Frequency-domain Manipulation for Face Obfuscation

Jintae Kim
Korea University
jtkim@mcl.korea.ac.kr

Keunsoo Ko
The Catholic University of Korea
ksko@catholic.ac.kr

Chang-Su Kim*
Korea University
changsukim@korea.ac.kr

Abstract

Facial image datasets have become essential resources for various face analysis tasks, but their use raises significant privacy concerns. To address this issue, face obfuscation has emerged as a practical approach to hide identity from humans while retaining cues decipherable by machines. However, existing methods often leave exploitable visual traces, making them vulnerable to reconstruction attacks that restore hidden identity. To address this issue, we propose a frequency-domain manipulation framework, called FreM, which adjusts frequency subbands differently to hide identity, retain machine-decipherable cues, and improve robustness against reconstruction attacks. Specifically, the proposed FreM first decomposes a facial image into frequency subbands and applies subband-adaptive modulation that regulates information according to the characteristics of each subband. The modulation parameters are then refined to yield the reliable obfuscated result. Extensive experiments across multiple face analysis benchmarks demonstrate that FreM achieves superior obfuscation quality and strong robustness against reconstruction attacks. The source codes are available at https://github.com/mcljtkim/FreM

1. Introduction

Large-scale facial image datasets [12, 16, 24, 30, 33, 39, 41, 46, 58, 59] have been widely used in various face analysis tasks, such as face recognition [10, 21, 41], age estimation [25, 47], facial expression recognition [43, 56], and attribute classification [42]. However, since facial data inherently includes personally identifiable information, the increasing accessibility of these datasets raises serious privacy concerns. In response, recent privacy regulations [13, 48, 49] have emphasized the need for stronger protection of personal identity, thereby motivating the development of various privacy-enhancing techniques.

Among these techniques, face obfuscation [1, 20, 27–29, 37, 38, 40, 51, 54] has recently attracted significant attention as it aims to conceal human-recognizable identity while leaving the traces useful for machine analysis. However, existing approaches suffer from an inherent trade-off between human indecipherability (*HI*) and machine decipherability (*MD*): increasing *HI* typically degrades *MD*, and vice versa. Although some methods [20, 51] mitigate this trade-off, they remain vulnerable to reconstruction attacks that can reveal obfuscated information, as illustrated in Figure 1. Since they mainly manipulate spatially adjacent pixel values, these methods may inadvertently leave structural information that reconstruction attacks can exploit.

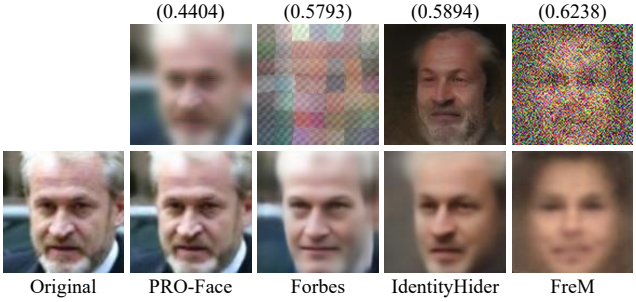

Figure 1. Comparison of obfuscated (top) and corresponding reconstructed (bottom) images. The scores in parentheses indicate feature similarities (ArcFace [10]) between the original and obfuscated images. Existing methods [20, 51, 54] leave identity cues that reconstruction models can exploit, whereas the proposed FreM suppresses such cues effectively.

To overcome these limitations of spatial-domain operations, prior studies in other fields have often adopted the frequency domain. As representative examples, robust image watermarking techniques [3, 23, 26] exploit frequency representations to conceal information in a perceptually imperceptible but tamper-resistant manner. Inspired by this, we perform face obfuscation by adaptively manipulating frequency subbands according to their contributions to *HI* and *MD*. As shown in Figure 1, this global processing improves the robustness against reconstruction attacks.

Building upon this foundation, we propose a novel frequency-domain manipulation framework, referred to as FreM, for face obfuscation. Specifically, an input facial im-

*Corresponding author.

age is first transformed into the frequency domain via the discrete cosine transform (DCT) in a block-wise manner. The frequency components are manipulated using subband-adaptive modulation and then converted back to the spatial domain via inverse DCT (IDCT), producing an initial obfuscated image. This image is iteratively refined by updating the modulation parameters based on the backpropagating refinement scheme [18] to balance *HI* and *MD*. Extensive experiments across multiple face analysis benchmarks demonstrate that the proposed FreM achieves superior obfuscation quality and strong robustness against reconstruction attacks.

The main contributions of this work are as follows:

- We propose FreM, a novel frequency-domain framework for face obfuscation.
- We introduce a subband-adaptive modulation strategy that independently controls subbands according to their contributions in *HI*, *MD*, and robustness against reconstruction attacks.
- Extensive experiments demonstrate that FreM outperforms existing methods across various benchmarks.

2. Related Work

2.1. Privacy-enhancing Techniques

Privacy-enhancing techniques can be broadly categorized into *face anonymization* and *face obfuscation*, depending on whether the results consider machine decipherability (*MD*). Face anonymization [2, 5, 7, 8, 15, 17, 32, 45] enhances personal privacy by removing all identity-related information recognizable by both humans and machines. This is typically achieved through simple operations such as masking or blurring [5, 15] or face replacement using generative models [2, 7, 8, 17, 32].

In contrast, face obfuscation [6, 19, 31, 34–36, 52, 53, 55] aims to conceal identifiable information by humans while preserving *MD*. For example, several methods [27–29, 38] have employed GANs to synthesize obfuscated faces with controlled attributes, thereby concealing original properties. Although these methods produce visually plausible results, they often fail to achieve a proper balance between *HI* and *MD*, either leaving residual identity cues or degrading *MD*.

