Privacy-Preserving Face Recognition in the Frequency Domain

Yinggui Wang*[1], Jian Liu[1], Man Luo[1], Le Yang[2], Li Wang[1]

[1]Ant Group, [2]University of Canterbury
wyinggui@gmail.com, {rex.lj, sankuai.luoman}@antgroup.com, le.yang@canterbury.ac.nz, raymond.wangl@antgroup.com

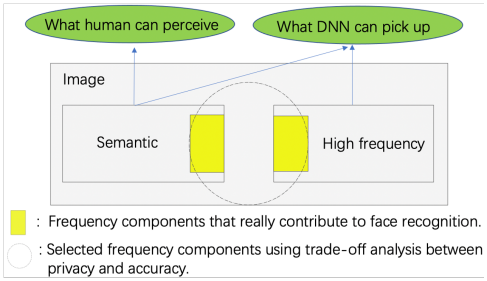
Abstract

Some applications require performing face recognition (FR) on third-party servers, which could be accessed by attackers with malicious intents to compromise the privacy of users' face information. This paper advocates a practical privacy-preserving frequency-domain FR scheme without key management. The new scheme first collects the components with the same frequency from different blocks of a face image to form component channels. Only part of the channels are retained and fed into the analysis network that performs an interpretable privacy-accuracy trade-off analysis to identify channels important for face image visualization but not crucial for maintaining high FR accuracy. For this purpose, the loss function of the analysis network consists of the empirical FR error loss and a face visualization penalty term, and the network is trained in an end-to-end manner. We find that with the developed analysis network, more than 94% of the image energy can be dropped while the face recognition accuracy stays almost undegraded. In order to further protect the remaining frequency components, we propose a fast masking method. Effectiveness of the new scheme in removing the visual information of face images while maintaining their distinguishability is validated over several large face datasets. Results show that the proposed scheme achieves a recognition performance and inference time comparable to ArcFace operating on original face images directly.

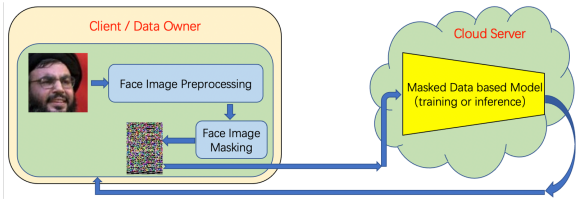
Introduction

The capability of accurately recognizing faces as passwords while preserving the privacy of users' face information is essential for the success of applications based on the client-server model. This paper presents a practical face image recognition scheme without key management. It can remove a significant part of the visual information in original images for privacy protection and increase the difficulty of face recovery, while retaining their distinguishability.

Fig. 1(a) illustrates image perception in the frequency domain. Human perception of images mainly depends on the semantic information of images (i.e., visualization or low-frequency components), while existing deep neural network (DNN) based face recognition (FR) systems rely on both low- and high-frequency components. In (Wang et al. 2020),



(a)



(b)

Figure 1: (a) Image perception in the frequency domain; (b) Framework of privacy-preserving FR.

a heuristic analysis of the impact of high-frequency information on image classification is conducted. But it does not identify which frequency components are crucial for image classification. In this paper, we shall propose a trade-off analysis network that can score the importance of each frequency component in images. This lays down the foundation for designing image processing techniques for new accurate FR systems that can explore processed face images with little visual information. Fig. 1(b) shows the framework of the privacy-preserving FR that we adopt. During model training, data owners perform some preprocessings, such as face detection, alignment and anti-spoofing, and then mask the face images. These masked images are utilized to training the FR model. During inference, the built-in camera at the client side captures the image of the user's face and then sends it after masking to the cloud server for face recognition. After FR is done, the encrypted identification result will be transferred to the client side.

The model used for FR is one trained using the masked

*Corresponding author. All authors contribute equally.

Method	ACC	Time cost
ArcFace (Deng et al. 2019)	**99.82**	2.86ms
FR-FD	**99.82**	2.82ms
FR-FD-noDC	99.69	**2.01ms**
PPFR-FD-DC	**99.82**	2.92ms
PPFR-FD	99.68	2.23ms
DP (Chamikara et al. 2020)	95.33	2.88ms
InstaHide (Huang et al. 2020)	98.76	2.99ms
CPGAN (Tseng and Wu 2020)	98.91	21.37ms

Table 1: Comparison of accuracy (ACC) and computational efficiency among methods with and without privacy protection in the face recognition process for the LFW dataset.

data instead of the original face images. This makes the proposed system satisfy the regulatory requirements that original face images cannot be directly utilized in both training and inference, or recovered easily from masked data. On the basis of deep convolutional neural networks (DCNNs), multiple FR methods have become available, including the softmax loss-based algorithms (Cao et al. 2018), triplet-loss-based algorithms (Schroff, Kalenichenko, and Philbin 2015), Sphereface (Liu et al. 2017), CosFace (Wang et al. 2018) and ArcFace (Deng et al. 2019). In this paper, we adopt ArcFace to illustrate the use and performance of our privacy-preserving scheme (see Table 1 for part of the experiment results over the LFW dataset).

