

Privacy-Preserving Face Recognition with Learnable Privacy Budgets in Frequency Domain

Jiazhen Ji¹, Huan Wang², Yuge Huang¹, Jiaxiang Wu¹, Xingkun Xu¹,
Shouhong Ding¹, ShengChuan Zhang², Liujuan Cao², and Rongrong Ji²

¹ Youtu Lab, Tencent

² Xiamen University

{royji, yugehuang, willjxwu, xingkunxu, ericding}@tencent.com,
hanawh@stu.xmu.edu.cn, {zsc_2016, caoliujuan, rrji}@xmu.edu.cn

Abstract. Face recognition technology has been used in many fields due to its high recognition accuracy, including the face unlocking of mobile devices, community access control systems, and city surveillance. As the current high accuracy is guaranteed by very deep network structures, facial images often need to be transmitted to third-party servers with high computational power for inference. However, facial images visually reveal the user’s identity information. In this process, both untrusted service providers and malicious users can significantly increase the risk of a personal privacy breach. Current privacy-preserving approaches to face recognition are often accompanied by many side effects, such as a significant increase in inference time or a noticeable decrease in recognition accuracy. This paper proposes a privacy-preserving face recognition method using differential privacy in the frequency domain. Due to the utilization of differential privacy, it offers a guarantee of privacy in theory. Meanwhile, the loss of accuracy is very slight. This method first converts the original image to the frequency domain and removes the direct component termed DC. Then a privacy budget allocation method can be learned based on the loss of the back-end face recognition network within the differential privacy framework. Finally, it adds the corresponding noise to the frequency domain features. Our method performs very well with several classical face recognition test sets according to the extensive experiments. Code will be available at <https://github.com/Tencent/TFace/tree/master/recognition/tasks/dctdp>.

Keywords: Privacy-Preserving, Face Recognition, Differential Privacy

1 Introduction

With the rapid development of deep learning, face recognition models based on convolutional neural networks have gained a remarkable breakthrough in recent years. The extremely high accuracy rate has led to its application in many daily life scenarios. However, due to the privacy sensitivity of facial images and the unauthorized collection and use of data by some service providers, people have become increasingly concerned about the leakage of their face privacy. In

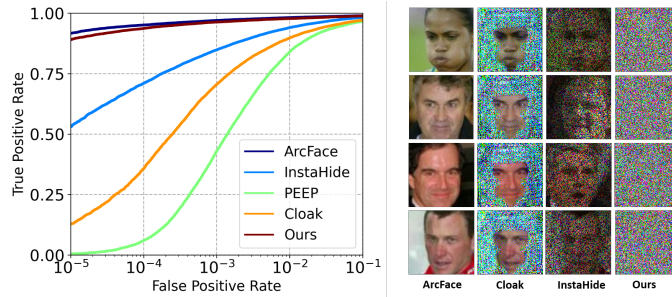


Fig. 1. Comparison with different privacy-preserving methods. Left: ROC curves under IJB-C dataset, the closer to the upper left area, the better the performance. Right: Visualization of processed images using LFW dataset as examples. Compared with other methods, our method has good privacy-preserving performance.

addition to these unregulated service providers, malicious users and hijackers pose a significant danger to privacy leakage. It is necessary to apply some privacy-preserving mechanisms to face recognition.

Homomorphic encryption [11] is an encryption method that encrypts the original data and performs inference on the encrypted data. It protects data privacy and maintains a high level of recognition accuracy. However, the introduction of the encryption process requires a great additional computation. Therefore, it is not suitable for large-scale and interactive scenarios.

PEEP [6] is a very typical approach to privacy protection that makes use of differential privacy. It first converts the original image to a projection on the eigenfaces. Then, it adds noise to it by utilizing the concept of differential privacy to provide better privacy guarantees than other privacy models. PEEP has a low computational complexity, not to slow down the inference speed. However, it significantly reduces the accuracy of face recognition. As shown in Fig. 1, the current face recognition privacy-preserving methods all perform poorly with large data sets and have relatively mediocre privacy-preserving capabilities.

This paper aims to limit the face recognition service provider to learn only the classification result (e.g., identity) with a certain level of confidence but does not have access to the original image (even by some automatic recovery techniques). We propose a privacy-preserving framework to tackle the privacy issues in deep face recognition. Inspired by frequency analysis performing well in selecting the signal of interest, we deeply explored the utility of frequency domain privacy preservation. We use block discrete cosine transform (DCT) to transfer the raw facial image to the frequency domain. It is a prerequisite for separating information critical to visualization from information critical to identification. Next, we remove the direct component (DC) channel because it aggregates most of the energy and visualization information in the image but is not essential for identification. Meanwhile, we consider that elements at different frequencies of the input image do not have the same importance for the identification task. Therefore it is not reasonable to set precisely the same privacy budget for all

elements. We propose a method taking into account the importance of elements for identification. It only needs to set the average privacy budget to obtain the trade-off between privacy and accuracy. Then the distribution of privacy budgets over all the elements will be learned according to the loss of the face recognition model. Compared with PEEP, our approach to switching to the frequency domain is simpler, faster, and easier to deploy. Moreover, because we learn different privacy budgets for different features from the loss of face recognition, our recognition accuracy is also far better than SOTAs. The contributions of this paper are summarized as follows:

