Visible Yet Unrecognizable: Frequency-Selective Facial Privacy via Attention

Atul Kumar[1] and Akshay Agarwal[1]

Trustworthy BiometraVision Lab, IISER Bhopal, India
{atulk23, akagarwal}@iiserb.ac.in

Abstract. Facial privacy has become a critical concern as unauthorized large-scale image scraping enables malicious actors to collect personal face images without consent and exploit them for training unauthorized recognition systems. Protecting facial privacy requires modifying a face image such that automated recognition systems fail to identify the individual, while the visual appearance remains unchanged to human observers, ensuring the image retains its social utility. However, existing methods fail to satisfy both requirements; they either irreversibly degrade visual quality, rendering images impractical, or alter facial identity to such an extent that the protected image bears no resemblance to the original subject. To address this fundamental privacy-utility tradeoff, we propose MIRAGE *(Multi-scale Identity Removal via Attention-Guided Encoding)*, a frequency-selective facial privacy framework grounded in the observation that identity-discriminative features reside predominantly in low-frequency image components, while visual appearance is encoded in high-frequency components. We further introduce *FAC (Frequency-Aware Consistency) Loss*, which jointly enforces identity separation and visual fidelity during training. A comprehensive evaluation across three benchmark datasets and nine state-of-the-art (SOTA) deep face recognition (DFR) models demonstrates that MIRAGE achieves robust facial privacy protection while preserving visual appearance, successfully bridging the privacy-utility gap left unresolved by existing methods. The code is available at: https://github.com/atulkr05/MIRAGE.git.

Keywords: Facial Privacy · Frequency-Selective Protection · Face Recognition · Identity Protection · Visual Fidelity

1 Introduction

The rapid growth of deep learning and large-scale image datasets [8, 12, 37] has raised serious concerns about facial privacy for individuals whose images are publicly accessible online [3]. Personal face images are routinely scraped from social media without consent and exploited to train unauthorized models for surveillance, identity fraud, and biometric extraction [4, 11, 19, 34]. A prominent real-world example is Clearview AI [10], which scraped billions of facial images from the internet to build a commercial face recognition system without the knowledge or consent of the individuals involved. To protect facial privacy,

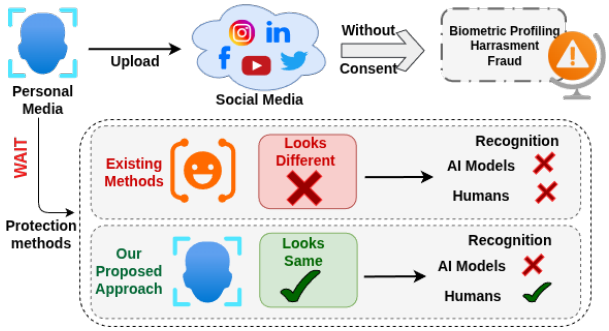

Fig. 1: A new perspective for privacy preservation.

researchers have developed various methods for facial privacy protection. *We define facial privacy protection as modifying a face image such that it fools automated AI recognition systems while keeping the visual appearance unchanged to human observers*, ensuring the image retains its social utility for everyday use. To the best of our knowledge, no existing method satisfies both requirements simultaneously. Traditional approaches, such as blurring and pixelation [33, 42], as well as deliberate facial occlusion using objects like hands or masks [21], severely distort the visual appearance, rendering the resulting images unsuitable for practical use. More recent generative methods [2, 9, 24] synthesize an entirely new facial identity to replace the original identity, which fools recognition systems but produces images that look entirely different from the original subject and cannot be shared on social platforms [29, 44, 46]. As illustrated in Fig. 1, this is the *central research gap* our work addresses: the lack of a method that can simultaneously fool AI recognition systems while preserving the natural visual appearance of the face for human observers.

Key Insight: Frequency-Selective Identity Decomposition
Our approach is motivated by a fundamental observation about how identity information is encoded within a facial image. Prior studies reveal that face recognition models exhibit differential sensitivity to various frequency components [14, 15], motivating a foundational question: *What frequency components play a critical role in face recognition?* To investigate this, we decompose facial images using the Discrete Wavelet Transform (DWT) into a low-frequency approximation subband (LF: LL) and three high-frequency detail subbands (HF:(LH, HL, HH)). As illustrated in Fig. 2, we evaluate recognition performance across nine SOTA DFR models using different frequency configurations for the gallery and probe sets. The results reveal two clear findings: *LF_Gallery* and *LF_Probe* achieve recognition accuracy remarkably close to *All_freq*, confirming that the LL subband alone carries the identity-discriminative information that recognition systems rely upon. In contrast, *HF_Gallery* and *HF_Probe* produce significantly lower accuracy across all models, confirming that high-frequency components carry visual appearance attributes.

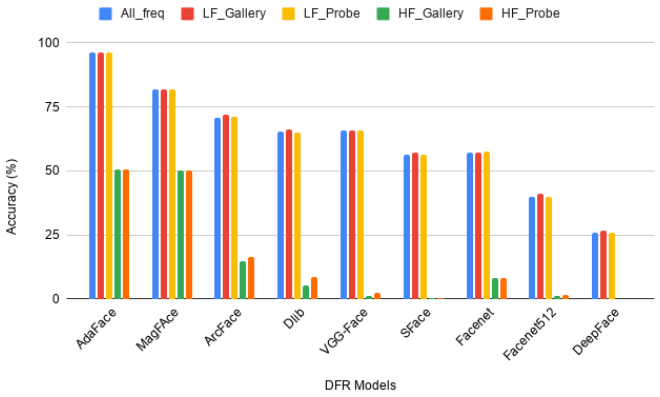

Fig. 2: Recognition accuracy across nine DFR models under different frequency configurations. *All_freq*: all frequency components used for both gallery and probe; *LF_Gallery*: only the LL (low-frequency) component used for the gallery; *LF_Probe*: only the LL component used for the probe; *HF_Gallery*: only high-frequency components (LH, HL, HH) used for the gallery; *HF_Probe*: only high-frequency components used for the probe.

