Hiding Visual Information via Obfuscating Adversarial Perturbations

Zhigang Su[1,*] Dawei Zhou[1,*] Nannan Wang[1,†] Decheng Liu[1,†]
Zhen Wang[2] Xinbo Gao[3]
[1]Xidian University, [2]Zhejiang Lab
[3]Chongqing University of Posts and Telecommunications
{zgsu, dwzhou}.xidian@gmail.com, {nnwang, dchliu}@xidian.edu.cn
wangzhen@zhejianglab.com, gaoxb@cqupt.edu.cn

Abstract

*Growing leakage and misuse of visual information raise security and privacy concerns, which promotes the development of information protection. Existing adversarial perturbations-based methods mainly focus on the de-identification against deep learning models. However, the inherent visual information of the data has not been well protected. In this work, inspired by the Type-I adversarial attack, we propose an Adversarial Visual Information Hiding (AVIH) method to protect the visual privacy of data. Specifically, the method generates **obfuscating adversarial perturbations** to obscure the visual information of the data. Meanwhile, it **maintains the hidden objectives to be correctly predicted by models**. In addition, our method does not modify the parameters of the applied model, which makes it flexible for different scenarios. Experimental results on the recognition and classification tasks demonstrate that the proposed method can effectively hide visual information and hardly affect the performances of models. The code is available at https://github.com/suzhigangssz/AVIH.*

1. Introduction

Deep neural networks (DNNs) have been widely applied in the computer vision [26, 10, 18]. However, the increasing leakage and misuse of visual information has raised serious concerns [45, 20], especially in fields such as face [27, 28, 51] and medicine [5, 1, 12]. A representative case is the security issue of data stored in the cloud environment [19, 4, 35]. Due to potential vulnerabilities in cyberspace [3], uploaded private images can be easily stolen and used maliciously [38]. Therefore, it is meaningful and urgent to explore effective strategies to protect visual information.

A classic strategy is visual information hiding, which mainly consists of two types [40]: Homomorphic Encryption (HE)-based methods [2, 46, 32], Perceptual Encryption (PE)-based methods [42, 40, 39, 8, 16] and steganography method [24, 50]. Affected by the nonlinear activation functions in DNNs, HE-based methods are difficult to perform well on advanced DNNs [40]. Although PE-based methods apply to DNNs, existing methods typically require retraining with data in the encrypted domain to guarantee accuracy on encrypted data [42, 40]. This affect the performance of service model on raw data and cause additional resource consumption (especially for large models). The steganography method can hide the visual information of sensitive data into another image, but it does not guarantee that the service model can correctly recognize the protected image.

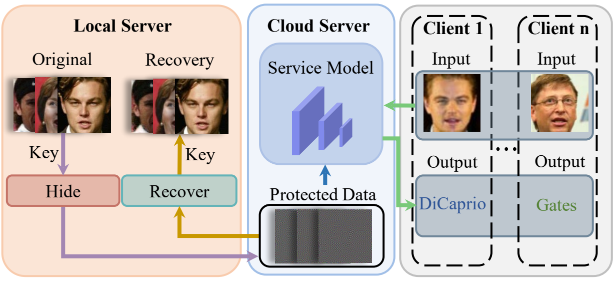


Figure 1. Illustration of visual information hiding for face recognition systems in cloud environments. The gallery set is protected and provided to the DNN in the cloud. Protected image has quite different visual information from the original image, but it is still correctly identified. The protected gallery set can be recovered by a key model by the owner.

To alleviate these negative effects, we expect to hide the visual information without making any modifications to service model. Namely, we hide sensitive visual information only by varying the input image. Previous researches have made some explorations in this regard. The transformation network-based methods [16], which try to protect the original image via a transformation function parameterized by

[*]Equal contributions. † Corresponding author.

a neural network, share the same philosophy. However, the method cannot easily recover the original image from the protected image for other purposes. They may suffer from adversarial vulnerability [9, 14] because the introduced neural network may be destroyed by adversarial attacks.

In fact, the negative effects of adversarial attacks can be utilized positively to protect privacy. Some works have exploited adversarial attacks for de-identification [36, 34, 48]. These methods add imperceptible perturbations or non-suspicious patches generated by adversarial attacks to original images, hindering DNNs to extract effective features and recognize identities, thus protecting the identity privacy in the image. However, in this work, we focus on visual information hiding, which means that the protected image is entirely different from the original image visually but can still be correctly predicted by DNNs (see Figure. 1). Fortunately, we observe a special type of adversarial attack (called Type-I attack) [43] quite different from the type used in previous methods. This type of attack guides DNNs to make consistent predictions on two distinct samples.

Inspired by the Type-I adversarial attack, in this paper, we propose an *Adversarial Visual Information Hiding* (AVIH) method. The proposed method hides the visual information in images while preserving their functional (*e.g.*, recognition and classification) features for service model. It can recover original images from protected images for their owners. Specifically, we reduce the visual correlation between the protected and original image while minimizing their distance in the feature space of service model, to generate the protected image. Meanwhile, we exploit a generative model pre-trained in a private training setting as the key model, then optimize the protected image based on it so that the recovered image is similar to the original image. Furthermore, to break through the tough trade-off between the capability of privacy protection and the quality of restored image, we design the variance consistency loss to enhance privacy protection without compromising image recovery (see Section. 3.3). Note that the protected image generated by our method can only be accurately recovered by the own key model, other models (even if the model architecture is the same) are difficult to recover well (see Section. 4.3).

AVIH can significantly improve the security and flexibility of image storage, which is extremely obvious in the protection of gallery sets for metric learning. Take the cloud-based face recognition system as an example. According to the service face recognition model, the face database manager can generate protected images locally or in the cloud and save the key model locally. The protected image contains no visual information and can be used by the service model to extract features correctly. Moreover, these protected images cannot be recognized by other models. Then, these protected images can be stored in the gallery set in the cloud for normal face recognition tasks. These protected images can be recovered using the key model when needed (such as maintaining the dataset or using face images for other tasks, etc.). The process is shown in Figure. 1.

