Veil Privacy on Visual Data: Concealing Privacy for Humans, Unveiling for DNNs

Shuchao Pang[1(✉)], Ruhao Ma[1], Bing Li[2(✉)], Yongbin Zhou[1], and Yazhou Yao[3]

[1] School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing, China {pangshuchao,maruhao,zhouyongbin}@njust.edu.cn
[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China bing_li@uestc.edu.cn
[3] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China yazhou.yao@njust.edu.cn

Abstract. Privacy laws like GDPR necessitate effective approaches to safeguard data privacy. Existing works on data privacy protection of DNNs mainly concentrated on the model training phase. However, these approaches become impractical when dealing with the outsourcing of sensitive data. Furthermore, they have encountered significant challenges in balancing the utility-privacy trade-off. How can we generate privacy-preserving surrogate data suitable for usage or sharing without a substantial performance loss? In this paper, we concentrate on a realistic scenario, where sensitive data must be entrusted to a third party for the development of a deep learning service. We introduce a straightforward yet highly effective framework for the practical protection of privacy in visual data via veiled examples. Our underlying concept involves segregating the privacy information present in images into two distinct categories: the privacy information perceptible at the human visual level (*i.e.*, Human-perceptible Info) and the latent features recognized by DNN models during training and prediction (*i.e.*, DNN-perceptible Info). Compared with the original data, the veiled data conserves the latent features while obfuscating the visual privacy information. Just like a learnable veil that is usable for DNNs but invisible for humans, the veiled data can be used for training and prediction of models. More importantly, models trained with the veiled data can effectively recognize the original data. Extensive evaluations of various datasets and models show the effectiveness and security of the Veil Privacy framework.

1 Introduction

Deep Neural Networks (DNNs) have achieved remarkable success across various scientific domains. The rapid development of DNNs [15, 39, 50] has greatly benefited from the availability of openly accessible datasets (*e.g.*, ImageNet [5]).

✉ Corresponding author. Shuchao Pang and Ruhao Ma contribute equally.

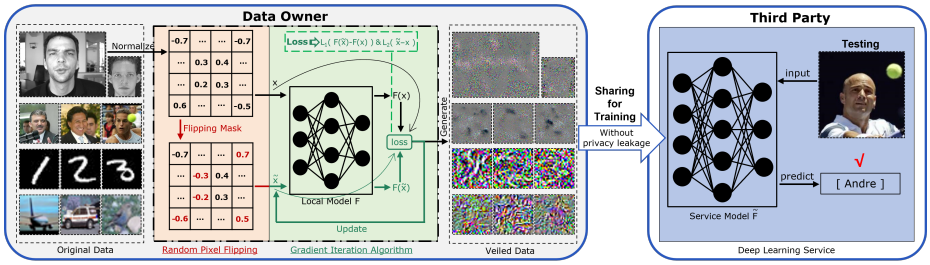

Fig. 1: The Veil Privacy framework and its application for visual privacy protection. (1) *Veiled Data Generation.* The data owner generates the veiled data from the original data via Veil Privacy and then shares them with the third party without visual privacy leakage. (2) *Application.* The third party trains the DNN model using the shared veiled data to provide deep learning services. Then the trained service model can effectively predict not only the veiled data but also the original data during the testing phase.

However, datasets may contain users' private information (*e.g.*, biomedical features). Privacy regulations (*e.g.*, GDPR [13] and CCPA [33]) mandate the adoption of effective measures to safeguard data privacy. Therefore, it's imperative to propose a framework that empowers service providers to leverage the data collected from users in a privacy-preserving manner. Images play a pivotal role in training DNNs to deliver computer vision services [20, 27, 49]. Nevertheless, most images include sensitive privacy-related content, *e.g.*, individuals' faces.

There have been some works on privacy protection for images [29, 35, 36]. However, they have generally overlooked the model's capacity to recognize the original data. Our focus is directed toward a realistic scenario, where sensitive data have to be outsourced to build deep learning services. To illustrate this, consider a situation in which a company, in the context of implementing a face recognition-based access control system, gathers facial images of its employees and delegates the training of a face recognition model to a third party. Given the personal nature of facial data and the uncertainty surrounding the third party's future data handling, the company is reluctant to share the data directly. This underscores the challenge of ensuring data security while facilitating its use and sharing. Existing works, including dataset distillation [41], have encountered substantial difficulties in striking a balance between utility and privacy. Therefore, the central question posed is: *How can we generate privacy-preserving surrogate data suitable for usage or sharing without a significant loss in performance?*

In this paper, we propose a practical approach *Veil Privacy* to safeguard the privacy of visual data (*e.g.*, images) via veiled examples. Our objective is to render the data suitable for training and inference of DNNs while imperceptible to human observers—akin to a learnable veil. We categorize the privacy information present in images into two distinct dimensions: human-perceptible info which means privacy info discernible at the human visual level, and DNN-perceptible info, *i.e.*, latent features utilized by DNN models during training and inference. As illustrated in Fig. 1, the Veil Privacy framework introduces a highly

effective methodology to generate a privacy-preserving variant of unprotected data, called veiled data. The primary aim is to conceal the human-perceptible info while minimally affecting the DNN-perceptible info. The framework comprises two main steps. Initially, we devise random pixel flipping to magnify the visual distinction between the veiled data and the original data. Subsequently, we propose a gradient iteration algorithm to preserve latent features that DNN models can recognize from the original data. The resultant veiled data no longer discloses visual privacy. Due to the preservation of similar latent features, models trained using the veiled data not only recognize the veiled data but also the original data. In summary, we make the following key contributions:

- This study represents the first instance in which privacy information within images is systematically classified into two distinct categories: privacy information discernible through human visual perception and latent features recognized by models during the training and testing phases.
- We provide a comprehensive overview of prior research on the privacy-preserving of publicly accessible images. Additionally, we identify distinct attributes and limitations associated with these strategies. As last, we address a compelling question: *How can we generate surrogate data that safeguards privacy during utilization or sharing, while minimizing the impact on performance?*
- To render data suitable for learning while concurrently rendering it imperceptible, we propose a practical framework, Veil Privacy, for safeguarding the privacy of visual data via veiled examples. This framework generates the veiled data by developing a combination of random pixel flipping and the gradient iteration algorithm. In contrast to the original data, the veiled data retains latent features while concealing visual privacy information.
- Comprehensive experiments conducted across various datasets and diverse models substantiate the effectiveness and security of the Veil Privacy framework. More importantly, models trained with the veiled data can effectively recognize the original data. Overall, as a pioneering work, this work provides the key technical support for customized AI security services on visual data.

