Privacy-Preserving Face Recognition Using Trainable Feature Subtraction

Supplementary Material

This supplementary material provides additional details on the following about the proposed MinusFace method:

- Experimental setup and implementation details;
- Further methodological and experimental discussions;
- Further ablation studies;
- Additional image visualization;
- Ethics discussion.

A. Detailed experimental setup

This section provides information about our experimental setup. We first further introduce the employed datasets and backbones, then discuss our detailed implementations.

A.1. Datasets

Training datasets. We train our FR models on the widely-used MS1Mv2 dataset [13], which consists of 5.8M face images from 85K distinct individuals, mainly celebrities. In accordance with CVPR guidelines, we provide further ethical discussion in Sec. E. Additionally, we utilize the smaller BUPT-BalancedFace dataset [54] (BUPT) for recovery attacks, comprising 1.3M images from 28K identities.

Test datasets. We compare MinusFace and SOTA methods on 7 datasets. (1) We benchmark on five widely-used, regular-sized datasets and report results as test accuracy (with a lower bound of 50%): LFW [17], 13K web-collected images from 5.7K individuals; CALFW [68] and CPLFW [67], reorganized version of LFW that enhances cross-age and cross-pose variations, respectively; AgeDB [40] and CFP-FP [44], similar in size and varied in age and pose. Notably, LFW is often viewed as a saturated dataset, where the highest recognition accuracy is anticipated. Conversely, CFP-FP and CPLFW pose greater challenges to SOTA FR methods due to their increased pose variations. (2) We extend our study to two large-scale benchmarks: IJB-B [59] and IJB-C [32], which provide 80K and 150K still images and video frames, respectively. Results are reported as TPR@FPR(1e-4), *i.e.*, the true positive rate (TPR) at a specific false positive rate (FPR) of 1e-4.

A.2. Backbones

We employ an adapted ResNet50 model with an improved residual unit (IR-50) [14] as the FR backbone for both f and f_p. The only modification to the backbone is changing the input channels of f to 192 to match the number of channels in x. For the generative model g and the attacker's recovery model f^{-1}, we utilize U-Net [42], a popular architecture for image generation and segmentation, which can be considered an autoencoder with skip connections. We employ a full-scale U-Net for f^{-1} to enhance the attacker's capability and a smaller one for g. The latter helps us to reduce MinusFace's local storage and inference time.

A.3. Implementation details

Image preprocessing. We preprocess the training and test datasets using standard methods: We crop faces from the images and align their positions based on the 5-point landmarks of the faces (positions of eyes, nose, and lips). To improve the FR model's generalization, we apply random horizontal flips to the training images.

Encoding and decoding mappings. We opt for DCT and its inverse IDCT as our concrete encoding and decoding mappings. DCT is a popular spatial-frequency transformation first employed by the JPEG compression standard [52]. It converts a $(3, H, W)$ spatial image into $(192, H, W)$ frequency channels. Specifically, the spatial image first undergoes an 8-fold up-sampling to obtain a shape of $(3, 8H, 8W)$. As DCT later divides H and W by 8, this makes sure the resulting frequency channels have consistent shapes. Subsequently, each channel of the image is split into $(8, 8)$-pixel blocks. DCT turns each block into a 1D array of 64 frequency coefficients and reorganizes all coefficients from the same frequency across blocks into an (H, W) frequency channel, that is spatially correlated to the original X. This conversion produces 64 frequency channels from each of the 3 spatial channels. These channels are then stacked to form the final shape of $(192, H, W)$. We further discuss the properties of DCT in Sec. B.1.

Training. Our training involves two stages: First to produce r, we train the generative model g and the FR model f in an end-to-end manner. Then, we freeze g and train f_p on X_p, which is generated from the decoding of randomly shuffled r. For both stages, we train the model from scratch for 24 epochs with a stochastic gradient descent (SGD) optimizer. We choose 64, 0.9, and 1e-4 for batch size, momentum, and weight decay, respectively. The training of f, f_p starts with an initial learning rate of 1e-2, which is successively divided by a factor of 10 at epochs 10, 18, and 22. The learning rate of g is further halved (*i.e.*, starting at 5e-3) since the generative model requires a smaller learning rate to facilitate convergence. We choose the weights for our training objective as $\alpha = 5, \beta = 1$. To help f_p generalize on randomized representation, we augment the dataset 3 times, similar to [36], meaning each X generates three X_p from distinct random shuffling. We apply the same training settings for SOTAs. For the recovery attacker, we train its model until convergence using an initial learning rate of 1e-3. Experiments are carried out in parallel on 8 NVIDIA Tesla V100

GPUs with PyTorch 1.10 and CUDA 11. We use the same random seed for all experiments.