Our contributions are as follows:

- We propose a network that performs an interpretable privacy-accuracy trade-off analysis to identify channels important for face image visualization/perception but not crucial for attaining a high FR accuracy.
- A new loss function is devised for the analysis network, which consists of the empirical FR error loss and a face visualization quality penalty term. The network is trained in an end-to-end manner.
- We propose a practical privacy-preserving FR scheme, which achieves satisfactory trade-off between privacy and accuracy, has a fast inference time, and can be incorporated into existing face recognition algorithms without significant network structure modifications.

Related Work

Learning in the Frequency Domain

Representing images in the frequency domain provides rich patterns for various tasks. In (Xu, Zhang, and Ren 2018; Wu et al. 2018), autoencoder-based networks are trained to jointly perform image compression and inference. In (Ehrlich and Davis 2019), a model conversion algorithm is developed to convert the spatial-domain CNNs into frequency-domain CNNs. In (Gueguen et al. 2018), image classification is carried out using frequency-domain features. A learning-based method is established in (Xu et al. 2020) to identify trivial frequency components that can be discarded without degrading classification performance. This observation is utilized in the development of our face masking method.

Privacy-Preserving CNNs

Currently available privacy-preserving CNNs (PP-CNNs) can be roughly categorized into CNNs with cryptography theories and the ones without using cryptography. PP-CNNs with cryptography theories are built using privacy-preserving deeping learning techniques that incorporate knowledge of cryptography. The secure multiparty computation (SMC) (Mohassel and Zhang 2017; Makri et al. 2019; Ma et al. 2019; Wagh et al. 2021; Wagh, Gupta, and Chandran 2019; Mohassel and Rindal 2018) and homomorphic encryption (HE) (Acar et al. 2018; Naresh Boddeti 2018) are typical examples. But this kind of methods generally have high computation cost. PP-CNNs without using cryptography operates on images whose pixels and color channels have been 'randomly' perturbated (Tanaka 2018; Sirichotedumrong, Kinoshita, and Kiya 2019; Madono et al. 2020). One disadvantage of these networks is that their recognition accuracy is evidently lower that their counterparts using original images. Morphed learning (Mole) developed in (Shen et al. 2019) is another PP-CNN method without using cryptography. This approach is similar to transfer learning (Yosinski et al. 2014). But MoLe exposes the augmented convolutional layer to the developer who can deduce the morphing matrix by designing images of specific structures. With the morphing matrix, the original images can be restored. The algorithms in (Xiang et al. 2019; Zhang et al. 2020) belong to the privacy protection mode of encryption and decryption, but key management is not an easy task. (Mireshghallah et al. 2019) and (Tseng and Wu 2020; Liu et al. 2019) protect the privacy of the inference process by learning noise distribution with generative adversarial networks (GAN). The privatizers are composed of multiple CNNs, which leads to huge increase in computation. InstaHide (Huang et al. 2020) encrypts each training image with the key obtained through mixing a number of randomly chosen images and applying a random pixel-wise mask, but it may still be hacked (Carlini et al. 2020).

Privacy-Preserving Face Recognition

This subsection reviews some privacy-preserving techniques for FR. (Chamikara et al. 2020) presents the privacy using EigEnface Perturbation (PEEP) method. It perturbates eigenfaces using differential privacy (Gong et al. 2020) and only stores perturbed data at the third-party servers that run the eigenface algorithm. But the FR accuracy of PEEP is much worse than that using original face images. In (Guo, Xiang, and Li 2019), a scheme employing encryption and decryption operations similar to those in (Xiang et al. 2019; Zhang et al. 2020) to achieve privacy-preserving FR in the cloud is proposed. (Ma et al. 2019) improves existing additive secret sharing-based functions and establishes a lightweight privacy-preserving ensemble classification algorithm for FR. This method has a relatively high complexity and long execution time.

Methodology

This section describes the proposed frequency-domain privacy-preserving FR scheme, referred to as PPFR-FD.

Block Discrete Cosine Transform (BDCT)

As the very first step of the proposed method, BDCT is carried out on the face image obtained after converting it from a color image to a gray one. Specifically, in a way similar to the convolution operation in CNN, BDCT is performed on image blocks with a size of $a \times b$ pixels according to the stride s. For each image block, an $a \times b$ BDCT coefficient matrix is generated, in which every element represents a particular frequency component. We collect, from all the image blocks, the frequency components that have the same position in the BDCT coefficient matrix to form one frequency component channel. The leftmost part of Figs. 2(a) and (b) gives an illustration of the above process. As an example, suppose the size of an original face image is 112×112 pixels. We set the block size to be 8×8 pixels and stride to be $s = 8$ pixels. In total, we can obtain 64 frequency component channels, each of which has a dimensionality of 14×14.