This core observation motivates our frequency-selective framework, MIRAGE (Multi-scale Identity Removal via Attention-Guided Encoding). Unlike conventional approaches that abruptly modify facial images [6, 16], MIRAGE applies targeted, attention-driven manipulations selectively in the frequency domain. The framework first decomposes the input face image into low-frequency (LL) and high-frequency (LH, HL, HH) subbands via a single-level DWT. Two distinct operations are then applied: (i) a cross-attention module perturbs the identity-rich LL subband by conditioning it on a randomly sampled protected identity embedding, disrupting recognition; and (ii) self-attention modules refine the high-frequency subbands to preserve fine-grained visual details essential for natural appearance and social utility. The modified subbands are reconstructed via IDWT and passed through a lightweight refinement network to produce a visually coherent protected face. MIRAGE supports two practical protection modes: (1) *proactive protection*, where a user uploads a protected version of their images as the gallery set; and (2) *reactive protection*, where a user whose images are *already publicly available* protects their probe set to prevent unauthorized recognition. We validate MIRAGE on three benchmark datasets: CelebA-HQ, 105-Pins[1], and LFW, across nine SOTA DFR models under both white-box and black-box settings. The main contributions of this work are:

– We propose *MIRAGE*, a frequency-selective facial privacy protection framework that applies cross-attention for identity perturbation in low-frequency subbands and self-attention for visual appearance preservation in high-frequency

[1] `https://www.kaggle.com/datasets/hereisburak/pins-face-recognition`

subbands, generating high-fidelity protected faces that robustly disrupt unauthorized face recognition.
- *FAC (Frequency-Aware Consistency) Loss* is introduced as a novel composite loss function comprising a frequency-aware triplet loss for multi-scale identity separation and a frequency-consistency focal loss with SSIM-based adaptive weighting for visual texture preservation.
- A dual-mode privacy pipeline is presented that supports both proactive gallery protection and reactive probe protection, providing end-to-end facial privacy coverage across diverse real-world threat scenarios.
- We provide comprehensive empirical validation across three datasets and nine SOTA DFR models in gender-agnostic, seen-gender, and unseen-gender settings, under both white-box and black-box threat scenarios, demonstrating the robustness and practical utility of our proposed approach.

2 Related Work

Research in facial privacy protection has progressed along several interconnected trajectories, each attempting to address the fundamental challenge of balancing privacy and visual utility. We organize this review across three directions: traditional and adversarial protection methods, generative and diffusion-based approaches, and frequency-domain methods.

Traditional and Adversarial Protection Methods: Early facial privacy methods rely on spatial techniques such as blurring and pixelation [25, 33, 36], which produce visually degraded outputs with prominent artifacts, severely limiting practical utility. To overcome these limitations, adversarial perturbation-based methods add carefully crafted imperceptible noise to face images to prevent the extraction of identity and facial attributes [5]. Majumdar et al. [31] extend this by proposing partial face tampering through the replacement or morphing of specific facial regions, demonstrating that localized modifications are sufficient to disrupt recognition. Agarwal et al. [1] further highlight the transferability of adversarial perturbations across different network architectures, revealing fundamental vulnerabilities in recognition systems that can be exploited to protect privacy. Kuang et al. [20] propose an attention-based approach that disrupts recognition by distracting both intrinsic and extrinsic attention mechanisms, showing that attention manipulation can effectively conceal identity while maintaining reasonable visual coherence. Sun et al. [41] take a complementary direction by using diffusion-based adversarial makeup transfer, demonstrating that appearance-level modifications guided by adversarial objectives can simultaneously protect identity while maintaining perceptual plausibility.

Generative and Diffusion-Based Protection: Generative models have emerged as a powerful tool for protecting facial privacy. Li et al. [27] develop an identity-obfuscation framework that suppresses visual attributes while maintaining discriminability for downstream recognition tasks. Zhai et al. [45] design A3GAN, a generative adversarial network that adaptively conceals identity features through attribute-aware disentanglement, enabling selective suppression of identity cues.

Barattin et al. [2] propose identity removal via latent code optimization in StyleGAN's latent space, demonstrating that careful manipulation of latent codes can eliminate identity information while retaining non-identity attributes. Beyond identity removal, synthetic face data itself has been explored as a privacy-preserving substitute for real images in training biometric classifiers [22]. He et al. [9] demonstrate diffusion-based identity suppression using multi-scale inversion embeddings to remove identity cues and synthesize protected faces without requiring facial landmarks or masks. Salar et al. [38] address the practical vulnerability of adversarially protected images to purification attacks by proposing to weaken diffusion-based purification, an important consideration for real-world deployment. Most recently, Kung et al. [24] introduce a simple yet effective approach that performs identity concealment directly in the latent space of pre-trained generative models, achieving competitive privacy protection with minimal architectural complexity. While these methods successfully disrupt recognition, they commonly alter facial appearance so significantly that the protected images lose visual resemblance to the original subject, limiting their practical social utility — the key limitation our work directly addresses.

Frequency-Domain Approaches: The frequency domain has emerged as an important frontier for both attack and defense in face recognition. Prior work establishes that face recognition models exhibit differential sensitivity to different frequency components [14, 15]. Frequency-based adversarial attacks exhibit superior imperceptibility compared to their spatial-domain counterparts [30], as frequency-domain perturbations are harder to detect visually. These findings collectively provide the theoretical foundation for frequency-selective privacy protection. Unlike prior methods that operate in spatial or latent domains, or those that apply attention manipulation without explicit frequency control, MIRAGE directly decomposes facial images into frequency subbands and applies cross-attention for identity manipulation in low-frequency bands and self-attention for appearance preservation in high-frequency bands. It enables principled facial privacy protection that simultaneously conceals identity from recognition systems while preserving visual appearance for human observers.