- We propose a framework for face privacy protection based on the differential privacy method. The method is fast and efficient and adjusts the privacy-preserving capability according to the choice of privacy budget.
- We design a learnable privacy budget allocation structure for image representation in the differential privacy framework, which can protect privacy while reducing accuracy loss.
- We design various privacy experiments that demonstrate the high privacy-preserving capability of our approach with marginal loss of accuracy.
- Our method can transform the original face recognition dataset into a privacy-preserving dataset while maintaining high availability.

2 Related Work

2.1 Face Recognition

The state-of-the-art (SOTA) research on face recognition mainly improve from the perspective of softmax-based loss function which aims to maximize the inter-class discrepancy and minimize the intra-class variance for better recognition accuracy [8,22,29,33,18]. However, existing margin-based methods do not consider potential privacy leakage. In actual applications, raw face images of users need to be delivered to remote servers with GPU devices, which raises the user’s concern about the abuse of their face images and potential privacy leakage during the transmission process. Our method takes masked face images rather than raw face images as the face recognition model’s inputs, which reduces the risk of misuse of user images.

2.2 Frequency Domain Learning

Frequency analysis has always been widely used in signal processing, which is a powerful tool for filtering signals of interest. In recent years, Some works that introduced frequency-domain analysis have been proposed to tackle the various aspects of the problem.

For instance, [13] trained CNNs directly on the blockwise discrete cosine transform (DCT) coefficients. [36] made an in-depth analysis of the selection of DCT coefficients on three different high-level vision tasks such as image classification, detection, and segmentation. [16] presented a frequency space domain

randomization technique that achieved superior segmentation performance. In the field of deepfake, frequency is an essential tool to distinguish real from synthetic images (or videos) [10]. As for face recognition, [20] presents a new approach through a feature-level fusion of face and iris traits extracted by polar fast fourier transform. [34] splits the frequency domain channels of the image and selects only some of them for subsequent tasks. On the basis of these works, we firstly and deeply explored the utility of privacy-preserving in the frequency domain.

2.3 Privacy Preserving

Privacy-preserving research can be broadly categorized based on how to process input data, i.e., encryption or perturbation. The majority of data encryption methods fall under the Homomorphic Encryption (HE) and Secure Multiparty Computation (SMC) [2,12,19,21]. However, these data encryption methods are unsuitable for current SOTA face recognition systems due to their prohibitive computation cost. Data perturbation, in contrast, avoids the high cost of encryption by applying a perturbation to raw inputs. Differential privacy (DP) is a common-used perturbation approach. [6] applied perturbation to eigenfaces utilizing differential privacy that equally split the privacy budget to every eigenface resulting in a great loss of accuracy. In addition to DP, there are other methods for data disturbance. For instance, [17] used the Mixup [37] method to perturb data while it has been hacked successfully [5]. [25] presented a Gaussian noise disturbance method to suppress unimportant pixels before sending them to the cloud. Recently, some privacy-preserving methods have also appeared in the field of face recognition. [1] proposed federated face recognition to train face recognition models using multi-party data via federated learning to avoid privacy risks. K-same [27] is a de-identification approach by using K-anonymity for face images. Our method maintains accuracy to the greatest extent based on a learnable privacy budget.

3 Method

In this section, we describe the framework of our proposed privacy-preserving face recognition method. Our method consists of three main modules, a frequency domain transformation module, a perturbation module that utilizes differential privacy, and a face recognition module, as shown in Fig.2. Each input image is first converted to frequency domain features by the frequency domain transformation module. Furthermore, the differential privacy perturbation module will generate the corresponding noise and add it to the frequency domain features. Finally, the perturbed frequency domain features will be transferred to the face recognition model. Because the size of the perturbed frequency domain features is $[H, W, C]$, we only need to change the input channels of the face recognition model from 3 to C to suit our input. In Section 3.1, we describe in detail the specific process of frequency-domain conversion. In Section 3.2, we first introduce

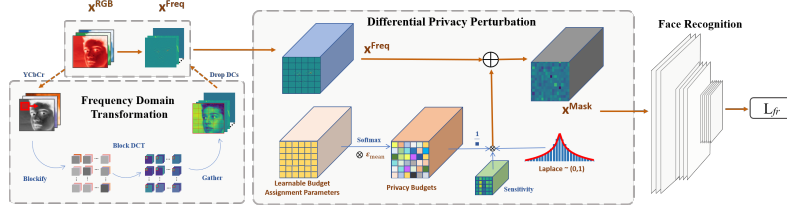


Fig. 2. Overview of our proposed method. It consists of three modules: frequency-domain transformation, differential privacy perturbation, and face recognition. The frequency-domain transformation module transforms the facial image to frequency domain features. The differential privacy perturbation module adds perturbations to frequency domain features. The face recognition module takes the perturbed features as input and performs face recognition.

the background knowledge of differential privacy and then describe the specific improvements in our differential privacy perturbation module.