2 Related Work

There has been a line of work has emerged concerning the protection of privacy in images [19, 23, 40]. Leveraging our defined privacy categories and considering a delicate balance between utility and privacy, we classify these prior works into three distinct categories: adversarial perturbation, encryption computing and privacy elimination. Further, we also regard dataset distillation as an approach for intuitive comparison. As shown in Tab. 1, these works offer privacy protection from various angles, employing a diverse range of methodologies.

Adversarial perturbation disrupts training or prediction of DNN models by subtle perturbations. In order to avoid disclosing personal privacy, it prevents the unauthorized acquisition of latent features. Consequently, adversarial perturbation techniques damage DNN-perceptible info within the origi-

Table 1: Key differences between Veil Privacy and existing methodologies. Veil Privacy presents some desired merits: it ensures the veiled data unperceptible to humans while perceptible to DNNs, and the feature space closely aligns with that of the original data. Functionally, Veil Privacy enhances model performance and remains compatible with the original data, allowing seamless switching from both veiled and original data.

Methodology	Characteristic			Functionality	
	Human-perceptible Info	DNN-perceptible Info	Feature Space	Model Performance	Original Data
Adversarial Perturbation	Unprotected	Damaged	Unaligned	✗	✗
Encryption Computing	Protected	Intact	Unaligned	✓	✗
Privacy Elimination	Protected	Damaged	Unaligned	✗	✗
Dataset Distillation	Protected Partially	Intact	Aligned Partially	✓	✓↖
Veil Privacy	Protected	Intact	Aligned	✓	✓

nal data, thereby enhancing privacy protection, while concurrently preserving human-perceptible info due to the minimal magnitude of adversarial perturbations. Wang *et al.* [37] secures private latent features by fooling the recognition algorithm, where well-designed masks make it difficult for adversaries to discern their presence. They also introduce an adversarial fusion algorithm [38], which alters the feature distribution of original images through perturbations without visually significant differences. Yang *et al.* [44] design a pivot pixel noise generator (PPNG), which generates tiny noise on images to provide privacy protection, thwarting DNN models from correctly labeling perturbed images. All these works [21, 22, 28, 37, 38, 43, 44] consider the potential features that DNN models can utilize while overlooking the preservation of visual privacy information, which is the most intuitive privacy for individuals.

Encryption computing achieves training and prediction of DNN models on encrypted data directly. Sensitive data pose a distinct challenge due to their inherent privacy considerations. Consequently, encryption computing simultaneously transforms both human-perceptible info and DNN-perceptible info of the original data. However, models trained on encrypted data remain challenging for inferring the original unencrypted data. Lee *et al.* [18] apply the bootstrapping technique from RNS-CKKS to DNN models, which enables the evaluation of arbitrary models on encrypted data while preserving privacy. Helena *et al.* [24] mention that privatize the identity for the human eyes but retain the utility for clinical. Pixel-based encryption [32] effectively facilitates training and inference for classification models with covert images. However, these approaches [2, 7, 12, 14, 18, 24, 30, 32] share a common limitation: they focus solely on training and inference with encrypted data. Yet, a practical concern arises as most post-deployment models finally encounter original data in real scenarios.

Privacy elimination focuses on how to identify and eliminate privacy in data without considering the training and prediction of DNN models. It eradicates both human-perceptible info and DNN-perceptible info within the original data. During the collection of public images, certain elements such as license plates, faces and other private information are considered sensitive. This information needs to be eliminated by some methods. Uittenbogaard *et al.* [34] introduce an alternative-blurring framework designed to automatically remove and

rectify specific sensitive objects in street-view imagery. Iprivacy [47] is devised to accomplish swiftly and accurately detecting privacy-related information, implementing privacy protection measures through obscuration. Frome *et al.* [10] combine standard sliding-window ways with rapid post-processing, actively identifying and blurring individuals' faces in Google Street View. However, it's worth noting that all of these approaches are primarily focused on eliminating privacy-related information from data, not suitable for integrating into DNN models.

Dataset distillation aims to distill a large dataset into a synthetic smaller dataset for improving the data efficiency when training DNN models, not aiming for privacy protection. The synthesized small dataset achieves similar effects to the original dataset for training and prediction of DNNs. Dataset distillation keeps DNN-perceptible info at the dataset level and changes human-perceptible information to a certain extent at the sample level. There have been some excellent works of dataset distillation, such as [3, 25, 26]. Further, Dong *et al.* [8] apply dataset distillation for privacy protection of images. However, these works inevitably reduce predictive performance on the original data. And there are still issues of high similarity between the synthetic data by distillation and the original data, as well as high time complexity from the distillation process. Dataset distillation is performed at the dataset level. If there is a need for adding a new sample, it requires incorporating a new sample into the original dataset and re-distilling the entire dataset. Therefore, dataset distillation has not truly achieved the balance between utility and privacy.

Summary. The existing works grapple with significant challenges in balancing between utility and privacy. *How can we generate privacy-preserving surrogate data for utilization or sharing without substantial performance degradation?* As presented in Tab. 1, we introduce an innovative approach Veil Privacy, which effectively preserves potential features for DNNs while concealing visual privacy information for human observers, achieving a favorable utility-privacy trade-off.

3 Veil Privacy for Visual Data

3.1 Overview

Our goal is that human-perceptible info remains well concealed, while DNN-perceptible info remains unaffected. Here, we provide the problem formulation at the data owner level. Let u be a data owner who possesses data samples (x_1, y_1),...,(x_n, y_n), where each sample has data $x \in X$ and label $y \in Y$. Then u generates privacy-preserving surrogate data with the same label for sharing. The third party collects these surrogate data samples from u to train the service model $\tilde{F}(\cdot)$. It should be noted that $F(\cdot)$ and $\tilde{F}(\cdot)$ can be different.