Taking visual information hiding of gallery set images for cloud-deployed face recognition systems as the basic task, we conduct a comprehensive evaluation of our proposed method in terms of both effectiveness and security. In order to compare with existing methods suitable for information hiding tasks, we extend our method to classification tasks. Experimental results on multiple service models and datasets show that the proposed method is effective. In addition, to prove the effectiveness of our proposed loss, we conduct an ablation study about it to further present the advantages of our proposed method.

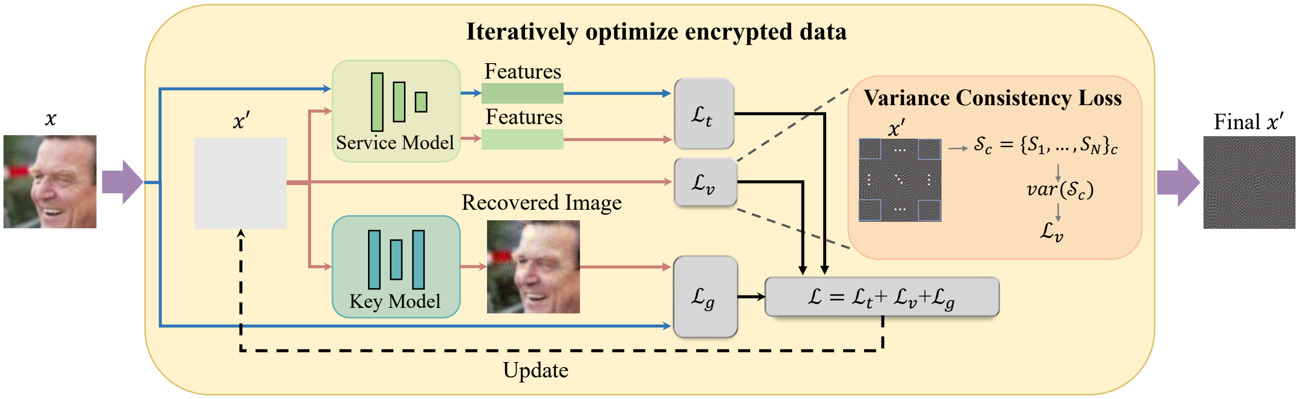
Our main contributions are as follows:

- Inspired by Type-I adversarial attacks, we propose a visual information hiding method AVIH. To alleviate the difficult trade-off between capability of information hiding and quality of recovered image, we design a variance consistency loss.
- Our proposed method has following properties: 1) The visual information in the image is clearly obfuscated. 2) Our method does not require retrain. 3) The protected image can be recovered by the own key model, but external models are difficult to recover.
- We validate the effectiveness of the proposed method on the face recognition task and the classification task. In addition, we conduct qualitative and quantitative ablation studies to show efficiency of the proposed loss.

2. Related Work

2.1. Visual Information Hiding

The visual information hiding of images is from the perspective of human vision. It is most directly manifested by the protected images being visually unrecognizable. Since Homomorphic Encryption (HE) only supports additive and multiplicative operations, HE are not suitable for nonlinear computations. Most DNNs contain a large number of nonlinear computations, so HE-based methods are hardly applicable to state-of-the-art DNNs [40]. The steganography methods hide sensitive information in a cover image and require as few changes to the cover image as possible. It does not need to guarantee that the service model can perform normal predictions on hidden information, which is very different from our work. Therefore, these two types of methods are not discussed in this paper. For the methods based on Perceptual Encryption (PE), some works [42, 40, 39, 8] focus on finding an encrypted domain and training the model directly using the encrypted images. However, this has a significant impact on the accuracy of



Iteratively optimize encrypted data
x
x'
Features
Features
Service Model
Recovered Image
Key Model
$\mathcal{L}_t$
$\mathcal{L}_v$
$\mathcal{L}_g$
$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_v + \mathcal{L}_g$
Update
Variance Consistency Loss
x'
$\mathcal{S}_c = \{S_1, ..., S_N\}_c$
$var(\mathcal{S}_c)$
$\mathcal{L}_v$
Final x'

Figure 2. Overview of adversarial attack-based visual information hiding (AVIH) method. Taking the face recognition as an example, given a service model and a pre-trained key model, we protect the original image x and obtain the protected image x'.

the model. To improve the accuracy of the classification model for protected images, Ito *et al.* [16] trained a transformation network to keep the classifier correctly classified while hiding visual information. However, the weakness of this method is that the protected image cannot be recovered.

Unlike the current work used in classification and segmentation tasks, we mainly focus on metric learning tasks represented by face recognition, which are more prone to privacy leakage. To compare with existing methods that can achieve visual information hiding, we also extend the AVIH method to classification tasks. Our proposed method can also compensate for the drawbacks of the above methods. It generates protected images for a specific model that already exists. It takes advantage of the vulnerability in the model itself to provide strong privacy protection to the image. Since our method does not modify the service model, it does not affect the accuracy on the original image.

2.2. Adversarial Attack

DNNs are vulnerable to some adversarial examples [9, 14]. There are many adversarial attack methods [9, 33, 6] to find adversarial examples efficiently. Tang *et al.* [43] divided the adversarial attacks into adversarial attack Type-I and adversarial attack Type-II based on the statistical Type I and Type II error. We take an optimization perspective on adversarial attacks. Then the Type-II attack maximizes the difference in the model output while ensuring a slight difference with the original input samples. Mathematically, it can be formulated as follows:

$$\max_{x'} f(x) - f(x') \quad s.t. \quad \|x' - x\| < \epsilon \ , \tag{1}$$

where x is the original sample, x' is the adversarial sample, and $f(\cdot)$ is the model which is attacked. The Type-I attack, in contrast to the Type-II attack, looks for an input sample that differs the most from the original input sample but makes the model output as same as possible. Mathematically, it can be formulated as follows:

$$\max_{x'} \|x' - x\| \quad s.t. \quad f(x) = f(x') \ . \tag{2}$$

The Type-I attack on classification and generative models was implemented in the work of Tang *et al.* [43]. Then, Sun *et al.* [41] implemented a Type-I attack against variational autoencoder (VAE) [23]. In this work, we implement Type-I attacks on face recognition and classification tasks. Inspired by these attacks, we propose the AVIH method.