To overcome this issue, many methods [6, 19, 31, 34–36, 51–55] trained obfuscation models by introducing task-specific losses based on predefined face analysis models to ensure *MD* explicitly. However, their dependency on predefined face analysis models necessitates re-training whenever these underlying models are updated, thereby limiting their adaptability and scalability.

Recently, Kim *et al*. [20] proposed training-free face obfuscation based on the Backpropagating Refinement Scheme (BRS) [18]. Their method employs parameterized local spatial filters and iteratively updates them until the desired result is obtained. While it strikes a good balance between *HI* and *MD* without additional training, it remains vulnerable to reconstruction attacks that can recover obfuscated information. This limitation indicates that purely spatial-domain manipulations may inadvertently leave structural information that reconstruction attacks can exploit. In contrast, FreM performs obfuscation in the frequency domain, where identity-related cues, task-relevant information, and tamper-resistant components can be independently controlled.

2.2. DCT-based Approaches

The discrete cosine transform (DCT) decomposes an image into frequency components and has been widely used in various vision tasks, including image compression [50] and watermarking [26]. Several face obfuscation methods have also explored the frequency domain by applying DCT to facial images in a block-wise manner and performing simple operations such as channel selection, shuffling, or masking [19, 34–36, 52]. Although these methods can achieve *HI* by processing low-frequency visual details while preserving *MD*, they operate on DCT coefficients without considering the distinct roles of different frequency regions.

In contrast, watermarking techniques [3, 23, 26] demonstrate that subband-wise processing can achieve a better balance between imperceptibility and robustness to tampering. Inspired by this principle, the proposed FreM performs block-wise DCT and introduces subband-adaptive manipulation, assigning dedicated modules to each subband, allowing FreM to achieve a superior trade-off between *HI* and *MD* while improving robustness against reconstruction attacks.

3. Proposed Algorithm

The proposed FreM aims to generate an obfuscated image I_{out} from an input image I_{in} that ensures both human indecipherability (*HI*) and machine decipherability (*MD*) while maintaining robustness against reconstruction attacks. As shown in Figure 2, we first apply the discrete cosine transform (DCT) in a block-wise manner to map I_{in} into the frequency domain and group it into four subbands (LL, LH, HL, HH) according to their frequency ranges. The identity-related information in each subband is independently manipulated using three distinct modules, each parameterized by learnable parameters Θ:

(1) *Neutralization* module – falsifies the low-frequency (LL) subband to remove identity cues for HI with Θ_{neu}.
(2) *Perturbation* module – slightly modifies components of LH and HL subbands using Θ_{per} to enhance feature representations for *MD*.
(3) *Suppression* module – replaces high-frequency (HH) subband with Θ_{sup} for reconstruction robustness.

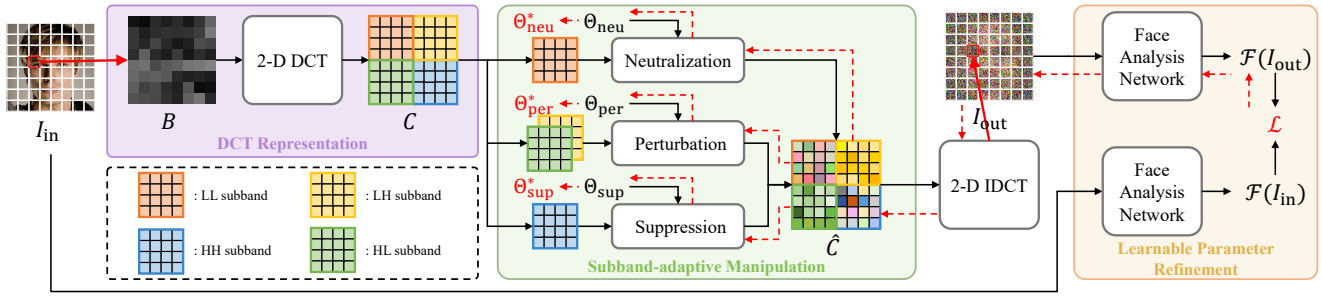

Figure 2. Overview of the proposed algorithm. The pipeline consists of three stages: block-wise DCT representation, subband-adaptive manipulation, and learnable parameter refinement, where the manipulation parameters Θ are updated using gradients of the objective function $\mathcal{L}$. Red dashed arrows indicate the gradient flow during the refinement process.

These modified results are transformed back into the spatial domain via block-wise inverse DCT (IDCT), producing I_{out}. The learnable parameters $(\Theta_{\text{neu}}, \Theta_{\text{per}}, \Theta_{\text{sup}})$ are optimized by minimizing an objective function defined by a pretrained face analysis network, which determines the *MD* criterion. Consequently, through this optimization, I_{out} is refined to achieve a balance between *HI* and *MD* while maintaining reconstruction robustness. Let us describe each stage subsequently.

3.1. DCT Representation

We employ the block transform with the DCT [14] to represent the input image in the frequency domain. Unlike a global DCT that transforms the entire image at once, this block-wise processing provides a localized frequency representation with a small number of significant coefficients, enabling efficient and interpretable manipulation.

Given an input image $I_{\text{in}} \in \mathbb{R}^{H\times W\times 3}$, it is first partitioned into non-overlapping blocks of equal size P. Using the 2-D DCT, we map each block $B \in \mathbb{R}^{P\times P}$ into its frequency coefficients $C \in \mathbb{R}^{P\times P}$ and then decompose them into four frequency subbands $(C_{\text{LL}}, C_{\text{LH}}, C_{\text{HL}}, C_{\text{HH}})$, corresponding to low- and high-frequency components along each axis as shown in Figure 2. This representation enables subband-wise modulation, where each subband, defined by a distinct frequency range, is adaptively manipulated according to its contribution to *HI* and *MD*.

