SecretGen: Privacy Recovery on Pre-Trained Models via Distribution Discrimination

Zhuowen Yuan[1], Fan Wu[1], Yunhui Long[1], Chaowei Xiao[2], and Bo Li[1]

[1] University of Illinois Urbana-Champaign
[2] Arizona State University

Abstract. Transfer learning through the use of pre-trained models has become a growing trend for the machine learning community. Consequently, numerous pre-trained models are released online to facilitate further research. However, it raises extensive concerns on whether these pre-trained models would leak privacy-sensitive information of their training data. Thus, in this work, we aim to answer the following questions: "Can we effectively recover private information from these pre-trained models? What are the sufficient conditions to retrieve such sensitive information?" We first explore different statistical information which can discriminate the private training distribution from other distributions. Based on our observations, we propose a novel private data reconstruction framework, SecretGen, to effectively recover private information. Compared with previous methods which can recover private data with the ground truth label of the targeted recovery instance, SecretGen does not require such prior knowledge, making it more practical. We conduct extensive experiments on different datasets under diverse scenarios to compare SecretGen with other baselines and provide a systematic benchmark to better understand the impact of different auxiliary information and optimization operations. We show that without prior knowledge about true class prediction, SecretGen is able to recover private data with similar performance compared with the ones that leverage such prior knowledge. If the prior knowledge is given, SecretGen will significantly outperform baseline methods. We also propose several quantitative metrics to further quantify the privacy vulnerability of pre-trained models, which will help the model selection for privacy-sensitive applications. Our code is available at: `https://github.com/AI-secure/SecretGen`.

Keywords: Privacy; Pre-trained models; Transfer learning

1 Introduction

As machine learning has achieved great successes in different domains, such as robotics [24], audio recognition [7], and face recognition [15], how to train the learning models efficiently given the available large-scale dataset becomes a timely problem. Transfer learning, which focuses on transferring knowledge across domains, is a promising learning paradigm [2]. In particular, many pre-trained models are available currently, such as TensorFlow Hubs [1] and PyTorch Hubs [22], which can be flexibly used for fine-tuning later for different

downstream tasks. As a result, the training paradigm with transfer learning has enabled efficient usage of the large-scale dataset without requiring training every model from scratch.

However, such an efficient transfer learning paradigm also leads to additional *privacy concerns*. For instance, if the training data of the pre-trained models contain privacy-sensitive information, an adversary who downloads the pre-trained models could potentially perform different privacy attacks to infer the private information. In particular, membership inference attacks [18,19] have been studied to infer whether a private instance is in the training set, and model inversion attacks have been studied to reconstruct the private training instances under certain assumptions [28,11,10,26], which raises more privacy and safety concerns.

To better understand the privacy vulnerabilities of such pre-trained models, a comprehensive analysis of different types of privacy attacks, especially the severe model inversion attacks, is required. Currently, there are several limitations of existing privacy model inversion attacks. First, the current *state-of-the-art* model inversion attack (i.e., GMI) [28] requires the ground truth label of the reconstructed instances, which is less practical. Furthermore, it is a known challenging problem to label the generated instances based on GANs [12]. Second, many existing model inversion attacks require whitebox access to the target pre-trained model, making it less practical in real-world applications. Thus, in this paper, we mainly aim to ask: *Can we reconstruct private sensitive training instances without requiring such information?*

To answer it, we propose a general private data recovery framework SecretGen, which consists of a generation backbone, a pseudo label predictor, and a latent vector selector. We first use a pseudo label predictor to generate a pseudo label for each private instance. Specifically, we randomly sample latent vectors and feed them into the generation backbone to get recovered instances. To stabilize prediction quality, we apply different transformations (*e.g.*cutouts) to such instances before feeding them into the targeted model to get the final predicted pseudo labels. We then propose a latent vector selector via a proposed selection algorithm to further optimize and constrain the recovery space. Finally, we perform joint optimization to train the end-to-end framework as shown in Fig. 1.