Privacy-Accuracy Trade-off Analysis

We propose an analysis network, as shown in Fig. 2(a) for analyzing the trade-off between privacy and FR accuracy. It can be seen that it takes the BDCT frequency component channels extracted from the original face image as input and first passes them through a channel selection module. The purpose of this module is to remove frequency component channels with amplitudes close to zero, which contribute little to the distinguishability and visualization of face images. The observation that discarding small-amplitude frequency components would not greatly degrade the classification accuracy was obtained when examining non-face image datasets such as ImageNet (Xu et al. 2020). The channels selected for different images for removal could be different if e.g., a trained SENet (Hu et al. 2020) is used. In this work, for simplicity, we choose to remove a pre-fixed subset of frequency component channels that span low to relatively high frequencies. Specifically, with the size of BDCT image blocks being 8×8 pixels, only 36 out of 64 channels are kept. The basis for selecting the number of channels will be discussed in the section of Security Evaluation.

Training the FR model can be guided by a loss function such as Arcface, but it is not easy to quantify face images' privacy protection level. To bypass this difficulty, we approximate the degree of privacy protection from the perspective of image visualization. If it is difficult to restore the visualization information from a processed image, it is thus considered that the degree of privacy protection is high. It is generally believed that the frequency components with high energy have large contribution to image visualization. Based on this observation, we calculate the absolute value of the elements in each frequency channel, perform averaging over each channel, and take the results as the channel energies. These values will be weighted by learnable parameters. Here, we use the absolute values of the frequency components, instead of the square of them, in order to avoid the penalty term being dominated by frequency channels with large energy levels. The loss function of privacy protection is thus given by

$$Loss_{pri} = \sum_{i=0}^{M} ReLu\left(a_i - \gamma\right).p_i \tag{1}$$

where a_i is the trainable weight coefficient for the ith channel, p_i is the energy of channel i, γ denotes a threshold, and M is the number of considered frequency channels. Clearly, if a_i is less than γ, the corresponding channel is considered unimportant in terms of its contribution to the loss function $Loss_{pri}$. This is realized by the use of $ReLu(y)$ function, which becomes zero when y is negative.

To constrain the value of a_i to the range between 0 and 1, we apply the transformation $a_i = 1/\left(1 + exp\left(-x_i\right)\right)$, where x_i is the parameter to be learned in the training process. The composite loss function is

$$Loss_{analysis-network} = Loss_{FR} + \lambda \cdot Loss_{pri} \tag{2}$$

where $Loss_{FR}$ is the loss function of FR and λ is a hyper parameter. By minimizing the loss function, the frequency channels with higher energy tend to be suppressed while attempting to maintaining a good FR accuracy using the remaining channels only (i.e., reducing $Loss_{FR}$ at the same time). This is desired, because we aim to remove frequency channels that contribute significantly to image visualization, which are mostly channels with large energy.

Note that this training process does not have dedicated protection for the image privacy. Here, we suggest two approaches to address this issue. The first method uses publicly available face data sets with distribution similar to that of data owners' images as training data for the analysis network. We can then carry out FR classifier training after removing the designated channels of data owners' images. The other method is to simply neglect the lowest-score channel directly. Through experiments, it is found that the lowest-score channel is the lowest frequency channel or the direct current (DC) component, which contributes greatly to image visualization, and accounts for more than 90% of the energy of images, but not much to the recognition task (see also Fig. 3 and Fig. 5(b) for an illustration). In order to obtain the best trade-off between privacy and accuracy, the following steps are based on the second method. To further protect the remaining frequency components, we propose a fast and effective masking method.

Fast Face Image Masking

The whole diagram of the face image masking method is shown in Fig. 2(b). It performs the BDCT and selects channels according the analysis network (note again that only the DC component is discarded for a high FR accuracy). Next, the remaining channels are shuffled two times with a channel mixing in between. After each shuffling operation, channel self-normalization is performed. The result of the second channel self-normalization is the masked face image that will be transmitted to third-party servers for face recognition. The proposed face masking method aims at further

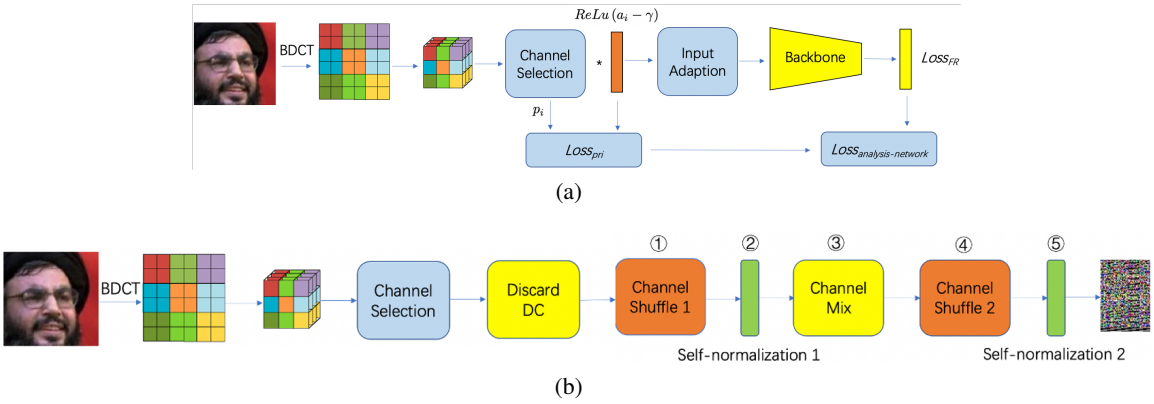


Figure 2: Schematic diagrams of (a) the proposed analysis network and (b) the proposed fast face image masking method.