3.2. Subband-adaptive Manipulation

Based on the distinct properties of frequency subbands, we introduce a subband-adaptive manipulation strategy with three frequency-specific modules: *Neutralization*, *Perturbation*, and *Suppression*, each of which is parameterized with its own learnable parameters Θ, as shown in Figure 3.

Neutralization module: This module is applied to the LL subband, which contains dominant identity-related information for *HI*. To neutralize the information, we compute the average facial image from the facial dataset and represent it in the DCT domain, where coefficients of each block are denoted as $\bar{C}_{\text{LL}}$. To further enhance *HI*, learnable parameters Θ_{neu} are added, yielding the neutralized coefficients $\hat{C}_{\text{LL}}$ as

$$\hat{C}_{\text{LL}} = \bar{C}_{\text{LL}} + \Theta_{\text{neu}} \tag{1}$$

where Θ_{neu} are initialized from a Gaussian distribution $\mathcal{N}(0, \sigma^2_{\text{neu}})$. This initialization preserves coarse facial structures while allowing small perturbations around the average face. During learnable parameter refinement, Θ_{neu} are optimized to balance *HI* and *MD*. As shown in Figure 3(b), the *Neutralization* module effectively reduces identity cues, leading to higher *HI*.

Perturbation module: This module is applied to the LH and HL subbands, which contain the components that are often imperceptible to humans but contain discriminative cues beneficial for *MD*. To leverage this property, we introduce learnable scaling parameters Θ_{LH} and Θ_{HL}, which slightly perturb the DCT coefficients C_{LH} and C_{HL}, respectively, as

$$\hat{C}_f = C_f \odot \Theta_f, \quad f \in \{\text{LH}, \text{HL}\} \tag{2}$$

where $\odot$ denotes element-wise multiplication. For convenience, we collectively denote the pair of parameters as $\Theta_{\text{per}} = \{\Theta_{\text{LH}}, \Theta_{\text{HL}}\}$. Each parameter of Θ_{per} is initialized to 1, which prevents the loss of *MD*-related cues inherently contained in LH and HL subbands. During learnable parameter refinement, the parameters Θ_{per} are adaptively updated to adjust the magnitude of the coefficients, thereby enhancing *MD* without compromising *HI*, as visualized in Figure 3(b).

Suppression module: The *Suppression* module targets the HH subband, whose components are almost imperceptible to humans but highly influential against reconstruction attacks. In this module, the original HH coefficients C_{HH} are replaced with learnable parameters $\Theta_{\text{sup}} \sim \mathcal{N}(0, \sigma^2_{\text{sup}})$, where the standard deviation σ_{sup} controls the strength of high-frequency suppression. These parameters are updated during learnable parameter refinement to yield the updated coefficients $\hat{C}_{\text{HH}}$. This replacement effectively disrupts the reconstruction of identity, as demonstrated in Figure 3(b).

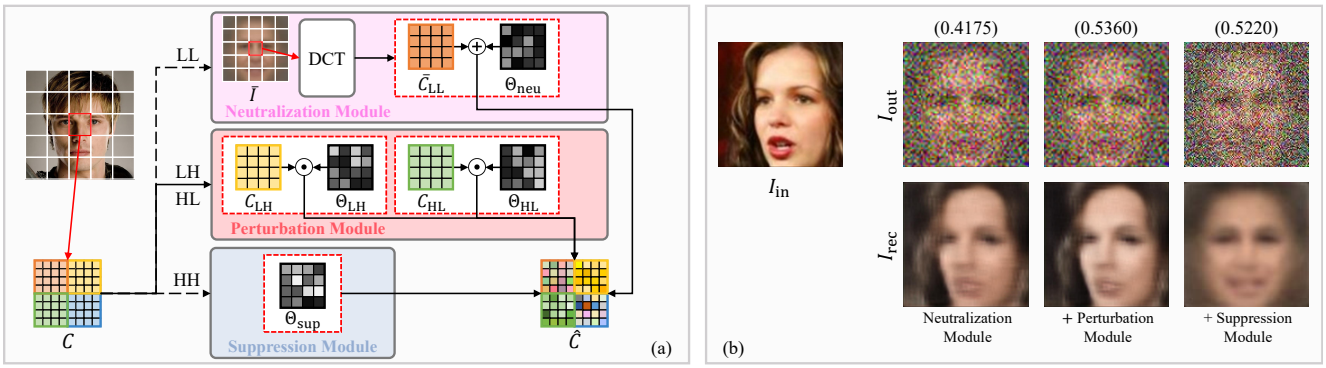

Figure 3. (a) Overall pipeline of the subband-adaptive manipulation strategy. Each subband of the input DCT coefficients is manipulated by its dedicated module. (b) Qualitative effects of individual modules. The top row shows obfuscated examples I_{out} illustrating the individual effect of each module. The bottom row presents the corresponding reconstructed images I_{rec}, and the number in parentheses denotes cosine similarities between the original and obfuscated images.

3.3. Learnable Parameter Refinement

Since *HI* and *MD* involve inherent trade-offs, it is crucial to balance them for reliable face obfuscation. To this end, we adopt the Backpropagating Refinement Scheme (BRS) [18] to perform per-image optimization under a pretrained face analysis network while freezing its weights. During optimization, the learnable parameters $(\Theta_{\text{neu}}, \Theta_{\text{per}}, \Theta_{\text{sup}})$ are iteratively refined by minimizing the objective functions. This process optimizes all learnable parameters according to the characteristics of each input image, instead of cumbersome training an obfuscation network or face analysis network as done in previous training-based approaches [6, 19, 31, 34–36, 51–55].