We conduct comprehensive experiments to evaluate the proposed SecretGen compared with multiple baselines. We show that SecretGen significantly outperforms baselines given the same ground truth label. Even without such information, SecretGen still achieves comparable performance compared to baselines which leverage the ground truth label information. In addition, to evaluate the performance of recovered data on downstream tasks, we propose different evaluation protocols considering different usage of the recovered private data, and we show that our observations are consistent for different protocols. We also evaluate the robustness of SecretGen against the purification defense [27]. Finally, we perform different ablation studies to show the effectiveness of our design choices. We make the following **contributions**:

- We propose a general private data recovery attack (i.e., model inversion) given a pre-trained model, SecretGen, without requiring the ground truth label as prior knowledge under both whitebox and blackbox settings.
- We propose a novel label predictor for the reconstructed instances considering different data transformations and latent vector selection, which can be flexibly used in other frameworks.
- We propose different evaluation protocols and metrics for evaluating the pre-trained models against general model inversion attacks.
- We conduct extensive experiments on different models, including the vision transformer and multiple datasets, to provide a benchmark on model inversion attacks. We show that SecretGen significantly outperforms baselines under different settings.

2 Related Work

Revealing privacy-sensitive information from a trained model has aroused extensive research interest. *Membership inference attacks* and *model inversion attacks* are two major categories of such attacks. In *membership inference attacks* [18,19], the adversary aims to decide whether a sample is a member of the training set, while in *model inversion attacks* [28,11,10,26], the adversary attempts to reconstruct the training set under certain assumptions.

[11] was the first to propose model inversion attacks aiming at recovering private training data. The authors demonstrated that personal genetic markers could be effectively recovered given the output of the model and auxiliary knowledge. [10] extended model inversion to more complex models, including shallow neural networks for face recognition. The recovered data with their proposed method are identified as the original person at a much higher rate than random guessing. However, the reconstructed images are blurry and not visually recognizable to humans. [26] proposed a training-based attack by training an auto-encoder on public data. The attack can be performed with *blackbox* accesses to the target model and partial (truncated) model predictions.

More recently, [28] proposed generative model inversion attack (GMI). The authors distill public knowledge by training a conditional GAN on public data and then solve an optimization problem to maximize the probability of the recovered image for the ground truth class label. GMI significantly outperforms previous methods in re-identification rate of the recovered data, as well as guaranteeing the recovered data are visually plausible. However, they still require the ground truth label for the target image and *whitebox* access to the victim model, which is often not accessible to the adversary. Another recent work distributional model inversion attack (DMI) [3] recovers the private data distribution for each target class by constructing representative samples. However, DMI does not support recovering every private instance given its non-sensitive version (*i.e.* instance-level model inversion), which is the adversary's goal in our setting.

3 Methodology

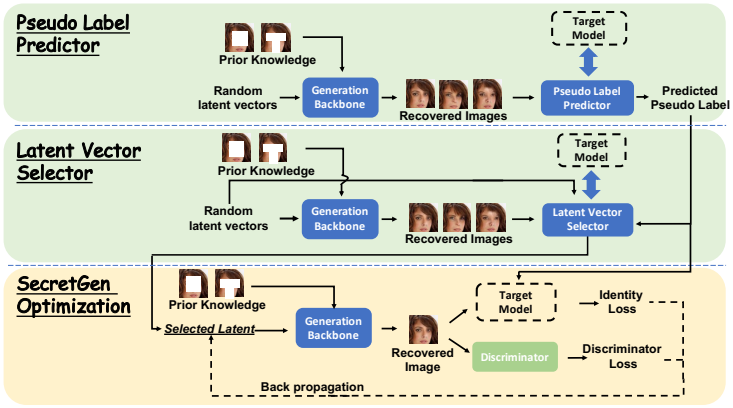

Fig. 1. Overview of the proposed SecretGen. The **blue modules** represent the proposed algorithms. The *Target Model* could allow either whitebox or blackbox access.