Figure 3: Two pictures (a)-(b) we randomly selected to extract the DC components (c)-(d).

increasing the difficulty in recovering the original face image from its masked version. We shall detail the operation of each module below. We denote the ith retained frequency component channel from image k in a dataset as $\mathbf{F}_i^k$, where $i = 0, 1, ..., 34$ (recall that the DC component has been discarded), $k = 1, 2, ..., N$ and N is the total number of face images in the dataset.

The remained 35 channels are shuffled in order to increase the difficulties in deducing their associated frequencies. Mathematically, after the first channel shuffling operation Shuffle1$(\cdot)$, the retained channels are

$$\mathbf{E}_j^k = \text{Shuffle1}\left(\mathbf{F}_i^k\right) \tag{3}$$

where $j = 0, 1, \ldots, 34$ and $\mathbf{E}_j^k$ is the jth channel of face image k after channel shuffling.

Next, we perform self-normalization on $\mathbf{E}_j^k$ via

$$\bar{\mathbf{E}}_j^k = (\mathbf{E}_j^k - \mu_1^k)./\sigma_1^k. \tag{4}$$

Note that the above operation is *element-wise* in the sense that each element in $\mathbf{E}_j^k$ is self-normalized individually. In particular, $./$ denotes element-wise division. μ_1^k and σ_1^k denote two matrices whose elements are the sample mean and sample standard deviation of all the entries in $\mathbf{E}_j^k$. In other words, each face image would have its own self-normalization parameters μ_1^k and σ_1^k. This is fundamentally different from the normalization process in e.g., (Xu et al. 2020), where the normalization parameters are calculated using all the images in the dataset (i.e., all images *share* the same normalization parameters). Clearly, the self-normalization procedure adopted in our face masking method is more secure, as the disclosure of the self-normalization parameters for some images will not affect the privacy of the remaining ones.

We then mix $\bar{\mathbf{E}}_j^k$ through linearly combining them using

$$\mathbf{M}_j^k = (\bar{\mathbf{E}}_j^k + \bar{\mathbf{E}}_{j+1}^k)/2 \tag{5}$$

where $j = 0, 1, \ldots, 33$. With the channel mixing operation, the number of frequency component channels is reduced by one. This further enlarges the difficulties in reconstructing the frequency-domain information of the original face image from the mixing output $\mathbf{M}_j^k$.

$\mathbf{M}_j^k$ will go through another channel shuffling Shuffle2 $(\cdot)$ to obtain

$$\mathbf{S}_l^k = \text{Shuffle2}\left(\mathbf{M}_j^k\right) \tag{6}$$

where $l = 0, 1, \ldots, 33$. Finally, $\mathbf{S}_l^k$ is self-normalized as

$$\bar{\mathbf{S}}_l^k = (\mathbf{S}_l^k - \mu_2^k)./\sigma_2^k. \tag{7}$$

Here, as in (4), μ_2^k and σ_2^k are two matrices whose elements are the sample mean and sample standard deviation of all the entries in $\mathbf{S}_l^k$. $\bar{\mathbf{S}}_l^k$ are the masked version of face image k and they will be transmitted to third-party servers for training and inference.

Note that Shuffle1 is pseudo-random in order to ensure that each channel is processed in a fixed order. However, Shuffl2 is completely random to enhance privacy protection, and it is random for all images in the dataset.

Privacy-Preserving Face Recognition

This subsection summarizes the proposed frequency-domain privacy-preserving face recognition (PPFR-FD) scheme that is based on the face masking method. As shown in Fig. 4(a), PPFR-FD adopts the existed FR model, such as ArcFace, as the face classifier. To protect the privacy of users' face data, both the training and recognition stages of PPFR-FD are carried out using the *masked* face images only.

In the training stage, all the original face images in the training dataset are converted to their masked version using the fast face masking method. Note that the order of the

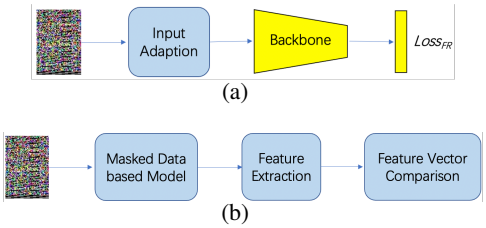


Figure 4: (a) Training process of the proposed PPFR-FD scheme. (b) Inference of the deployed PPFR-FD scheme.

combined frequency channels in each masked data is different. Before the masked data is input into the network, the channels are sorted according to a certain rule, such as the energy of channels. On the other hand, the masked face images have more than three frequency component channels (see (7)). Moreover, the dimensionality of each channel is smaller than the size of input image of ArcFace (e.g., 14×14 vs. 112×112). To address these aspects, we modify ArcFace by increasing the number of its input channels and upsample each frequency component channel of masked images to the size of 112×112 pixels. In this way, except for the input layer, the remaining part of ArcFace is kept intact, which avoids extra workload due to significantly changing the ArcFace network architecture. Another advantage is that the proposed PPFR-FD scheme can provide an inference time close to that of ArcFace operating on original face images.