Objective function: The objective function $\mathcal{L}$ consists of two complementary components: Machine decipherability loss ($\mathcal{L}_{\text{MD}}$) and coefficient energy constraint loss ($\mathcal{L}_{\text{CEC}}$). It is formally defined by

$$\mathcal{L} = \mathcal{L}_{\text{MD}} + \lambda_{\text{CEC}}\mathcal{L}_{\text{CEC}} \tag{3}$$

where λ_{CEC} is a balancing weight between *HI* and *MD*.

The machine decipherability loss $\mathcal{L}_{\text{MD}}$ is defined as

$$\mathcal{L}_{\text{MD}} = 1 - \mathcal{F}(I_{\text{in}})^T \mathcal{F}(I_{\text{out}}) \tag{4}$$

which encourages the feature representations of the input image I_{in} and the obfuscated image I_{out} to be similar in the pretrained face analysis network $\mathcal{F}$. This enables the obfuscated image to preserve identity-related information, thereby maintaining *MD*.

The coefficient energy constraint loss $\mathcal{L}_{\text{CEC}}$ is introduced to preserve the identity-neutral state established by the *Neutralization* module. During optimization, $\mathcal{L}_{\text{MD}}$ tends to amplify low-frequency coefficients in pursuit of higher *MD*, which may reintroduce human-perceptible cues. To counteract this effect, $\mathcal{L}_{\text{CEC}}$ regularizes the magnitude of the modified coefficients $\hat{C}$, as defined by

$$\mathcal{L}_{\text{CEC}} = \left| \|\hat{C}\|_1 - \|C\|_1 \right| \tag{5}$$

which prevents excessive signal amplification beyond the original coefficients C. This constraint also stabilizes the IDCT process and suppresses overflow artifacts in I_{out}.

Refinement process: The parameter refinement is performed independently for each test image, during which only the gradients of Θ_{neu}, Θ_{per}, and Θ_{sup} are computed and updated. This process continues until $\mathcal{L}_{\text{MD}}$ falls below a predefined threshold τ or the maximum iteration count $T_{\max}$ is reached. It ensures that I_{out} achieves at least the minimum required MD performance.

4. Experiments

4.1. Experimental Setup

Dataset: We validate the effectiveness of the proposed algorithm on 10 datasets covering four tasks: face recognition, age estimation, expression recognition, and binary attribute classification. Specifically, we use LFW [16], AgeDB [39], CALFW [59], CPLFW [58], and CFP-FP [46] for face recognition; MORPH II [44] and UTKFace [57] for age estimation; RAF-DB [9] and FERPlus [4] for expression recognition; and CelebA [30] for binary attribute classification. The details of the datasets are reported in the supplement.

Implementation details: For the learnable parameter refinement, we use the Adam optimizer [22] with a learning rate of 10^{-3}. The loss balancing parameter λ_{CEC} is set to 10^{-2} throughout all experiments. The maximum number of iterations is fixed to $T_{\max} = 50$, and the optimization process terminates when $\mathcal{L}_{\text{MD}}$ falls below the threshold $\tau = 0.4$ or $T_{\max}$ is reached. The block size is fixed to

Table 1. Quantitative comparison (XDR/ODR) of face obfuscation results on the LFW [16], AgeDB [39], CALFW [59], CPLFW [58], and CFP-FP [46] datasets. In each test, the best result is **boldfaced**, while the second best is underlined. The recognition accuracies are measured using the face recognition network: an IResNet50 [11] backbone trained with the ArcFace loss [10]. R_{rec} denotes the reconstruction robustness (PSNR).

	Dataset					R_{rec} (↓)	Runtime (ms)
	LFW	AgeDB	CALFW	CPLFW	CFP-FP		
Original	99.83 / -	97.55 / -	95.88 / -	91.80 / -	97.26 / -	-	-
PRO-Face (Blur) [54]	93.58 / 91.30	80.72 / 70.28	85.05 / 77.62	82.57 / 71.90	82.94 / 71.49	32.97	12.80
PRO-Face (Pixelate) [54]	92.67 / 87.17	76.25 / 65.03	82.82 / 75.40	78.98 / 66.15	84.91 / 68.37	29.59	9.27
PRO-Face (FaceShifter) [54]	96.48 / 95.78	84.27 / 80.30	87.87 / 85.10	82.70 / 72.72	91.83 / 77.39	36.12	32.41
PRO-Face (SimSwap) [54]	88.00 / 90.90	79.68 / 76.78	83.23 / 82.93	75.25 / 70.75	88.87 / 79.36	33.74	30.93
Forbes [20]	95.72 / 82.77	87.02 / 72.17	89.45 / 75.27	83.53 / 71.68	86.73 / 71.54	22.96	739.57
IdentityHider [51]	99.08 / 98.48	94.93 / **93.37**	94.65 / **93.17**	87.87 / 83.27	91.49 / 87.14	15.33	68.14
FreM	**99.53 / 98.67**	**95.95** / 92.45	**94.95** / 92.86	**90.91 / 86.88**	**94.41 / 90.74**	**13.59**	67.19

$P = 8$. For the learnable parameters Θ_{neu} and Θ_{sup}, we set $\sigma_{\text{neu}} = 0.5$ and $\sigma_{\text{sup}} = 1$. All experiments are conducted using a PC with an AMD Ryzen 9 3900X CPU and an NVIDIA RTX 3090 GPU.

