by batch normalization and LeakyReLU activation. This progressively reduces spatial resolution while increasing feature dimensionality, allowing the network to capture high-level semantic and identity-related information hidden within the protected image. The encoder produces a compact latent representation encoding both global facial structure and subtle local features. The decoder mirrors the encoder through transposed convolution layers (or upsampling followed by convolution), each followed by batch normalization and *ReLU* activation, progressively restoring the spatial resolution. Skip connections between encoder and decoder layers allow low-level visual information, such as edges, contours, and textures to be preserved. A final convolutional layer with *Tanh* activation outputs the reconstructed image in the pixel domain. The generator is trained using a combination of pixel-level reconstruction loss, feature-level identity loss, attribute-preserving loss, and adversarial loss. The adversarial loss encourages the generator to produce outputs that lie on the manifold of natural faces, improving perceptual quality and reducing artifacts. The generator loss is expressed as:

$$\mathcal{L}_G = \alpha\, \mathcal{L}_{\text{pixel}} + \beta\, \mathcal{L}_{\text{feat}} + \gamma\, \mathcal{L}_{\text{attr}} + \lambda\, \mathcal{L}_{\text{adv}}, \tag{1}$$

where α, β, γ, and λ control the relative importance of each term. The pixel-level loss ($\mathcal{L}_{\text{pixel}}$) enforces local texture and structure alignment, the feature-level loss ($\mathcal{L}_{\text{feat}}$) ensures identity preservation through a pretrained face recognition model, the attribute loss ($\mathcal{L}_{\text{attr}}$) maintains deep learning task relevant features on reconstructed image, and the adversarial loss ($\mathcal{L}_{\text{adv}}$) encourages realistic reconstructions.

The discriminator follows a *PatchGAN* architecture, evaluating local image patches instead of the entire image. It consists of convolutional layers with stride of 2 downsampling, each followed by batch normalization and LeakyReLU activation. This produces a spatial grid where each element corresponds to the discriminator's assessment of a 70×70 patch. By operating at the patch level, the discriminator ensures high-frequency detail consistency and forces the generator to produce locally realistic facial textures. It is a binary cross-entropy loss defined as:

$$\mathcal{L}_D = -\mathbb{E}_x \left[\log D(x)\right] - \mathbb{E}_{\hat{x}} \left[\log\left(1 - D(G(\hat{x}))\right)\right], \tag{2}$$

Through this adversarial training, the discriminator continually pushes the generator to improve realism at the patch level, while the generator balances realism with reconstruction of the original identity.

3.3 Loss Functions

To achieve high-quality reconstruction, *GAN-Invert* minimize a multi-objective loss function that balances pixel-level fidelity, semantic identity consistency, attribute preservation, and visual realism. Each component plays a distinct role in overcoming the limitations of bounded-distortion transformations.

Pixel-Level Reconstruction Loss: The pixel-level loss directly minimizes the ℓ_2 distance between the reconstructed and original images:

$$\mathcal{L}_{\text{pixel}} = \left[\|G_\theta(f(x)) - x\|_2^2\right]. \tag{3}$$

This term enforces alignment at the level of local textures and structural details. Since bounded-distortion transformations are constrained not to introduce drastic pixel changes, enforcing pixel-wise similarity allows the model to exploit the subtle residual traces of identity left behind. Although this loss often leads to overly smooth reconstructions when used in isolation, it forms a necessary foundation for accurate reconstruction.

Feature-Level Identity Loss: Pixel alignment alone does not guarantee reconstruction of high-level identity information, which is critical for reversing *DLPP* transformations. To address this, we

introduce a feature-based loss that compares representations extracted from a pretrained face recognition network ϕ:

$$\mathcal{L}_{\text{feat}} = \sum_{l} \left[\|\phi_l(G_\theta(f(x))) - \phi_l(x)\|_2^2 \right], \tag{4}$$

where $\phi_l(\cdot)$ denotes the feature map at layer l of a pre-trained face recognition network. This loss imposes constraints on the reconstructed image at multiple feature levels, where low-level layers ensure that textures, edges, and fine details are preserved, while high-level layers enforce that semantic and identity-related features are maintained. By operating across these layers, the loss ensures that the reconstructed image retains identity information and perceptual fidelity, even if some pixel-level alignment is not perfect, making it a powerful complement to the pixel loss.

Attribute-Preserving Loss: While identity reconstruction is the primary focus, *DLPP* transformations are often designed to maintain functionality for deep learning tasks such as age prediction, expression recognition, etc. To ensure that reconstructions remain useful for these purposes, we introduce an attribute-preserving loss:

$$\mathcal{L}_{\text{attr}} = \left[\|\mathcal{A}(G_\theta(f(x))) - \mathcal{A}(x)\|_2^2 \right], \tag{5}$$

where $\mathcal{A}$ is a pretrained attribute predictor. This term encourages the generator to preserve facial attributes such as gender, age, or expression, thereby demonstrating that reconstruction not only restores identity but also retains functionality for deep learning tasks.

Adversarial Loss: Pixel-based losses tend to produce blurred outputs due to the averaging effect of ℓ_2 minimization. To encourage sharper and more realistic reconstructions, we optionally incorporate an adversarial objective:

$$\mathcal{L}_{\text{adv}} = \min_{G_\theta} \max_{D_\theta} \left[\log D_\theta(x) + \log(1 - D_\theta(G_\theta(f(x)))) \right]. \tag{6}$$

The discriminator D_θ distinguishes between real and reconstructed images, forcing the generator to produce outputs that lie on the manifold of natural faces. This adversarial regularization significantly improves perceptual quality and reduces the visibility of artifacts.

3.4 Theoretical Analysis: Universality of Reconstruction

We now formalize the claim that a *DLPP* transformation constrained by bounded distortion is invertible. This theoretical result underscores the fundamental vulnerability of privacy-preservation transformation methods that our proposed *GAN-Invert* framework exploits to reconstruct the protected images.

Let $x \in X \subseteq \mathbb{R}^{H\times W\times C}$ denote an original image, where H, W, C are the height, width, and number of channels, respectively, and let $\hat{x} = f(x)$ be its protected image generated by a *DLPP* transformation method $f : X \to Z$, where $Z \subseteq \mathbb{R}^{H\times W\times C}$ is the set of protected images. The objective in f is to induce semantic distortion in the original image to get the protected image. In some *DLPP* methods the semantic distortion is directly performed [20, 43], and in other cases, the distortion is performed by performing pixel-level distortion [21, 36]. For our theoretical analysis, we consider the second case, which is more generic. The *DLPP* method f maintains a bounded-distortion property: whenever the pixel distortion $|x - \hat{x}| \le \delta$ ($\hat{x}$ lies within a δ bound neighborhood of X), it follows the semantic distortion $|\phi(x) - \phi(\hat{x})| \le \epsilon$, where $\epsilon > 0$ bounds the corresponding distortion in Z. This Lipschitz-like condition reflects the requirement that f introduces only controlled distortion, ensuring that deep learning tasks relevant information is preserved even as identity is obfuscated. The goal of a reconstruction model is to learn a mapping $G_\theta : Z \to X$ that reconstructs $\tilde{x} = G_\theta(\hat{x})$, where the reconstructed image $\tilde{x} \approx x$.

Theorem 3.1 (Universality of reconstruction). *For any bounded distortion transformation $f : X \to Z$, there exists a neural network $G_\theta : Z \to X$ such that for all $x \in X$ and $\hat{x} \in Z$:*

$$\|G_\theta(f(x)) - x\| \le \eta \tag{7}$$

where $\eta > 0$ is an arbitrarily small reconstruction error.