Specifically, Veil Privacy generates the veiled data $\tilde{X}$ from the original data X. As shown in Algorithm 1, it primarily comprises two steps, *i.e.*, (a) Random Pixel Flipping $RPF(\cdot)$ is devised to conceal the visual privacy, with a specific flipping ratio p; and (b) Gradient Iteration Algorithm $GIA(\cdot)$ is proposed to retain the latent features, with a local model $F(\cdot)$ of the data owner.

Algorithm 1 Veiled Data Generation

Input: Data $X = \{x_1, ..., x_n\}$, Label $Y = \{y_1, ..., y_n\}$, Flipping Ratio p and Local Model $F(\cdot)$
Output: Veiled Data $\tilde{X}$

begin
  $X = Normalize(X)$
  Initialize $\tilde{X}$ with X
  for every $\tilde{x}_i \in \tilde{X}$ **do**
    Initialize matrix $\tilde{m}$ with shape $C \times H \times W$ from $U(0, 1)$
    for i in range(0, $W * H * C$) **do**
      $\tilde{m}[i] = \tilde{m}[i] - p$
      $\tilde{m}[i] = 1$ if $\tilde{m}[i] \geq 0$ else $\tilde{m}[i] = -1$
    end
    $\tilde{x}_i = \tilde{x}_i \times \tilde{m}$ # Element-wise Multiplication
  end
  for every $\tilde{x}_i \in \tilde{X}$ **do**
    for t in range(0, T) **do**
      $L(\tilde{x}_i) = \alpha \cdot \|F(\tilde{x}_i) - F(x_i)\|_1 + \beta \cdot (-1) \cdot \|\tilde{x}_i - x_i\|_2$
      $\tilde{x}_i = \tilde{x}_i - lr * \bigtriangledown_{\tilde{x}_i} L(\tilde{x}_i)$
      if $L(\tilde{x}_i) < \xi$ **then** break
    end
  end
  $\tilde{X} = Renormalize(\tilde{X})$
  return $\tilde{X}$
end

3.2 Random Pixel Flipping

The primary objective of Veil Privacy for visual data is to conceal privacy information perceptible to human eyes effectively. To achieve this, we devise the random pixel flipping $RPF(\cdot)$, which significantly enhances the visual dissimilarity between the veiled data $\tilde{X}$ and the original data X.

Initially, we normalize $x_i \in X$ via dividing by 255, following implementing $\frac{x-\mu}{\delta}$ with mean $\mu = 0.5$ and standard deviation $\delta = 0.5$. This preprocessing transforms the range of pixel values from 0∼255 to -1∼1, for the convenience of subsequent operations. Then we initialize $\tilde{X}$ with X.

The matrix $\tilde{m}$ is created with the same dimensions $C \times H \times W$. The values within $\tilde{m}$ are randomly drawn from a uniform distribution $U(0, 1)$, where values are between 0 and 1. We define the flipping ratio p∼(0,1). As shown in Eq. (1), we construct the random flipping mask $\tilde{m}$ by subtracting p and subsequently transforming values less than zero to -1 and otherwise to 1. Afterward, we update $\tilde{X}$ through element-wise multiplication between $\tilde{x}_i \in \tilde{X}$ and $\tilde{m}$.

$$\tilde{m}[i] = \begin{cases} 1, & if\ \tilde{m}[i] - p \geq 0 \\ -1, & if\ \tilde{m}[i] - p < 0 \end{cases}. \tag{1}$$

The parameter p plays a crucial role in determining the extent of visual disparity between $\tilde{X}$ and X. Experimentally, $p = 0.5$ is a good choice. The detailed design choice for the selection of p can be found in Sec. 5.4.

3.3 Gradient Iteration Algorithm

Another objective is that the latent features recognized by DNN models during training and prediction remain unchanged. To achieve this, we design the

gradient iteration algorithm $GIA(\cdot)$, which constrains the latent features of the veiled data $\tilde{X}$ to resemble those of the original data X closely.

We define the optimization problem presented in Eq. (2), where $\tilde{x}_i \in \tilde{X}$, $x_i \in X$ and $F(\cdot)$ denotes the local model trained using (X, Y). $F(x_i) \subset \mathbb{R}^n$ is the output vector of $F(\cdot)$ for x_i and n denotes the number of classes in Y.

$$\min L(\tilde{x}_i) = \alpha \cdot \|F(\tilde{x}_i) - F(x_i)\|_1 + \beta \cdot (-1) \cdot \|\tilde{x}_i - x_i\|_2. \tag{2}$$

We opt for Stochastic Gradient Descent (SGD) to generate $\tilde{X}$. SGD iteratively updates $\tilde{x}_i \in \tilde{X}$ in the direction opposite to the gradient $\bigtriangledown L(\cdot)$ to identify an optimal solution. We consider $F(x_i)$ as the explicit representation of the latent features. By reducing $\|F(\tilde{x}_i) - F(x_i)\|_1$, the disparity of latent features between $\tilde{x}_i$ and x_i gradually diminishes, where L_1 norm makes it easier to converge. $\|\tilde{x}_i - x_i\|_2$ ensures $\tilde{x}$ inconsistent with x_i, where L_2 norm is more capable of enhancing pixel value differences (approximating random noise is also an effective method). And α and β are coefficients. The iteration process is shown in Eq. (3), with max iteration number T and termination threshold ξ in Algorithm 1.

$$\tilde{x}_i = \tilde{x}_i - lr * \bigtriangledown_{\tilde{x}_i} L(\tilde{x}_i). \tag{3}$$

Note that we have defined $F(\cdot)$ as the last layer of the local model. In fact, different layers can be selected, resulting in distinct $\tilde{X}$. The last layer is an optimal choice for $F(\cdot)$ indeed. The detailed comparison can be found in Sec. 5.4.