4.2. Comparison Results

We compare the results of the proposed FreM against conventional face obfuscation techniques: PRO-Face [54], Forbes [20], and IdentityHider [51]. Among them, Forbes is a training-free method based on the BRS, similar to FreM, while PRO-Face and IdentityHider are training-based methods that require dedicated task-specific obfuscation networks, each retrained to align with a particular face analysis task. All results are obtained by executing available source codes. Since PRO-Face and IdentityHider are designed for face recognition, comparisons with these methods are limited to the face recognition task. Forbes, which supports a wide range of face analysis tasks, is evaluated for all tasks to ensure a fair comparison.

Face recognition: For face recognition, we use an IResNet50 [11] backbone trained with the ArcFace loss [10] as the face analysis network. We evaluate the recognition accuracy of obfuscated images to measure *MD*. We analyze two scenarios for robust privacy assessment: Cross-domain recognition (XDR) and Obfuscated-domain recognition (ODR). XDR involves matching a pair consisting of one obfuscated image against an original (non-obfuscated) image, whereas ODR involves matching two obfuscated images. In Table 1, we present the recognition results of the face obfuscation methods. Across the five benchmarks and two protocols (XDR/ODR), the proposed FreM ranks first in 8 out of 10 cases and second in the remaining two, yielding the best overall performance.

In Table 2, we compare the average accuracies across ten different random seeds with Forbes [20], as both methods perform parameter refinement that can be affected by initialization. The results show that the proposed FreM delivers stable performance and consistent convergence regardless of different initializations. Despite being a training-free approach, FreM achieves a comparable runtime ($\approx$ 67 ms) to IdentityHider [51], which relies on a separately trained obfuscation network, and is more than ten times faster than Forbes. This demonstrates that FreM attains both accuracy and computational efficiency without the need for additional training.

Table 2. Quantitative comparison of face recognition under ten different random seeds on the LFW dataset [16]. $\pm$ values denote half the range between the minimum and maximum accuracies.

	XDR	ODR	Runtime
Original	99.6	-	-
Forbes [20]	95.6 ($\pm$0.4)	82.8 ($\pm$0.4)	740 ($\pm$5)
FreM	99.5 ($\pm$0.1)	98.7 ($\pm$0.2)	67 ($\pm$4)

Figure 4 compares the qualitative results on the LFW dataset. PRO-Face variants (Blur, Pixelate, FaceShifter, and SimSwap) [54] and IdentityHider [51] do not sufficiently distort the facial structure. Forbes [20] produces stronger obfuscation, but still retains recognizable features, such as skin texture and head shape. In contrast, the proposed FreM effectively distorts most perceptually relevant facial components, producing visually unrecognizable results to humans while ensuring *MD* through imperceptible frequency modification in the DCT domain.

Reconstruction robustness: In addition to recognition comparisons, we evaluate the reconstruction robustness under a black-box attack scenario. In this setup, the attacker processes the input image I_{in} through the face obfuscation algorithm to generate the obfuscated image I_{out}. The at-

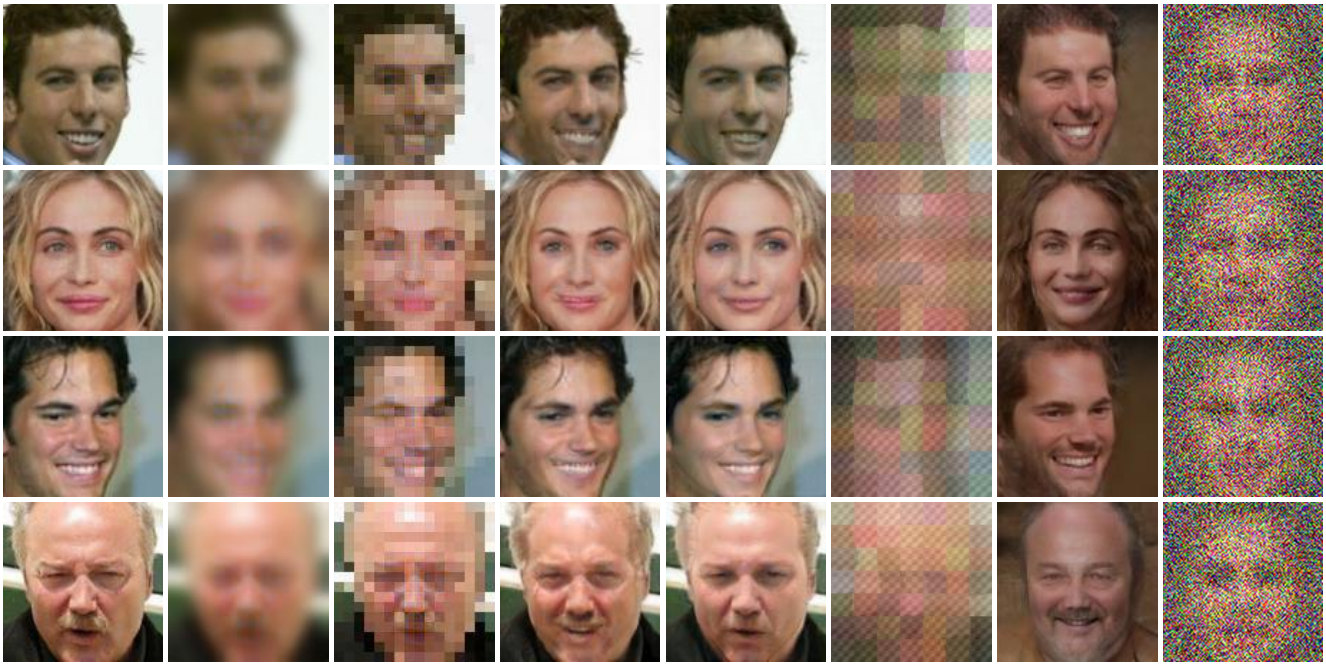

Figure 4. Qualitative comparison of obfuscated images in the LFW [16] dataset.

tacker then trains a U-Net encoder-decoder as a reconstruction network using the generated pairs. We measure the PSNR between the reconstructed image I_{rec} and the original image I_{in} on the LFW dataset, and report the results in Table 1. The proposed FreM achieves the lowest PSNR, which indicates that the obfuscated images are the most difficult to reconstruct to their original form. This result demonstrates that the robustness of FreM against reconstruction attacks.