3.1 Problem Formulation

We focus on recovering the privacy-sensitive training data based on the trained classification models. Throughout the paper, we will refer to the model that is subjective to attacks as the *target model* F, which is trained on private training data $D_{\text{pri}}^{\text{train}}$, aiming to perform evaluation on private test data $D_{\text{pri}}^{\text{test}}$. The target model returns a prediction vector $F(x)$ given an input instance x. The prediction vector represents a probability distribution over C classes, where C denotes the number of classes of the whole private dataset $D_{\text{pri}} = D_{\text{pri}}^{\text{train}} \cup D_{\text{pri}}^{\text{test}}$.

The adversary's **goal** is to recover the private training data $D_{\text{pri}}^{\text{train}}$ given the trained target model F and certain prior information, *e.g.*, partially corrupted images from $D_{\text{pri}}^{\text{train}}$. In particular, such corrupted images only contain non-sensitive background information (pixels) x_{ns} with the sensitive region x_s cropped out. These corrupted images are usually easy to obtain, given that such corruption is often applied to protect the privacy of individuals in practice [28]. Specifically, in our evaluation, we consider cropping the whole face using two face datasets, leaving only the non-sensitive background regions (Section 4).

Regarding the adversary's **ability**, we consider (1) *whitebox* access to the target model, where all parameters and intermediate computations of the target model are visible to the adversary, and (2) *blackbox* access to the target model, where the adversary can only obtain the final prediction from the target model F. Additionally, we assume that the adversary also has access to some public data D_{pub} from the similar distribution for general training purpose.

3.2 Method Overview

An overview of SecretGen is illustrated in Fig. 1, where SecretGen takes non-sensitive information x_{ns} as input and returns the recovered images that contain privacy-sensitive training information (*e.g.*, human faces). SecretGen is composed of *three* components: *generation backbone*, *pseudo label predictor*, and *latent vector selector*, which are jointly optimized under a unified framework. The **generation backbone** leverages a conditional GAN trained on public data as a backbone to generate realistic images based on the prior information (*e.g.*, cropped images), and the generation process is controlled by the latent vector z sent to the GAN's generator G. The **pseudo label predictor** predicts the most possible pseudo label for each recovered private image based on the distributional statistics of recovered images. The **latent vector selector** selects the optimal latent vector $\hat{z}$ which is the most likely to contain privacy-sensitive information based on the proposed selection algorithm. Finally, we perform joint optimization on the selected $\hat{z}$, taking the pseudo label provided by the pseudo label predictor as the prediction target, to reconstruct image $G(\hat{z}^*, x_{ns})$. In the next following sections, we will describe each component in detail.

3.3 Generation Backbone of SecretGen

To recover the privacy-sensitive training data, we train a generation backbone for conditional image recovery on public data D_{pub}. In particular, we will start from certain prior knowledge, such as the corrupted private data containing only the nonsensitive information x_{ns}. We then perform the same corruption operation *corr* on D_{pub} to construct the training set for the generation backbone: $D_{\text{pub\_corr}} = \{corr(x)|x \in D_{\text{pub}}\}$.

Next, we train a conditional GAN which is composed of two networks: generator G and discriminator D. G is conditioned on $x_{ns} \in D_{\text{pub\_corr}}$ and z is the latent vector which is sampled from a prior distribution during training. Throughout the paper, we use the prior distribution as standard Gaussian distribution. We leverage the Wasserstein-GAN loss [13] for GAN training:

$$\min_G \max_D \mathcal{L}_{\text{wgan}} = \mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z, x_{ns}))] \tag{1}$$