3. Methodology

The objective of our proposed method is to learn a protected image x' which satisfies: 1) x' is completely different from the original image x, 2) for the service model the output $f_s(x')$ is the same as the $f_s(x)$, and 3) x' can be recovered as the x by the key model. Mathematically,

$$\begin{aligned} &\text{From} \quad x \in \mathcal{X} \quad \text{Generate} \quad x' = \mathcal{A}(x) \\ &s.t. \qquad\quad \|x' - x\| > \epsilon \\ &\qquad\quad f_s(x') = f_s(x), \ key(x') = x. \end{aligned} \tag{3}$$

The pipeline of our method is shown in Figure. 2.

3.1. Image Visual Information Hiding

Exploiting the adversarial vulnerability of DNNs, we can perform Type-I attack on the service model to get images that are visually completely different from the original image but with extremely similar features. Suppose that $d(\cdot)$ measures the difference between the outputs of the model. For different tasks, it behaves as different functions. Then for a particular service model f_s, the difference between the output of the adversarial sample and the original sample is

$$\mathcal{L}_t(x', x) = d(f_s(x), f_s(x')). \tag{4}$$

We definite $\mathcal{L}_d$ as

$$\mathcal{L}_d(x', x) = \|x - x'\|_2^2, \tag{5}$$

then the loss of the Type-I attack on the service model is formulated as

$$\mathcal{L}_r(x', x) = \mathcal{L}_t(x', x) - \lambda \cdot \mathcal{L}_d(x', x), \tag{6}$$

where λ is a positive hyperparameter which balances image level differences and output level differences. Then with a certain number of iterations K, we can optimize $\mathcal{L}_r$ by the following operations to get the final adversarial sample.

$$g_{k+1} = \alpha \cdot g_k + \frac{\nabla \mathcal{L}(x'_k, x)}{\|\nabla \mathcal{L}(x'_k, x)\|_2^2}, \tag{7}$$

$$x'_{k+1} = x'_k - \beta \cdot g_{k+1}, \tag{8}$$

where

$$g_0 = \frac{\nabla \mathcal{L}(x'_0, x)}{\|\nabla \mathcal{L}(x'_0, x)\|_2^2}. \tag{9}$$

At the first iteration, x'_0 can be randomly initialized, or it can be made $x'_0 = x$. The former provides an easier way to get an adversarial example with more significant visual differences. The latter provides a faster way to get an adversarial example that meets the criteria.

3.2. Image Recovery

The protected image obtained by Equation. 6 can effectively hide visual information and keep the function of the sensitive information for the service model. However, the protected image is not recoverable. To achieve the goal of Equation. 3, we attach a recovery module that can recover the generated information hidden image. We refer to the generated images with perturbations as protected images. This perturbation has the function of protecting the visual information of the image.

To get the protected images, we first train the generative model G, which can generate the same images as the input. Then we perform Type-I attacks on both the service model and the generative model. We define $\mathcal{L}_g$ as follows:

$$\mathcal{L}_g(x', x) = \|x - G(x')\|_2^2. \tag{10}$$

The loss $\mathcal{L}_g$ can help to keep the protected images recoverable by the key model we have chosen. Thus, the loss becomes as follows:

$$\mathcal{L}_e(x', x) = \mathcal{L}_r(x', x) + \mu \cdot \mathcal{L}_g(x', x), \tag{11}$$

where μ is a hyperparameter that balances the protection quality and restore quality. With Equation. 7 and Equation. 8, we can optimize $\mathcal{L}_e$ to obtain the protected image x',which can satisfy the objective of Equation. 3.

Algorithm 1 Adversarial visual information hiding method

Input: Service model f_s; key model G; original image x; number of iterations K; gradient g; number of times the loss has increased num; maximum number of times the loss increase $maxnum$.
Output: Protected images x'.

```
$g_0 = 0$; $num = 0$;
Random initialization $x'_0$;
for $k = 0$ to $K$ do
    Compute $\mathcal{L}_{AVIH}(x'_k, x)$ via Equation. 15;
    $g_{k+1} = \alpha \cdot g_k + \frac{\nabla \mathcal{L}_{AVIH}(x'_k,x)}{\|\nabla \mathcal{L}_{AVIH}(x'_k,x)\|_2^2}$ (Equation. 7);
    $x'_{k+1} = x'_k - \beta \cdot g_{k+1}$ (Equation. 8);
    if $\mathcal{L}_{AVIH}(x'_{k-1}, x) < \mathcal{L}_{AVIH}(x'{}_k, x)$ then
        $num = num + 1$;
    end if
    if $num > maxnum$ then
        $\beta = 0.85 \cdot \beta$;
        $num = 0$;
    end if
end for
```

3.3. Variance Consistency Loss

The protected image obtained by the objective function of Equation. 6 satisfies the requirements of Equation. 3, but its protection quality is not high. The obtained protected images have the problem of the difficult trade-off between protection quality and recovery quality. That is, if we want to obtain an image that is difficult to be cracked successfully, the quality of the image recovered by the key model will be poor. Another problem is that the obtained protected image, although differing greatly from the original in color, has obvious visual information that the original image has, which negatively impacts visual information protection. To solve the above problem, we propose a variance consistency loss. It improves the quality of protection by limiting the differences between each part of the image to make the protected image visually more confusing.