3 Proposed Methodology

3.1 Problem Formulation

Given an input image $x \in \mathbb{R}^{3\times H\times W}$ with identity y_i and non-identity attributes a, we seek a generator G that produces a protected image $x' = G(x, z)$, where z is a randomly sampled target identity embedding. To achieve effective facial privacy protection while maintaining visual utility, G must satisfy two constraints:

1. **Identity Dissimilarity:** The protected image x' must be biometrically distinct from the original image x. Let $\mathcal{E}_i$ denote a pre-trained face recognition encoder. We require the identity embedding $e'_i = \mathcal{E}_i(x')$ to be sufficiently

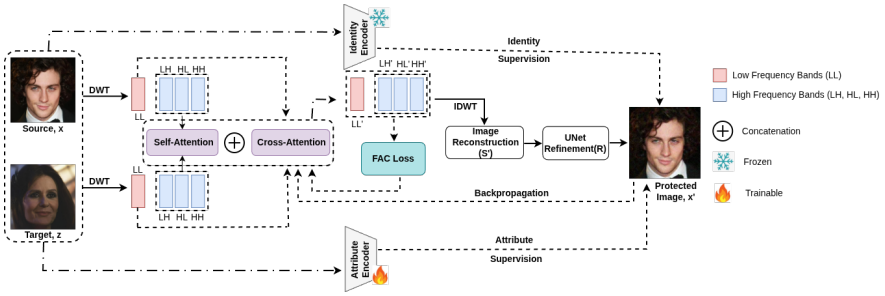

Fig. 3: Proposed architecture of MIRAGE with integration of self and cross-attention modules leveraging low and high-frequency components to preserve privacy.

distant from the original embedding $e_i = \mathcal{E}_i(x)$:

$$d_{\text{cosine}}(e_i, e_i') = 1 - \frac{e_i \cdot e_i'}{\|e_i\| \, \|e_i'\|} \geq \tau_i, \tag{1}$$

where $\tau_i = 0.85$ is the target dissimilarity threshold.

2. **Perceptual Quality:** The protected image x' must remain photorealistic and visually coherent to human observers. We enforce this using standard image quality metrics, targeting $\text{SSIM}(x, x') \geq 0.75$ and $\text{LPIPS}(x, x') \leq 0.20$, where SSIM denotes Structural Similarity Index and LPIPS [47] denotes Learned Perceptual Image Patch Similarity.

3.2 MIRAGE Framework Architecture

MIRAGE processes each input image through three sequential stages: (1) Dual-Feature Encoding and Frequency-Domain Decomposition, (2) Attention-Guided Frequency Manipulation, and (3) Reconstruction and Refinement. Fig. 3 illustrates the complete architecture of our proposed approach.

Dual-Feature Encoding The input image x is first processed by two parallel encoders to disentangle its identity and non-identity features:

- *Identity Encoder ($\mathcal{E}_i$):* A pre-trained InceptionResNetV1 model (trained on VGGFace2), kept *frozen* during training, extracts a 512-dimensional identity embedding $e_i = \mathcal{E}_i(x)$.
- *Attribute Encoder ($\mathcal{E}_a$):* A *trainable* lightweight CNN with residual blocks extracts a 512-dimensional attribute embedding $e_a = \mathcal{E}_a(x)$, capturing non-identity visual features.

Frequency-Domain Decomposition Simultaneously, the input image x is decomposed using a single-level 2D Discrete Wavelet Transform (DWT) with a Haar wavelet into four frequency subbands:

$$\{x^{LL}, x^{LH}, x^{HL}, x^{HH}\} = \text{DWT}(x), \tag{2}$$

where x^{LL} is the low-frequency approximation subband carrying identity-discriminative information, and $\{x^{LH}, x^{HL}, x^{HH}\}$ are the high-frequency detail subbands encoding visual appearance attributes.

Attention-Guided Frequency Manipulation This stage forms the core of MIRAGE, in which identity is selectively manipulated by processing low- and high-frequency subbands with two distinct attention mechanisms.
(1) Low-Frequency Identity Perturbation. The identity-rich x^{LL} subband is processed by an AdaIN-based cross-attention module [13], where the source low-frequency features serve as queries, and the target identity embedding z provides the keys and values. This cross-attention mechanism allows the low-frequency spatial features of the source image to selectively attend to the target identity embedding, while AdaIN simultaneously aligns the feature statistics of x^{LL} with the style distribution of z:

$$x'^{LL} = \text{AdaIN-CrossAttn}(x^{LL}, z). \tag{3}$$

This causes recognition systems to extract an identity embedding corresponding to the protected target identity rather than the original subject.
(2) High-Frequency Appearance Preservation. Each high-frequency subband x^{HF}, where $HF \in \{LH, HL, HH\}$, is independently processed by a CBAM [43] that applies self-attention purely within the subband itself with no external identity conditioning to selectively refine and preserve fine visual details:

$$x'^{HF} = \text{CBAM}(x^{HF}). \tag{4}$$

Channel attention recalibrates feature responses across frequency channels, while spatial attention focuses preservation on perceptually important regions. Since CBAM operates solely on the subband's own features, without any target identity influence, it ensures that high-frequency appearance details remain natural and consistent with the original subject, free from identity interference.