B. Further discussion

B.1. Properties of DCT

DCT is invertible. DCT and IDCT together are invertible as they, by design, map a spatial image to and from its frequency channels *losslessly*. Hence, it naturally holds:

$$d(e(X)) = X. \tag{10}$$

Importantly, it however does not guarantee

$$e(d(x)) = x, \tag{11}$$

unless x is a *valid frequency representation* that is directly encoded via DCT from a spatial image X as $x = e(X)$ (this case satisfies Eq. (6)). In MinusFace, since x', r are not encoded via DCT but derived from regeneration and feature subtraction, they are not frequency representations of X', R' but rather serve as generic high-dimensional representations. Therefore, we can expect

$$e(d(x')) \neq x', e(d(r)) \neq r. \tag{12}$$

This property benefits privacy, as an attacker cannot invert a plausible r from X', R' or X_p. We previously demonstrated this in Sec. 4.5.

DCT is homomorphic. DCT is a linear transformation. Any linear transformation is additively homomorphic by definition. Hence it satisfies

$$d(x_1 + x_2) = d(x_1) + d(x_2). \tag{13}$$

DCT is used differently by MinusFace and SOTAs. Notably, some SOTA PPFR methods [24, 35, 36, 56] also utilize DCT in their pipelines. Both MinusFace and SOTAs are likely inspired by [61], which demonstrates that an image recognition model can perform accurately on DCT components, making DCT an ideal choice for lossless transformation. However, their motivations differ: SOTAs utilize DCT to exploit specific properties of frequency representations (the perceptual disparity among frequency channels), which allows for heuristic channel pruning. On the other hand, MinusFace employs DCT for its invertible and homomorphic mapping properties, as well as its high-dimensional redundancy, which helps to produce identity-informative r and privacy-preserving R'.

DCT is replaceable with other mappings. As discussed in Sec. 3.3, DCT/IDCT can be replaced by other mapping algorithms that satisfy Eq. (3). We present an alternative option, discrete wavelet transform (DWT), in Sec. C.2. Nonetheless, we empirically find that DCT offers a better trade-off between accuracy and privacy.

Method	CFP-FP	AgeDB	CPLFW
IR-18, unprotected	92.31	94.65	89.41
IR-18, MinusFace	90.21	93.25	87.60
CosFace, unprotected	92.89	95.15	89.52
CosFace, MinusFace	89.23	94.77	86.97

Table 4. Compatibility of MinusFace. Combining MinusFace with different FR backbones (IR-18) or losses (CosFace) also sustains high accuracy, compared to their unprotected baselines.

B.2. Recovery result of SOTAs

Some prior studies [24, 35, 56, 57] claim that their proposed methods are resistant to recovery attacks. However, in Fig. 6(b), this paper finds that they can, to some extent, be recovered. We investigate the cause of these distinct experimental outcomes. For the attacker, successful recovery depends on various factors, such as training resources (*e.g.*, the volume of training data) and strategy (*e.g.*, choice of training objective, optimizer, learning rate, and batch size). In fact, this paper examines *a more advanced attacker* than those in SOTAs.

Our attacker exploits a larger training dataset. Among SOTAs, PPFR-FD and DuetFace employ a recovery model trained on just ≤100K face images. AdvFace trains its recovery model on 500K images. In contrast, our analyzed attacker can exploit the entire BUPT dataset, which consists of 1.3M images. The increased data volume can plausibly enhance the attacker's capability.

Our attacker employs an improved training strategy. For recovery, we empirically find that selecting appropriate learning rates and batch sizes is crucial. DCTDP opts for a learning rate of 1e-1 and a batch size of 512, which could be too large for the recovery model to stably converge. In this paper, we choose a learning rate of 1e-3 and a batch size of 64, which we find facilitate recovery.

Despite the advancement in the attacker's capability, MinusFace still effectively prevents the successful recovery of protected faces. This further demonstrates the robust protection provided by MinusFace.

C. Further ablation study

C.1. Choice of FR backbones and objectives

As discussed in Sec. 4.7, MinusFace is compatible with different SOTA FR backbones and training objectives. To illustrate this, we combine MinusFace with a distinct IR-18 FR model and CosFace [53] loss, using BUPT as the training dataset. Table 4 shows the recognition accuracy on CFP-FP, AgeDB, and CPLFW, comparing MinusFace with unprotected baselines. MinusFace maintains a stable performance that is close to the unprotected baseline.