Figure 5 visualizes the obfuscated results (odd row) and the corresponding reconstructions produced by the attacker (even row). Existing methods exhibit clear vulnerability to reconstruction attacks, as their reconstructed images retain noticeable facial structures and identity cues from the original faces. In contrast, the proposed FreM shows strong robustness, yielding reconstructions that fail to recover any discernible identity-related information. These results confirm the robustness of FreM against reconstruction attacks. Since *HI* is inherently subjective and difficult to quantify, we provide additional qualitative examples in the supplement to illustrate the obfuscation quality.

Other face analysis tasks: To further verify the generality of the proposed FreM, we evaluate FreM on additional face analysis tasks: age estimation, expression recognition, and binary attribute classification. For age estimation, we use the MWR [47] network as the face analysis network. Since MWR performs relative comparisons rather than absolute regression, evaluation is conducted under the XDR scenario, and we report the mean absolute error (MAE) and cumulative score (CS), where CS measures the percentage of images whose absolute prediction error falls within the tolerance level $l = 5$ [47]. As shown in Table 3, FreM achieves lower MAE and higher CS than Forbes, demonstrating minimal degradation in *MD* performance for age estimation.

Table 3. Comparison of the age estimation results on the MORPH II [44] and UTKFace [57] datasets. All methods use the same MWR [47] network as the age estimator.

	MORPH II		UTKFace	
	MAE	CS(%)	MAE	CS(%)
Original	2.24	94.6	4.49	71.0
Forbes [20]	3.38	77.4	6.28	51.7
FreM	2.41	91.8	4.77	68.5

Table 4. Comparison of the expression recognition results on the RAF-DB [9] and FERPlus [4] datasets, measured by classification accuracy. All methods use the same Faceptor [43] network.

	RAF-DB	FERPlus
Original	85.95	77.75
Forbes [20]	75.23	65.38
FreM	83.02	76.13

For both expression recognition and binary attribute classification, we use the Faceptor [43] network. Evaluation

Figure 5. Qualitative comparison of reconstructed images in the LFW [16] dataset.

Table 5. Comparison of the binary attribute classification results on the CelebA [30] datasets. All methods use the same Faceptor [43] network as the classifier.

	mAcc.	Runtime
Original	90.35	-
Forbes [20]	88.11	1062
FreM	88.80	68.99

is performed in a standard classification setting, where we report the classification accuracy (Acc.) for expression recognition and the mean attribute classification accuracy (mAcc.) across various attributes. As shown in Tables 4 and 5, FreM consistently outperforms Forbes across all tasks, demonstrating its robustness and adaptability in maintaining *MD* performance across face analysis tasks.

4.3. Ablation and Analysis

We conduct ablation studies to validate the contributions of each component and to analyze the robustness of the design choices. All ablation experiments are performed on the face recognition task.

Subband-adaptive manipulation strategy: We analyze the efficacy of the proposed subband-adaptive manipulation modules in Table 6 and Figure 6. Methods (1)–(4) are constructed by selectively enabling the proposed modules: *Neutralization* ($\bar{C}_{\mathrm{LL}}$ and Θ_{neu}), *Perturbation* (Θ_{per}), and

Table 6. Ablation study of the subband-adaptive manipulation modules.

Method	$\bar{C}_{\mathrm{LL}}$	Θ_{neu}	Θ_{per}	Θ_{sup}	Acc.	PSNR
(1)	✓	✓			99.15	16.45
(2)	✓	✓	✓		99.61	16.72
(3)		✓	✓	✓	98.13	13.23
(4)	✓	✓	✓	✓	99.50	13.59

Suppression (Θ_{sup}) modules, where Method (4) represents the proposed method. For a fair comparison, all methods perform learnable parameter refinement with the same number of refinement iterations.

Comparing (1) and (2), introducing the *Perturbation* module improves *MD* performance, but is still vulnerable to reconstruction attacks. In contrast, Methods (3) and (4), which incorporate the *Suppression* module, enhance robustness against reconstruction attacks by modifying high-frequency components. Finally, comparing (3) and (4), neutralizing the input face with $\bar{C}_{\mathrm{LL}}$ preserves general facial structures, resulting in improved *MD* performance.

Effect of block size: We analyze the effect of block size P on *MD* performance. Table 7 lists the accuracies for different block sizes. When P is too small, each block captures only limited local frequency information, restricting the representation capacity of the DCT representation. In

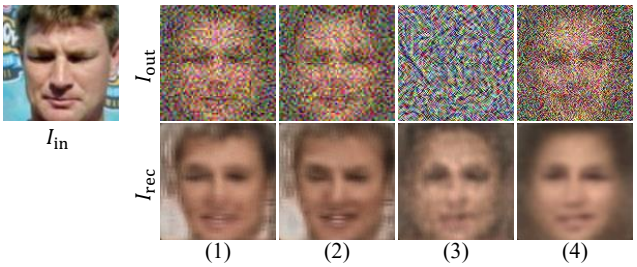

Figure 6. Examples of obfuscated images I_{out} and reconstructed results I_{rec} for the ablated methods, (1), (2), and (3), and the proposed one (4).

Table 7. Ablation study of block size P.