Reconstruction and Refinement The manipulated frequency subbands are recombined via an IDWT to produce a coarse protected image $\tilde{x}$:

$$\tilde{x} = \text{IDWT}\big(x'^{LL}, x'^{LH}, x'^{HL}, x'^{HH}\big). \tag{5}$$

To enable quality-aware refinement, $\tilde{x}$ is concatenated with the original image x to form a 6-channel input, which is passed through a U-Net refinement network $\mathcal{R}$ that synthesizes photorealistic details and removes any wavelet-induced artifacts:

$$x_{\text{refined}} = \mathcal{R}([\tilde{x}; x]), \tag{6}$$

where $[\cdot\,;\,\cdot]$ denotes channel-wise concatenation. Finally, a learned spatial face mask $M \in [0,1]^{H \times W}$, predicted by a lightweight mask network, performs spatially adaptive blending that focuses identity perturbation on facial regions while preserving the background:

$$x' = M \odot x_{\text{refined}} + (1 - M) \odot x, \tag{7}$$

where $\odot$ denotes element-wise multiplication. The resulting protected image x' is visually coherent and natural to human observers, yet systematically causes recognition systems to misclassify the subject's identity.

3.3 Loss Functions

The generator G is trained using a composite loss $\mathcal{L}_{\text{total}}$, comprising our proposed *Frequency-Aware Consistency (FAC) Loss* and four auxiliary losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{FAC}}\mathcal{L}_{\text{FAC}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \quad (8)$$

where $\lambda_{(\cdot)}$ are scalar weighting hyperparameters balancing the contribution of each loss term.

FAC Loss. The core training objective $\mathcal{L}_{\text{FAC}}$ enforces robust identity separation and high-fidelity visual consistency across frequency subbands through two complementary components:

$$\mathcal{L}_{\text{FAC}} = \mathcal{L}_{\text{f-margin}} + \mathcal{L}_{\text{f-focal}}. \quad (9)$$

(1) Frequency-Aware Margin Loss ($\mathcal{L}_{f\text{-}margin}$). Unlike triplet-based losses that require negative sample mining, this loss directly enforces identity separation in each frequency subband by penalizing cosine distances between the original and protected identity embeddings that fall below a band-specific margin m_b:

$$\mathcal{L}_{\text{f-margin}} = \sum_{b\in\{LL,LH,HL,HH\}} w_b \cdot \text{ReLU}\Big(m_b - d_{\cos}\big(e_i^b,\, e_i'^b\big)\Big)^2, \quad (10)$$

where $e_i^b = \mathcal{E}_i(x^b)$ and $e_i'^b = \mathcal{E}_i(x'^b)$ denote the identity embeddings extracted from frequency subband b of the original and protected images respectively, $d_{\cos}(\cdot,\cdot)$ is the cosine distance, $\text{ReLU}(x) = \max(0, x)$, and $w_b \in \{1.5, 0.5, 0.5, 0.3\}$ and $m_b \in \{0.7, 0.3, 0.3, 0.2\}$ are band-specific weights and margins assigned to $\{LL, LH, HL, HH\}$ respectively, with higher values assigned to the identity-critical LL subband.

(2) Frequency-Consistency Focal Loss ($\mathcal{L}_{f\text{-}focal}$). This loss preserves high-frequency visual texture by concentrating the reconstruction penalty on regions that are hardest to reconstruct, identified by low SSIM scores:

$$\mathcal{L}_{\text{f-focal}} = \sum_{b\in\{LH,HL,HH\}} \Big(1 - \text{SSIM}(x^b, x'^b)\Big)^{\gamma} \cdot \|x^b - x'^b\|_2^2, \quad \gamma = 2.0, \quad (11)$$

where $\|\cdot\|_2$ denotes the L2 norm.

Auxiliary Losses. Four standard auxiliary losses complement FAC Loss to preserve visual attributes and photorealism.

(1) Attribute Loss ($\mathcal{L}_{attr}$): Preserves non-identity attributes from the original image to the protected output:

$$\mathcal{L}_{\text{attr}} = \|e_a - e'_a\|_2^2, \quad (12)$$

where $e'_a = \mathcal{E}_a(x')$ is the attribute embedding of the protected image.
(2) Reconstruction Loss ($\mathcal{L}_{recon}$): Enforces pixel-level similarity between input and output to preserve overall structure and background, using a combination of L1 and L2 norms:

$$\mathcal{L}_{\text{recon}} = \|x - x'\|_1 + 0.5\,\|x - x'\|_2^2, \tag{13}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L1 and L2 norms respectively.
(3) Perceptual Loss ($\mathcal{L}_{perc}$): Maintains high-level perceptual similarity by minimizing the distance between VGG-19 [40] feature representations at multiple layers:

$$\mathcal{L}_{\text{perc}} = \sum_l \|\phi^l_{\text{VGG}}(x) - \phi^l_{\text{VGG}}(x')\|_2^2, \tag{14}$$

where ϕ^l_{VGG} denotes the VGG-19 feature map at layer l.
(4) Adversarial Loss ($\mathcal{L}_{adv}$): Encourages the generator to produce photorealistic outputs that are indistinguishable from real images by a PatchGAN discriminator D:

$$\mathcal{L}_{\text{adv}} = -\log D(x'). \tag{15}$$

4 Experimental Setup

4.1 Datasets and Evaluation Models

We train MIRAGE on the *CelebA-HQ* dataset [26], which comprises 30,000 high-quality images, of which 5,000 are used for testing. We evaluate the performance of our framework and all baselines on three datasets: the CelebA-HQ (test set), the 105-pin dataset, and the LFW dataset, having a subset of 400 identities, except the 105-pin dataset, which has 105 identities. To measure the impact of our privacy-preserving methods on face recognition, we use nine SOTA models: AdaFace [17], MagFace [32], ArcFace [7], Dlib [18], VGG-Face [35], FaceNet [39], FaceNet-512 [39], SphereFace [28], and DeepFace [35].