For the input image, we divide the image in each channel (R, G, B) into N blocks respectively, *i.e.*, $\{b_1, b_2, \ldots, b_N\}_c$, where $c \in \{R, G, B\}$, $b_n \in \mathbb{R}^{h \times w}$ and h, w denote the height and width of the block. The pixels of the blocks are allowed to have overlapping parts between them. Let $p_{i,j}^{b_n} \in [0, 1]$ denote the normalized pixel value at (i, j) in block b_n. Then, we calculate the sum of each block:

$$S_n = \sum_{i=1}^{h} \sum_{j=1}^{w} p_{i,j}^{b_n}. \tag{12}$$

We use $\mathcal{S}$ denote the set of sum of the blocks for each channel, *i.e.*, $\mathcal{S}_c = \{S_1, S_2, \ldots, S_N\}_c$. In our practice, we convolve the image with a convolution kernel of size $h \times w$ to obtain blocks. Then, we calculate the variance of $\mathcal{S}_c$ and obtain $\sigma_c^2 = \mathrm{var}(\mathcal{S}_c)$. Finally, we define the variance consistency loss as:

$$\mathcal{L}_v(x') = \sigma_R^2 + \sigma_G^2 + \sigma_B^2, \tag{13}$$

where σ_R, σ_G and σ_B denote the variances for R, G, B channels, respectively. By minimizing $\mathcal{L}_v$, we can get a protected image with a more uniform distribution of pixel values. Proof-of-concept experiments (see Section. 4.4) show that $\mathcal{L}_v$ can eliminate the visual semantics in the protected images which are similar to the original images, and can help obtain protected images with high quality of protection and recovery. More details on the variance consistency loss can be found in Supplementary Material D.

Based on the variance consistency loss, the loss function in Equation. 6 is modified as:

$$\mathcal{L}_r(x', x) = \mathcal{L}_t(x', x) + \lambda \cdot \mathcal{L}_v(x'), \tag{14}$$

and the Equation. 11 is reformulated as:

$$\mathcal{L}_{AVIH}(x', x) = \mathcal{L}_t(x', x) + \lambda \cdot \mathcal{L}_v(x') + \mu \cdot \mathcal{L}_g(x', x), \tag{15}$$

where μ are hyperparameters. This is our proposed AVIH framework for visual information hiding. The algorithm is summarized in Algorithm. 1.

3.4. AVIH Method for Specific Tasks

Our method can provide effective protection for the stored image and can be applied to a wide range of tasks. In this work, we take face recognition and classification as examples to illustrate the ability of our method.

For the face recognition task, we want the features extracted by the service model f_s from the protected and original images to be as same as possible. Thus, we modify $\mathcal{L}_t$ as follows:

$$\mathcal{L}_t(x', x) = \|f_s(x) - f_s(x')\|_2^2. \tag{16}$$

For the classification task, compared with directly minimizing the mean square error (MSE) of $f_s(x)$ and $f_s(x')$, we find that maximizing $f_s^c(x')$ where c is the prediction class of the service model on the original image can more effectively reduce the impact of visual information hiding on the accuracy. We thus first obtain the logit output of the original image $f_s(x)$ and convert it to one-hot format $\delta(f_s(x))$. Then, we reify $\mathcal{L}_t$ as follows:

$$\mathcal{L}_t(x', x) = -\delta(f_s(x)) \cdot f_s(x'). \tag{17}$$

4. Experiments

In this section, we first verify the effectiveness of the AVIH method for face recognition tasks. Then, taking the face recognition system as an example, the security of the AVIH method and the effectiveness of the variance consistency loss are explored. Finally, we verify the effectiveness of the AVIH method for classification tasks and compare it with other methods.

4.1. Experimental Settings

Dataset and service models. Our experiments were mainly evaluated based on the *Labeled Face in the Wild (LFW)* [13] dataset. We choose four face recognition models ArcFace [7], CosFace [44], SphereFace [30] used in [48], and AdaFace [22] to fully evaluate the performance of our method. Among them, the input size of ArcFace and AdaFace [22] is 112 × 112, and the input size of CosFace and SphereFace is 112 × 96. Therefore, we first use the MTCNN [49] to align and crop the face images to the corresponding face recognition model's input size.

Key models. We choose the pix2pix framework [15] to train our key model. To pre-train the key model, we randomly selected 1,2878 images from the *CelebA* [31] dataset as the training set. Then we set the input and output of the model to the same image, set the batch size to 1, and train for 4 epochs. Finally, the trained generator is used directly as the key model. Unless otherwise stated, we use only one key model to protect the entire dataset.

Evaluation metrics. To evaluate the effectiveness of our method more realistically and inspired by the evaluation method in MegaFace [21], we modified the evaluation method of *LFW*. We randomly selected 12 persons from the *LFW* dataset as the probe set. Each contains more than 12 facial images, comprising 355 images. The other 12878 images, we use as the gallery set. In the testing phase, we take one face image of a person in the probe set and put it into the gallery set. Then use the remaining images of this person as the test set. Next, we use the above-divided dataset to test the accuracy of the face recognition model. In this way, we put each person's image in the probe set into the gallery set in turn to measure the average accuracy. This metric can well demonstrate the impact of our method on face recognition models in practical applications.

We use the Structural Similarity Index Measure (SSIM) [47] and Learned Perceptual Image Patch Similarity (LPIPS) to quantitatively measure the quality of the protected images. In this work, we use SSIM$_e$ / LPIPS$_e$ to represent the average of the SSIM / LPIPS values between the protected image x' and the original image x, and use SSIM$_d$ / LPIPS$_d$ to represent the average of the SSIM / LPIPS values between x and the recovered image $G(x')$. A higher SSIM indicates that the two images are more similar and a lower LPIPS indicates that the two images are more similar.