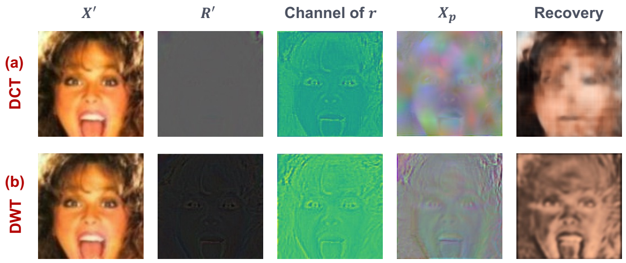

Figure 8. Comparison between DCT and DWT. We replace our mappings with DWT and its inverse and visualize X', R', sample channel of r, X_p, and its recovery for DCT and DWT, respectively. DWT also succeeds in creating an almost blank R' and visually indiscernible X_p, illustrating the effectiveness of MinusFace. However, the attacker can achieve a clearer recovery on DWT, making DCT a more secure choice.

Method	CFP-FP	AgeDB	CPLFW
MinusFace (default)	90.21	93.25	87.60
DWT	90.47	93.53	87.92
DCT, masking	81.40	88.03	82.82

Table 5. Ablation studies of MinusFace. Replacing DCT with DWT achieves comparable recognition accuracy. Yet, replacing shuffling with random masking significantly degrades accuracy.

C.2. Choice of encoding and decoding mappings

To demonstrate a generic pair of encoding and decoding satisfying Eq. (3) may also serve as our mappings e, d, We replace DCT/IDCT with an alternative choice: discrete wavelet transform (DWT). By default, DWT converts the $(3, H, W)$ image into a $(12, H, W)$ representation. We compare its recognition accuracy, concealment of visual images, and protection against recovery to default DCT.

Figure 8 demonstrates the outcomes using DWT and its inverse (IWT) as the mappings, compared with those using DCT/IDCT, respectively. It can be observed that MinusFace effectively produces almost blank R' and visually uninformative X_p for both DCT and DWT. Table 2 further demonstrates that replacing DCT with DWT can achieve on-par recognition accuracy. However, we find that X_p generated via DWT is less resistant to recovery, as face contours can still be observed in the recovered image. We attribute this relative deficiency to two reasons: (1) X_p generated from DWT is less obfuscated in texture details, as its shuffling is less randomized, and (2) DWT produces a significantly smaller random space of 12! (192! as of DCT), which eases the attacker's learning of consistent representations.

The results indicate that DCT can indeed be replaced with other mappings. However, the specific choice of d, e should also be more carefully considered based on the experimental context.

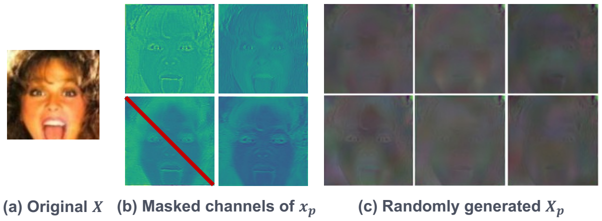

Figure 9. Replacing random shuffling with (b) masking also results in (c) X_p revealing slightly recognizable features at a marginal cost to privacy, which aligns with our expectations.

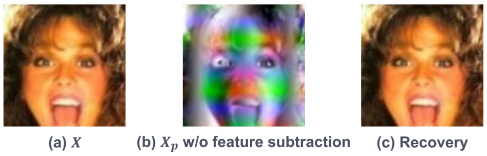

Figure 10. Ablation study of feature subtraction. (b) X_p is directly produced from the high-dimensional representation of (a) X. It reveals clear visual features and is easy to (c) be recovered.

C.3. Choice of perturbation on r

In Sec. 3.4, we opt for random channel shuffling as our perturbation. We here demonstrate it achieves better privacy and accuracy trade-off than other options. Specifically, we generate r as usual yet replace $s(r;\theta)$ with random *masking* of channels $m(r;\theta)$. We produce $X_p = d(m(r;\theta))$, where θ denotes the random seed. We choose a masking ratio of 25%. Figure 9 (b) shows its process.

As per the derivation in Eq. (8), $X_p = d(m(r;\theta))$ should also reveal recognizable features of X at marginal costs to privacy. We observe that the produced X_p in Fig. 9 aligns with our expectations. However, we find that it suffers from a downgrade in recognition accuracy, as shown in Tab. 5, since the features it reveals could be too subtle for FR models to effectively leverage. In this sense, random channel shuffling better balances privacy and accuracy.