3.1 Frequency-Domain Transformation Module

[32] discovered that humans rely only on low-frequency information for image recognition while neural networks use low-and high-frequency information. Therefore, using low-frequency information as little as possible is very effective for image privacy protection. DCT transformation can be beneficial in separating the low-frequency information that is important for visualization from the high-frequency information that is important for identification. In the frequency domain transformation module, inspired by the compression operation in JPEG, we utilized block discrete cosine transform (BDCT) as our basis of frequency-domain transformation. For each input image, we first convert it from RGB color spaces to YCbCr color spaces. We then adjust its value range to $[-128, 127]$ to meet the requirement of BDCT input and then split it into $\frac{H}{8} \times \frac{W}{8}$ blocks with a size of 8×8 . For a fairer comparison and as little adjustment as possible to the structure of the recognition network, we perform an 8-fold up-sampling on the facial images before BDCT. Then a normalized, two-dimensional type-II DCT is used to convert each block into 8×8 frequency-domain coefficients.

At this time, we can see that the element in the upper left corner in Fig.3(b) has an extreme value. It defines the basic tone of the entire block, which we call the DC component. For elements in the same position in each block, we collect them together in the same channel and arrange them according to the relative position of the block in the original image. Here we have converted the original image with size $[H, W, 3]$ into a frequency domain representation with size $[H, W, 8, 8, 3]$. As we can see in Fig.3(c), the most energy (91.6%) of facial images is concentrated in the DC channel. According to our experiment results in Tab.3, DC is crucial for visualization, while it has little impact on recognition. Thus it is removed before it enters the next module.

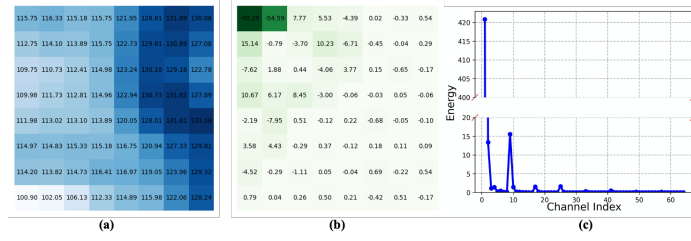


Fig. 3. (a) One block of the Original image with size 8×8 . (b) Frequency-domain features of the image block after DCT transformation. (c) Energies among different channels. DC channels take up most of the energy

3.2 Differential Privacy (DP)

Definition. DP [24] is known to be a privacy model that provides maximum privacy by minimizing the possibility of personal records being identified. By adding noise to the query information of the database, DP helps to ensure the security of personal information while obtaining comprehensive statistical information. For any two databases that differ by only one individual, we say they are adjacent.

A randomized algorithm \mathcal{A} satisfies ϵ -DP if for every adjacent pairs of database D_1 and D_2 and all $\mathcal{Q} \subseteq \text{Range}(\mathcal{A})$, Equation 1 holds.

$$\Pr(\mathcal{A}(D_1) \in \mathcal{Q}) \leq e^\epsilon \Pr(\mathcal{A}(D_2) \in \mathcal{Q}) \quad (1)$$

DP for Generative Models. Since our goal is to protect facial privacy by scrambling the images after BDCT, we need to move from the database domain to the generative model representations domain. Under the generative model representation domain, query sensitivity and adjacency in the database domain are no longer applicable. Each input image should correspond to a separate database. Some methods [7] apply DP generalization to arbitrary secrets, where a secret is any numerical representation of the data.

In our approach, we consider the BDCT representation of the facial image as a secret. The distance between secrets replaces the notion of adjacency between databases. We can control the noise by the distance metric to make similar (in visualization) secrets indistinguishable while keeping very different secrets that remain distinguishable. Thus, the recoverability is minimized while ensuring as much identifiability as possible. Therefore the choice of distance metric for secrets is critical. In our approach, each image is transformed to a BDCT representation with size $[H, W, C]$. Let $R_{i,j,k} = [r_{min}^{i,j,k}, r_{max}^{i,j,k}]$ be the sensitivity of the element in the $[i, j, k]$ position of representation. Then we define an element-wise distance as follows:

$$d_{i,j,k}(x_1, x_2) = \frac{|x_1 - x_2|}{r_{max}^{i,j,k} - r_{min}^{i,j,k}} \quad \forall x_1, x_2 \in R_{i,j,k} \quad (2)$$