Proof. Since X is compact, it can be covered by δ-balls $B(x_i, \delta)$ centered at $x_i \approx x$, defined as

$$B(x_i, \delta) = \{\hat{x} \in X \mid \|\hat{x} - x_i\| \le \delta\}.$$

For any $\hat{x} \in B(x_i, \delta)$, the bounded-distortion property implies that $\phi(\hat{x}) \in B(\phi(x_i), \epsilon)$. Hence, all obfuscated samples $\hat{x} = f(x)$ corresponding to inputs near x_i are mapped within the ϵ-bound neighborhood of $f(x_i)$. Consider an ideal inverse mapping $G^* : Z \to X$ by assigning each obfuscated neighborhood back to its representative center, i.e.,

$$G^*(\hat{x}) = x_i \quad \text{whenever } \hat{x} \in f(B(x_i, \delta)).$$

By construction, for any $x \in B(x_i, \delta)$,

$$\|G^*(f(x)) - x\| = \|x_i - x\| \approx 0.$$

Although G^* is piecewise constant and not realizable by standard learning architectures, the universal approximation theorem [5, 25] ensures that for any $\eta > 0$, there exists a neural network G_θ such that

$$\|G_\theta(\hat{x}) - G^*(\hat{x})\| \le \eta, \quad \forall \hat{x} \in Z.$$

Finally, by adding and subtracting the $G^*(f(x))$ in intermediate terms of Eq. 7, we obtain:

$$\begin{aligned} \|G_\theta(f(x)) - x\| &= \|G_\theta(f(x)) - G^*(f(x)) + G^*(f(x)) - x\| \\ &\le \|G_\theta(f(x)) - G^*(f(x))\| + \|G^*(f(x)) - x\| \\ &\le \|G_\theta(\hat{x}) - G^*(\hat{x})\| + \|G^*(\hat{x} - x\| \\ &\le \eta. \end{aligned}$$

This completes the proof. □

The above theorem establishes that bounded-distortion transformations are universally invertible with negligible error by a sufficiently expressive model. In section 6.4, we experimentally determine the bounded distortion limits δ, ϵ corresponding to target *DLPP* methods and also the value η corresponding to *GAN-Invert*. In practice, this theoretical result is implemented via our *GAN-Invert* framework. To experimentally validate the theorem, we train G_θ using a combination of pixel-level loss, feature-level loss, attribute-preserving loss, and adversarial loss, which demonstrates high-fidelity reconstruction while preserving identity-relevant information.

4 Experimental Setup

4.1 Datasets

We evaluate *GAN-Invert* on multiple widely-used face datasets to comprehensively assess reconstruction performance across diverse demographics, image qualities, and real-world conditions. Our primary benchmark is the *CelebA-HQ* [14] dataset, which contains 30,000 high-resolution celebrity images spanning 6,217 unique identities with variations in age, gender, ethnicity, and facial expression. The dataset is split by identity into 80% for training and 20% for testing, ensuring that no individual identities overlap between train and test sets. To further validate the effectiveness and generalizability of *GAN-Invert*, we also include the *FFHQ* [15] dataset and the *LFW* [12] dataset. *FFHQ* dataset provides 70,000 high-fidelity face images captured under controlled conditions with a wide variety of identities, lighting, and poses. *LFW* dataset offers unconstrained, real-world images that reflect the challenges of in-the-wild face recognition, including varied backgrounds, occlusions, and head poses. Importantly, both *FFHQ* and *LFW* share no overlapping identities with *CelebA-HQ*, ensuring complete disjointness across all datasets.

For our experiments, *GAN-Invert* is primarily trained on the *CelebA-HQ* dataset (80% training split of the dataset). The training data is first processed to generate the corresponding protected images $\hat{x}$ using the *DLPP* transformations under study (*e.g.*, pixel-level Manipulation, Generative Adversarial Makeup, feature disentanglement). For each original image x in the training set, the protected image $\hat{x} = f(x)$ is generated and paired with its original image as $(\hat{x}, x)$, which are used to train the generator G_θ. For evaluation, the model is tested on the held-out 20% split of the *CelebA-HQ* dataset, ensuring that none of the test images are seen during training. Additional model evaluation experiments are performed on the *FFHQ* dataset and the *LFW* dataset to assess generalization under different image distributions and uncontrolled conditions.

4.2 Target Deep learning Privacy-Preserving Transformation Methods

We evaluate our attacks against a representative set of *DLPP* transformations spanning the following four categories. To ensure consistency and fair comparison across all *DLPP* methods, we use the same set of four tasks: age estimation, gender prediction, race classification, and emotion recognition. We reported the mean accuracy across these four tasks. Furthermore, in order to ensure uniform evaluation of all *DLPP* methods, we fixed their task accuracy to approximately $\approx 89\%$ by calibrating their corresponding bounded-distortion levels. This way, once the *DLPP* methods are configured, corresponding protected images are generated for subsequent attack evaluation.

Pixel-Level Manipulation: *VisualMixer* [21] and *AVIH* [36] apply pixel shuffling and adversarial perturbations to obscure identity while retaining structural cues.

Generative Adversarial Makeup: *CLIP2Protect* [32], *Diff-Privacy* [10], and *AMT-GAN* [11] generate naturalistic transformations using GANs or diffusion models, balancing privacy and visual realism.

Feature Disentanglement: *Feature Subtraction* [23], *Disentangled De-ID* [43], and *PRO-Face C* [42] operate in latent space to suppress identity while preserving deep learning task-relevant features.

Key-Based Reversible Encryption: *RiDDLE* [20] and *Identity Transformers* [9] allow controlled reversible *DLPP* transformation method conditioned on secret keys.

These methods collectively provide a robust benchmark across pixel-level, generative, feature-based, and key-based defenses. Our evaluation emphasizes how *GAN-Invert* can exploit residual information preserved by each method, as guided by pixel-level, feature-level, and attribute-preserving losses.

4.3 Evaluation Metrics

In order to evaluate the performance of our attack, we used two main metrics: reconstruction quality and the extent of privacy leakage. Please note that these metrics together represents η in Theorem 3.4.

Reconstruction Quality: Reconstruction quality is measured based on Peak Signal-to-Noise Ratio (*PSNR*) [7], Structural Similarity Index Measure (*SSIM*) [40], and Learned Perceptual Image Patch Similarity (*LPIPS*) [45] scores. *PSNR* evaluates pixel-level similarity between reconstructed and original images, where higher values indicate finer detail preservation in reconstructed images. It ranges from 0 to ∞ dB. *PSNR* value between 25dB and 40dB is considered to be a good reconstruction quality. Whereas *PSNR* more than 40dB indicates that the reconstructed image is almost identical to the original image. *SSIM* measures structural similarity in luminance, contrast, and texture, ranging from 0 to 1, where values above 0.7 denote good perceptual quality. *LPIPS* quantifies perceptual similarity using deep features, with lower scores indicating higher semantic consistency between reconstructed and original images. *LPIPS* values below 0.3 are generally considered good reconstruction quality. Additional details, along with a visual example, are provided in Appendix A.

Privacy Leakage: Please note that the job of the privacy preservation transformation methods is to alter the original image into a protected image such that face recognition models fail to classify the identity of the protected image. Therefore, the objective of an attack on privacy preservation transformation methods is to reconstruct the original image from the protected, such that face recognition models will classify the reconstructed image as the original image. To assess the degree of privacy leakage, we measure face recognition accuracy and identity preservation score of the reconstructed image. Face recognition accuracy evaluates whether state-of-the-art models (such as *ArcFace* [6], *CosFace* [39]) can correctly classify the reconstructed image as the original image, thus leaking the privacy by disclosing the facial identity. Therefore, face recognition accuracy quantifies potential privacy leakage. Identity preservation score quantifies the cosine similarity between embeddings of the original image and the reconstructed image, reflecting how effectively feature-level identity cues are suppressed in the reconstructed image. Together, these metrics offer a comprehensive evaluation of privacy leakage while capturing both identity and feature-level vulnerabilities.

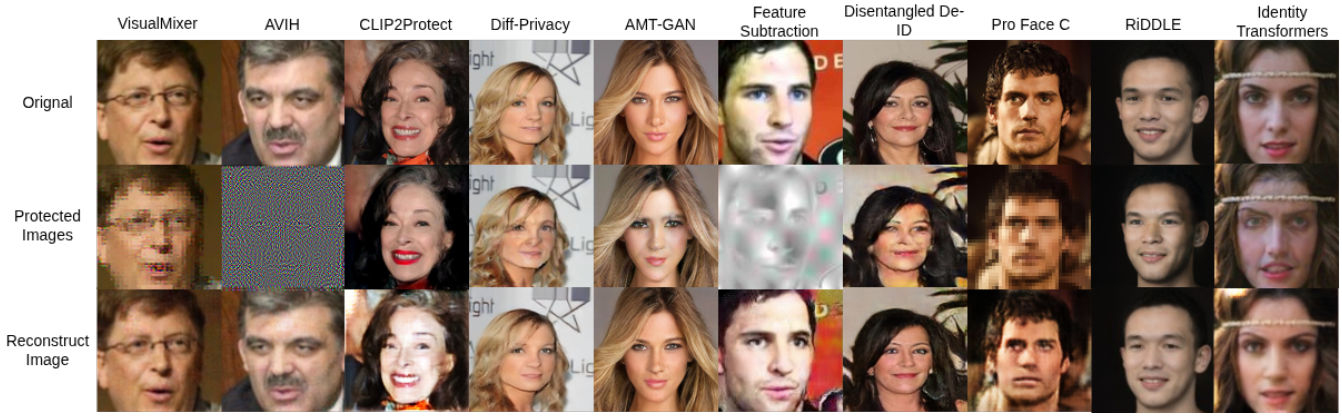


Figure 2: Visualization quality of original image reconstructions by GAN-Invert from protected images produced through DLPP transformations.