P	4	8	16	28	112
Acc.	99.42	99.52	99.40	99.42	99.27

Table 8. Ablation study of subband partition.

P_{L}	1	2	3	4	5	6	7
Acc.	98.68	99.20	99.32	99.52	99.27	99.23	99.28

contrast, large blocks ($P \geq 16$) reduce the effectiveness of localized subband manipulation. Therefore, $P = 8$ is adopted as the default block size, providing stable and consistent *MD* performance.

Subband partition: To determine the optimal subband partition, we evaluate seven configurations by varying the size of the LL region P_{L}. Note that per-block DCT coefficients $C \in \mathbb{R}^{P \times P}$ are partitioned into $C_{\text{LL}} \in \mathbb{R}^{P_{\text{L}} \times P_{\text{L}}}$, $C_{\text{LH}} \in \mathbb{R}^{P_{\text{L}} \times P_{\text{H}}}$, $C_{\text{HL}} \in \mathbb{R}^{P_{\text{H}} \times P_{\text{L}}}$, and $C_{\text{HH}} \in \mathbb{R}^{P_{\text{H}} \times P_{\text{H}}}$, where $P_{\text{H}} = P - P_{\text{L}}$. As shown in Table 8, the best accuracy is achieved when $P_{\text{L}} = 4$, which is adopted as the default.

Loss functions: We introduce the coefficient energy constraint loss $\mathcal{L}_{\text{CEC}}$ to preserve the identity-neutral state established by the *Neutralization* module by constraining the overall change in coefficient energy. As shown in Figure 7, removing $\mathcal{L}_{\text{CEC}}$ makes the obfuscated images more vulnerable to reconstruction attacks, since its absence tends to retain information that can reintroduce human-perceptible identity cues in reconstructed images.

To determine an appropriate balancing weight, we vary λ_{CEC} from 0 to 0.1 and measure the *MD* performance, as shown in Table 9. The best *MD* performance is obtained at $\lambda_{\text{CEC}} = 0.01$, which is adopted as the default.

Multi-task applications: To demonstrate the flexibility and scalability of the proposed algorithm, we analyze its performance when optimizing for multiple tasks simultaneously. We employ a single multi-task backbone network, Faceptor [43]. Since co-annotated datasets covering all four tasks are scarce, we instead measure the feature distance in

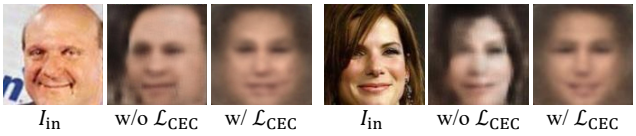

Figure 7. Examples of reconstructed results I_{rec} without (w/o) and with (w/) $\mathcal{L}_{\text{CEC}}$.

Table 9. Ablation study of $\mathcal{L}_{\text{CEC}}$.

λ_{CEC}	0	0.005	0.01	0.05	0.1
Acc.	99.37	99.43	99.53	99.47	99.45

Table 10. Analysis of the multi-task utility preservation.

	Identity	Age	Expression	Attribute
Identity-only	**0.86**	2.74	2.98	4.06
Age-only	1.02	**1.81**	3.17	4.26
Expression-only	1.28	2.65	**1.57**	3.72
Attribute-only	0.95	2.68	3.13	**3.03**
Multi-task	<u>0.88</u>	<u>1.82</u>	<u>1.77</u>	<u>3.09</u>

the Faceptor's embedding space between the original and the obfuscated images, where the task-specific features are extracted at the task-specific layers. Table 10 compares single-task and multi-task optimization. The single-task setting minimizes one *MD* loss for each task independently, whereas the multi-task setting jointly optimizes all losses. The results demonstrate that the proposed method maintains stable *MD* performance in the multiple-task scenario, indicating its strong scalability and general applicability.

5. Conclusions

We presented a novel frequency-domain manipulation framework, called FreM, that adaptively adjusts frequency subbands to balance *HI* and *MD* while enhancing robustness against reconstruction attacks. By applying subband-adaptive neutralization, perturbation, and suppression modules, FreM effectively suppresses identity cues while preserving machine-decipherable cues. Extensive experiments across multiple face analysis tasks and benchmarks demonstrate that FreM consistently outperforms existing methods in both obfuscation quality and robustness against reconstruction attacks, without requiring any additional training.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) (No. RS-2024-00397293, RS-2022-NR068986, RS-2026-25490591), and by the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by MSIT (No. RQT-25-090187).

References

[1] Mohammed Talha Alam, Fahad Shamshad, Fakhri Karray, and Karthik Nandakumar. FaceAnonyMixer: Cancelable faces via identity consistent latent space mixing. In *IJCB*, 2025. 1

[2] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *CVPR*, 2023. 2

[3] Mauro Barni, Franco Bartolini, and Alessandro Piva. Improved wavelet-based watermarking through pixel-wise masking. *TIP*, 10(5):783–791, 2001. 1, 2

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, 2016. 4, 6

[5] Karla Brkić, Ivan Sikirić, Tomislav Hrkać, and Zoran Kalafatić. I Know That Person: generative full body and face de-identification of people in images. In *CVPRW*, 2017. 2

[6] Mahawaga Arachchige Pathum Chamikara, Peter Bertok, Ibrahim Khalil, Dongxi Liu, and Seyit Camtepe. Privacy preserving face recognition utilizing differential privacy. *Computers & Security*, 97:101951, 2020. 2, 4

[7] Durkhyun Cho, Jin Han Lee, and Il Hong Suh. CLEANIR: controllable attribute-preserving natural identity remover. *Applied Sciences*, 10(3):1120, 2020. 2

[8] Umur A Ciftci, Gokturk Yuksek, and Ilke Demir. My Face My Choice: privacy enhancing deepfakes for social media anonymization. In *WACV*, 2023. 2