4.2. Effectiveness of Face Privacy Protection

Effectiveness and impact on service model. To the best of our knowledge, our method is the first (PE)-based

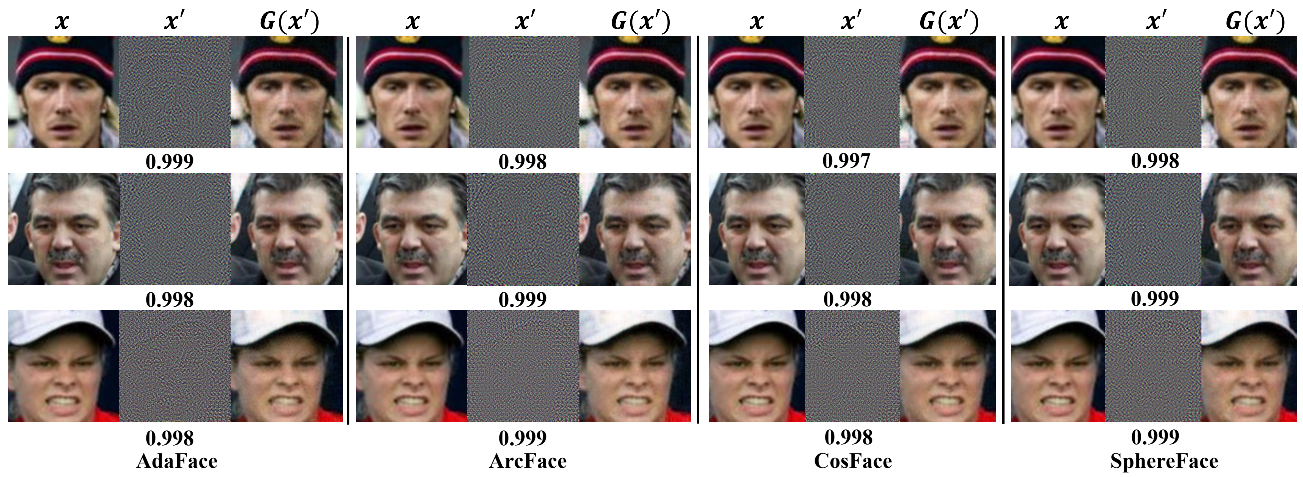


Figure 3. Adversarial Visual Information Hiding (AVIH) method protect face images for different service models. We mark below each protected image the cosine similarity between its feature vector and the original image's feature vector.

Table 1. Image visual quality metrics and cosine similarity (cos-sim) between the output features of the service model for original and protected images on *LFW*.

	SSIM$_e$ ↓	LPIPS$_e$↑	SSIM$_d$ ↑	LPIPS$_d$↓	cos-sim ↑
AdaFace	0.028	0.886	0.893	0.144	0.997
ArcFace	0.030	0.888	0.903	0.128	0.994
CosFace	0.028	0.891	0.899	0.128	0.998
SphereFace	0.029	0.888	0.906	0.120	0.996

Table 2. Accuracy (percentage) of face recognition models for original image and different protected image on *LFW*. The results of AVIH-ONE and AVIH-ALL are expected to be close to the results of original.

Models	Original	AVIH-ONE (%)	AVIH-ALL (%)
AdaFace	98.6	98.6	98.6
ArcFace	96.5	96.5	96.5
CosFace	89.4	89.2	89.3
SphereFace	80.3	80.0	80.6

method for face recognition. There is no closely related method yet to compare the performance of protected images. Therefore, we conduct qualitative and quantitative evaluations of protected images. High visual metric scores and a very slight impact on accuracy demonstrate the effectiveness of our method.

We use the AVIH method to protect the original face image x in the probe set to test its effectiveness on face visual information hiding. Then, we used the obtained protected image x' as the input of the key model $G(\cdot)$. Finally get the recovered image $G(x')$. The results are shown in Figure. 3. From Figure. 3, it is evident that the protected image generated by our method is significantly different from the original image. Since no useful visual information is obtained from the protected image, it has the effect of visual information hiding for the original image. In Table. 1, the average SSIM value between the protected and original images for each service model is less than 0.04, while the cosine similarity between their features is higher than 0.99. Therefore, the protected image generated by AVIH can completely replace the original image while hiding the visual information.

To evaluate the impact of AVIH method on the accuracy of the face recognition models, we first generate a protected face image in the probe set before putting it into the gallery set. Then put the protected image into the gallery set instead of the original image. We tested the accuracy of the face recognition models in this way. The result AVIH-ONE is shown in Table. 2. Compared with the original accuracy of the models, it can be concluded that the AVIH method's impact on the models' accuracy is very slight, which can achieve the same accuracy in the ArcFace model.

Considering the actual situation, all the images stored in the gallery set are protected images. So we replaced all the images in the gallery set with protected images and then retested the accuracy of the models, as AVIH-ALL in Table. 2. In this case, our method has a very slight impact on the accuracy of the models and has a beneficial effect on the accuracy of the SphereFace model.

We test the average of SSIM values, as shown in Table. 1. Combined with Figure. 3 shows that the recovered image has good quality and can recover most of the visual information of the original image. Since the recovered images

Table 3. The performances of our method when the pre-trained model is based on non-face image (*e.g.*, *COCO*). The numbers in parentheses represent the difference from the original accuracy of the model. For example, (-0.2) means the accuracy is 0.2 lower than the original.

Model	AVIH-ONE (%)	SSIM$_e$ ↓	SSIM$_d$ ↑
AdaFace	98.6(-0.0)	0.026	0.891
ArcFace	96.5(-0.0)	0.027	0.909
CosFace	89.2(-0.2)	0.031	0.882
SphereFace	79.9(-0.4)	0.031	0.890

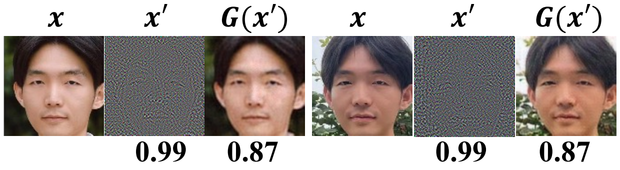


Figure 4. Visual information hiding in real-world scenes. The cosine similarity between recognition features of protected images (x') and original images (x) is marked below x'. The SSIM between recovered images ($G(x')$) and original images is marked below $G(x')$.

are generated by the generative model, a better-trained generative model can achieve better recovery quality. We also explore the time spent in protecting images of AVIH in Supplementary Materials A.