4.4 Implementation Details

Architecture: The generator in *GAN-Invert* follows a *U-Net* [28] style encoder-decoder architecture with skip connections, specifically designed for conditional image-to-image translation from protected to reconstructed face images. The network consists of 7 encoder layers, 1 latent layer, and 7 decoder layers. Each encoder layer employs a 4×4 convolution with a stride of 2, progressively reducing the spatial resolution while increasing feature dimensionality. The number of feature channels across the encoder is $\{64, 128, 256, 512, 512, 512, 512, 512\}$. All layers except the first are followed by batch normalization and LeakyReLU activations with a negative slope of 0.2. The outputs of the encoder layers are stored for skip connections to the corresponding decoder layers, enabling the retention of low-level spatial information. The encoder produces a latent feature map of size $1 \times 1 \times 512$, which encodes both the global facial structure and fine-grained identity-specific details.

The decoder mirrors the encoder using transposed convolutional layers (or upsampling followed by convolution) with kernel size 4×4 and stride of 2. The decoder channels are $\{512, 512, 512, 512, 256, 128, 64, 3\}$. Each layer is followed by batch normalization and ReLU activation. Skip connections concatenate the corresponding encoder outputs at each stage, ensuring that edges, contours, textures, and other low-level details are preserved and incorporated into the reconstruction. The final layer is a 4×4 convolution with a stride of 1, followed by a Tanh activation, which produces the reconstructed image in the pixel domain. During training, the generator is optimized with a multi-objective loss combining pixel-level reconstruction loss, feature-level identity loss, attribute-preserving loss, and an adversarial loss provided by a *PatchGAN* discriminator. This training strategy enables the generator to produce reconstruction images that are both visually realistic and semantically match the original identity, effectively reversing *DLPP* transformations while preserving the relevant features of the deep learning task in the reconstruction images.

Feature Extraction Model: For identity preservation, we utilize a pretrained *ArcFace* [6] model. The feature-level loss is computed by extracting the feature map of multiple layers pretrained *ArcFace*. These feature maps capture both low-level texture details and high-level semantic identity information. This multi-layer supervision ensures that reconstructed images maintain discriminative facial features.

Attribute Prediction Model: In our *GAN-Invert* framework, facial attributes are estimated using the pretrained *DeepFace* [30], which provides reliable predictions for age, gender, race, and facial expression. *DeepFace* achieves 97.35% accuracy for gender classification and a mean absolute error of ±4.65 years for age estimation, ensuring high-quality attribute supervision. These predictions are used to compute the attribute-preserving loss, guiding the generator to maintain deep learning task-relevant information in reconstructed images while still revealing identity-related cues.

Training: *GAN-Invert* is trained using the Adam optimizer with a learning rate $\eta_G = \eta_D = 0.001$ and a batch size of 16. The training objective combines four losses with corresponding hyperparameters: pixel-level reconstruction ($\alpha = 0.3$), feature-level identity ($\beta = 0.2$), attribute-preserving ($\gamma = 0.4$), and adversarial guidance ($\lambda = 0.1$). Pixel-level reconstruction loss enforces alignment of local textures and structural details. Feature-level loss uses embeddings from a pretrained face recognition network (*ArcFace*) to preserve identity cues. Attribute-preserving loss relies on a pretrained attribute classifier (*DeepFace*) to maintain the visual appearance of the reconstructed images and facial attributes such as age, gender, and expression. Adversarial loss, applied via a discriminator network, encourages reconstructions to appear realistic and reduces blurriness. Models are trained for 120 epochs with early stopping based on validation performance to ensure high-fidelity outputs. We train the attack model separately for each defense method, using the corresponding original image and protected image pairs as training data. The trained model is then evaluated on disjoint and unseen identities with protected images to measure attack generalization and privacy leakage under realistic conditions.

Hardware and Preprocessing: Experiments are conducted on NVIDIA *A*6000 *GPU*s with 48*GB* memory. All images are resized to 256×256 pixels and normalized to the range $[-1, 1]$. Random horizontal flips are applied during training for data augmentation.

Table 1: Reconstruction quality of a GAN-Invert model across all datasets. Higher PSNR/SSIM and lower LPIPS indicate better reconstruction quality. Values reported in $\mu \pm \sigma$ format, where μ is mean, and σ is standard deviation.

Method	CelebA-HQ			FFHQ			LFW		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VisualMixer	27.9 ± 0.04	0.81 ± 0.02	0.39 ± 0.03	27.2 ± 0.03	0.81 ± 0.03	0.40 ± 0.02	24.9 ± 0.04	0.75 ± 0.03	0.46 ± 0.03
AVIH	29.2 ± 0.03	0.89 ± 0.01	0.18 ± 0.02	28.8 ± 0.04	0.87 ± 0.02	0.19 ± 0.03	26.0 ± 0.03	0.83 ± 0.02	0.26 ± 0.03
CLIP2Protect	27.6 ± 0.04	0.85 ± 0.02	0.17 ± 0.02	27.0 ± 0.03	0.85 ± 0.03	0.18 ± 0.02	26.0 ± 0.04	0.80 ± 0.02	0.24 ± 0.03
Diff-Privacy	30.2 ± 0.02	0.89 ± 0.03	0.16 ± 0.02	29.8 ± 0.03	0.84 ± 0.03	0.18 ± 0.02	27.0 ± 0.02	0.83 ± 0.03	0.20 ± 0.02
AMT-GAN	29.8 ± 0.03	0.87 ± 0.02	0.18 ± 0.03	28.0 ± 0.02	0.85 ± 0.03	0.18 ± 0.03	27.8 ± 0.03	0.82 ± 0.02	0.21 ± 0.02
Feature Subtraction	26.1 ± 0.05	0.75 ± 0.03	0.40 ± 0.04	25.5 ± 0.04	0.73 ± 0.04	0.42 ± 0.03	23.8 ± 0.05	0.70 ± 0.03	0.47 ± 0.04
Disentangled De-ID	26.9 ± 0.04	0.80 ± 0.03	0.29 ± 0.03	25.0 ± 0.03	0.78 ± 0.03	0.31 ± 0.03	24.8 ± 0.04	0.75 ± 0.03	0.38 ± 0.04
PRO-Face C	27.8 ± 0.03	0.85 ± 0.02	0.18 ± 0.02	27.5 ± 0.02	0.83 ± 0.03	0.21 ± 0.03	26.0 ± 0.03	0.80 ± 0.02	0.25 ± 0.03
RiDDLE	27.4 ± 0.04	0.79 ± 0.03	0.39 ± 0.04	26.7 ± 0.03	0.75 ± 0.03	0.40 ± 0.03	26.0 ± 0.04	0.73 ± 0.03	0.43 ± 0.04
Identity Transformers	27.5 ± 0.03	0.87 ± 0.02	0.21 ± 0.03	28.5 ± 0.03	0.86 ± 0.02	0.23 ± 0.02	27.0 ± 0.03	0.82 ± 0.03	0.30 ± 0.03

5 Results

5.1 Reconstruction Quality Evaluation

Our *GAN-Invert* framework demonstrates consistently high-quality reconstructions across a wide range of *DLPP* transformations. Table 1 presents quantitative metrics, including *PSNR*, *SSIM*, and *LPIPS*, which collectively evaluate reconstruction quality, structural similarity, and perceptual quality. For pixel-level manipulation methods, *GAN-Invert* achieves strong results across all datasets. On *AVIH* protected images, *GAN-Invert* reconstructs with a *PSNR* of 29.2 *dB* on *CelebA-HQ*, 28.8 *dB* on *FFHQ*, and 26.0 *dB* on *LFW*. The corresponding *SSIM* values are 0.89, 0.87, and 0.83, while *LPIPS* scores are 0.18, 0.19, and 0.26, respectively. *VisualMixer* protected images are reconstructed with a *PSNR* of 27.9 *dB* on *CelebA-HQ*, 27.2 *dB* on *FFHQ*, and 24.9 *dB* on *LFW*. The *SSIM* values for these datasets are 0.81, 0.81, and 0.75, and the *LPIPS* scores are 0.39, 0.40, and 0.46. These results indicate that even randomized or adversarial pixel manipulations leave sufficient information for accurate reconstruction of the original image.

For generative adversarial transformations, *GAN-Invert* achieves moderate reconstruction quality. *CLIP2Protect* protected images are reconstructed with a *PSNR* of 27.6 *dB* on *CelebA-HQ*, 27.0 *dB* on *FFHQ*, and 26.0 *dB* on *LFW*. The corresponding *SSIM* values are 0.85, 0.85, and 0.80, with *LPIPS* scores of 0.17, 0.18, and 0.24. *Diff-Privacy* images reach a *PSNR* of 30.2 *dB* on *CelebA-HQ*, 29.8 *dB* on *FFHQ*, and 27.0 *dB* on *LFW*. Their *SSIM* values are 0.89, 0.84, and 0.83, and the *LPIPS* scores are 0.16, 0.18, and 0.20. *AMT-GAN* images are reconstructed with a *PSNR* of 29.8 *dB* on *CelebA-HQ*, 28.0 *dB* on *FFHQ*, and 27.8 *dB* on *LFW*. The associated *SSIM* values are 0.87, 0.85, and 0.82, respectively, while *LPIPS* scores are 0.18, 0.18, and 0.21. Despite these moderate reconstruction metrics, *GAN-Invert* consistently preserves recognizable identity features, demonstrating the limitations of generative obfuscations in fully concealing facial identity.