Moreover, the distance between the whole representations are defined as follow:

$$\begin{aligned} d(X_1, X_2) &= \max_{i,j,k} (d_{i,j,k}(x_1, x_2)) \\ \forall X_1, X_2 &\in \mathbb{R}^{H,W,C} \end{aligned} \quad (3)$$

Thus, the ϵ -DP protection can be guaranteed for mechanism \mathcal{K} if for any representation X_1, X_2 and $E \in \mathbb{R}^{H,W,C}$. Equation 4 satisfied:

$$\begin{aligned} Pr(\mathcal{K}(X_1) = E) &\leq e^{\epsilon d(X_1, X_2)} Pr(\mathcal{K}(X_2) = E) \\ \forall X_1, X_2, E &\in \mathbb{R}^{H,W,C} \end{aligned} \quad (4)$$

With the definition of distance between representations and DP for representations, we can claim

Lemma 1. *Any BDCT representation of image $X \in \mathbb{R}^{H,W,C}$ can be protected by ϵ -DP though the addition of a vector $Y \in \mathbb{R}^{H,W,C}$ where each $Y_{i,j,k}$ is an independent random variable following a Laplace distribution with a scaling parameter*

$$\sigma_{i,j,k} = \frac{r_{max}^{i,j,k} - r_{min}^{i,j,k}}{\epsilon_{i,j,k}}, \quad \text{where } \sum_{i,j,k} \epsilon_{i,j,k} = \epsilon \quad (5)$$

With the setting in Lemma 1, Equation 4 can be guaranteed. The proof has been attached in Appendix.

Differential Privacy Perturbation Module. The size of the output of the frequency domain transformation module is $[H, W, C]$. We need a noise matrix of the same size to mask the frequency features. Due to the properties of DCT, most energies are gathered in small areas. Thus, the frequency features have very different importance for face recognition. In order to protect face privacy with the best possible face recognition accuracy, we use a learnable privacy budgets allocation method in the module. It allows us to assign more privacy budgets to locations that are important for face recognition.

To achieve the idea of learnable privacy budgets, we utilize the setting in Lemma 1. We first initialize learnable privacy budget assignment parameters with the same size as the frequency features and put them into the softmax layer. Then, we multiply each element of the output of the softmax layer by ϵ . Because the sum of the output after the softmax layer is 1, no matter how the learnable budget assignment parameters are changed, the total privacy budget always equals ϵ . According to Lemma 1, if we add a vector $Y \in \mathbb{R}^{H,W,C}$ to the frequency features where each $Y_{i,j,k}$ is an independent random variable following a Laplace distribution with a scaling parameter $\sigma_{i,j,k} = \frac{r_{max}^{i,j,k} - r_{min}^{i,j,k}}{\epsilon_{i,j,k}}$, then we can ensure that this feature is protected by $\epsilon - DP$. To get the sensitivities of the frequency features of facial images, we transferred all the images in

VGGFace2 [4] and refined MS1MV2 into the frequency domain. Then we obtained the maximum values and minimum values at each position. Sensitivities will equal the value of MAX - MIN. By now, we have prepared all the scaling parameters of Laplace noise. We sample the Laplace distribution according to the parameters and add them to the frequency features. The masked frequency features will be the output of the differential privacy perturbation module and transmitted to the face recognition model. The learnable budget assignment parameters are learned based on the loss function of the face recognition model and get the best allocation scheme that guarantees recognition accuracy. For face recognition module, we use ArcFace [8] as loss function and ResNet50 [14] as backbone.

4 Experiment

4.1 Datasets

We use VGGFace2 [4] that contains about 3.31M images of 9131 subjects for training. We extensively test our method on several popular face recognition benchmarks, including five small testing datasets and two general large-scale benchmarks. LFW [15] contains 13233 web-collected images from 5749 different identities. CFP-FP [30], CPLFW [38], CALFW [38] and AgeDB [26] utilize the similar evaluation metric of LFW to test face recognition with various challenges, such as cross pose, cross age. IJB-B [35] and IJB-C [23] are two general large-scale benchmarks. The IJB-B dataset contains 12,115 templates with 10,270 genuine matches and 8M impostor matches. The IJB-C dataset is a further extension of IJB-B, having 23,124 templates with 19,557 genuine matches and 15,639K impostor matches.

4.2 Implementation Details

Each input face is resized to 112×112 . After the frequency domain transformation, the input channel C is set to 189. We set the same random seed in all experiments. Unless otherwise stated, we use the following setting. We train the baseline model on ResNet50 [14] backbone. For our proposed model, we first convert the raw input RGB image to BDCT coefficients using some functions in TorchJPEG [9]. Secondly, we process the BDCT coefficients using the proposed method. We calculate the sensitivity among the whole training dataset. The initial values of the learnable budget allocation parameters are set to be 0 so that the privacy budget of each pixel is equal in the initial stage. The whole model is trained from scratch using the SGD algorithm for 24 epochs, and the batch size is set to be 512. The learning rate of the learnable budget allocation parameters and the backbone parameters is 0.1. The momentum and the weight-decay are set to be 0.9 and $5e-4$, respectively. We conducted all the experiments on 8 NVIDIA Tesla V100 GPU with the PyTorch framework. We divide the learning rate by 10 at 10, 18, 22 epochs. For ArcFace [8], we set $s = 64$ and $m = 0.4$. For CosFace [31], we set $s = 64$ and $m = 0.35$.