3.4 Visual Similarity Measurement

Structural Similarity Index Measure (SSIM) [42] is frequently applied to measure the similarity between any two images x_1 and x_2. SSIM performs better than other metrics (*e.g.*, RMSE [6] and PSNR [1]) due to its alignment with human visual standards, consisting of luminance similarity, contrast similarity and structure similarity. SSIM yields values within range of 0 to 1, with higher values indicating greater similarity between x_1 and x_2. The definition of SSIM is shown in Eq. (4), where a, b and c represent the important degree, with mean μ and variance σ. C_1, C_2 and C_3 are tiny constants to prevent division by zero.

$$\begin{gathered} SSIM(x_1, x_2) = [L(x_1, x_2)]^a [C(x_1, x_2)]^b [S(x_1, x_2)]^c, \\ L(x_1, x_2) = \frac{2\mu_{x_1}\mu_{x_2}+C_1}{\mu_{x_1}^2+\mu_{x_2}^2+C_1}, C(x_1, x_2) = \frac{2\sigma_{x_1}\sigma_{x_2}+C_2}{\sigma_{x_1}^2+\sigma_{x_2}^2+C_2}, S(x_1, x_2) = \frac{\sigma_{x_1x_2}+C_3}{\sigma_{x_1}\sigma_{x_2}+C_3}. \end{gathered} \tag{4}$$

Generally, a=b=c=1 and C_3=$C_2/2$ are used to further simplify SSIM with Eq. (5). In this way, we measure SSIM between different images of the same person. Experimentally, SSIM for various pictures of the same one remains around 0.3.

$$SSIM(x_1, x_2) = \frac{(2\mu_{x_1}\mu_{x_2} + C_1)(2\sigma_{x_1x_2} + C_2)}{(\mu_{x_1}^2 + \mu_{x_2}^2 + C_1)(\sigma_{x_1}^2 + \sigma_{x_2}^2 + C_2)}. \tag{5}$$

4 Security Analysis

4.1 Threat Model

Attack Scenario. The data owner u generates the veiled data $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_n\}$ from the original data $X = \{x_1, ..., x_n\}$ with label $Y = \{y_1, ..., y_n\}$. This process involves a flipping ratio p and the local model $F(\cdot)$. Then, $(\tilde{X}, Y)$ is shared with the third party for training a classification model $\tilde{F}(\cdot)$. We adopt the black-box assumption on the access to $\tilde{F}(\cdot)$, allowing anyone to obtain $\tilde{F}(x)$ for input x. Additionally, there is also a risk of the leakage of $\tilde{X}$ and Y during data sharing.
Attacker Capability. Generally, the attacker is limited to acquiring $\tilde{F}(\cdot)$ in a black-box way. Meanwhile, when data sharing is not protected by u, attackers can also steal $\tilde{X}$ and Y. Consequently, the attacker's capabilities contain $(\tilde{X}, Y)$ and $\tilde{F}(\cdot)$. The goal of attackers is to reconstruct the original data X.

4.2 Hardness of Attack

Naive Attack. Due to a lack of knowledge of $RPF(\cdot)$ and $GIA(\cdot)$, it is hard for attackers to restore the original data X from the veiled data $\tilde{X}$.

For ORL, each image $x_i \in X$ has a size of 112×92, totaling 10,304 pixels. Other datasets may have larger dimensions. When the flipping ratio $p = 0.5$, the number of flipped pixels is 5,152. Potential variations in the flipping process result in an extremely large number of possibilities, *i.e.*, C_{10304}^{5152}. Additionally, distinct random flipping masks $\tilde{m}$ are constructed for each x_i. What's more, the specific structure of $F(\cdot)$ is important to generate $\tilde{X}$, even subtle variations in hyperparameters lead to distinct outputs. We have attempted some common denoising methods (*e.g.*, filtering), which are ineffective in restoring X from $\tilde{X}$. Consequently, achieving a successful naive attack is unattainable.
Adaptive Attack. Adaptive attack assumes attackers possess implementation of $RPF(\cdot)$ and $GIA(\cdot)$, as well as the flipping ratio p and the structure $\tilde{F}(\cdot)$.

Theoretically, $x_i \in X$ could be restored through reverse methods. However, random flipping masks $\tilde{m}$ are inherently non-identical, even using the same random pixel flipping $RPF(\cdot)$. Attempting to use $\tilde{m}$ corresponding to x_j to restore x_i results in more chaos rather than restoration. Additionally, $\tilde{F}(\cdot)$ may has a different structure from $F(\cdot)$. The optimization standard X for the gradient iteration algorithm $GIA(\cdot)$ can't be obtained and it typically finds local optimal solutions. Therefore, X is hard to be restored from $\tilde{X}$. In summary, the adaptive attack proves to be challenging in practice.

5 Experimental Evaluation

5.1 Setup

Datasets. For comprehensive experiments, we utilize many widely-used image datasets, including BioID, ORL, LFW and CelebA for face recognition tasks while MNIST, CIFAR10 and CIFAR100 for image classification tasks. Among

Table 2: Datasets' details and experimental results on their corresponding testing data. For each dataset, the baseline model is trained with the original training data, and the veiled model is trained with the generated veiled training data.

	#Dimensionality	#Class	#Instance	Backbone	Baseline Model	Veiled Model	
					Original Data	Veiled Data	Original Data
BioID	1*384*286	20	1,488	VGG16	99.34%	98.68%	98.01%
ORL	1*92*112	11	110	LeNet-5	96.97%	96.97%	96.97%
LFW	3*250*250	5,749	13,233	VGG16	88.89%	88.89%	85.19%
CelebA	3*178*218	10,177	202,599	VGG16	87.96%	87.09%	83.87%
MNIST	1*28*28	10	60,000	AlexNet	99.13%	98.58%	96.60%
CIFAR10	3*32*32	10	60,000	AlexNet	85.99%	84.09%	81.73%
CIFAR100	3*32*32	100	60,000	VIT	87.35%	83.22%	81.86%
RSS	3*512*512	45	197,121	ResNet18	88.50%	86.62%	85.36%

them, BioID comprises 20 identities, totaling 1,488 grayscale facial images, with dimensions of 384 $\times$ 286. ORL contains 11 identities, each contributing 10 grayscale facial images, with dimensions of 92 $\times$ 112. LFW includes 5749 persons with 13,233 RGB facial images of 250 $\times$ 250 size. CelebA comprises 10,177 identities, including 202,599 RGB images with 178 $\times$ 218 size. MNIST contains 10 categories, a total of 60,000 grayscale images, with 28 $\times$ 28 size. CIFAR10 involves 10 classes, a total of 60,000 RGB images, with 32 $\times$ 32 size, while CIFAR100 involves 100 classes. RSS is considered for remote sensing scene classification, with 512 $\times$ 512 sizes, 45 classes, and a total of 197,121 RGB images.
Networks. Various commonly used DNN models are employed as the baseline model for different datasets. We choose VGG16 [31], LeNet-5 [17], AlexNet [16], VIT [9] and ResNet18 [11] for corresponding different datasets. In addition, we have verified that other networks are also feasible.
Training Settings. CrossEntropy is chosen as the loss function and Stochastic Gradient Descent (SGD) is applied as the optimizer. Flipping ratio $p = 0.5$. Hyper-parameters $\alpha = 0.999$ and $\beta = 0.001$. Max iteration number $T = 300$ and termination threshold $\xi = 0.3$. Details of setup are available in Tab. 2.