For feature disentanglement transformations, *GAN-Invert* exhibits mixed reconstruction performance. *PRO-Face C* protected images are reconstructed with a *PSNR* of 27.8 *dB* on *CelebA-HQ*, 27.5 *dB* on *FFHQ*, and 26.0 *dB* on *LFW*. Their *SSIM* values are 0.85, 0.83, and 0.80, and *LPIPS* scores are 0.18, 0.21, and 0.25. *Feature Subtraction* protected images are reconstructed with *PSNR* of 26.1 *dB* on *CelebA-HQ*, 25.5 *dB* on *FFHQ*, and 23.8 *dB* on *LFW*, with *SSIM* values of 0.75, 0.73, and 0.70, and *LPIPS* scores of 0.40, 0.42, and 0.47. *Disentangled De-ID* protected images are reconstructed with *PSNR* of 26.9 *dB* on *CelebA-HQ*, 25.0 *dB* on *FFHQ*, and 24.8 *dB* on *LFW*. The corresponding *SSIM* values are 0.80, 0.78, and 0.75, while *LPIPS* scores are 0.29, 0.31, and 0.38. This variability reflects the degree of information loss induced by latent disentanglement or feature suppression. Finally, for key-based reversible encryption transformations, *GAN-Invert* shows consistent performance. *RiDDLE* protected images are reconstructed with a *PSNR* of 27.4 *dB* on *CelebA-HQ*, 26.7 *dB* on *FFHQ*, and 26.0 *dB* on *LFW*. Their *SSIM* values are 0.79, 0.75, and 0.73, with *LPIPS* scores of 0.39, 0.40, and 0.43. In contrast, *Identity Transformers* protected images are reconstruct with higher fidelity, achieving a *PSNR* of 27.5 *dB* on *CelebA-HQ*, 28.5 *dB* on *FFHQ*, and 27.0 *dB* on *LFW*. The corresponding *SSIM* values are 0.87, 0.86, and 0.82, while *LPIPS* scores are 0.21, 0.23, and 0.30. These results and Figure 2 demonstrate that *GAN-Invert* can effectively reconstruct the original image from its protected image even without access to secret keys. Overall, *GAN-Invert* consistently recovers identity and facial attributes across diverse transformation methods and datasets, highlighting the inherent vulnerability of current *DLPP* transformation methods.

5.2 Privacy Leakage Analysis

In addition to reconstruction quality, we evaluate the privacy implications of *GAN-Invert* by measuring how effectively reconstructed images reveal identity information.

Face Recognition Accuracy: We evaluate the effectiveness of *GAN-Invert* in breaching privacy by testing whether state-of-the-art face recognition systems (*ArcFace*, *CosFace*) can correctly match reconstructed images to disjoint images of the same identity rather than the exact instance used for reconstruction. The gallery used for identification contains 1,000 distinct identities. Figure 3 presents the recognition accuracies for all evaluated methods. The protected images achieve an average recognition accuracy of 19.5% with *ArcFace* and 21.2% with *CosFace*, showing that the applied protection methods effectively conceal identity information. To

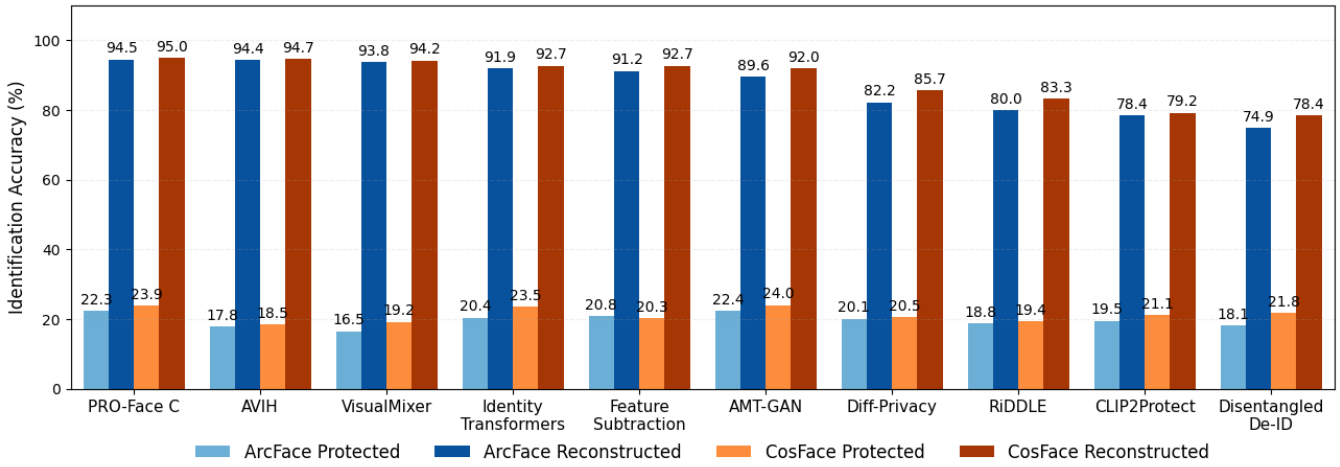


Figure 3: Face recognition accuracy of ArcFace and CosFace models on protected images and GAN-Invert reconstructed images across different *DLPP* transformation methods.

further validate the overlap between the task-relevant and identity features, we analyze two practical scenarios within the *DLPP* framework. In the first scenario of high task performance, we conduct an experiment where the task accuracy on the original image is approximately 97%. We then gradually add perturbations to the image until the task accuracy begins to decrease. Under a high task performance setup, we observe that the baseline privacy leakage remains around 80% when task-relevant features are preserved to maintain high task accuracy. Theoretical analysis of high task transformation is also shown in the appendix C. In the second scenario of a high *DLPP* transformation, we progressively increase the perturbation strength until the protected identity no longer matches the original identity. In this case, we observe a 40% drop in task accuracy. These results indicate that task and identity features exhibit substantial overlap, making it challenging to achieve high levels of privacy without sacrificing deep learning task accuracy.

Building upon these observations, we next examine how the *GAN-Invert* attack exploits this overlap to recover hidden identity information. In order to evaluate the performance of *GAN-Invert* on all the target *DLPP* methods, we need to consider a common evaluation parameter. In this case, we fix the task accuracy of $\approx 89\%$ all the target *DLPP* methods by calibrating the corresponding bounded distortion level. Once the distortion level is fixed, we acquired the corresponding protected image to launch *GAN-Invert*. The results are shown in Figure 3. The results indicate substantial privacy leakage across all target transformations. Pixel-level manipulations (*Visual Mixer*, *AVIH*) offer minimal protection, achieving recognition accuracies above 93.8% and 94.4% with *ArcFace* and above 94.2% and 94.7% with *CosFace*, confirming the vulnerability of simple shuffling and perturbation methods. Generative adversarial makeup methods (*CLIP2Protect*, *Diff-Privacy*, *AMT-GAN*) show moderate resistance, but reconstructed images remain largely recognizable. *Disentangled De-ID* provides the lowest recognition rates (74.9% and 78.4%) but still significantly above random chance. Key-based reversible methods (*RiDDLE*, *Identity Transformers*) are designed for secure reconstruction with an authentic key, but they become highly vulnerable when the protected image is reconstructed using *GAN-Invert*, leading to recognition accuracies of 80% and 91.9% with *ArcFace*, and 83.3% and 92.7% with *CosFace*. Overall, these results underscore the effectiveness of *GAN-Invert* in recovering sensitive identity information, demonstrating that even sophisticated *DLPP* transformations are vulnerable to reconstruction attacks.

We further verified that increasing the gallery size to 5,000 identities reduces recognition accuracy by approximately 13% on average. This behavior is consistent with standard face recognition models, where identification accuracy decreases as the gallery size increases, confirming the effectiveness of our evaluation framework under realistic privacy conditions. We also note that fine-tuned face recognition models can reveal even higher privacy leakage. On average, fine-tuning increases recognition accuracy from 87.13% to 94.11% for *ArcFace* and from 88.79% to 94.86% for *CosFace* across all defense methods. However, this scenario does not align with our threat model, where the attacker lacks access to the original images or identity labels. For real-world applicability, we rely on pre-trained models for evaluation, which provide a conservative and realistic estimate of privacy leakage.