4.3 Comparisons with SOTA Methods

Settings for Other Methods. To evaluate the effectiveness of the model, we compare it with five baselines: **(1) ArcFace** [8]: The model is ResNet50 equipped with ArcFace, which is the simplest baseline with the original RGB image as input and introduces an additive angular margin inside the target cosine similarity. **(2) CosFace** [31]: The model is ResNet50 equipped with CosFace, which is another baseline with the original RGB image as input and subtracts a positive value from the target cosine similarity. **(3) PEEP** [6]: This method is the first to use DP in privacy-preserving face recognition. We reproduce it and run it on our benchmarks. Due to a large amount of training data, half of the data is selected for each ID to calculate the eigenface. The privacy budget ϵ is set to 5. **(4) Cloak** [25]: We run its official code. Note that when training large datasets, the privacy-accuracy parameter is adjusted according to our experiment setting, which is set to 100. **(5) InstaHide** [17]: This method incorporates mix-up to solve the privacy-preserving problem. We adapt it to face privacy-preserving problems. We set k to 2 and adopt an inside-dataset scheme that mixes each training image with random images within the same private training dataset.

Method (%)		Privacy-Preserving LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	IJB-B(1e-4)	IJB-C(1e-4)
ArcFace (Baseline)	No	99.60	98.32	95.88	94.16	92.68	91.02	93.25
CosFace (Baseline)	No	99.63	98.52	95.83	93.96	93.30	90.79	93.14
PEEP	Yes	98.41	74.47	87.47	90.06	79.58	5.82	6.02
Cloak	Yes	98.91	87.97	92.60	92.18	83.43	33.58	33.82
InstaHide	Yes	96.53	83.20	79.58	86.24	81.03	61.88	69.02
Ours, Arcface ($\epsilon_{mean} = 0.5$)	Yes	99.48	97.20	94.37	93.47	90.6	89.33	91.22
Ours, Cosface ($\epsilon_{mean} = 0.5$)	Yes	99.47	97.16	94.13	93.36	90.88	89.37	91.21

Table 1. Comparison of the face recognition accuracy among different privacy-preserving face recognition methods.

Results on LFW, CFP-FP, CPLFW, CALFW, AgeDB. The results of the comparison with other SOTA methods can be seen in Tab.1. For our method, the accuracy on all five datasets is close. Our method has a similar performance to the baselines on LFW and CALFW, with only an average drop of 0.14% and 0.65%. For CFP-FP, AgeDB, and CPLFW, our method has an average drop of 2.5% compared with the baseline. We believe this is because the images in these datasets have more complex poses and are therefore inherently less robust and more susceptible to interference from noise. However, our method still has a considerable lead in performance on these datasets compared to other SOTA privacy-preserving methods. In particular, on the CFP-FP dataset, other privacy-preserving methods have accuracy losses of more than 10%, but we still perform well.

Results on IJB-B and IJB-C. We also compare our method with baseline and other SOTA privacy-preserving methods over IJB-B and IJB-C. As shown

in Tab.1 and Fig.1, our method has a very similar performance compared to the baseline. Under different false positive rates, the true positive rate of our method is still at a high level. However, the other SOTA methods do not perform well in this area. Their true-positive rate is much lower than the baseline with the lower false-positive rate.

4.4 Privacy Attack

White-box Attacking Experiments. We assume that the attacker already knows all our operations in the white-box attack section. Therefore he will perform an inverse DCT (IDCT) operation on the transmitted data. Since our operation to remove DCs is practically irreversible, we assume he fills all DCs to 0. We set the privacy budget to 0.5 and 100, respectively, to better demonstrate the effect. However, in practice, we do not recommend setting the privacy budget as large as 100. Further, we denoise the images after IDCT. Here we use non-local means denoising [3] as the denoising method. As shown in Fig.4, after losing the information of DC, it is difficult to show the original facial information in the recovered figure of the white box attack. Even if the privacy budget is as large as 100, we cannot obtain valid information about the user’s facial features from the recovered image. Moreover, at this time, the denoising method also cannot work effectively because the IDCT image is full of noise.