5.2 Veiled Data Generation

Firstly, we train baseline models on the original data for different datasets. In Tab. 2, the baseline models achieve 99.34%, 96.97%, 88.89%, 87.96%, 99.13%, 85.99%, 87.35% and 88.50% accuracy for BioID, ORL, LFW, CelebA, MNIST, CIFAR10, CIFAR100 and RSS datasets respectively.

Afterwards, we generate the veiled data $\tilde{X}$ from the original data X via the Veil Privacy framework. It results in a significant visual distinction between $\tilde{X}$ and X, while ensuring that $\tilde{X}$ and X remain similar latent features to $F(\cdot)$. Examples of X and $\tilde{X}$ are shown in Fig. 2. For different datasets, $\tilde{X}$ has effectively hidden the visual privacy of X. As shown in Tab. 3, we measure the SSIM between $\tilde{X}$ and X. All SSIMs almost remain below 0.05, far below 0.3 which indicates the SSIM between different images of the same class. As an efficient framework, Veil Privacy has a low time complexity (See the supplement C).

Fig. 2: Examples of the original data X and the veiled data $\tilde{X}$ for different datasets.

Table 3: SSIM between X and $\tilde{X}$. Examples 1, 2 and 3 are shown in Fig. 2 respectively.

	BioID	ORL	LFW	CelebA	MNIST	CIFAR10	CIFAR100
Example 1	0.02903	0.03130	0.06522	0.03024	0.02130	0.00331	0.08573
Example 2	0.05010	0.04346	0.05702	0.04417	0.02835	0.01100	0.02004
Example 3	0.05238	0.03576	0.05366	0.01773	0.04311	0.02144	0.02988

5.3 Main Experiments

Testing on Veiled Data and Original Data. We substitute X with $\tilde{X}$ to train the veiled model and the results are presented in Tab. 2. The veiled model for BioID achieves 98.68% accuracy on $\tilde{X}$ and 98.01% accuracy on X. The veiled model for ORL achieves 96.97% accuracy on $\tilde{X}$ and 96.97% accuracy on X. The veiled model for LFW achieves 88.89% accuracy on $\tilde{X}$ and 85.19% accuracy on X. The veiled model for CelebA achieves 87.09% accuracy on $\tilde{X}$ and 83.87% accuracy on X. The veiled model for MNIST achieves 98.58% accuracy on $\tilde{X}$ and 96.60% accuracy on X. The veiled model for CIFAR10 achieves 84.09% accuracy on $\tilde{X}$ and 81.73% accuracy on X while the veiled model for CIFAR100 achieves 83.22% accuracy on $\tilde{X}$ and 81.86% accuracy on X. The veiled model for RSS achieves 86.62% accuracy on $\tilde{X}$ and 85.36% accuracy on X. $\tilde{X}$ has almost maintained the same utility for DNNs as X.

It validates the effectiveness of Veil Privacy in balancing the utility-privacy trade-off. The veiled model not only predicts the veiled data but also efficiently recognizes the original data. As a pioneering work for this practical problem, we establish our own standard baselines for comparison (details seen in *Supplementary Material D*). Furthermore, we further investigate the utility of the veiled data $\tilde{X}$ for combination and fine-tuning training on facial datasets.

Table 4: Results of combination&fine-tuning training, transferability and comparison.

	Combination Training			Fine-tuning Training			Transferability	
	Total	Original Data	Veiled data	Total	Original Data	Veiled data	Veiled Data	Original Data
BioID	99.34%	99.33%	100.00%	99.35%	99.34%	100.00%	98.01%	95.36%
ORL	96.97%	96.88%	100.00%	97.22%	96.97%	100.00%	96.97%	96.97%
LFW	88.89%	86.37%	100.00%	89.29%	88.89%	100.00%	85.19%	85.07%
CIFAR10				[3]	[25]	[26]	[8]	Veil Privacy
Accuracy on Original Data				64.3±0.7%	66.3±0.2%	64.7±0.2%	57.5±0.5%	81.73%

Combination Training. We substitute one class of the original data X with its the corresponding veiled data $\tilde{X}$. Consequently, the combination dataset comprises one class of $\tilde{X}$ and other classes of X. Subsequently, the combination model is trained using the combination dataset.

The results are presented in Tab. 4. For BioID, the combination model achieves 99.34% accuracy totally, while 99.33% accuracy on X and 100% accuracy on $\tilde{X}$. For ORL, the combination model achieves 96.97% accuracy totally, while 96.88% accuracy on X and 100% accuracy on $\tilde{X}$. For LFW, the combination model achieves 88.89% accuracy totally, while 86.37% accuracy on X and 100% accuracy on $\tilde{X}$. More importantly, all the combination models achieve 100% accuracy on the one corresponding substituted class of the original data. It indicates the effectiveness of the veiled data $\tilde{X}$ for combination Training.

Fine-tuning Training. We have reserved one class in advance for fine-tuning. The veiled data $\tilde{X}$ of the new class is added to the original dataset to build the fine-tuning dataset. Therefore, the fine-tuning dataset comprises one new class of $\tilde{X}$ and other classes of X. We employ the baseline model as the pre-trained model and fix all parameters while only retraining the last classification layer.