The face recognition process of the trained PPFR-FD is shown in Fig. 4(b). A user's face image is first masked and transmitted to ArcFace to perform template matching-based recognition. A successful match will be declared if the distance between the features extracted from the masked data and pre-stored template is less than a threshold.

Security Goal and Evaluation

The goal of PPFR-FD is to provide a light-weight masking method to make it difficult for attackers to recover the training and inference face images in the FR system (Fig. 1 (b)). Similar to comparison algorithms in Tabs. 1 and 2, it is not designed to provide the level of security strength as encryption methods such as RSA (Rivest, Shamir, and Adleman 1978). It is also an initial step towards exploring better privacy preservation while maintaining data utility. Here, we shall establish the security of the PPFR-FD scheme under the condition that the flow of the face masking process is known but detailed knowledge on channel shuffling, mixing and self-normalization parameters is unavailable. Results obtained under different attacking experiments on the face masking algorithm are given in Section 'Experiments'.

The first challenge for an attacker to recover the original face image is that we discard more than 90% of the energy of images, which makes it difficult to reconstruct the original image in the attack mode like GAN. The second challenge is that the number of frequency component channels in the face masking algorithm output is smaller than that of the retained BDCT channels. This is owing to the use of channel mixing between the two shuffling operations (see (5)). Reconstructing data from its lower-dimensional version is difficult without knowledge on e.g., the data structure. Moreover, face image recovery requires reversing the two channel shuffling operations. Consider the typical setup with a BDCT block size of 8×8 pixels and 35 retained frequency component channels. There would be a total number of $35! \times 34!$ possible shuffling operations. In fact, since the randomized channels are sorted according to the channel energy in Shuffle2, the space for brute force search caused by two shuffles would be reduced, but it is still much larger than $35!$. That is, the size of the search space is much greater than that of the 128-bit AES encryption algorithm. This greatly increases the difficulty of deducing the shuffling operations via the method of brute force search. The shuffle operation is only a part of the masking algorithm, and the security of the masking algorithm is constructed by multiple steps.

Here, we explain why the number of channels selected is 36. In terms of accuracy, too many frequency channels will not significantly improve the accuracy. In terms of data size, too many frequency channels will increase the size of desensitized data. In terms of security, 36 channels are selected so that the brute force cracking space from two shuffles is much greater than 2^{128}; from the perspective of cryptography, the size of the brute force space is safe enough under the existing computing power. So this number is a tradeoff choice based on some aspects, and not the only choice.

Note that the mask method proposed is a part of the FR process as in Fig.1(b). Inputs must undergo some preprocessing operations (face anti-spoofing, face detection) to reach the masking stage. So it's hard to probe masking steps by using specially designed or chosen images. Even if all steps of masking are provided to the attacker (i.e., in a white-box attack), it is still difficult to recover original images, which is justified in details in the following section and Part 1 of the appendix. Meanwhile, the sorting operation can be used as a part of the white box attack to recover the raw image from the masked data. See Section 'White-box Attacking Experiments' and the appendix for specific operations.

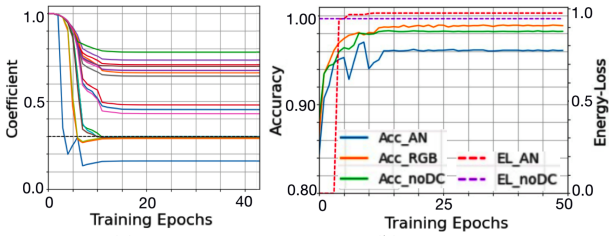
Experiments

In this section, we first evaluate the proposed analysis network for trade-off analysis between privacy and accuracy. Performance comparisons of different algorithms over standard face datasets are carried out, following by attacking experiments and discussions for PPFR-FD.

Experiments on Analysis Network

In order to simplify the process and gain insights, we train the baseline model on MobileNetV2 (Sandler et al. 2019) backbone using the ArcFace loss. The head of the baseline model is: BackBone-Flatten-FC-BN with embedding dimensions of 512 and dropout probability of 0.4 to output the embedding feature. All models are trained for 50 epochs using the SGD optimizer with the momentum of 0.9, weight decay of 0.0001. For the threshold γ in (1), we set it to 0.3. λ in (2) is set to 1. We use CASIA (Yi et al. 2014) and LFW (Zhang and Deng 2016) as the training and test datasets.