Evaluation of Identity Preservation: To capture the variability of the attack, we split the original test set into three random, non-overlapping subsets. Each subset was independently used for a full *GAN-Invert* reconstruction. The results are reported in the $\mu \pm \sigma$ format, where μ represents the mean and σ denotes the standard deviation across these splits, showing consistent identity leakage across *DLPP* methods based on *GAN-Invert*. Pixel-level manipulations (*VisualMixer*, *AVIH*) exhibit high similarity scores (0.87–0.92), indicating that pixel-level transformations preserve identity features. Feature disentanglement methods are particularly vulnerable, with *PRO-Face C* reaching 0.93 (*ArcFace*) and 0.94 (*CosFace*), and the *Feature Subtraction* method achieves 0.87–0.89, highlighting the privacy risks of manipulating feature representations under bounded distortion. Generative methods (*Diff-Privacy*, *AMT-GAN*) achieve moderate similarity

Table 2: Cosine similarity between embeddings of original and reconstructed images. Higher values indicate more identity information leakage. Values reported in $\mu \pm \sigma$ format, where μ is mean, and σ is standard deviation.

Method	ArcFace	CosFace
VisualMixer	0.87 ± 0.06	0.89 ± 0.06
AVIH	0.90 ± 0.05	0.92 ± 0.07
CLIP2Protect	0.80 ± 0.14	0.81 ± 0.11
Diff-Privacy	0.82 ± 0.18	0.86 ± 0.14
AMT-GAN	0.90 ± 0.10	0.92 ± 0.08
Feature Subtraction	0.87 ± 0.09	0.89 ± 0.07
Disentangled De-ID	0.75 ± 0.17	0.78 ± 0.16
PRO-Face C	0.93 ± 0.05	0.94 ± 0.05
RiDDLE	0.80 ± 0.08	0.82 ± 0.13
Identity Transformers	0.88 ± 0.08	0.91 ± 0.07

(0.82–0.92), retaining partial identity despite intended obfuscation. The methods (*Disentangled De-ID*, *CLIP2Protect*, *RiDDLE*) attain lower similarity scores (0.75–0.82), yet still preserve enough cues to pose recognition risks. High cosine similarity values ($\geq$ 0.8) indicate critical privacy failures, while lower scores do not guarantee full protection. These results underscore the inherent limitation of bounded-distortion defenses, preserving deep learning task-relevant information inherently constrains privacy, leaving a persistent vulnerability exploitable by *GAN-Invert*.

6 Ablation Studies

To better understand the factors contributing to the reconstruction success of *GAN-Invert*, we conduct ablation studies examining network architecture choices, training data requirements, loss function components and empirical validation of bounded distortion.

6.1 Architecture Ablation

To evaluate the impact of generator design on reconstruction quality, we conduct an architecture ablation study comparing three representative generator architectures: a standard CNN, a ResNet-based encoder-decoder, and a *U-Net*. Table 3 summarizes the results in terms of *PSNR*, *SSIM*, and training time. The standard CNN exhibits the lowest reconstruction fidelity (*PSNR* 23.1 *dB*, *SSIM* 0.76), primarily due to the lack of multi-scale feature propagation, which limits its ability to reconstruct the fine-grained identity cues. The *ResNet* encoder-decoder improves both *PSNR* (25.8 *dB*) and *SSIM* (0.81) by incorporating residual connections that facilitate gradient flow and feature reuse. It still struggles to capture all the fine spatial details needed to accurately reconstruct the original image from the protected images.

In contrast, the *U-Net* architecture achieves the highest reconstruction quality (*PSNR* 29.1 *dB*, *SSIM* 0.91), demonstrating a 3.3 *dB* improvement over the ResNet variant. This improvement underscores the importance of skip connections for multi-scale feature preservation, enabling the generator to reconstruct both global facial structures and subtle local textures effectively. Although the *U-Net* requires longer training time (5.2 hours), the trade-off is justified by the significant gain in reconstruction quality, making it the preferred architecture for *GAN-Invert*.

Table 3: Impact of network architecture on reconstruction quality and training efficiency. Values reported in $\mu \pm \sigma$ format, where μ is mean, and σ is standard deviation.

Architecture	PSNR	SSIM	Training Time
Standard CNN	23.1±1.02	0.76±0.03	2.3 hours
ResNet Encoder-Decoder	25.8±0.8	0.81±0.03	3.1 hours
U-Net	29.1±0.52	0.91±0.02	5.2 hours

6.2 Training Data Requirements

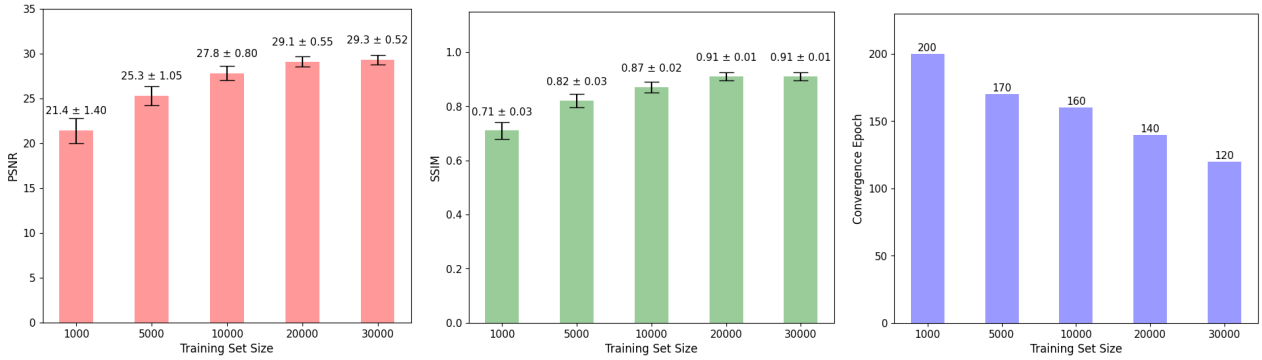
Next, we analyze how the amount of training data affects *GAN-Invert*'s reconstruction performance and convergence speed. Figure 4 demonstrates *PSNR*, *SSIM*, and the number of epochs required for convergence across different training set sizes. With only 1,000 training images, the model achieves limited reconstruction quality (*PSNR* 21.4 *dB*, *SSIM* 0.61) and requires 200 epochs to converge, highlighting the difficulty of learning the reconstruction (of the original image from its protected image) from small datasets. As the training dataset size increases to 5,000 and 10,000 images, both *PSNR* and *SSIM* improve significantly (25.3 *dB* / 0.72 and 27.8 *dB* / 0.87, respectively), while convergence occurs earlier, indicating more efficient learning.

Performance gain begins to stabilize around 20,000 images training set, where *PSNR* reaches 29.1 *dB* and *SSIM* 0.91, with convergence in 140 epochs. Increasing the dataset further to 30,000 images offers negligible improvement (*PSNR* 29.3 *dB*, *SSIM* 0.91), suggesting that *GAN-Invert* can achieve high-quality reconstruction with a moderately large, feasible dataset. These results underscore the data-dependent nature of identity reconstruction and provide insight into the resources required for adversaries for effective reconstruction attacks.

Computational Requirements: The *GAN-Invert*, used *U-Net* architecture as backbone with input image size 256×256, requires approximately 5.2 hours of training per *DLPP* defense method on a single *NVIDIA A6000 GPU*. Once trained, the *GAN-Invert* performs reconstruction in $\approx 17ms$ per protected image. For identity recognition (*ArcFace* and *CosFace*), on a 1,000 identity gallery, it takes $\approx 14.3ms$ per reconstructed image, and for a 5,000 identity gallery, it takes $\approx 21.5ms$ per reconstructed image. This level of resource is easily accessible through cloud service providers such as Amazon and Google, where a *A100 GPU* (better than *A6000*) can be rented for approximately $2.3/hour. This shows the practical feasibility of the *GAN-Invert* attack.

6.3 Loss Function Analysis

Since *GAN-Invert* is trained in an adversarial setting, the adversarial loss $\mathcal{L}_{\text{adv}}$ is a compulsory component of the objective. To assess the contribution of the auxiliary objectives, we evaluate different combinations of pixel-level reconstruction loss, feature-level identity loss, and attribute-preserving losses alongside the adversarial



(a) (b) (c)

Figure 4: Effect of training set size on reconstruction quality and convergence: (a) PSNR, (b) SSIM, (c) Convergence Epoch.

Table 4: Loss function comparison for efficient GAN-Invert. Values reported in $\mu \pm \sigma$ format, where μ is mean, and σ is standard deviation.