The results are shown in Tab. 4. For BioID, the fine-tuning model achieves 99.35% accuracy totally, while 99.34% accuracy on X and 100% accuracy on the new $\tilde{X}$. For ORL, the fine-tuning model achieves 97.22% accuracy totally, while 96.97% accuracy on X and 100% accuracy on the new $\tilde{X}$. For LFW, the fine-tuning model achieves 89.29% accuracy totally, while 88.89% accuracy on X and 100% accuracy on the new $\tilde{X}$. And all the fine-tuning models achieve 100% accuracy on the one corresponding new class of the original data. It validates the effectiveness of the veiled data $\tilde{X}$ for fine-tuning training.

Transferability of Veiled Data. Inspired by Demontis *et al*. [4] which demonstrates the transferability of adversarial examples by the intrinsic adversarial vulnerability of the transfer model and the surrogate model's complexity, we also explore the transferability of the veiled data $\tilde{X}$. Here, we consider the baseline model as the surrogate model, with AlexNet serving as the transfer model. For three facial datasets, we train the transfer model using veiled data $\tilde{X}$ that is generated with the baseline model.

The results are shown in Tab. 4. For BioID, the transfer model achieves 98.01% accuracy on $\tilde{X}$ and 95.36% accuracy on X. For ORL, the transfer model achieves 96.97% accuracy on $\tilde{X}$ and 96.97% accuracy on X. For LFW, the transfer model achieves 85.19% accuracy on $\tilde{X}$ and 85.07% accuracy on X. It proves a notable degree of transferability exhibited by the veiled data $\tilde{X}$.

Method Comparisons. As a pioneering work and the first complete solution, it is difficult to compare Veil Privacy with existing works on the same task. As stated in Section II, if dataset distillation is regarded as a privacy technique, our framework preserves the utility of the original data more effectively. Advanced dataset distillation techniques (*e.g.*, [3], [25], [26] and [8]) respectively achieved only around 64.3%, 66.3%, 64.7% and 57.5% accuracy (See Tab. 4) on CIFAR10 with AlexNet, compared to our 81.73%. Furthermore, we also conduct more basic explorations (See supplement A) for a more comprehensive comparison.

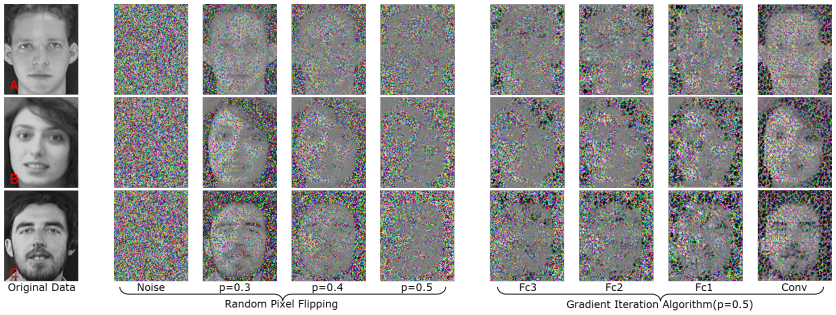

Fig. 3: Ablation. Examples for $\tilde{X}$ that are generated with different p and $F(\cdot)$

Table 5: Ablation Study. Images A, B and C are shown in Fig. 3.

		Random Pixel Flipping				Gradient Iteration Algorithm($p = 0.5$)			
		$Noise$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$Fc3$	$Fc2$	$Fc1$	$Conv$
SSIM	Image A	0.01031	0.14702	0.08653	0.03478	0.03130	0.04429	0.05292	0.10403
	Image B	0.00735	0.17094	0.09527	0.03607	0.04346	0.05021	0.06085	0.11821
	Image C	0.00723	0.18134	0.09172	0.01760	0.03576	0.04818	0.05899	0.11647
ACC	$\tilde{X}$	93.94%	96.97%	96.97%	96.97%	96.97%	96.97%	96.97%	96.97%
	X	55.76%	96.97%	96.97%	96.97%	96.97%	96.97%	96.97%	96.97%

5.4 Ablation Study

Random Pixel Flipping. Random pixel flipping $RPF(\cdot)$ is applied to create the visual difference between the veiled data $\tilde{X}$ and the original data X. Here, we explore the optimal setting of the flipping ratio p. For comparison, we replace random flipping with generating random noise directly. With the same gradient iteration algorithm, we generate different $\tilde{X}$. Fig. 3 shows some examples and SSIM between $\tilde{X}$ and X are shown in Tab. 5. It's observed that the visual difference between $\tilde{X}$ and X scales up gradually with the increase of p.

We train the veiled model with different $\tilde{X}$. The results are shown in Tab. 5. All the veiled models with different p achieve high accuracy both on $\tilde{X}$ and X. Notably, the veiled model with noise achieves 96.94% accuracy on $\tilde{X}$ but only 55.76% accuracy on X. We also find that if one initializes $GIA(\cdot)$ with X (*i.e.* abandon $RPF(\cdot)$), it will completely fail to result in a significant visual distinction. It indicates the random pixel flipping $RPF(\cdot)$ not only hides visual privacy information but also preserves latent features of the original data to a certain extent. Overall, $p = 0.5$ has a suitable balance between concealing visual privacy and retaining latent features.

Gradient Iteration Algorithm. Gradient iteration algorithm $GIA(\cdot)$ makes the veiled data $\tilde{X}$ retain similar latent features to the original data X. Here, we explore the best choice among different layers of the local model $F(\cdot)$. Specifically, three fully connected layers $Fc3(\cdot)$, $Fc2(\cdot)$, $Fc1(\cdot)$ and the last convolutional layer $Conv(\cdot)$ are assayed. With the same random pixel flipping $RPF(\cdot)$ ($p = 0.5$), we generate different $\tilde{X}$. Fig. 3 shows some examples and SSIM between $\tilde{X}$

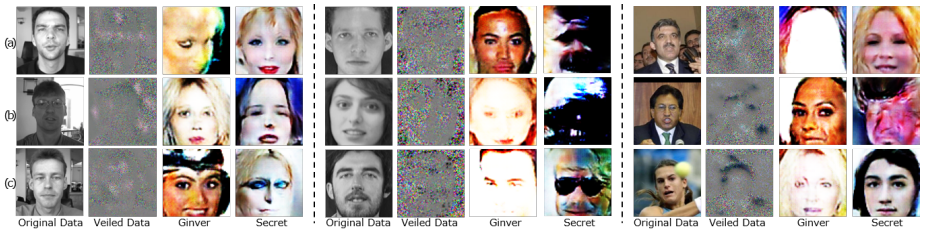

Fig. 4: Advanced MIAs (Ginver and Secret) on BioID, ORL and LFW

Table 6: SSIM compared to X. Images a, b and c are shown in Fig. 4.