Fig 5. (a) shows the evolution of channel weighting coefficients a_i with training epochs. For clarity, we only show the



(a) (b)

Figure 5: Schematic diagrams of the evolution of (a) channel coefficients, (b) accuracy of face recognition and energy loss with training epochs.

weights of the first 16 channels. The considered 16 channels are indexed in a descending order according to their energies. It can be seen from the figure that the channel coefficient is set to 1 at the beginning of the training. After 50 epochs, there are 6 channels whose coefficient values are less than 0.3. The indexes of these 6 channels are 0, 1, 2, 8, 9 and 15. Note that although the DC component is given the lowest score, not all channels with large energies have scores lower than the threshold. Fig 5. (b) compares recognition accuracy and energy loss with training epochs. Acc_{AN}, Acc_{RGB} and Acc_{noDC} represent accuracy of the analysis network without constraining the number of discarded channels, for RGB images without removing any channels and for frequency domain learning with DC component removed only. EL_{AN} and EL_{noDC} represent the energy loss of the analysis network and using frequency domain learning without DC. It can be seen that by abandoning 6 channels, the energy loss of the image has exceeded 97%, but the accuracy drops about 3%. With only the DC component discarded, the energy loss of the image is over 94%, but the FR accuracy is marginally affected. This reflects the effectiveness of the analysis network in identifying channels not crucial for face classification. More importantly, it reveals that the best trade-off between accuracy and privacy is to discard the lowest-score channel, i.e., the DC channel.

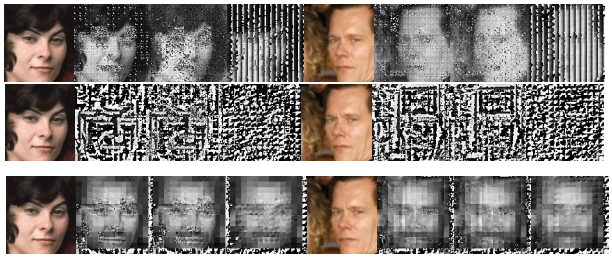
Experiments of PPFR-FD and Others

We use the MS-Celeb-1M dataset with 3,648,176 images from 79,891 subjects as the training set. 7 benchmarks including LFW (Zhang and Deng 2016), CFP-FP (Sengupta et al. 2016), AgeDB (Moschoglou et al. 2017), CALFW (Zhang and Deng 2016), CPLFW (Zhang and Deng 2016) and Vggface2 (Cao et al. 2018)are used to evaluate the performance of PPFR-FD following the standard evaluation protocols. We train the baseline model on ResNet50 (He et al. 2016) backbone with SE-blocks (Hu et al. 2020) and batch size of 512. Other settings are the same as that in Section 'Experiments on Analysis Network'.

Results are reported in Table 2. ArcFace and FR-FD denote the face recognition algorithms for RGB images and in the frequency domain without privacy protection, respectively. In FR-FD and FR-FD-noDC, the network inputs are data generated by the operation Choose Channels before "Discard DC" and "Channel Shuffle 1" in Fig. 2(b). PPFR-FD-DC denotes PPFR-FD without discarding DC. For fair comparison, the classification algorithms in the three considered methods are replaced by the FR algorithm in this section. In DP, ϵ is set to 5. For InstaHide, we set k to 3, that is, only one public image used for encryption. For the convenience, we use a fixed public image for image synthesis. The accuracy obtained by using the same public image for encryption in training and testing is better than that in the large-scale dataset mentioned in InstaHide. In Table 1, we also calculate the total running time of each algorithm, including masking and inference time, on the LFW dataset. In Table 2, although PPFR-FD reduces more than 94% energy of images and is masked, its performance is comparable to that of the baseline model and is better than that of other similar algorithms. The performance of DP is the worst. Although CPGAN also has better performance, it takes longer time to execute and can only protect privacy in inference.

Attacking Experiments for PPFR-FD

In this section, the privacy protection reliability of PPFR-FD is analyzed based on white-box and black-box attack experiments. Black-box means that the attacker does not know the structure and parameters of the model, but can access the results of the model. White-box means that the attacker can not only access the results of the model, but also know the structure of the model.The purpose of attack experiments is to reconstruct the original image.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6: White-box attack experiments for PPFR-FD-DC (the 1st row), PPFR-FD (the 2nd row), PPFR-FD utilizing the DC component of Fig. 3(a) (the 3rd row). (a)(e): Raw images; settings of (b)-(d) and (f)-(h) are listed in Table 3.

White-box Attacking Experiments We first do white-box attack experiments for PPFR-FD with DC, i.e., PPFR-FD-DC. Some results are showed in the 1st row of Fig. 6. The detailed method and some tricks used by white-box reconstruction in Fig. 6 is shown in the appendix. Table 3 gives the experiment settings for recovering face images. Figures 6(a)(e) shows the original gray face images of two different persons. The images in the 2nd row of Fig. 6 show white-box attack experiments for PPFR-FD. Due to the lack of the DC component, the reconstructed images are rather fuzzy. In Fig. 6 (the 3rd row), we also consider replacing the missing DC with the DC components from other pictures. We add the DC components of Fig. 3(a) to the reconstructed frequency

Method	Energy loss	Mask	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Vggface2
ArcFace (Deng et al. 2019)	0	No	99.82	97.78	97.85	96.02	92.77	95.12
FR-FD	0	No	99.82	96.59	97.83	95.93	91.35	94.62
FR-FD-noDC	94.28	No	99.69	95.97	97.82	95.73	91.02	94.36
PPFR-FD-DC	>0	Yes	99.82	96.76	97.88	96.02	91.58	94.70
PPFR-FD	**$>$94.28**	**Yes**	**99.68**	**95.04**	**97.37**	**95.72**	**90.78**	**94.08**
DP (Chamikara et al. 2020)	–	Yes	95.33	92.28	93.61	90.01	85.44	89.13
InstaHide (Huang et al. 2020)	–	Yes	98.76	94.15	95.76	93.49	88.90	93.17
CPGAN (Tseng and Wu 2020)	–	Yes	98.91	94.56	97.07	94.82	90.47	93.29

Table 2: Comparison of the face recognition accuracy among methods with and without privacy protection in the face recognition process for different datasets. The 2nd row shows results of methods without privacy protection, the 3rd-5th rows are results of the ablation experiments for PPFR-FD, and the 6th-9th rows show results of methods with privacy protection.