[9] Alessandro Conti, Paolo Rota, Yiming Wang, and Elisa Ricci. Cluster-level pseudo-labelling for source-free cross-domain facial expression recognition. *arXiv preprint arXiv:2210.05246*, 2022. 4, 6

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 5

[11] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *ICPR*, 2021. 5

[12] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.*, 9(12):2170–2179, 2014. 1

[13] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. 1

[14] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Pearson Education, 2018. 3

[15] Eman T. Hassan, Rakibul Hasan, Patrick Shaffer, David Crandall, and Apu Kapadia. Cartooning for enhanced privacy in lifelogging and streaming videos. In *CVPRW*, 2017. 2

[16] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. 1, 4, 5, 6, 7

[17] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A generative adversarial network for face anonymization. In *ISVC*, 2019. 2

[18] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *CVPR*, 2019. 2, 4

[19] Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, and Rongrong Ji. Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *ECCV*, 2022. 2, 4

[20] Jintae Kim, Seungwon Yang, Seong-Gyun Jeong, and Chang-Su Kim. Forbes: Face obfuscation rendering via backpropagation refinement scheme. In *ECCV*, 2024. 1, 2, 5, 6, 7

[21] Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *CVPR*, 2022. 1

[22] Diederik P. Kingma and Jimmy L. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[23] Deepa Kundur and Dimitrios Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *ICASSP*, 1998. 1, 2

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1

[25] Seon-Ho Lee, Nyeong-Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation. In *NeurIPS*, 2022. 1

[26] Sung Ju Lee and Nam Ik Cho. Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity. In *ICCV*, 2025. 1, 2

[27] Jingzhi Li, Lutong Han, Ruoyu Chen, Hua Zhang, Bing Han, Lili Wang, and Xiaochun Cao. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *ACM MM*, 2021. 1, 2

[28] Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. Learning disentangled representations for identity preserving surveillance face camouflage. In *ICPR*, 2021.

[29] Jingzhi Li, Hua Zhang, Siyuan Liang, Pengwen Dai, and Xiaochun Cao. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Trans. Inf. Forensics Secur.*, 18:3632–3646, 2023. 1, 2

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 4, 7

[31] Yunlong Mao, Shanhe Yi, Qun Li, Jinghao Feng, Fengyuan Xu, and Sheng Zhong. A privacy-preserving deep learning approach for face recognition with edge computing. In *Proc. USENIX Workshop Hot Topics Edge Comput. (HotEdge)*, 2018. 2, 4

[32] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: conditional identity anonymization generative adversarial networks. In *CVPR*, 2020. 2

[33] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler

Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA Janus Benchmark - C: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, 2018. 1

[34] Yuxi Mi, Yuge Huang, Jiazhen Ji, Hongquan Liu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. DuetFace: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In *ACM MM*, 2022. 2, 4

[35] Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using random frequency components. In *ICCV*, 2023.

[36] Yuxi Mi, Yuge Huang, Zhizhou Zhong, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *CVPR*, 2024. 2, 4

[37] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, 2018. 1

[38] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. PrivacyNet: Semi-adversarial networks for multi-attribute face privacy. *IEEE TIP*, 29:9400–9412, 2020. 1, 2

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *CVPRW*, 2017. 1, 4, 5

[40] Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *ECCVW*, 2014. 1

[41] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 1

[42] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE TCSVT*, 34(4):2223–2234, 2024. 1

[43] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. 2024. 1, 6, 7, 8

[44] Karl Ricanek and Tamirat Tesafaye. MORPH: a longitudinal image database of normal adult age-progression. In *FGR*, 2006. 4, 6

[45] Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, and Fernando Alonso-Fernandez. FIVA: facial image and video anonymization and anonymization defense. In *ICCVW*, 2023. 2

[46] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 1, 4, 5

[47] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *CVPR*, 2022. 1, 6

[48] South Korea, National Assembly. Personal Information Protection Act (PIPA), Law No. 16930. 1

[49] State of California. California Privacy Rights Act of 2020 (CPRA). 1

[50] Gregory K. Wallace. The jpeg still picture compression standard. *IEEE Trans. Consum. Electron*, 38(1):xviii–xxxiv, 1992. 2

[51] Tao Wang, Yushu Zhang, Zixuan Yang, Xiangli Xiao, Hua Zhang, and Zhongyun Hua. Seeing is not believing: An identity hider for human vision privacy protection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(2): 170–181, 2025. 1, 2, 4, 5

[52] Yinggui Wang, Jian Liu, Man Luo, Le Yang, and Li Wang. Privacy-preserving face recognition in the frequency domain. In *AAAI*, 2022. 2

[53] Zhibo Wang, He Wang, Shuaifan Jin, Wenwen Zhang, Jiahui Hu, Yan Wang, Peng Sun, Wei Yuan, Kaixin Liu, and Kui Ren. Privacy-preserving adversarial facial features. In *CVPR*, 2023. 2

[54] Lin Yuan, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao. PRO-Face: A generic framework for privacy-preserving recognizable obfuscation of face images. In *ACM MM*, 2022. 1, 5

[55] Lin Yuan, Wu Chen, Xiao Pu, Yan Zhang, Hongbo Li, Yushu Zhang, Xinbo Gao, and Touradj Ebrahimi. PRO-Face C: Privacy-preserving recognition of obfuscated face via feature compensation. *IEEE Trans. Inf. Forensics Secur.*, 19: 4930–4944, 2024. 2, 4

[56] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *ECCV*. Springer, 2022. 1

[57] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 4, 6

[58] Tianyue Zheng and Weihong Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018. 1, 4, 5

[59] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 1, 4, 5