	BioID			ORL			LFW		
	Veiled Data	Ginver	Secret	Veiled Data	Ginver	Secret	Veiled Data	Ginver	Secret
Image a	0.02870	0.08195	0.02068	0.03130	0.09019	0.02905	0.06522	0.05163	0.07567
Image b	0.08073	0.09552	0.05144	0.04346	0.04801	0.01224	0.05702	0.05223	0.00331
Image c	0.05563	0.00128	0.06491	0.03576	0.07975	0.01070	0.05428	0.05996	0.00566

and X are reported in Tab. 5. As $F(\cdot)$ moves closer to the input layer, $\tilde{X}$ and X exhibit greater visual similarity.

We train the veiled model with different $\tilde{X}$. The results are presented in Tab. 5. Each veiled model achieves high accuracy on both $\tilde{X}$ and X. Notably, the veiled model with $Fc3(\cdot)$ achieves 96.97% accuracy on $\tilde{X}$ and 96.97% accuracy on X, offering the most effective concealment of visual privacy. And we also find that keeping $RPF(\cdot)$ with p=0.5 while removing $GIA(\cdot)$ will destroy the utility of data. Therefore, selecting the last layer of the local model as $F(\cdot)$ is a well-considered design choice.

5.5 Security Evaluation

Model Inversion Attack (MIA). MIA aims to reconstruct training data through the prediction vector of the attacked model. In our attack scenario, a successful attack means the restoration of the veiled data $\tilde{X}$ through the veiled $\tilde{F}(\cdot)$. However, we explore the risk of whether the original data X are possible to be restored through MIA on $\tilde{F}(\cdot)$ due to similar latent features with $\tilde{X}$.

We conducted two advanced MIAs, *i.e.*, Ginver [46] and Secret [48], with labels to attack the classification model $\tilde{F}(\cdot)$. As shown in Fig. 4, the reconstructed images are visually completely different from the original ones. We measure SSIM between original images and restored images. As shown in Tab. 6, SSIMs between the original data and the restored samples are almost below 0.1. It indicates that our framework has a defensive capability against MIAs.

Image Restoration Method. We further examine the robustness of the framework against advanced image restoration methods. Specifically, we implement GAN Prior Embedded Network [45] (GPEN) to restore the original data X from the veiled data $\tilde{X}$. As a comparison, we add gaussian noise into the original images and employ GPEN on the noised data.

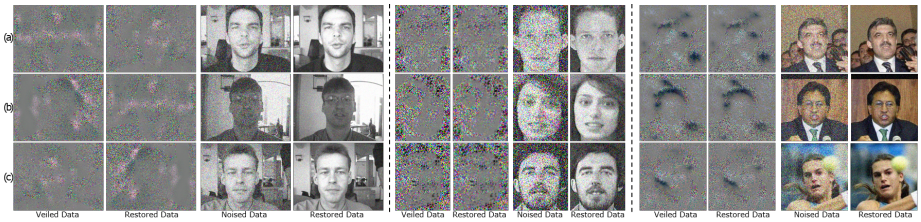

Fig. 5: Image Restoration Method (GPEN) on BioID, ORL and LFW

Table 7: SSIM compared to X. Images a, b and c are shown in Fig. 5.

	BioID				ORL				LFW			
	Veiled	Restored	Noised	Restored	Veiled	Restored	Noised	Restored	Veiled	Restored	Noised	Restored
Image a	0.02903	0.06315	0.09249	0.33262	0.03130	0.08289	0.10282	0.36875	0.06522	0.06357	0.11898	0.43874
Image b	0.05010	0.05621	0.08740	0.34118	0.04346	0.08676	0.11860	0.41322	0.05702	0.06323	0.12707	0.40150
Image c	0.05238	0.08083	0.09153	0.32008	0.03576	0.04736	0.09712	0.48241	0.03576	0.04393	0.11516	0.40930

The results are shown in Fig. 5. GPEN effectively restores the noised images but fails to restore the original images from the veiled ones. We measure SSIM between the original images and others, and the results are shown in Tab. 7. SSIM between the original data and the restored data remains below 0.1, which means reconstructed images are almost completely different from the original ones. It further proves the security of the Veil Privacy framework. More classical security experiments can be found in the supplement B.

6 Conclusion

Existing works on data privacy protection have encountered numerous difficulties in balancing the trade-off between utility and privacy. *How can we generate privacy-preserving surrogate data for utilization or sharing without a significant performance degradation?* In this paper, we introduce a remarkably effective framework, Veil Privacy, for safeguarding the privacy of visual data. This framework generates the veiled data by designing two key techniques: the random pixel flipping and the gradient iteration algorithm. In contrast to the original data, the veiled data retains latent features while expunging visual privacy information. Extensive evaluations conducted across various datasets and models demonstrate the effectiveness and security of the framework. In future research, we will further explore the balance between utility and privacy for adjusting hyper-parameters and different customized AI security services, *e.g.*, the preservation of privacy attributes such as gender and age within the veiled data.

Acknowledgements. This work is supported by the National Key R&D Program of China (No.2023YFB2703900), National Natural Science Foundation of China (No.62206128), National Key R&D Program of China (No.2022YFB3103800) and National Natural Science Foundation of China (No.U2336205).