4.2 Evaluation Protocol

This paper follows the standard biometric protocol, in which one image per identity is used as the gallery set and three images per identity as the probe set. We evaluate under two protection scenarios:

- **Proactive protection:** The user uploads protected versions of their images as the gallery set. We protect the gallery set in this scenario and keep the probe set unprotected.
- **Reactive protection:** The user gives his/her image as a testing or probe set. The gallery images remain clean (unprotected), but all probe images are protected before recognition.

For implementation details, please refer to the supplementary material.

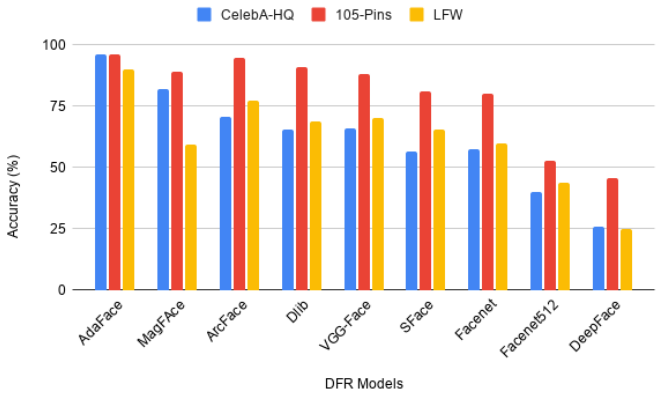

Fig. 4: Performance visualization of baseline clean accuracy on three different datasets over nine SOTA DFR models.

Table 1: Performance (%) for MIRAGE under proactive (gallery) protection across gender-agnostic, seen-gender, and unseen-gender settings, evaluated for three different datasets. The green color shows the best privacy-preserving model across different settings. The lower the accuracy, the better the privacy.

	Gender-Agnostic			Seen Gender			Unseen Gender		
Model	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	70.24	71.63	65.42	69.72	74.71	65.02	71.54	67.53	70.35
MagFace	57.87	61.75	44.13	58.11	59.28	44.86	56.56	57.77	46.32
ArcFace	45.93	49.31	41.06	46.02	50.47	44.35	42.88	44.26	39.87
Dlib	39.36	31.40	36.64	42.87	35.16	36.88	37.05	41.24	33.93
VGG-Face	38.31	29.68	38.93	39.91	36.16	35.74	37.80	39.61	34.79
SFace	37.65	38.69	34.09	37.96	41.33	33.47	36.24	37.03	33.26
Facenet	45.04	50.28	43.89	46.66	51.43	44.36	45.91	47.12	43.25
Facenet512	22.87	25.39	20.04	28.06	28.63	20.57	26.81	27.47	25.78
DeepFace	24.64	27.19	23.05	25.66	28.20	24.45	23.47	26.72	22.23

5 Experimental Results and Analysis

In this section, we validate the facial privacy protection strength of MIRAGE under both proactive and reactive settings, isolate the impact of pure frequency-domain perturbations using a DWT sub-band swapper, and present an ablation study validating the contribution of each component. Fig. 4 shows the baseline clean accuracy of nine SOTA DFR models across three datasets. To achieve privacy protection, we aim to reduce this baseline accuracy; *a greater reduction in accuracy indicates a stronger level of privacy protection.*

5.1 Effectiveness of our proposed MIRAGE approach

We evaluate MIRAGE's facial privacy protection under both proactive and reactive protocols across nine SOTA face recognition models and three datasets. Tables 1 and 2 present results for proactive and reactive settings, respectively. In the

Table 2: Performance (%) for MIRAGE under reactive (probe) protection across different gender settings, evaluated for three different datasets. The green color shows the best privacy-preserving model across different settings. The lower the accuracy, the better the privacy.

	Gender-Agnostic			Seen Gender			Unseen Gender		
Model	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	68.56	69.48	65.63	68.72	70.75	67.73	72.00	71.38	69.48
MagFace	56.67	64.21	44.89	60.59	58.18	45.45	58.36	59.04	45.68
ArcFace	46.02	49.60	41.05	46.91	49.86	45.52	43.95	43.35	38.41
Dlib	40.30	30.29	34.96	43.37	35.21	35.57	37.32	39.78	32.86
VGG-Face	39.81	31.10	37.47	40.12	37.82	35.24	38.78	38.20	34.43
SFace	38.88	36.77	32.54	37.42	41.49	34.02	35.04	36.16	32.95
Facenet	46.85	49.20	45.98	47.78	50.06	44.66	46.98	47.75	43.04
Facenet512	23.09	26.40	19.38	28.40	29.63	20.23	26.69	27.34	26.93
DeepFace	24.78	28.36	24.15	26.39	28.29	23.33	24.24	27.67	21.31

Table 3: Performance (%) of sub-band swapping across different datasets and models on proactive setting. The green color shows the best privacy-preserving model across different settings.

	Gender-Agnostic			Seen Gender			Unseen Gender		
Model	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	95.47	95.90	86.17	95.79	95.30	86.30	95.24	95.47	86.51
MagFace	77.18	85.25	58.29	77.54	84.59	57.84	76.67	82.89	58.32
ArcFace	55.73	40.48	54.33	57.15	46.64	56.83	52.64	38.36	51.17
Dlib	50.13	33.91	44.00	53.89	40.99	43.00	48.45	24.32	41.33
VGG-Face	49.87	34.60	46.50	50.96	42.76	48.42	48.20	27.05	44.42
SFace	44.10	13.15	49.50	46.28	21.91	49.08	42.59	13.01	48.50
Facenet	27.36	17.65	24.67	29.96	21.91	25.50	26.11	14.38	22.42
Facenet512	9.62	3.46	10.67	12.38	4.24	12.33	9.96	4.11	10.33
DeepFace	25.02	28.72	24.25	23.93	30.74	23.75	24.77	30.82	23.42