C.4. Without feature subtraction

We discuss the importance of feature subtraction in achieving privacy protection. The primary outcome of feature subtraction is to create a recognizable r that precisely maps to a blank R'. By perturbing r, we factually *restore identity features* in R' to produce X_p with minimum privacy cost. Here, we emphasize the critical role of first achieving a blank R'.

To ablate feature subtraction, we directly perform random channel shuffling on x (*i.e.*, the high-dimensional representation of X) to obtain X_p. Figure 10 illustrates the resulting X_p and its recovery, where privacy is completely undermined as X_p contain a wealth of visual features. Recall that Eq. (8) is based on the premise that $d(r) \to 0$. Without

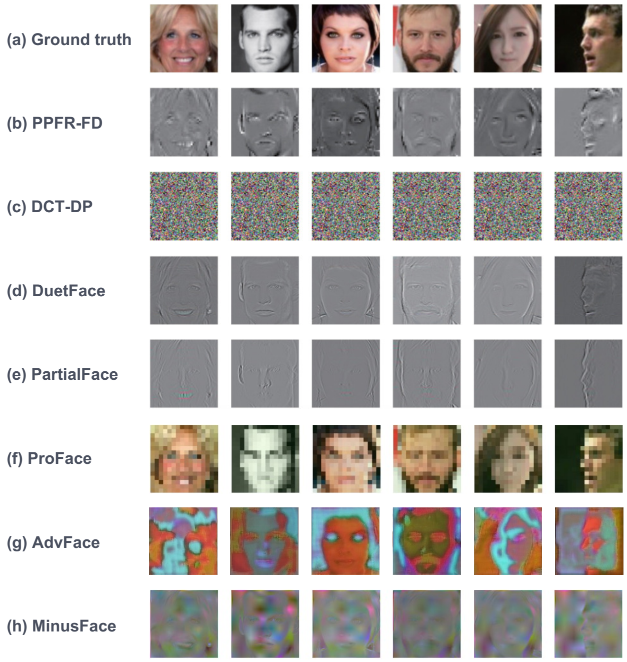

Figure 11. Additional sample images for the protective X_p.

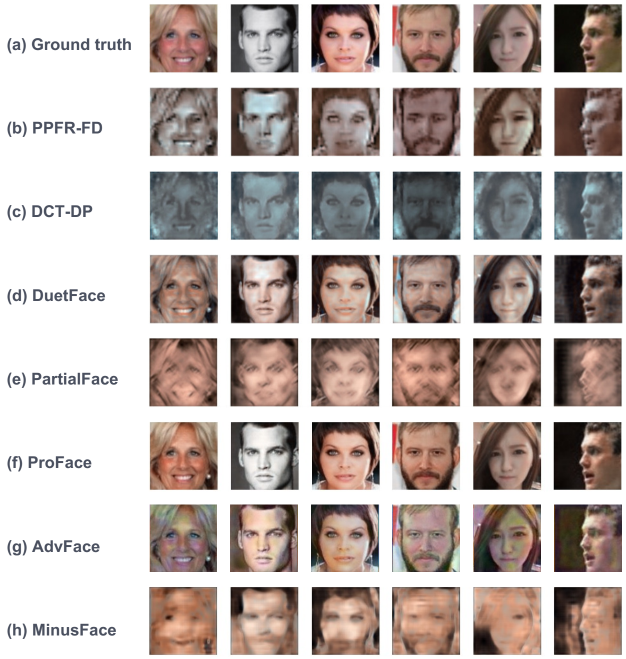

Figure 12. Additional sample images for the recovery from X_p.

feature subtraction, this condition would not be satisfied, as $R' = d(r)$ would not be blank.

D. Additional example images

To demonstrate the generality of MinusFace, we further supplement Fig. 6 with additional example images for the protective representation X_p and its recovery, comparing MinusFace with SOTAs. Specifically, Fig. 11 illustrates X_p and Fig. 12 illustrates the recovery.

E. Ethics discussion

Our models primarily utilize the MS1Mv2 dataset, a modified version of the MS-Celeb-1M (MS1M) dataset provided by the InsightFace project[2], containing celebrity face images. As personal characteristics such as facial semantics may be inferred from the dataset, we are obliged to justify its use per CVPR ethics guidelines. The reasons for using MS1Mv2 include its essential role in ensuring fair comparisons: It is one of the *de facto* standard training datasets in face recognition, and is employed by the majority of the methods [4, 8, 24, 35–37, 50] we compare.

[2] https://github.com/deepinsight/insightface/