Loss Configuration	PSNR	SSIM	LPIPS
$\mathcal{L}_{\text{adv}}$	23.7 ± 0.8	0.70 ± 0.07	0.50 ± 0.09
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{pixel}}$	26.8 ± 0.7	0.75 ± 0.06	0.40 ± 0.07
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{feat}}$	27.2 ± 0.6	0.85 ± 0.05	0.24 ± 0.06
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{attr}}$	25.9 ± 0.7	0.81 ± 0.06	0.30 ± 0.06
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{pixel}}$ + $\mathcal{L}_{\text{feat}}$	28.6 ± 0.4	0.89 ± 0.03	0.20 ± 0.04
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{pixel}}$ + $\mathcal{L}_{\text{attr}}$	27.9 ± 0.4	0.87 ± 0.03	0.22 ± 0.03
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{feat}}$ + $\mathcal{L}_{\text{attr}}$	28.1 ± 0.3	0.88 ± 0.02	0.18 ± 0.03
$\mathcal{L}_{\text{adv}}$ + $\mathcal{L}_{\text{pixel}}$ + $\mathcal{L}_{\text{feat}}$ + $\mathcal{L}_{\text{attr}}$	**30.1 ± 0.2**	**0.91 ± 0.01**	**0.16 ± 0.02**

Table 5: Evaluation of pixel and semantic distortion in relation to a fixed task accuracy of ≈ 89% for different *DLPP* on the *CelebA-HQ* dataset (sorted by ϵ).

Method	δ	ϵ
AMT-GAN	0.21	0.23
Diff-Privacy	0.20	0.26
Feature Subtraction	0.41	0.26
PRO-Face C	0.26	0.30
Identity Transformers	0.28	0.32
VisualMixer	0.30	0.35
AVIH	0.49	0.36
CLIP2Protect	0.30	0.39
RiDDLE	0.21	0.41
Disentangled De-ID	0.22	0.48

loss. Table 4 reports reconstruction quality with *PSNR*, *SSIM* score and perceptual similarity reported with *LPIPS* score. The results highlight that adversarial loss alone does not ensure sharpness and realism in the reconstructed images of the original images from the protected images. Adding pixel-level reconstruction loss substantially improves *PSNR* and *SSIM*, enforcing low-level alignment with ground-truth images. Feature-level identity loss further enhances perceptual similarity by preserving high-level identity cues, while attribute-preserving loss supports deep learning tasks functionality. The combination of pixel-level reconstruction loss and feature-level identity loss with adversarial training yields strong perceptual quality, and the full objective achieves the best trade-off across all metrics. This confirms that adversarial loss is indispensable for training stability and realism, while the pixel loss and attribute losses act as complementary terms to jointly optimize for high fidelity, identity preservation, and perform deep learning tasks.

6.4 Empirical Validation of Bounded-Distortion

We conducted an ablation study to evaluate the practical limits of bounded-distortion (δ and ϵ) as well as the reconstruction error η in target *DLPP* methods, which also validates Theorem 3.4. The reconstruction error η is quantified using *PSNR*, *LPIPS*, and *SSIM*. For uniform comparison across all *DLPP* transformations, we fixed task accuracy to ≈ 89% by calibrating the corresponding distortion level. Table 5 summarizes the distortion characteristics (δ and ϵ limits) according to the fixed task accuracy across all *DLPP* methods on the *CelebA-HQ* dataset. The corresponding *PSNR*, *LPIPS*, and *SSIM* values are listed in the first column of Table 1. The bounded distortion δ, ϵ, and task accuracy are measured using normalized ℓ_2 distance, cosine distance in the *ArcFace* feature space, and facial attribute classification accuracy, respectively.

The results in Table 1 and Figure 3 show that our attack succeeds even when the target *DLPP* methods are operating at their highest allowable bounded distortion limits, validating Theorem 3.4. A visible pattern emerges across *DLPP* methods. For example, methods such as *VisualMixer*, *Diff-Privacy*, *AMT-GAN*, *PRO-Face C*, and *Identity Transformers* show low semantic distortion ($\epsilon \approx 0.23$–0.35), and high reconstruction quality with PSNR values between 27.5–30.2 *dB*, *SSIM* 0.85–0.89, and *LPIPS* 0.16–0.39. Their reconstructed identification accuracy is also high (*ArcFace*: 82.2–94.5%, *CosFace*: 85.7–95.0%). These methods preserve visual realism and task performance but remain vulnerable to identity recovery due to limited distortion.

In contrast, *CLIP2Protect*, *Disentangled De-ID*, and *RiDDLE* show higher semantic distortion ($\epsilon \approx 0.39$–0.55), resulting in lower reconstruction quality with *PSNR* 26.9–27.8 *dB*, *SSIM* 0.78–0.82, and *LPIPS* 0.29–0.39, along with lower reconstructed identification accuracy (*ArcFace*: 74.9–80.0%, *CosFace*: 78.4–83.3%). This trend aligns with the theoretical observation that reconstruction performance drops sharply as ϵ approaches the threshold ($\epsilon \approx 0.60$).

Overall, approximately 70% of evaluated methods fall under the strong and moderate bounded-distortion region, demonstrating that real-world *DLPP* systems prefer low distortion to preserve task accuracy. However, increasing distortion for stronger privacy leads to a severe decline in task performance (dropping below 60%), limiting practical usability. The *GAN-Invert* attack effectively exploits low-distortion defenses; however, when distortion exceeds the threshold ($\epsilon > 0.6$), protected images drift far from the original manifold, making inversion unstable. Finally, an ideal *DLPP* transformation would eliminate all identity information without altering pixel-level appearance requiring $\delta = 0$ and $\epsilon = 1$. However, as demonstrated across existing methods, complete semantic distortion is unattainable because δ and ϵ are tightly coupled, changing one inevitably affects the other.

7 Comparison

We compare the proposed *GAN-Invert* framework with existing attack models to evaluate its reconstruction quality and generalization across multiple *DLPP* transformations. The most relevant prior work, *Master Key GAN* [17], was designed specifically for the *AVIH* pixel-level manipulation method and cannot be generalized across generative, disentanglement-based, or key-driven reversible methods. On the *CelebA-HQ* dataset, *Master Key GAN* achieves an *SSIM* of 0.59, *LPIPS* of 0.45, and 70% recognition accuracy, whereas *GAN-Invert* attains 0.89 and 0.18, 94.4% respectively, with an improvement of 0.3 in *SSIM*, 0.27 in *LPIPS* and 24.4% recognition accuracy. To further validate its effectiveness, we retrained *Fantomas* [38] using the same data split as *GAN-Invert* for a fair comparison. *Fantomas* achieves an average recognition accuracy of 10.15%, substantially lower than *GAN-Invert* which achieves 87.09% recognition accuracy. Across defense categories, GAN-Invert demonstrates consistently stronger performance: 94.45% for Pixel-Level Manipulation, 85.63% for Generative Methods, 85.55% for Feature Disentanglement, and 90.3% for Key-Based Reversible Encryption, compared to 4%, 11%, 15.67%, and 6.75% obtained by *Fantomas*, respectively. For a fair evaluation, *GAN-Invert* was also trained and tested on all defense methods originally attacked by *Fantomas*. Under these identical conditions, GAN-Invert achieves an overall recognition accuracy of 93.6%, surpassing 88.3% achieved by *Fantomas*. These results confirm that *GAN-Invert* consistently achieves higher reconstruction fidelity, stronger identity recovery, and broader adaptability across *DLPP* methods, establishing it as a reliable benchmark for assessing the security of modern *DLPP* transformation methods.

8 Conclusion

In this work, we introduced *GAN-Invert*, a framework that systematically exposes the inherent vulnerabilities of bounded-distortion *DLPP* transformation methods. Through both theoretical reasoning and extensive experiments, we demonstrated that transformations designed to obscure identity while maintaining deep learning task functionality inevitably leave recoverable cues that can be exploited by reconstruction attacks. Our ablation studies highlight the factors that make reconstruction effective, architectural choices such as *U-Net*, the availability of large-scale training data, and carefully balanced loss functions combining pixel, feature, attribute, and adversarial supervision. Across ten representative *DLPP* transformation methods, across four categories visual mixing, adversarial perturbations, feature manipulations and disentangled de-identification, *GAN-Invert* consistently achieves high-fidelity reconstructions that retain critical identity cues.

Face recognition tests confirm that these reconstructed images remain identifiable, underscoring the fundamental weakness of bounded-distortion *DLPP* methods. These results suggest that reliance on such methods provides only an illusion of privacy, as sophisticated adversaries can circumvent them with relative ease. In summary, our findings show that current bounded-distortion *DLPP* methods cannot guarantee real privacy.