proactive setting, MIRAGE achieves substantial accuracy degradation across all evaluated models. The most recent and robust model, AdaFace, exhibits a significant performance drop on CelebA-HQ from a clean baseline of 96.16% to 70.24% in the gender-agnostic scenario, representing a 25.92% reduction. This degradation remains consistent across seen-gender (69.72%, –26.44%) and unseen-gender (71.54%, –24.62%) settings, demonstrating that MIRAGE's frequency-selective protection is robust regardless of target identity gender. Comparable drops are observed on 105-Pins (71.63%, –24.53%) and LFW (65.42%, –30.74%), validating effectiveness on both controlled and unconstrained real-world images. Stronger protection is observed on classical architectures: ArcFace drops sharply from 70.88% to 45.93% (-24.95%) on CelebA-HQ, while Facenet512 and DeepFace show the strongest protection, with recognition accuracy reduced to 22.87% and 24.64%, respectively, in a gender-agnostic scenario. These results confirm that cross-attention manipulation of identity-bearing low-frequency components effectively disrupts the biometric features that recognition systems rely upon. Consistent trends are observed in the reactive setting (Table 2), demonstrating that MIRAGE provides robust facial privacy protection.

Table 4: Performance (%) of sub-band swapping across different datasets and models on the reactive setting. The green color shows the best privacy-preserving model across different settings.

	Gender-Agnostic			Seen Gender			Unseen Gender		
Model	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	95.02	96.88	86.28	95.27	96.33	86.58	95.33	96.54	85.07
MagFace	77.67	86.90	57.99	76.52	85.22	57.99	76.79	85.31	57.15
ArcFace	56.15	55.56	58.25	59.16	54.13	58.00	55.48	48.67	51.92
Dlib	52.47	46.73	48.00	53.72	47.19	46.08	49.29	33.00	41.33
VGG-Face	48.62	43.46	50.25	50.71	40.59	47.58	48.37	31.33	46.00
SFace	44.18	18.63	52.67	44.44	21.78	49.83	41.17	12.67	47.25
Facenet	27.45	25.49	25.33	29.71	27.72	26.83	25.94	23.00	21.08
Facenet512	10.71	5.23	12.33	10.46	7.92	10.75	8.79	5.67	7.83
DeepFace	24.44	29.74	23.67	23.77	37.95	23.75	24.02	26.67	23.50

Table 5: Performance analysis of baseline methods, FPP [38], DiffAM [41], FAS [24], and FALCO [2], against our proposed approach on both proactive and reactive settings over two DFR models under a gender-agnostic setting. The green color indicates the best privacy-preserving method for each dataset and DFR model.

DFR Models	**Method**	Proactive (↓)			Reactive (↓)		
		CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	FPP [38]	96.19	94.45	90.52	96.18	94.05	89.54
	DiffAM [41]	96.39	96.16	88.05	96.45	96.25	87.78
	FAS [24]	68.35	70.22	71.70	68.45	71.11	72.03
	FALCO [2]	56.77	54.46	60.05	59.70	56.22	56.92
	MIRAGE (OURS)	70.24	71.63	65.42	68.56	69.48	65.63
MagFace	FPP [38]	77.77	84.09	60.80	77.82	83.69	61.01
	DiffAM [41]	83.49	88.25	58.67	82.25	90.94	58.42
	FAS [24]	57.06	61.76	56.37	58.02	59.68	56.89
	FALCO [2]	56.83	53.42	52.61	56.39	55.15	52.08
	MIRAGE (OURS)	57.87	61.75	44.13	56.67	64.21	44.89

5.2 Can DWT sub-band swapping provide sufficient privacy?

To assess whether simple frequency swapping alone can meaningfully protect facial privacy, we evaluate a DWT sub-band swapper. Tables 3 and 4 report the frequency swapping results across nine DFR models and three datasets in proactive and reactive scenarios, respectively. In the proactive scenario, AdaFace on CelebA-HQ drops only marginally from 96.16% to 95.47% in the gender-agnostic setting, a negligible 0.69% reduction compared to MIRAGE's 25.92% drop under the same configuration. This clearly demonstrates that naive frequency swapping is insufficient to disrupt robust modern face recognition systems. The substantial gap between sub-band swapping and MIRAGE highlights the critical role of learned attention mechanisms: the combination of cross-attention on the LL subband, self-attention on high-frequency subbands, and FAC loss supervision transforms a fundamentally insufficient baseline into the robust privacy protection demonstrated in Section 5.1.

5.3 Trade-offs between Privacy and Utility: Comparison with Benchmark

To quantify the privacy-utility tradeoff, we compare MIRAGE against four SOTA facial privacy protection methods: FPP [38], DiffAM [41], FAS [24], and

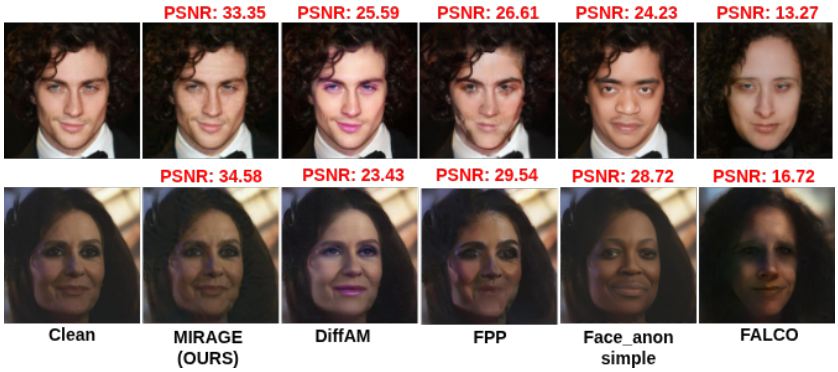

Fig. 5: Perceptual quality comparison of MIRAGE against existing SOTA generative approaches, our proposed approach MIRAGE achieves superior visual quality while maintaining strong privacy protection.