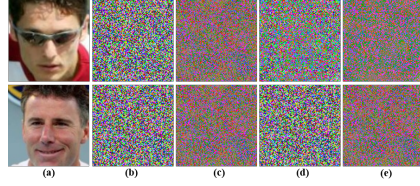


Fig. 4. White-box attack for our method. (a) Raw image. (b) IDCT with $\epsilon=0.5$. (c) Denoising image of (b). (d) IDCT with $\epsilon=100$. (e) Denoising image of (d).

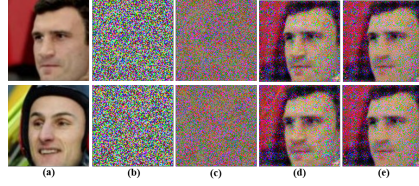


Fig. 5. White-box attack for our method with guessed DC. For both line, we add the DC of the upper user. (a) Raw image. (b) IDCT with $\epsilon=0.5$. (c) Denoising image of (b). (d) IDCT with $\epsilon=100$. (e) Denoising image of (d).

To further demonstrate our approach’s inability to provide valid information about the user’s facial features, we assume that the attacker knows this data originates from a specific dataset. He makes certain guesses about the user to whom the data belongs and adds the DC of the guessed user to the data. The corresponding results are shown in Fig.5. We assumed that the attacker guessed exactly the corresponding user in the above row. However, with the privacy budget set to 0.5, the recovery image still does not reveal any facial information. In the following row, we assume that the data belong to another user. However,

the attacker still guesses that the data belongs to the same user as in the above row. In other words, the attacker guesses the wrong data source and adds the wrong DC to it. The recovery image does not give any information to the attacker about whether he guesses the right user to whom the data belongs. In this way, we can say that our method can protect the user’s privacy well.









	Org	Cloak	InstaHide	Ours	Metric	Method	LFW	CFP-FP	CPLFW	AgeDB	CALFW
					PSNR(db)	Cloak	20.292	20.069	19.705	19.946	20.327
						InstaHide	23.519	22.807	22.546	23.271	23.010
						Ours	14.281	13.145	13.515	13.330	13.330
					Similarity	Cloak	0.564	0.464	0.578	0.574	0.526
						InstaHide	0.729	0.649	0.732	0.737	0.693
						Ours	0.214	0.175	0.264	0.250	0.202

Fig. 6. Visualization of different privacy-preserving methods under black-box attack.

Fig. 7. PSNRs and similarities between the original images and the ones recovered from the output of different privacy-preserving methods. The lower the value is, the better the privacy-preserving method is.

Black-box Attacking Experiments. In this section, we analyze the privacy-preserving reliability of the proposed method from the perspective of a black-box attack. A black box attack means that the attacker does not know the internal structure and parameters of the proposed model. However, attackers can collect large-scale face images from websites or other public face datasets. They can obtain the processed inputs by feeding those data to the model. Subsequently, they can train a decoder to map the processed inputs to the original face images. Finally, attackers can employ the trained decoder to recover the user’s face image. Under these circumstances, we use UNet [28] as our decoder to reconstruct original images from processed images. In the training phase, we use an SGD optimizer with a learning rate of 0.1 with 10 epochs, and the batch size is set to 512. As shown in Fig. 6, we compare our method with other methods on reconstructed images. For Cloak, the face is still evident since the added noise only affects the face’s background. For InstaHide, most encrypted images can be recovered under our setting. For our method, the reconstructed images have been blurred. The facial structure of recovered images has been disrupted with an average privacy budget of 0.5.

In addition, Fig. 7 shows some quantitative results to illustrate the effectiveness of our method further. The reported results correspond to the results in Tab. 1. We first compare the PSNR between the original image and the one reconstructed. PSNR is often used to measure the reconstruction quality of lossy compression. It is also a good measure of image similarity. As we can find in the first row of Fig. 7, compared with other methods, our method has lower PSNR and higher recognition accuracy. Furthermore, we also evaluate the average Feature Similarity across five small data sets in order to show the privacy-preserving

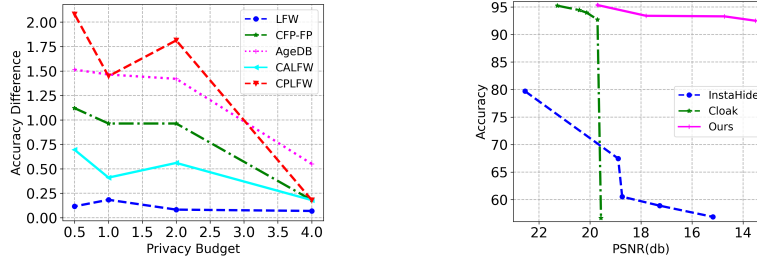


Fig. 8. **Left:** The recognition accuracy among different settings of privacy budget under different data set. The lower the value is, the less impact the method has on recognition accuracy. **Right:** Trade-off abilities of different approaches. The position in the upper right-hand corner indicates greater resilience to recovery while maintaining accuracy.

ability of the model at the feature level, as shown in the second row of Fig. 7. Precisely, we feed recovered RGB images to a pre-trained Arcface backbone. Then we can get new feature embedding to perform feature similarity calculation with origin embedding. The smaller the similarity, the more recovery-resistant the privacy-preserving method is at the feature level. As we can see, when the average privacy budget is chosen to be lower than 1, the similarities are much smaller than the other methods.