Randomness of the dataset for training the key model. In this work, we use *CelebA*, which is also a face dataset, to train the key model. However, in real scenarios, large-scale face images are often difficult to obtain due to the privacy of each individual involved. So we changed *CelebA* to *COCO* [29] containing various objects to test the effectiveness of the new key model. We choose the test set of *COCO* as the train set and train 6 epochs on the key model. The results are shown in Table. 3. It can be concluded that the impact of protected images on the model accuracy and the quality of recovered images are slightly different from the key model trained with *CelebA*. Therefore, we can use different types of datasets to train the key model, not just limited to face images.

Real world face privacy protection. We randomly selected real-world face photos taken with phones and then protected them to test the performance of the AVIH method for using real-world scenarios. Part of the results is shown in Figure. 4. In realistic scenarios, the cosine similarity of the features between the protected image generated by the AVIH method and the original image is still higher than 0.99. The recovered image can still be recovered well.

Robustness to noise. We tested the quality of the recovery after adding Gaussian noise $X_{noise} \sim \mathrm{N}(0, \sigma^2)$ to the protected image, as shown in Table 4. It can be concluded that our method is robust to small noise.

Table 4. Robustness of protected images against noise.

σ	0	3/255	5/255	8/255	10/255
SSIM$_d$	0.93	0.92	0.89	0.84	0.81

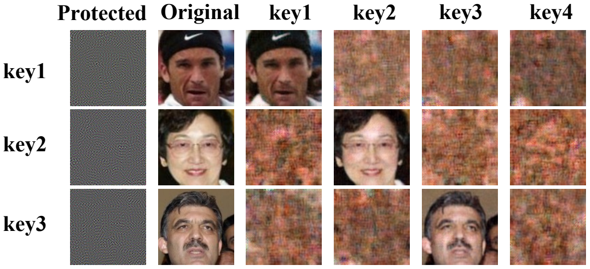


Figure 5. Partial results of the key model randomness analysis. The left column is the key used for protection, and the top row is the key used for recovery. Every key model is trained with different initialization only. More images of the results are shown in Supplementary Material B.

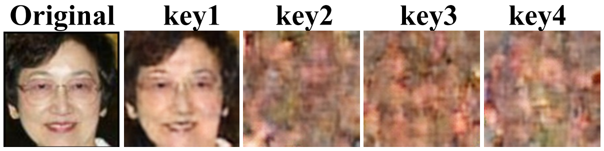


Figure 6. Crack the protected image using the attack model. The training dataset for the attack model is obtained using key1 protection. An attacker can recover images protected with his own key (key1), but cannot recover images protected with other keys.

4.3. Security Analysis

Key model randomness analysis. To test the randomness of the key model, we use the same strategy as Ding *et al.* [8]. We first trained 16 key models using the same settings but with different initialization values. Different initialization values may result in significantly different key models. Then, we choose one of them in turn as the private key model to participate in the protection of the image, and use the others as the external key model to try to recover the protected image. From the results shown in Figure. 5, it can be concluded that the protected image obtained by one private key model cannot be recovered by other external key models, demonstrating that our proposed method can greatly improve the security of the face visual information. In addition, it is convenient that we can obtain mutually independent keys by simply changing the initialization without changing factors such as the structure of the key model and the training dataset. We suppose that one reason for this phenomenon is the instability of GAN training, where dif-

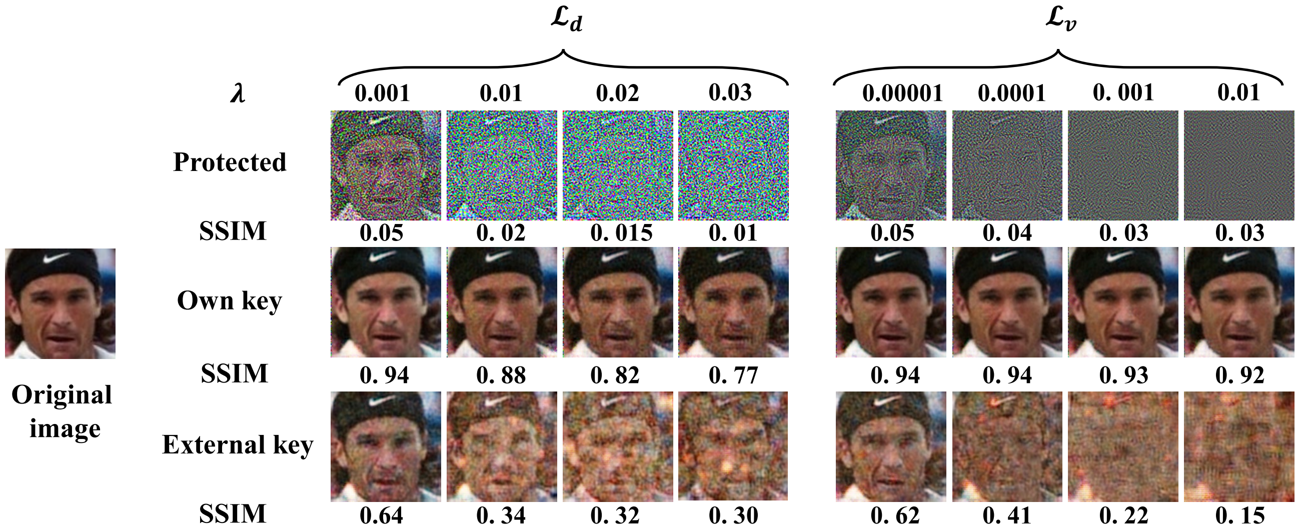


Figure 7. Results of the ablation study. The SSIM value between the processed image and the original image is shown under each image. The own key represents the key model used for generating the protected image, and the external key represents other key models that are initialized differently during training.

ferent initialization values lead to different locally optimal solutions. Another reason could be that our method tends to find the boundary points of the input field corresponding to the output error allowed by the key model. However, the instability of GAN training causes this boundary to change significantly when the initialization value is changed.