Future Direction: Our analysis shows that the strength of *DLPP* transformation methods mainly depends on how their latent features are structured and how much semantic distortion they add to hide identity features. *DLPP* transformation methods, such as *Disentangled De-ID* and *RiDDLE*, reduce identity leakage by swapping an individual's identity with a nearby one using password-based identity mixing. This swapping leads to high semantic distortion while keeping low pixel-level distortion. Low pixel-level distortion helps them to achieve high task accuracy, whereas high semantic distortion protects them from reconstruction attacks. Therefore, one of the future directions should be to focus on such *DLPP* methods. However, complete privacy cannot be achieved without sacrificing some task accuracy, marking the natural limit of all bounded-distortion methods.

8.1 Ethical Considerations

Datasets: We only use publicly available benchmark datasets in our experiments. All tests are done in a controlled research setting, and we do not use any private, sensitive, or personal data. This keeps our work aligned with ethical research standards and ensures reproducibility.

Responsible Research Conduct: GAN-Invert is presented only as a proof-of-concept attack to test how strong current *DLPP* transformation methods really are. We understand that reconstructing faces from protected images could be misused for re-identification or surveillance. To avoid this, we use the method strictly for research, with the goal of exposing weaknesses so that stronger and safer *DLPP* transformation methods can be developed.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Wei Ao and Vishnu Naresh Boddeti. 2025. CryptoFace: End-to-End Encrypted Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19197–19206.

[2] Zubin C Bhaidasna, Priya R Swaminarayan, and Hetal Z Bhaidasna. 2024. Implementing Deep Learning: A Novel Approach in CNNs for Face Recognition. *SSRG International Journal of Electrical and Electronics Engineering* 11, 8 (2024), 295–308.
[3] Jingyi Cao, Xiangyi Chen, Bo Liu, Ming Ding, Rong Xie, Li Song, Zhu Li, and Wenjun Zhang. 2024. Face de-identification: State-of-the-art methods and comparative studies. *arXiv preprint arXiv:2411.09863* (2024).
[4] Chen Chen, Mengyuan Sun, Xueluan Gong, Yanjiao Chen, and Qian Wang. 2025. A Survey on Facial Image Privacy Preservation in Cloud-Based Services. *arXiv preprint arXiv:2501.08665* (2025).
[5] George Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* 2, 4 (1989), 303–314.
[6] 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 (CVPR)*. 4690–4699.
[7] Fernando A Fardo, Victor H Conforto, Francisco C de Oliveira, and Paulo S Rodrigues. 2016. A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms. *arXiv preprint arXiv:1605.07116* (2016).
[8] Ian Fischer. 2020. The conditional entropy bottleneck. *Entropy* 22, 9 (2020), 999.
[9] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. 2020. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 715–731.
[10] Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. 2024. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
[11] 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 (CVPR)*. 12212–12222.
[12] 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*.
[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*. 565–578.
[14] 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).
[15] 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 (CVPR)*. 4401–4410.
[16] 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 (CVPR)*. 8110–8119.
[17] Umesh Kashyap, Sudev Kumar Padhi, and Sk Subidh Ali. 2025. Is Perceptual Encryption Secure? A Security Benchmark for Perceptual Encryption Methods. *IEEE Transactions on Artificial Intelligence* (2025).
[18] Lamyanba Laishram, Muhammad Shaheryar, Jong Taek Lee, and Soon Ki Jung. 2025. Toward a privacy-preserving face recognition system: A survey of leakages and solutions. *Comput. Surveys* 57, 6 (2025), 1–38.
[19] Minh-Ha Le and Niklas Carlsson. 2025. DiffPrivate: Facial Privacy Protection with Diffusion Models. *Proceedings on Privacy Enhancing Technologies* (2025).
[20] Dongze Li, Wei Wang, Kaihao Zhao, Jing Dong, and Tieniu Tan. 2023. RiDDLE: Reversible and Diversified De-identification with Latent Encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[21] Qiuling Li, Yan Zhang, Jie Ren, Qinghua Li, and Yisen Zhang. 2024. You Can Use But Cannot Recognize: Preserving Visual Privacy in Deep Neural Networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.
[22] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5447–5456.
[23] Yuxi Mi et al. 2024. Privacy-Preserving Face Recognition Using Trainable Feature Subtraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12781–12790.
[24] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. 2021. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*. 669–680.
[25] Dennis Rochau, Robin Chan, and Hanno Gottschalk. 2024. New advances in universal approximation with neural networks of minimal width. *arXiv preprint arXiv:2411.08735* (2024).
[26] Marcos Rodrigo, Carlos Cuevas, and Narciso García. 2024. Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks. *Scientific reports* 14, 1 (2024), 21392.
[27] Felipe Romero-Moreno. 2021. AI facial recognition and biometric detection: balancing consumer rights and corporate interests. In *International Carnahan Conference on Security Technology (ICCST)*. 1–5.
[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 234–241.
[29] Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. 2025. Enhancing Facial Privacy Protection via Weakening Diffusion Purification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8235–8244.
[30] Sefik Serengil and Alper Ozpinar. 2024. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Journal of Information Technologies* 17, 2 (2024), 95–107.
[31] Novendra Setyawan, Chi-Chia Sun, Mao-Hsiu Hsu, Wen-Kai Kuo, and Jun-Wei Hsieh. 2025. FaceLiVT: Face Recognition using Linear Vision Transformer with Structural Reparameterization For Mobile Device. *arXiv preprint arXiv:2506.10361* (2025).
[32] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 2023. CLIP2Protect: Protecting Facial Privacy Using Text-Guided Makeup via Adversarial Latent Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20595–20605.
[33] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya. 2019. Pixel-based image encryption without key management for privacy-preserving deep neural networks. *IEEE Access* 7 (2019), 177844–177855.
[34] Warit Sirichotedumrong, Takahiro Maekawa, Yuma Kinoshita, and Hitoshi Kiya. 2019. Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain. In *IEEE International Conference on Image Processing (ICIP)*. 674–678.
[35] Theresa Stadler, Bogdan Kulynych, Michael C Gastpar, Nicolas Papernot, and Carmela Troncoso. 2024. The fundamental limits of least-privilege learning. *arXiv preprint arXiv:2402.12235* (2024).
[36] Zhigang Su, Dawei Zhou, Nannan Wang, Decheng Liu, Zhen Wang, and Xinbo Gao. 2023. Hiding Visual Information via Obfuscating Adversarial Perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4356–4365.
[37] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep Learning Face Representation from Predicting 10,000+ Classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1891–1898.
[38] Julian Todt, Simon Hanisch, and Thorsten Strufe. 2022. Fant\^ omas: Understanding Face Anonymization Reversibility. *arXiv preprint arXiv:2210.10651* (2022).
[39] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5265–5274.
[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.
[41] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3676–3684.
[42] Lin Yuan, Wu Chen, Xiao Pu, Yan Zhang, Hongbo Li, Yushu Zhang, Xinbo Gao, and Touradj Ebrahimi. 2024. PRO-face C: Privacy-preserving recognition of obfuscated face via feature compensation. *IEEE Transactions on Information Forensics and Security* 19 (2024), 4930–4944.
[43] Lin Yuan, Kai Liang, Xiong Li, Tao Wu, Nannan Wang, and Xinbo Gao. 2025. iFA-DIT: Invertible face anonymization via disentangled identity transform. *Pattern Recognition* (2025), 111807.
[44] Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe Jin, and Xuejun Li. 2024. Validating Privacy-Preserving Face Recognition under a Minimum Assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.
[46] Yaoyao Zhong and Weihong Deng. 2021. Face Transformer for Recognition. *arXiv preprint arXiv:2103.14803* (2021).

A Visual Interpretation of PSNR, SSIM, and LPIPS Metrics

This section provides an intuitive analysis of how *PSNR*, *SSIM*, and *LPIPS* reflect changes in visual quality by examining the reference image and a range of distorted variants shown in Figure 5. As distortion increases, using various distortion methods such as Gaussian