We further experimented with the trade-off ability of different approaches for privacy and accuracy, as shown in Fig. 8. Cloak does not protect the privacy of the face to a high degree, so the PSNR of the recovered image under black-box attack and the original image can hardly be lower than 19. In contrast, InstaHide can protect face privacy quite well as the number of mixed images increases. However, the resulting loss of accuracy is very heavy. For our approach, depending on the choice of the average privacy budget, the degree of face privacy protection can be freely chosen, and the loss of accuracy can be controlled to a small extent. In summary, our method is more resistant to recovery than other methods, which means a more robust privacy-preserving ability.

4.5 Ability to protect the privacy of training data.

Unlike other privacy-preserving methods, we can protect privacy during the inference and training stages. After the first training, we can get fixed privacy budget allocation parameters. We perform a DCT operation on the existing raw face recognition dataset and add noise according to the privacy budget allocation parameters. Thus, the original privacy leaked dataset is transformed into a privacy-protected dataset. Moreover, we experimentally demonstrate that the transformed dataset can still be used to train face recognition tasks with high accuracy. As shown in Tab. 2, the model trained using privacy-preserving datasets still has high accuracy. Notably, even though we trained fixed privacy

Method (%)	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
VGGFace2	99.60	98.32	95.88	94.16	92.68
MS1M	99.76	97.94	98.00	96.10	92.30
Privacy-Preserved VGGFace2	99.68	97.88	95.85	93.97	92.11
Privacy-Preserved MS1M	99.73	96.94	97.96	95.96	91.73

Table 2. Comparison of recognition accuracy training with different datasets. **VGGFace2/MS1M:** Training with the original VGGFace2/MS1M. **Privacy-Preserved VGGFace2/MS1M:** Training with the dataset that transformed from the original VGGFace2/MS1M. They used the same fixed privacy budget allocation parameters learned from VGGFace2 and $\epsilon = 2$.

assignment parameters using VGGFace2 and transformed MS1M using it, the transformed privacy-preserving MS1M retained high usability. It demonstrated the versatility and separability of the privacy budget allocation parameters. The corresponding results are shown in the fourth row.

4.6 Ablation Study

Effects of transformation to the frequency domain. To demonstrate the effects of transformation to the frequency domain, we directly input the raw RGB facial image to the differential privacy perturbation module. Here we choose $\epsilon_{mean} = 0.5$. As shown in Tab.3, the model without transformation to frequency domain has a much lower accuracy even with the same setting for other parts.

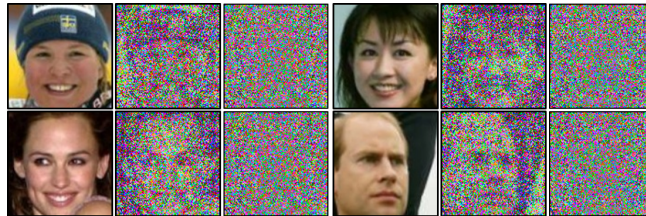


Fig. 9. From left to right, the original image, the IDCT image without DC removal and the IDCT image with dc removal.

Effects of removing DCs. To show the effects of removing DCs, we train models with and without removing DCs for comparison. We chose a larger privacy budget to make the difference between removing DC and not removing DC more visible. Here we set the budget equal to 20. By comparing the IDCT visualization of masked images without removing DCs shown in Fig.9, we can see that the removal of DCs has a good effect on facial information protection.

Method (%)	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
ArcFace (Baseline)	99.60	98.32	95.88	94.16	92.68
BDCT-DC	99.68	98.3	95.93	94.25	93.28
BDCT-NoDC	99.48	98.39	95.73	94.35	92.92
RGB ($\epsilon_{mean} = 0.5$)	84.78	63.87	73.66	76.85	63.08
Ours ($\epsilon_{mean} = 0.5$)	99.48	97.20	94.37	93.47	90.6

Table 3. Comparison of the face recognition accuracy among methods with or without privacy protection. **BDCT-DC/BDCT-NoDC**: The baseline model trained on BDCT coefficients with/without DC. **RGB**: The model trained on the original image with learnable privacy budgets.

Effects of using learnable DP budgets. To demonstrate the effects of using learnable DP budgets, we compare the accuracy of using learnable DP budgets and using the same DP budget for all elements. We set the privacy budget to be 0.5 at every element and test its accuracy using a pre-trained model with noiseless frequency domain features as input. The accuracy over LFW is only 65.966%, which is much lower than the one with learnable DP budgets, shown in Tab.3.