Responding to a possible attack. For the key model, even if the key model structure and its training settings are leaked, the protection is still difficult to break if the initialization point of the key training model is not leaked. As for the image, if pairs of original and protected images are leaked in large quantities, then an attacker can directly use these image to train a generative model to directly map the protected domain to the original domain. Ito *et al*. [17] exploited this method to evaluate the robustness of protection. So, we protect all the face images in the gallery set using the same key model, producing pairs of protected and original images. Then we use these paired images as a training set, and train a model using the same structure as the attack model. Using such a model, we attempt to recover the protected images in the probe set. The results are shown in Figure. 6. When the training set images and the protected image to be cracked are protected from the same key model, the attack model can recover the protected image. However, when they are protected with different key models, the trained model cannot crack the protected image. Therefore, we can increase the frequency of changing key models to effectively defend against such attacks.

The protected images obtained by our method also have good statistical properties. The correlation analysis and histogram analysis are shown in Supplementary Material B. We also verify that the protected image generated by a service face recognition model cannot be recognized by other models with the same function.

4.4. Ablation Study

Effectiveness of the variance consistency loss. We compare the variance consistency loss $\mathcal{L}_v$ with the distance loss $\mathcal{L}_d$ (in Equation. 5) to verify the effectiveness of $\mathcal{L}_v$. For each loss, we choose a set of suitable weights λ separately to represent its trade-off between the protection quality and recovery quality. Its results are shown in Figure. 7. The protected image generated using $\mathcal{L}_d$ has a smaller SSIM value compared to the protected image using $\mathcal{L}_v$, but a clear outline can be seen. So the visual information of the original image still exists. If the protected image is to be made free of such visual information, there is a significant loss in the quality of the recovered image. In contrast, using $\mathcal{L}_v$, it is possible to generate protected images with both no residual visual information and high recovery quality. We also use another key to recover the protected image to further investigate the effect of the two different losses on the protection quality. From Figure. 7, it can be concluded that the protected image using $\mathcal{L}_v$ is more difficult to be recovered by similar key models with different initialization while ensuring the quality of the recovered image. We compare $\mathcal{L}_v$ with more other losses in Supplementary Material D to further demonstrate the effectiveness of $\mathcal{L}_v$.

As shown in Figure. 5, the difference between the results when $\lambda = 0.001$ and $\lambda = 0.01$ is very slight for the protected and recovery quality of the protected images. This means that our method requires no carefully adjustment of

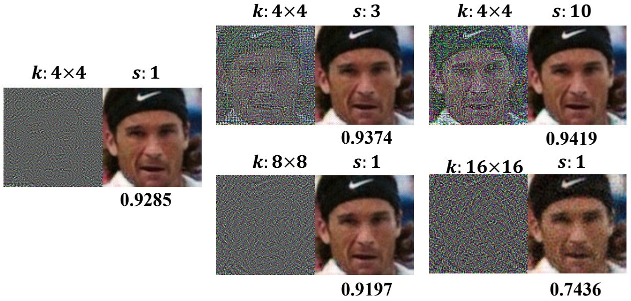


Figure 8. Impact of parameter settings on protection quality and recovery quality. k represents the size of the convolution kernel and s represents the step size. For each pair of images, the former is the protected image and the latter is the recovery image. Below the recovery image is marked the SSIM value between it and the original image.

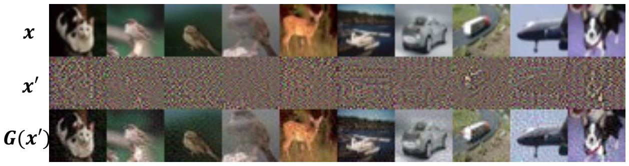


Figure 9. Results of AVIH method for Resnet50 as the service model. Each column represents a different sample.

the weight λ for $\mathcal{L}_v$, which is very convenient.

Parameters of variance consistency loss. In the variance consistency loss calculation, the process of blocking the image and calculating the internal sum of each block can be regarded as a convolution kernel with all values of 1 to convolve the image. The convolution kernel size represents the size of the block, and the step size of the convolution kernel represents the degree of overlap of the block. We explored the effect of block size and degree of overlap on performance, as shown in Figure 8. When the overlap degree of blocks is certain (especially when it is highly overlapping), the quality of the recovery image decreases as the size of the blocks increases. When the size of blocks is certain, the protection quality decreases as the overlap degree of blocks decreases. In this work, we divide the image into 4×4 blocks, with 12 pixels overlapping between adjacent blocks (step size is 1). This setting is applicable to most tasks and there is no need to change it for different samples.

4.5. Privacy protection for classification models

We trained ResNet50 [11], VGG19 [37] using the train set of *CIFAR-10* [25] and implemented AVIH method for the images of the test set. The results are shown in Figure. 9. We also compared two existing methods suitable for visual information hiding and the results are shown in Table. 5, where the ITP [16] do not recover protected images. Since the LIE [42] uses protected images to train the model, the prediction accuracy on the original data is low. We show the visual metrics, the results of the two comparative methods, and the results on Imgnet in Supplementary Material C. Our method has minimal impact on the accuracy of the models, and the average SSIM value of the recovered images can reach above 0.9 for both models after our tests.

Table 5. Impact of privacy protection methods on the accuracy (percentage) of classification models on *CIFAR-10*.