We also incorporate a diversity loss term $\mathcal{L}_{\text{div}}$ [25] for training the generator to prevent mode collapsing by sampling different latent vectors, say, z_1 and z_2:

$$\mathcal{L}_{\text{div}} = -\mathbb{E}_{z_1,z_2}\left[\frac{\|f(G(z_1, x_{ns})) - f(G(z_2, x_{ns}))\|}{\|z_1 - z_2\|}\right] \tag{2}$$

where f is the feature extractor of the target model, which returns the feature embeddings of the input images in the *whitebox* setting. In the *blackbox* setting, we use a feature extractor trained on public data f_{pub} for this process. The overall loss term for the generator is as following:

$$\mathcal{L}_G = \mathcal{L}_{\text{wgan}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} \tag{3}$$

After the generation backbone is trained, we freeze the parameters for both G and D before we enter the next stage. We denote $\hat{x}$ as the recovered image, *i.e.*, $\hat{x} = G(z, x_{ns})$.

3.4 Pseudo Label Predictor

The main challenge in this data reconstruction process is that we have no knowledge about the ground truth label of the private images (related work assumes that they have access to the ground truth label [28,26,11,10], while we do not). To tackle this problem, we propose a *pseudo label predictor* which infers the label prediction with proposed discrimination metrics. We will first introduce the design of our discrimination metric, and then we elaborate on how the pseudo label predictor is optimized.

Discrimination Metric. Given the certain prior knowledge x_{ns}, we randomly sample n latent vectors $\{z_i\}_{i=1}^n$ from the prior distribution. We generate n recovered images using our generation backbone: $\{\hat{x}_i\}_{i=1}^n$, where $\hat{x}_i = G(z_i, x_{ns})$. In order to improve the prediction stability, we consider prediction under different transformations. Concretely, let the list of considered transformation functions be $\mathcal{T} = \{t_i\}_{i=1}^m$. On each recovered image $\hat{x}_i$, we perform m transformations independently to obtain m transformed images $\{\tilde{x}_i^j\}_{j=1}^m$, where $\tilde{x}_i^j = t_j(\hat{x}_i)$. We additionally define $\tilde{x}_i^0 = \hat{x}_i$. Let $F_c(\cdot)$ denote the model's prediction confidence for class label c based on target model F. We define the discrimination metric $\mathcal{M}$ on label c as follows:

$$\mathcal{M}(c; n, m) \triangleq \frac{1}{n(m+1)} \sum_{i=1}^{n} \sum_{j=0}^{m} F_c(\tilde{x}_i^j), \quad \forall c \in [1, C]. \tag{4}$$

The discrimination metric returns a score indicating how likely it is for a label c to be the consistent prediction across different transformations. Based on existing studies of contrastive learning [4], we will select the class c with the highest discrimination metric score as the final label prediction.

In particular, we define the list of transformations as a sequence of fix-sized cutouts. We split an image into fix-sized patches and define t_j as the transformation that cuts out the j-th patch of the given image, as illustrated in Fig. 2.

Intuitively, the discrimination metric $\mathcal{M}(c; n, m)$ should preserve the following properties. First, $\mathcal{M}(c; n, m)$ is likely to have a higher score when c equals the label associated with the corrupted image x_{ns} since the model has learned some correlation between the non-sensitive background information in x_{ns} and the label of the original image. Such correlation should be stronger if the target model is more overfitted to private training data. Second, when the recovered image $\tilde{x}_i^j$ is close to the training data, $F_c(\tilde{x}_i^j)$ should be *consistently* higher on the correct label because training data are often more resistant to transformations than non-training data [6]. Based on these intuitions, we use the discrimination metric as the foundation of the pseudo label predictor in SecretGen.

Pseudo Label Predictor. Given the discrimination metric $\mathcal{M}$, we next describe in detail how we leverage $\mathcal{M}$ to infer the pseudo prediction label considering different sampled latent vectors, which aims to approximate the ground truth. We first sample a set of n latent vectors randomly and compute $\mathcal{M}$ for all class labels. The pseudo label predictor chooses the label with the maximum discrimination metric score as the predicted label $\hat{c}$:

$$\hat{c} = \arg\max_{c \in [1,C]} \mathcal{M}(c) \tag{5}$$

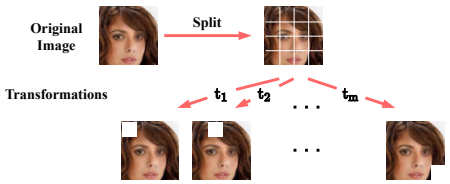

Fig. 2. Sequential cutout for the recovered image as transformations. The image is first split into m fix-sized patches. Operations of cutting out each patch are viewed as transformations respectively.