Effects of choosing different DP budgets. To show the effect of choosing different privacy budgets, we chose different privacy budgets and tested their loss of accuracy relative to the baseline. The results of the tests are presented in Fig.8. As we have seen, the smaller the privacy budget is set, the higher the accuracy loss will be. This result is in line with the theory of differential privacy that a smaller privacy budget means more privacy-preserving and has a correspondingly higher accuracy loss. In our approach, the smaller privacy budget also provides stronger privacy protection, which is proven in Section 4.4.

5 Conclusions

In this paper, we propose a privacy-preserving approach for face recognition. It provides a privacy-accuracy trade-off capability. Our approach preserves image privacy by transforming the image to the frequency domain and adding random perturbations. It utilizes the concept of differential privacy to provide strong protection of face privacy. It regulates the ability of privacy protection by changing the average privacy budget. It is lightweight and easily compatible to be easily added to existing face recognition models. Moreover, it is shown experimentally that our method can fully protect face privacy under white-box attacks and maintain similar accuracy as the baseline. The masked images can defend against the black-box recovery attack of UNet. These show that our method performs far better than other SOTA face recognition protection methods.

References

1. Bai, F., Wu, J., Shen, P., Li, S., Zhou, S.: Federated face recognition. arXiv preprint arXiv:2105.02501 (2021)
2. Boemer, F., Cammarota, R., Demmler, D., Schneider, T., Yalame, H.: Mp2ml: a mixed-protocol machine learning framework for private inference. In: Proceedings of the 15th International Conference on Availability, Reliability and Security. pp. 1–10 (2020)
3. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 60–65. IEEE (2005)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
5. Carlini, N., Deng, S., Garg, S., Jha, S., Mahloujifar, S., Mahmood, M., Song, S., Thakurta, A., Tramer, F.: An attack on instahide: Is private learning possible with instance encoding? arXiv preprint arXiv:2011.05315 (2020)
6. Chamikara, M.A.P., Bertók, P., Khalil, I., Liu, D., Camtepe, S.: Privacy preserving face recognition utilizing differential privacy. *Computers & Security* **97**, 101951 (2020)
7. Croft, W.L., Sack, J.R., Shi, W.: Obfuscation of images via differential privacy: from facial images to general images. *Peer-to-Peer Networking and Applications* **14**(3), 1705–1733 (2021)
8. 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)
9. Ehrlich, M., Davis, L., Lim, S.N., Shrivastava, A.: Quantization guided jpeg artifact correction. Proceedings of the European Conference on Computer Vision (2020)
10. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. pp. 3247–3258. PMLR (2020)
11. Gentry, C., Halevi, S.: Implementing gentry’s fully-homomorphic encryption scheme. In: Annual international conference on the theory and applications of cryptographic techniques. pp. 129–148. Springer (2011)
12. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International conference on machine learning. pp. 201–210. PMLR (2016)
13. Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., Yosinski, J.: Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems* **31**, 3933–3944 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: 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)
16. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsd: Frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6891–6902 (2021)

17. Huang, Y., Song, Z., Li, K., Arora, S.: Instahide: Instance-hiding schemes for private distributed learning. In: International Conference on Machine Learning. pp. 4507–4518. PMLR (2020)
18. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricular-face: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
19. Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: GAZELLE: A low latency framework for secure neural network inference. In: 27th USENIX Security Symposium (USENIX Security 18). pp. 1651–1669. USENIX Association, Baltimore, MD (Aug 2018), <https://www.usenix.org/conference/usenixsecurity18/presentation/juvekar>
20. Kagawade, V.C., Angadi, S.A.: Fusion of frequency domain features of face and iris traits for person identification. Journal of The Institution of Engineers (India): Series B pp. 1–10 (2021)
21. Liu, J., Juuti, M., Lu, Y., Asokan, N.: Oblivious neural network predictions via minionn transformations. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp. 619–631 (2017)
22. 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)
23. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)
24. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). pp. 94–103. IEEE (2007)
25. Mireshghallah, F., Taram, M., Jalali, A., Elthakeb, A.T.T., Tullsen, D., Esmaeilzadeh, H.: Not all features are equal: Discovering essential features for preserving prediction privacy. In: Proceedings of the Web Conference 2021. pp. 669–680 (2021)
26. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
27. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE transactions on Knowledge and Data Engineering **17**(2), 232–243 (2005)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
29. 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)
30. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
31. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)

32. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8681–8691. IEEE (2020)
33. Wang, X., Zhang, S., Wang, S., Fu, T., Shi, H., Mei, T.: Mis-classified vector guided softmax loss for face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12241–12248 (2020)
34. Wang, Y., Liu, J., Luo, M., Yang, L., Wang, L.: Privacy-preserving face recognition in the frequency domain (2022)
35. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 90–98 (2017)
36. Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K., Ren, F.: Learning in the frequency domain. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
37. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
38. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)