Known operations	All operations in Fig. 2(b)
Fig. 6(b)(f): 1,2,3,4,5	1 or 4: Channel Shuffle 1, 2
Fig. 6(c)(g): 2,3,4,5	2 or 5: Self-normalization 1, 2
Fig. 6(d)(h): 2,3,5	3: Channel Mix

Table 3: Ablation experiment settings for recovering images.

channels. From Fig. 6 (the 3rd row), we can see that the reconstructed image contour is close to that of the face image that provides the DC component. This illustrates that the algorithm sacrifices a little degradation in the face recognition accuracy, but it has high security.

The reason of poor reconstruction performance is that the distinguishability of face images exists in higher frequency, these frequency values are small, and small errors in them will lead to great impact on the visualization. The above image attack/reconstruction experiments are also some security analysis experiments with some prior information of face images and the face image processing algorithm. From another point of view, the PPRF-FD method can protect the privacy of face images to a certain extent, which can make Client/ Server applications more secure.

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 7: Face images recovered by GAN. (a)(e): Original images; (b)(f) are recovered images from masked data generated by PPFR-FD with 'Shuffle2' using a pseudo-random sequence; (c)(g) are ones generated by PPFR-FD with 'Shuffle2' using fully-random sequences; (d)(h) are recovered images from sorted-channel masked data generated by PPFR-FD with 'Shuffle2' using fully-random sequences.

Black-box Attacking Experiments Here, we use GAN, which has a strong fitting ability, to carry out black-box attack. In Fig. 7, we adopt the combined model of Pix2Pix (Isola et al. 2017) and StyleGAN2 (Karras et al. 2020) as the black-box attack GAN model to reconstruct original images from the masked images. The generetor in Pix2Pix is replaced by the one from StyleGAN2 (Karras et al. 2020) which is pretrained with the Flickr-Faces-HQ dataset (Karras, Laine, and Aila 2019), and the discriminator is the 6-layer PatchCNN (Isola et al. 2017). In the training phase, we use Adam optimizer with a learning rate begins at 0.0002. We carry out the GAN attack experiment on PPFR-FD. The results show that it is difficult to reconstruct face images even using GAN. More results are shown in the appendix.

Limitations

This paper is similar to the work in literatures (Tseng and Wu 2020; Huang et al. 2020) that solves the problem of face privacy protection in the process of training and/or inference. From the perspective of biological template protection (Nandakumar and Jain 2015), it also needs to address the revocability and unlinkability of biological template, which is beyond the scope of this work. In addition, the design of a special network for frequency-domain data is also worthy of investigations. This paper does not consider the privacy protection for the preprocessing steps, which also needs further research. However, we can alleviate this problem with the hardware-level protection, such as TEE (Mo et al. 2020), which can build a secure computing environment.

Conclusion

In this paper, we proposed an analysis network that performs an interpretable privacy-accuracy trade-off analysis to identify channels important for face image visualization but not crucial for maintaining a high FR accuracy. Based on trade-off analysis bewteen privacy and accuracy, a PPFR-FD scheme was developed and its security was established analytically. The proposed scheme has the advantage of being able to achieve the best trade-off between privacy and accuracy, have a fast inference time, and be incorporated into existing face recognition algorithms without significant network structure modifications.

References

Acar, A.; Aksu, H.; Uluagac, A. S.; and Conti, M. 2018. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Comput. Surv.*, 51(4).

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 67–74.

Carlini, N.; Deng, S.; Garg, S.; Jha, S.; Mahloujifar, S.; Mahmoody, M.; Song, S.; Thakurta, A.; and Tramer, F. 2020. An Attack on InstaHide: Is Private Learning Possible with Instance Encoding? arXiv:2011.05315.

Chamikara, M. A.; Bertok, P.; Khalil, I.; Liu, D.; and Camtepe, S. 2020. Privacy Preserving Face Recognition Utilizing Differential Privacy. *Computers and Security*, 97.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 4685–4694.

Ehrlich, M.; and Davis, L. 2019. Deep residual learning in the JPEG transform domain. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October: 3483–3492.

Gong, M.; Xie, Y.; Pan, K.; Feng, K.; and Qin, A. K. 2020. A Survey on Differentially Private Machine Learning [Review Article]. *IEEE Computational Intelligence Magazine*, 15(2): 49–64.

Gueguen, L.; Sergeev, A.; Liu, R.; and Yosinski, J. 2018. Faster neural networks straight from JPEG. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, (NeurIPS): 1–12.