We defer the detailed algorithm for label prediction with $\mathcal{M}$ to the appendix. Note that there are various design choices for the discrimination metric $\mathcal{M}$, *e.g.*, the average confidence on only the recovered images without including their transformed versions. It is clear that more advanced $\mathcal{M}$ will provide more accurate pseudo label predictors. We will analyze the performance of the pseudo label predictor given different designs of $\mathcal{M}$ in Section 4.5.

3.5 Latent Vector Selector

In addition to the availability of ground truth labels, another challenge during private data recovery is that we may not have *whitebox* access to the target model. In systems where machine learning is used as a service (MLaaS), the adversary can only query the model and the prediction vector is returned from the service provider. All internal computations and model parameters are unknown to the adversary. In previous work [28], the adversary can directly optimize the latent vector z to maximize the target model's confidence given a known ground truth label, which is less practical. Without the whitebox access, performing back-propagation with the target model is infeasible in our practical case.

To tackle this problem, we design a *latent vector selector* to first randomly sample n random latent vectors, and then select the ones which lead to their recovered data classified as the predicted label from the pseudo label predictor. If there is no latent vector that leads to the recovered images which can be classified as the predicted label consistently, the selector returns a randomly sampled latent vector from the prior distribution. Otherwise, it returns the latent vector which has the highest confidence of the predicted label. We omit the detailed algorithm to the appendix.

3.6 SecretGen Optimization

To put every proposed component within SecretGen together, we perform joint optimization to maximize the consistent label prediction likelihood of recovered images indicated by the discrimination metric (*i.e.*, identity loss), while keeping the recovered images realistic (*i.e.*, discriminator loss). In the *whitebox* setting, we perform backpropagation on the target model with identity loss $\mathcal{L}_{\text{id}}$. $\mathcal{L}_{\text{id}}$

encourages the generated images to achieve consistently high label prediction likelihood given the target model for class label c.

$$\mathcal{L}_{\text{id}} = -\log[F_c(G(z, x_{ns}))] \tag{6}$$

We utilize discriminator loss as regularization to penalize unrealistic images.

$$\mathcal{L}_{\text{disc}} = -D(G(z, x_{ns})) \tag{7}$$

Then we initialize z with $\hat{z}$ returned by our latent vector selector and optimize z with the following objective function:

$$\hat{z}^*_{\text{whitebox}} = \arg\min_z \mathcal{L}_{\text{disc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \tag{8}$$

In the *blackbox* setting, we perform the latent vector selection optimization only with the discriminator loss since the target model is not locally accessible:

$$\hat{z}^*_{\text{blackbox}} = \arg\min_z \mathcal{L}_{\text{disc}} \tag{9}$$

Note that in the blackbox setting where we have the ground truth labels, the identity loss is still minimized by the latent vector selector through random sampling based on the prediction vector of the target model, guaranteeing that the recovered image is close to the density region of the ground truth identity.

3.7 Discussion

Our proposed SecretGen works under a wide range of scenarios regarding different types of prior knowledge. See Table 1 for the scenarios under which SecretGen and existing methods can be applied. Although EMI theoretically works in *blackbox* cases with ground truth labels, its performance and efficiency dramatically suffer against deep models. Although SecretGen still requires non-sensitive private data as prior knowledge, such assumption is realistic as image corruption is often leveraged for privacy protection by individuals [28]. Furthermore, if the knowledge of ground truth labels is available, it can be incorporated into SecretGen conveniently. More details are deferred to the appendix.