References

1. Almohammad, A., Ghinea, G.: Stego image quality and the reliability of psnr. In: 2010 2nd International Conference on Image Processing Theory, Tools and Applications. pp. 215–220. IEEE (2010)
2. Boddeti, V.N.: Secure face matching using fully homomorphic encryption. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–10. IEEE (2018)
3. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4750–4759 (2022)
4. Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., Roli, F.: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th USENIX security symposium (USENIX security 19). pp. 321–338 (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dennison, P.E., Roberts, D.A.: Endmember selection for multiple endmember spectral mixture analysis using endmember average rmse. Remote sensing of environment **87**(2-3), 123–135 (2003)
7. Dimiccoli, M., Marín, J., Thomaz, E.: Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(4), 1–18 (2018)
8. Dong, T., Zhao, B., Lyu, L.: Privacy for free: How does dataset condensation help privacy? In: International Conference on Machine Learning. pp. 5378–5396. PMLR (2022)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in google street view. In: 2009 IEEE 12th international conference on computer vision. pp. 2373–2380. IEEE (2009)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hesamifard, E., Takabi, H., Ghasemi, M.: Cryptodl: Deep neural networks over encrypted data. arXiv preprint arXiv:1711.05189 (2017)
13. Hoofnagle, C.J., van der Sloot, B., Borgesius, F.Z.: The european union general data protection regulation: what it is and what it means. Information & Communications Technology Law **28**(1), 65–98 (2019)
14. Huang, Y., Song, Z., Li, K., Arora, S.: Instahide: Instance-hiding schemes for private distributed learning. In: International conference on machine learning. pp. 4507–4518. PMLR (2020)
15. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)

16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
18. Lee, J.W., Kang, H., Lee, Y., Choi, W., Eom, J., Deryabin, M., Lee, E., Lee, J., Yoo, D., Kim, Y.S., et al.: Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. IEEE Access **10**, 30039–30054 (2022)
19. Li, F., Li, H., Niu, B., Chen, J.: Privacy computing: concept, computing framework, and future development trends. Engineering **5**(6), 1179–1192 (2019)
20. Liang, D., Xu, W., Bai, X.: An end-to-end transformer model for crowd localization. In: European Conference on Computer Vision. pp. 38–54. Springer (2022)
21. Liu, B., Ding, M., Zhu, T., Xiang, Y., Zhou, W.: Using adversarial noises to protect privacy in deep learning era. In: 2018 IEEE Global Communications Conference (GLOBECOM). pp. 1–6. IEEE (2018)
22. Liu, B., Xiong, J., Wu, Y., Ding, M., Wu, C.M.: Protecting multimedia privacy from both humans and ai. In: 2019 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB). pp. 1–6. IEEE (2019)
23. Mahesh, R., Meyyappan, T.: Anonymization technique through record elimination to preserve privacy of published data. In: 2013 international conference on pattern recognition, informatics and mobile engineering. pp. 328–332. IEEE (2013)
24. Montenegro, H., Silva, W., Gaudio, A., Fredrikson, M., Smailagic, A., Cardoso, J.S.: Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy. IEEE Access **10**, 28333–28347 (2022). `https://doi.org/10.1109/ACCESS.2022.3157589`
25. Nguyen, T., Chen, Z., Lee, J.: Dataset meta-learning from kernel ridge-regression. arXiv preprint arXiv:2011.00050 (2020)
26. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. Advances in Neural Information Processing Systems **34**, 5186–5198 (2021)
27. Peng, L., Wu, X., Yang, Z., Liu, H., Cai, D.: Did-m3d: Decoupling instance depth for monocular 3d object detection. In: European Conference on Computer Vision. pp. 71–88. Springer (2022)
28. Sanchez-Matilla, R., Li, C.Y., Shamsabadi, A.S., Mazzon, R., Cavallaro, A.: Exploiting vulnerabilities of deep neural networks for privacy protection. IEEE Transactions on Multimedia **22**(7), 1862–1873 (2020)
29. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)
30. Shen, W., Wu, Z., Zhang, J.: A face privacy protection algorithm based on block scrambling and deep learning. In: Cloud Computing and Security: 4th International Conference, ICCCS 2018, Haikou, China, June 8–10, 2018, Revised Selected Papers, Part III 4. pp. 359–369. Springer (2018)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Sirichotedumrong, W., Maekawa, T., Kinoshita, Y., Kiya, H.: 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). pp. 674–678. IEEE (2019)
33. de la Torre, L.: A guide to the california consumer privacy act of 2018. Available at SSRN 3275571 (2018)

34. Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., Gavrila, D.M., et al.: Privacy protection in street-view panoramas using depth and multi-view imagery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10581–10590 (2019)
35. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2116–2127 (2023)
36. Vyas, N., Kakade, S.M., Barak, B.: On provable copyright protection for generative models. In: International Conference on Machine Learning. pp. 35277–35299. PMLR (2023)
37. Wang, H., Bai, Y., Sun, G., Liu, J.: Privacy protection based on mask template. arXiv preprint arXiv:2202.06250 (2022)
38. Wang, H., Sun, G., Zheng, K., Li, H., Liu, J., Bai, Y.: Privacy protection generalization with adversarial fusion. Math. Biosci. Eng **19**, 7314–7336 (2022)
39. Wang, H., Tian, H., Chen, J., Wan, X., Xia, J., Zeng, G., Bai, W., Jiang, J., Wang, Y., Chen, K.: Towards {Domain-Specific} network transport for distributed {DNN} training. In: 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24). pp. 1421–1443 (2024)
40. Wang, X., Li, J., Kuang, X., Tan, Y.a., Li, J.: The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing **130**, 12–23 (2019)
41. Wang, X., Asif, H., Vaidya, J.: Preserving missing data distribution in synthetic data. In: Proceedings of the ACM Web Conference 2023. pp. 2110–2121 (2023)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
43. Xue, H., Liu, B., Din, M., Song, L., Zhu, T.: Hiding private information in images from ai. In: ICC 2020-2020 IEEE International Conference on Communications (ICC). pp. 1–6. IEEE (2020)
44. Yang, J.: Privacy protection for visual data against deep learning based computer vision models (2021)
45. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 672–681 (2021)
46. Yin, Y., Zhang, X., Zhang, H., Li, F., Yu, Y., Cheng, X., Hu, P.: Ginver: Generative model inversion attacks against collaborative inference. In: Proceedings of the ACM Web Conference 2023. pp. 2122–2131 (2023)
47. Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J.: iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Transactions on Information Forensics and Security **12**(5), 1005–1016 (2016)
48. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: Generative model-inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 253–261 (2020)
49. Zhao, C., Hu, Y., Salzmann, M.: Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects. In: European Conference on Computer Vision. pp. 106–122. Springer (2022)
50. Zhou, H., Zhang, H., Deng, H., Liu, D., Shen, W., Chan, S.H., Zhang, Q.: Explaining generalization power of a dnn using interactive concepts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17105–17113 (2024)