Guo, S.; Xiang, T.; and Li, X. 2019. Towards efficient privacy-preserving face recognition in the cloud. *Signal Processing*, 164: 320–328.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December: 770–778.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8): 2011–2023.

Huang, Y.; Song, Z.; Li, K.; and Arora, S. 2020. Instahide: Instance-hiding schemes for private distributed learning. In *International Conference on Machine Learning*, 4507–4518. PMLR.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Liu, S.; Shrivastava, A.; Du, J.; and Zhong, L. 2019. Better accuracy with quantified privacy: representations learned via reconstructive adversarial network. arXiv:1901.08730.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep hypersphere embedding for face recognition. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January: 6738–6746.

Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; and Ren, K. 2019. Lightweight privacy-preserving ensemble classification for face recognition. *IEEE Internet of Things Journal*, 6(3): 5778–5790.

Madono, K.; Tanaka, M.; Onishi, M.; and Ogawa, T. 2020. Block-wise Scrambled Image Recognition Using Adaptation Network. *arXiv*, (Lowe 1999).

Makri, E.; Rotaru, D.; Smart, N. P.; and Vercauteren, F. 2019. EPIC: Efficient Private Image Classification (or: Learning from the Masters). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11405 LNCS: 473–492.

Mireshghallah, F.; Taram, M.; Ramrakhyani, P.; Tullsen, D. M.; and Esmaeilzadeh, H. 2019. Shredder: Learning Noise to Protect Privacy with Partial DNN Inference on the Edge. *CoRR*, abs/1905.11814.

Mo, F.; Shamsabadi, A. S.; Katevas, K.; Demetriou, S.; Leontiadis, I.; Cavallaro, A.; and Haddadi, H. 2020. Darknetz: towards model privacy at the edge using trusted execution environments. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 161–174.

Mohassel, P.; and Rindal, P. 2018. ABY3: A mixed protocol framework for machine learning. *Proceedings of the ACM Conference on Computer and Communications Security*, 35–52.

Mohassel, P.; and Zhang, Y. 2017. SecureML: A System for Scalable Privacy-Preserving Machine Learning. *Proceedings - IEEE Symposium on Security and Privacy*, 19–38.

Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; and Zafeiriou, S. 2017. AgeDB: The First Manually Collected, In-the-Wild Age Database. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July: 1997–2005.

Nandakumar, K.; and Jain, A. K. 2015. Biometric Template Protection: Bridging the performance gap between theory and practice. *IEEE Signal Processing Magazine*, 32(5): 88–100.

Naresh Boddeti, V. 2018. Secure Face Matching Using Fully Homomorphic Encryption. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–10.

Rivest, R. L.; Shamir, A.; and Adleman, L. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2): 120–126.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015: 815–823.

Sengupta, S.; Chen, J. C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*.

Shen, J.; Liu, J.; Chen, Y.; and Li, H. 2019. Towards efficient and secure delivery of data for deep learning with privacy-preserving. *arXiv*.

Sirichotedumrong, W.; Kinoshita, Y.; and Kiya, H. 2019. Pixel-Based Image Encryption without Key Management for Privacy-Preserving Deep Neural Networks. *IEEE Access*, 7(Ml): 177844–177855.

Tanaka, M. 2018. Learnable image encryption. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 1–2. IEEE.

Tseng, B.; and Wu, P. 2020. Compressive Privacy Generative Adversarial Network. *IEEE Transactions on Information Forensics and Security*, 15: 2499–2513.

Wagh, S.; Gupta, D.; and Chandran, N. 2019. SecureNN: 3-Party Secure Computation for Neural Network Training. *Proceedings on Privacy Enhancing Technologies*, 2019(3): 26–49.

Wagh, S.; Tople, S.; Benhamouda, F.; Kushilevitz, E.; Mittal, P.; and Rabin, T. 2021. F: Honest-Majority Maliciously Secure Framework for Private Deep Learning. *Proceedings on Privacy Enhancing Technologies*, 2021(1): 188–208.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8681–8691.

Wu, C. Y.; Zaheer, M.; Hu, H.; Manmatha, R.; Smola, A. J.; and Krahenbuhl, P. 2018. Compressed Video Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6026–6035.

Xiang, L.; Ma, H.; Zhang, H.; Zhang, Y.; and Zhang, Q. 2019. Complex-valued neural networks for privacy protection. *arXiv preprint arXiv:1901.09546*.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y. K.; and Ren, F. 2020. Learning in the frequency domain. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1737–1746.

Xu, K.; Zhang, Z.; and Ren, F. 2018. LAPRAN: A scalable laplacian pyramid reconstructive adversarial network for flexible compressive sensing reconstruction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11214 LNCS: 491–507.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning Face Representation from Scratch. arXiv:1411.7923.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 4(January): 3320–3328.

Zhang, H.; Chen, Y.; Ma, H.; Cheng, X.; Ren, Q.; Xiang, L.; Shi, J.; and Zhang, Q. 2020. Rotation-Equivariant Neural Networks for Privacy Protection. *arXiv preprint arXiv:2006.13016*.

Zhang, N.; and Deng, W. 2016. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *2016 International Conference on Biometrics, ICB 2016*, 1–11.