In conclusion, SecretGen is more practical without requiring whitebox access to the target model or the ground truth label. In addition, SecretGen is very efficient and applicable to high-dimensional image data considering deep models as the target model as shown in our evaluation (Section 4).

Table 1. Comparison with existing methods on the information required by the adversary to recover private training data. The symbol ✘ means that in theory, the method can work without the information, but the actual performance on deep models is bad.

Methods	Non-sensitive Data?	Whitebox Access?	Ground Truth Label?
PII [25]	✓	✗	✗
EMI [10]	✗	✘	✓
GMI [28]	✘	✓	✓
SecretGen	✓	✗	✗

4 Experiments

In this section, we first present the experimental setup. Then, we introduce the evaluation protocols and report the attack performance respectively. We also evaluate the robustness of SecretGen against the purification defense [27]. In the end, we describe some ablation studies to better understand our method.

4.1 Experimental Setup

Datasets. We evaluate SecretGen on two face datasets: (1) CelebA [20] which contains 202,599 face images of 10,177 identities. We filter out those identities with 25 or fewer images and randomly select 25,000 images of 1,000 identities as private data D_{pri}. We also randomly select 50,000 images of 2,000 identities from the rest as adversary's public data D_{pub}. There exist no overlapped identities between D_{pub} and D_{pri}. (2) FaceScrub [21] which consists of 106,863 face images of 530 identities. We use the images of 250 identities as D_{pri} and images of another 250 identities as D_{pub}. We further split D_{pri} into $D_{\text{pri}}^{\text{train}}$ and $D_{\text{pri}}^{\text{test}}$ for training and testing. All the images are cropped and resized to 64×64.

Prior Information. We consider two types of prior information that the adversary has access to: corrupted images by center mask and face T mask following the standard setting in [28]. Center mask blocks the center part of the private image, but the mouth information may still be exposed. Face T mask completely hides the identity revealing features of the face image.

Model Architectures. We perform evaluation on *target models* with various architectures: (1) `VGG16` [23]; (2) `ResNet152` [15]; (3) `face.evoLVe` [5] with an `IR50` backbone; (4) `ViT-B_16` [9] We utilize `IR152` [5] as the *evaluation model* to predict the identity of input images. Both `VGG16` and `ResNet152` are pre-trained on `ImageNet` [8]. `face.evoLVe` and the evaluation model are pre-trained on `MS-Celeb-1M` [14]. `ViT-B_16` is pre-trained on `ImageNet21k` [8]. The architecture of SecretGen generation backbone is adopted from [28].

Baselines. We compare SecretGen with the *state-of-the-art* model inversion attack GMI [28]. GMI assumes the adversary has access to the ground truth labels and performs optimization with identity loss and discriminator loss. We also compare our results with pure image inpainting (PII) [25], which only optimizes the discriminator loss for generating realistic images. Latent vectors of both GMI and PII are sampled randomly from Gaussian distribution. We do not compare with EMI [10], since it has been demonstrated in [28] that the effectiveness of EMI is quite limited against deep models. We defer additional details regarding model training and attack to the appendix.

4.2 Evaluation Protocols

We consider two principles for evaluating the privacy attack performance: "how much privacy sensitive identity information can be recovered" and "how well the recovered data can perform in downstream tasks".

Corresponding to the two principles, we evaluate the privacy attack performance by *attack accuracy* under the following two protocols:

- Protocol 1: Train the evaluation model on the private data, and evaluate on the recovered data.
- Protocol 2: Train the evaluation model on the recovered data, and evaluate on the private data.

Protocol 1 was introduced in [28], which evaluates *instance-level* privacy recovery. However, we demonstrate that even if some instances are not recovered correctly, the recovered data can be used for downstream tasks, *e.g.*, training another classification model. The adversary can potentially use the trained evaluation model for malicious purposes, *e.g.*, performing unauthorized face recognition on private identities with significantly higher accuracy than the target model itself. Thus, we propose Protocol 2, which aims to evaluate *distribution-level* privacy recovery. In addition, a common goal of the adversary to reconstruct the private data is to leverage such data for other downstream tasks, and therefore Protocol 2 explicitly reflects the utility of the recovered data.