Table 6: Ablation evaluating the impact of removing MIRAGE components under proactive and reactive settings across DFR models. Lower values (↓) indicate stronger privacy protection.

DFR Models	Method	Proactive (↓)			Reactive (↓)		
		CelebA-HQ	105-Pins	LFW	CelebA-HQ	105-Pins	LFW
AdaFace	w/o cross-attention	96.30	94.66	87.61	95.77	95.84	88.51
	w/o self-attention	92.20	94.05	86.53	93.86	95.56	86.88
	w/o FAC	92.98	94.17	87.02	92.74	95.58	86.64
	MIRAGE (Ours)	**70.24**	**71.63**	**65.42**	**68.56**	**69.48**	**65.63**
MagFace	w/o cross-attention	80.20	93.61	58.35	79.81	85.19	58.86
	w/o self-attention	77.14	86.38	55.20	77.66	84.21	56.54
	w/o FAC	78.01	86.55	57.41	77.68	83.55	56.66
	MIRAGE (Ours)	**57.87**	**61.75**	**44.13**	**56.67**	**64.21**	**44.89**

FALCO [2] under both proactive and reactive settings. Table 5 presents these comparisons in a gender-agnostic setting across two SOTA face recognition models, AdaFace and MagFace. Among the facial protection methods, FPP and DiffAM provide the weakest privacy protection, with recognition accuracies remaining close to the clean baseline across all datasets and both models, indicating that these methods offer insufficient disruption of identity features. FALCO achieves the strongest numerical privacy protection, consistently yielding the lowest recognition accuracies across most datasets and model combinations. MIRAGE demonstrates competitive performance overall, with particularly strong results on the LFW dataset under MagFace (44.13% proactive, 44.89% reactive), where it outperforms all compared methods, including FALCO.

However, recognition accuracy alone does not fully capture the effectiveness of a facial privacy protection method. Fig. 5 reveals the decisive factor: *visual appearance preservation*. While FALCO achieves strong identity disruption, it fundamentally alters the subject's facial appearance, producing images that look entirely different from the original person and are therefore impractical for social uses. FPP and DiffAM, despite preserving appearance, fail to provide meaningful

Table 7: Robustness of MIRAGE under compression and resizing.

Model	JPEG Quality Factor ↓				Resizing ↓
	QF=50	QF=75	QF=85	QF=95	112×112
AdaFace	64.02	64.38	64.94	64.58	63.98
MagFace	48.78	49.42	50.16	49.85	50.25

privacy protection. In contrast, MIRAGE achieves an optimal balance between the two requirements, providing strong identity protection while preserving the natural visual appearance of the face, with an average PSNR and SSIM value of 32.85 and 0.89, respectively. Complete quantitative comparisons across all nine DFR models, as well as qualitative comparisons, including seamless clone face-swapping results at the pixel level, are provided in the supplementary material.

5.4 Robustness on compression, resizing and adaptive fine-tuning

We evaluate MIRAGE-protected images under common post-processing operations, namely JPEG compression at varying quality factors and spatial resizing. As shown in Table 7, rather than weakening privacy protection, recognition accuracy further *decreases* under compression (e.g., ∼6% drop for AdaFace at QF=50 compared to the uncompressed setting in Table 1). A plausible explanation is that these preprocessing operations introduce artifacts that further perturb the already disrupted low-frequency identity signal, compounding the recognition system's vulnerability [23]. This indicates that MIRAGE's identity perturbation, embedded predominantly in the low-frequency subband, is resilient to, and even reinforced by, common transformations in real-world deployment pipelines. To assess whether a defender could adapt by retraining on MIRAGE-protected images, we fine-tune AdaFace and MagFace using such samples. Accuracy improves only marginally (e.g., +1.84% on AdaFace) and remains well below clean-image performance, confirming that MIRAGE's protection is robust even against adaptive, white-box retraining.

6 Ablation Study

To validate the contribution of each component in MIRAGE, we conduct an ablation study by systematically removing individual components and evaluating the impact on privacy protection. Table 6 reports recognition accuracy under both proactive and reactive settings across two recent face recognition models, where lower accuracy indicates stronger privacy protection. The results clearly demonstrate that every component contributes meaningfully to MIRAGE's overall effectiveness. Removing cross-attention causes the most severe degradation, with recognition accuracy approaching the clean baseline across all datasets and both models; for example, AdaFace accuracy increases from 70.24% to 96.30%

on CelebA-HQ in the proactive setting, essentially eliminating privacy protection entirely. This confirms that cross-attention is the core component that imposes the protected identity signal on the low-frequency subband. Removing self-attention or FAC also leads to a consistent increase in recognition accuracy, though less pronounced. Self-attention preserves high-frequency appearance details, stabilizing the output, while FAC provides frequency-aware supervision that jointly enforces identity separation and visual fidelity during training. Together, all three components work in concert to achieve MIRAGE's optimal privacy-utility balance, and the absence of any single component measurably weakens the overall framework. Complete ablation results across all nine DFR models are provided in the supplementary material.

6.1 Broader Impact

The rapid expansion of AI-driven face recognition has intensified facial privacy risks, enabling the misuse of personal face images for surveillance, identity profiling, and unauthorized biometric extraction. MIRAGE addresses these concerns by disrupting automated recognition systems while preserving natural visual appearance for everyday use. We hope this work encourages further exploration of frequency-aware and attention-driven defenses against the misuse of facial biometric data.