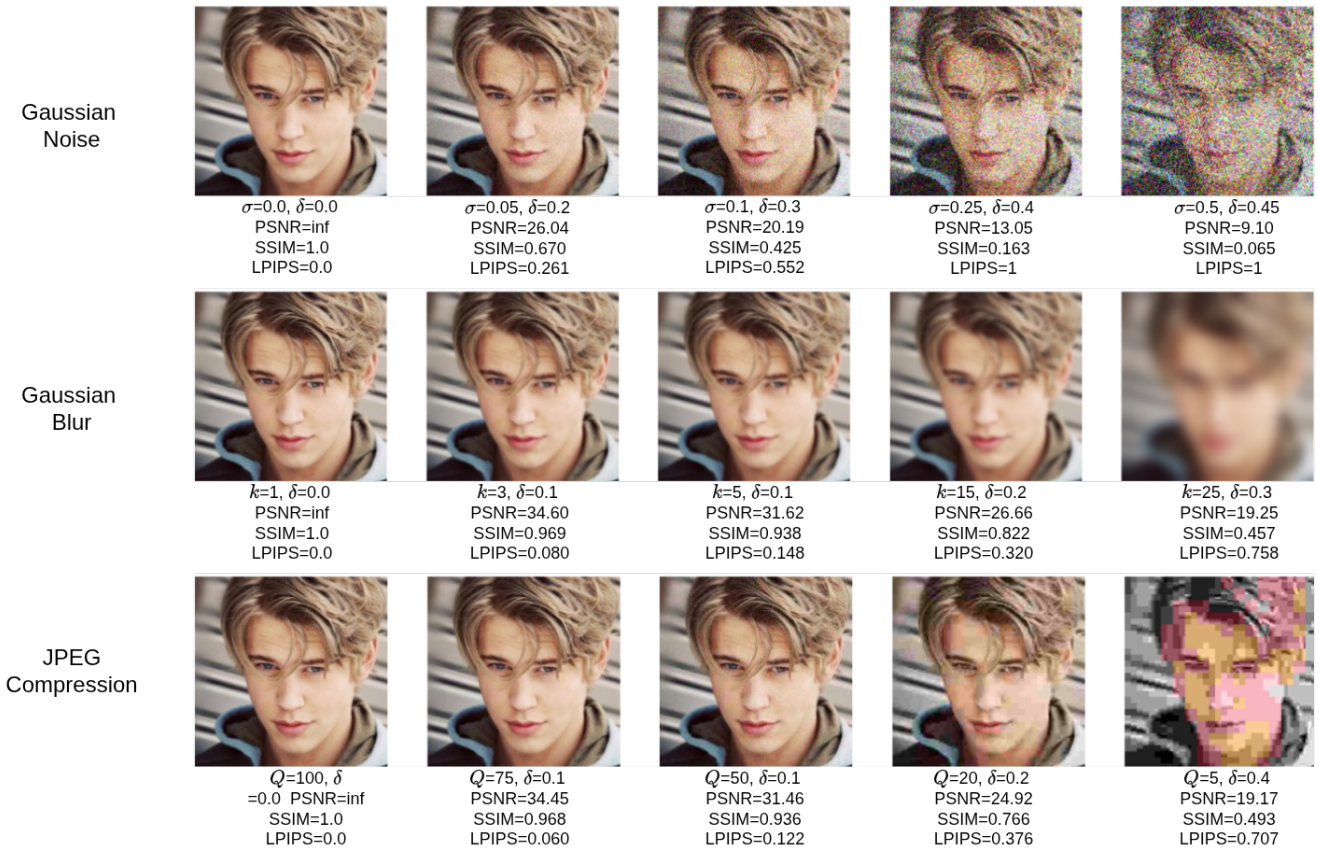


Figure 5: Visual example of PSNR, SSIM and LPIPS values on different distortion levels.

Table 6: Performance comparison of protection and reconstruction across ArcFace and CosFace models ($\mu \pm \sigma$).

Method	ArcFace Protected	ArcFace Reconstructed	CosFace Protected	CosFace Reconstructed
VisualMixer	16.5 ± 1.15	93.8 ± 0.89	19.2 ± 1.23	94.2 ± 0.70
AVIH	17.8 ± 1.15	94.4 ± 0.61	18.5 ± 1.17	94.7 ± 0.61
CLIP2Protect	19.5 ± 1.17	78.4 ± 1.42	21.1 ± 1.20	79.2 ± 1.31
Diff-Privacy	20.1 ± 1.33	82.2 ± 1.61	20.5 ± 1.34	85.7 ± 1.54
AMT-GAN	22.4 ± 1.49	89.6 ± 1.19	24.0 ± 1.53	92.0 ± 1.13
Feature Subtraction	20.8 ± 1.29	91.2 ± 0.90	20.3 ± 1.28	92.7 ± 0.85
Disentangled De-ID	18.1 ± 1.54	74.9 ± 1.69	21.8 ± 1.59	78.4 ± 1.64
PRO-Face C	22.3 ± 1.22	94.5 ± 0.43	23.9 ± 1.24	95.0 ± 0.35
RiDDLE	18.8 ± 1.38	80.0 ± 1.42	19.4 ± 1.38	83.3 ± 1.33
Identity Transformers	20.4 ± 1.18	91.9 ± 0.65	23.5 ± 1.22	92.7 ± 0.58

noise with standard deviation values $\sigma = (0.0, 0.05, 0.1, 0.25, 0.5)$, Gaussian blur with kernel sizes $k = (1, 3, 5, 15, 25)$, and JPEG compression with quality factors $Q = (100, 75, 50, 20, 5)$, the numerical values of the *PSNR*, *SSIM*, and *LPIPS* metrics change, allowing them to closely align with human perception. When *PSNR* drops below 20 dB, the images appear heavily distorted and structural details become difficult to recognize. In the 20–30 dB range, the overall content remains visible, but artifacts, noise, and blur are clearly noticeable. Once the *PSNR* exceeds 30 dB, the visual quality becomes substantially better, and distortion has no effect on the original images. For very high *PSNR* values, typically above 40 dB, the images look almost identical to the original image.

The *SSIM* demonstrates a similar pattern. Images with *SSIM* above 0.90 maintain strong structural consistency and appear visually stable, while values below 0.70 correspond to noticeable structural loss. *LPIPS* exhibits the opposite trend, values below 0.20 indicate close perceptual similarity, while values 0.20–0.40 reflect visible changes and values above 0.50 capture cases where perceptual distortion is strong and clearly noticeable. These observations highlight how each metric responds to common forms of

distortion and how its numerical scales map to visible changes in image quality.

Building upon the visual trends observed in the examples, we summarize practical ranges for interpreting reconstruction quality using *PSNR*, *SSIM*, and *LPIPS*. High-quality reconstructions generally exhibit *PSNR* above 30 dB, *SSIM* above 0.90, and *LPIPS* below 0.20, reflecting strong consistency in both structure and perceptual appearance. Very high *PSNR* values exceeding 40 dB typically correspond to reconstructions that are visually indistinguishable from the original image.

Moderate quality reconstructions fall within *PSNR* = 20–30 dB, *SSIM* = 0.70–0.90, and *LPIPS* = 0.20–0.40, where the images remain clear and recognizable with visible artifacts. Low-quality reconstructions lie outside these boundaries, particularly when *PSNR* drops below 20 dB, *SSIM* falls under 0.70, or *LPIPS* rises above 0.50. These conditions produce the heavily distorted outputs visible in the examples, reflecting significant perceptual and structural changes. These ranges provide a straightforward way to interpret reconstruction metrics and demonstrate how numerical values correspond to actual visual quality across common types of distortion.

B Standard Deviation Values of Results

Table 6 shows the standard deviation corresponding to all three split test datasets of Figure 3 results.

C Theoretical Analysis of Privacy Leakage

We analyzed the *CelebA-HQ* containing 30, 000 images corresponding to 6, 217 unique identities. Using Shannon entropy as a measure of uncertainty, we first computed the identity entropy $H(ID)$, which quantifies the uncertainty of predicting an individual's identity without any additional information:

$$H(ID) = -\sum_{i=1}^{K} P(id_i) \log_2(P(id_i))$$

Where $K = 6,217$ is the number of unique identities, and $P(id_i)$ is the probability of each identity in the dataset, calculated as the fraction of images belonging to that identity. For this dataset, the computed entropy is:

$$H(ID) = 12.4723 \text{ bits}$$

This relatively high value indicates a considerable level of uncertainty in predicting identities when no other information is available. Next, we calculated the conditional entropy $H(ID \mid Attr)$, which measures the remaining uncertainty about the identity when attributes of the images are known:

$$H(ID \mid Attr) = \sum_{a \in Attr} P(a) \sum_{i=1}^{K} P(id_i \mid a) \log_2(P(id_i \mid a))$$

Here, $Attr$ represents all unique attribute combinations in the dataset, and $P(id_i \mid a)$ is the probability of identity i given attribute combination a. Where CelebA-HQ dataset has 40 attributes and all have binary values. Therefore a total of 2^{40} combinations of attributes are possible. The computed value is:

$$H(ID \mid Attr) = 0.5613 \text{ bits}$$

This substantial reduction in entropy demonstrates that knowing the functional attributes of an image greatly reduces uncertainty about the corresponding identity. Finally, the privacy leakage, defined as the reduction in identity uncertainty due to attribute knowledge, was calculated as:

$$\text{Leakage} = H(ID) - H(ID \mid Attr) = 11.9119 \text{ bits}$$

This corresponds to a privacy leakage percentage of 95.50%, indicating that functional attributes carry nearly all the information necessary to re-identify individuals. These results highlight a significant privacy risk, even sharing attribute information could allow an adversary to accurately infer identities, underlining the importance of privacy-preserving methods in facial datasets.