Method	Model	Original (%)	Protect (%)
LIE [42]	VGG19	10.59	87.78
	Resnet50	11.00	91.53
ITP [16]	VGG19	93.95	90.70
	Resnet50	95.53	90.16
AVIH (Our)	VGG19	93.95	**93.95**
	Resnet50	95.53	**95.28**

5. Conclusion

In this paper, we propose a visual information hiding method AVIH based on Type-I attack. We evaluate our method by image protection of face recognition systems in the cloud. Experiments show that the AVIH method can protect images while preserving their functionality for the service model. We also propose a variance consistency loss to solve the problem of the difficult trade-off between protection quality and recovery quality. Finally, we use the AVIH method in classification tasks with satisfactory results. Our work provides a new perspective on image visual information protection, which has beneficial implications for adversarial learning communities and private protection communities. The AVIH method requires complete service model information. While this is feasible in most image storage situations, it limits the applicability of the AVIH method to a wider range of image protection scenarios. This is also a problem we will solve in future work.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096 and 62036007; in part by the Fundamental Research Funds for the Central Universities under Grants QTZX23042 and XJS221502; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01; in part by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2022JQ-696; in part by the Fundamental Research Funds for the Central Universities and in part by the Innovation Fund of Xidian University under Grant YJSJ23012.

References

[1] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M Taha, and Vijayan K Asari. Recurrent residual u-net for medical image segmentation. *Journal of Medical Imaging*, 6(1):014006–014006, 2019. 1

[2] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017. 1

[3] L Arockiam and S Monikandan. Efficient cloud storage confidentiality to ensure data security. In *2014 International Conference on Computer Communication and Informatics*, pages 1–5. IEEE, 2014. 1

[4] Deyan Chen and Hong Zhao. Data security and privacy protection issues in cloud computing. In *2012 International Conference on Computer Science and Electronics Engineering*, volume 1, pages 647–651. IEEE, 2012. 1

[5] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Drinet for medical image segmentation. *IEEE Transactions on Medical Imaging*, 37(11):2453–2462, 2018. 1

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[8] Yi Ding, Guozheng Wu, Dajiang Chen, Ning Zhang, Linpeng Gong, Mingsheng Cao, and Zhiguang Qin. Deepedn: a deep-learning-based image encryption and decryption network for internet of medical things. *IEEE Internet of Things Journal*, 8(3):1504–1518, 2020. 1, 2, 7

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 9

[12] Lei Hu, Da Wei Zhou, Cai Xia Fu, Thomas Benkert, Chun Yu Jiang, Rui Ting Li, Li Ming Wei, and Jun Gong Zhao. Advanced zoomed diffusion-weighted imaging vs. full-field-of-view diffusion-weighted imaging in prostate cancer detection: a radiomic features study. *European Radiology*, 31:1760–1769, 2021. 1

[13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 2, 3

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5

[16] Hiroki Ito, Yuma Kinoshita, Maungmaung Aprilpyone, and Hitoshi Kiya. Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks. *IEEE Access*, 9:64629–64638, 2021. 1, 3, 9

[17] Hiroki Ito, Yuma Kinoshita, and Hitoshi Kiya. Image transformation network for privacy-preserving deep neural networks and its security evaluation. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 822–825. IEEE, 2020. 8

[18] He Kaiming, Gkioxari Georgia, Dollar Piotr, and Girshick Ross. Mask r-cnn. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP:1–1, 2017. 1

[19] Seny Kamara and Kristin Lauter. Cryptographic cloud storage. In *International Conference on Financial Cryptography and Data Security*, pages 136–149. Springer, 2010. 1

[20] Manjit Kaur and Vijay Kumar. A comprehensive review on image encryption techniques. *Archives of Computational Methods in Engineering*, 27(1):15–43, 2020. 1

[21] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 5

[22] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 5

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[24] Varsha Kishore, Xiangyu Chen, Yan Wang, Boyi Li, and Kilian Q Weinberger. Fixed neural network steganography: Train the images, not the network. In *International Conference on Learning Representations*, 2021. 1

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 9

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[27] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4166–4175, 2022. 1

[28] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Unconstrained facial expression recognition with no-reference de-elements learning. *IEEE Transactions on Affective Computing*, 2023. 1

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 5

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[32] Qian Lou and Lei Jiang. She: A fast and accurate deep neural network for encrypted data. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 3

[34] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017. 2

[35] Siani Pearson and Azzedine Benameur. Privacy, security and trust issues arising from cloud computing. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pages 693–702. IEEE, 2010. 1

[36] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9

[38] Ashish Singh and Kakali Chatterjee. Cloud security issues and challenges: A survey. *Journal of Network and Computer Applications*, 79:88–115, 2017. 1

[39] Warit Sirichotedumrong and Hitoshi Kiya. A gan-based image transformation scheme for privacy-preserving deep neural networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 745–749. IEEE, 2021. 1, 2

[40] Warit Sirichotedumrong, Takahiro Maekawa, Yuma Kinoshita, and Hitoshi Kiya. Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 674–678. IEEE, 2019. 1, 2

[41] Chengjin Sun, Sizhe Chen, Jia Cai, and Xiaolin Huang. Type i attack for generative models. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 593–597. IEEE, 2020. 3

[42] Masayuki Tanaka. Learnable image encryption. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2018. 1, 2, 9

[43] Sanli Tang, Xiaolin Huang, Mingjian Chen, Chengjin Sun, and Jie Yang. Adversarial attack type i: Cheat classifiers by significant changes. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1100–1109, 2019. 2, 3

[44] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 5

[45] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14920–14929, 2022. 1

[46] Yizhi Wang, Jun Lin, and Zhongfeng Wang. An efficient convolution core architecture for privacy-preserving deep learning. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018. 1

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[48] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021. 2, 5

[49] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 5

[50] Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. Steganogan: High capacity image steganography with gans. *arXiv preprint arXiv:1901.03892*, 2019. 1

[51] Dawei Zhou, Nannan Wang, Chunlei Peng, Yi Yu, Xi Yang, and Xinbo Gao. Towards multi-domain face synthesis via domain-invariant representations and multi-level feature parts. *IEEE Transactions on Multimedia*, 24:3469–3479, 2021. 1