7 Conclusion

We present MIRAGE, a frequency-selective facial privacy protection framework that addresses the fundamental privacy-utility tradeoff that existing methods leave unresolved. By combining DWT-based frequency decomposition with cross-attention for low-frequency identity perturbation and self-attention for high-frequency appearance preservation, together with our proposed FAC loss, MIRAGE achieves robust identity concealment while maintaining the natural visual appearance of the face. Comprehensive experiments across three benchmark datasets and nine SOTA face recognition models demonstrate that MIRAGE consistently reduces recognition accuracy while preserving visual quality, as evidenced by high PSNR values. Unlike existing methods that achieve privacy only at the cost of complete facial appearance alteration, MIRAGE generates protected images that are both resistant to unauthorized recognition and visually suitable for everyday social use. In future work, we aim to extend this dual defense into a unified framework that simultaneously addresses multiple facial privacy threats.

Acknowledgements

A. Kumar is partially supported through the JRF fellowship of the UGC, India.

References

1. Agarwal, A., Ratha, N., Vatsa, M., Singh, R.: Crafting adversarial perturbations via transformed image component swapping. IEEE Transactions on Image Processing **31**, 7338–7349 (2022)
2. Barattin, S., Tzelepis, C., Patras, I., Sebe, N.: Attribute-preserving face dataset anonymization via latent code optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8001–8010 (2023)
3. Birhane, A., Prabhu, V.U.: Large image datasets: A pyrrhic win for computer vision? In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1536–1546. IEEE (2021)
4. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2003–2011 (2020)
5. Chhabra, S., Singh, R., Vatsa, M., Gupta, G.: Anonymizing k-facial attributes via adversarial perturbations. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 656–662 (2018)
6. Ciftci, U.A., Yuksek, G., Demir, I.: My face my choice: Privacy enhancing deepfakes for social media anonymization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1369–1379 (2023)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
8. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
9. He, X., Zhu, M., Chen, D., Wang, N., Gao, X.: Diff-privacy: Diffusion-based face privacy protection. IEEE Transactions on Circuits and Systems for Video Technology (2024)
10. Hill, K.: The secretive company that might end privacy as we know it. In: Ethics of Data and Analytics, pp. 170–177. Auerbach Publications (2022)
11. Hsu, G.S., Tsai, C.H., Wu, H.Y.: Dual-generator face reenactment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 642–650 (2022)
12. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: 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 (2008)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
14. Huber, M., Damer, N.: Beyond spatial explanations: Explainable face recognition in the frequency domain. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1016–1026. IEEE (2025)
15. Jia, S., Ma, C., Yao, T., Yin, B., Ding, S., Yang, X.: Exploring frequency adversarial attacks for face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4103–4112 (2022)
16. Kato, G., Fukuhara, Y., Isogawa, M., Tsunashima, H., Kataoka, H., Morishima, S.: Scapegoat generation for privacy protection from deepfake. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3364–3368. IEEE (2023)

17. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18750–18759 (2022)
18. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research **10**, 1755–1758 (2009)
19. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 3677–3685 (2017)
20. Kuang, Z., Yang, X., Shen, Y., Hu, C., Yu, J.: Facial identity anonymization via intrinsic and extrinsic attention distraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12406–12415 (2024)
21. Kumar, A., Agarwal, A., Ratha, N.: Your face, your privacy: Combating unauthorized usage. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2025)
22. Kumar, A., Agarwal, A., et al.: On which data distribution (synthetic or real) we should rely for soft biometric classification. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 6238–6247 (2025)
23. Kumar, V., Shukla, S., Agarwal, A.: Robustness benchmarking of convolutional and transformer architectures for image classification. IEEE Transactions on Big Data (2025)
24. Kung, H.W., Varanka, T., Saha, S., Sim, T., Sebe, N.: Face anonymization made simple. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1040–1050. IEEE (2025)
25. Lander, K., Bruce, V., Hill, H.: Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition **15**(1), 101–116 (2001)
26. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5549–5558 (2020)
27. Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., Cao, X.: Identity-preserving face anonymization via adaptively facial attributes obfuscation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3891–3899 (2021)
28. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
29. Liu, Y., Jia, C., Xiao, R., Jia, X., Wei, H., Jiang, K., Wang, Z.: Local features meet stochastic anonymization: Revolutionizing privacy-preserving face recognition for black-box models. CoRR (2024)
30. Luo, X., Wang, Y.: Frequency-domain masking and spatial interaction for generalizable deepfake detection. Electronics **14**(7), 1302 (2025)
31. Majumdar, P., Agarwal, A., Singh, R., Vatsa, M.: Evading face recognition via partial tampering of faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
32. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14234 (2021)
33. Mittal, S., Chowdhury, R.D., Vatsa, M., Singh, R.: Decordface: A framework for degraded and corrupted face recognition. Journal of Data-centric Machine Learning Research (2025)

34. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7184–7193 (2019)
35. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association (2015)
36. Pulfer, E.M.: Different approaches to blurring digital images and their effect on facial detection (2019)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
38. Salar, A., Liu, Q., Tian, Y., Zhao, G.: Enhancing facial privacy protection via weakening diffusion purification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8235–8244 (2025)
39. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
41. Sun, Y., Yu, L., Xie, H., Li, J., Zhang, Y.: Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24584–24594 (2024)
42. Vishwamitra, N., Knijnenburg, B., Hu, H., Kelly Caine, Y.P., et al.: Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 39–47 (2017)
43. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
44. Xu, S., Chang, C.C., Nguyen, H.H., Echizen, I.: Reversible anonymization for privacy of facial biometrics via cyclic learning. EURASIP Journal on Information Security **2024**(1), 24 (2024)
45. Zhai, L., Guo, Q., Xie, X., Ma, L., Wang, Y.E., Liu, Y.: A3gan: Attribute-aware anonymization networks for face de-identification. In: Proceedings of the 30th ACM international conference on multimedia. pp. 5303–5313 (2022)
46. Zhang, H., Dong, X., Lai, Y., Zhou, Y., Zhang, X., Lv, X., Jin, Z., Li, X.: Validating privacy-preserving face recognition under a minimum assumption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12205–12214 (2024)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)