For Protocol 1, we train the evaluation model on $D_{\text{pri}}^{\text{train}}$ and the resulting evaluation model achieves 98.0% classification accuracy over the private identities on $D_{\text{pri}}^{\text{test}}$. For Protocol 2, we first perform the attack on all corrupted private images D_{pri}—for each corrupted image $x_{ns} \in D_{\text{pri}}$, we recover an image $\hat{x} = G(\hat{z}^*, x_{ns})$ via SecretGen, with label $\hat{c} = \arg\max_{c \in [1,C]} F_c(\hat{x})$. We then compose the recovered images into a recovered private set D_{rec}, which is separated into $D_{\text{rec}}^{\text{train}}$ and $D_{\text{rec}}^{\text{valid}}$ by 4:1. We train the evaluation model on $D_{\text{rec}}^{\text{train}}$ with $D_{\text{rec}}^{\text{valid}}$ as the validation set. We then evaluate the model performance on $D_{\text{pri}}^{\text{test}}$.

We also report Peak Signal-to-Noise Ratio (PSNR) [16] between original and recovered private data, which reflects the *pixel-level* reconstruction quality of our attack. Note that the recovered data can still reveal identity information even if the generated image is not close to the ground truth image pixel-wise. For example, the recovered images can exhibit variations in pose and light condition while keeping the identity.

4.3 Attack Performance

Whitebox Attacks. Table 2 compares the performance of SecretGen with baseline methods on CelebA. See the appendix for results on FaceScrub.

We can see that SecretGen significantly outperforms GMI under both Protocol 1 and Protocol 2 if the ground truth label is given. Without such information, with the proposed pipeline especially the pseudo label predictor, SecretGen still achieves comparable performance with GMI under Protocol 1. Under Protocol 2, GMI with ground truth label performs better than SecretGen without ground truth label. The reason is that if the predicted pseudo label is incorrect, our pseudo label predictor and optimization push the recovery to be closer to the wrong identity. However, we still outperform PII by a large margin.

We also observe that attack accuracy under Protocol 2 is much higher than that under Protocol 1. The reason is that Protocol 1 and 2 work at different

Table 2. *Whitebox* attack performance on CelebA. See the Ground Truth Label column for whether ground truth label is provided for each attack method.

Target Model	Methods	Ground Truth Label	Center Mask			Face T Mask		
			Protocol 1	Protocol 2	PSNR	Protocol 1	Protocol 2	PSNR
VGG16	PII	✗	0.423	0.561	27.583	0.166	0.363	26.276
	GMI	✓	0.569	0.955	27.587	0.305	0.928	26.240
	SecretGen	✗	0.584	0.928	27.955	0.312	0.793	26.632
	SecretGen	✓	**0.639**	**0.965**	**28.071**	**0.377**	**0.931**	**26.821**
ResNet152	PII	✗	0.403	0.719	26.892	0.170	0.555	26.117
	GMI	✓	0.556	0.965	27.177	0.295	**0.946**	26.482
	SecretGen	✗	0.595	0.948	27.506	0.324	0.884	26.821
	SecretGen	✓	**0.618**	**0.971**	**27.587**	**0.349**	0.945	**26.967**
face.evoLVe	PII	✗	0.267	0.455	27.317	0.122	0.343	26.356
	GMI	✓	0.595	0.946	27.444	0.467	0.935	26.563
	SecretGen	✗	0.551	0.841	27.613	0.274	0.630	26.562
	SecretGen	✓	**0.788**	**0.963**	**27.781**	**0.695**	**0.954**	**26.827**
ViT	PII	✗	0.380	0.389	26.698	0.173	0.306	26.377
	GMI	✓	0.482	0.893	24.907	0.214	0.715	24.624
	SecretGen	✗	0.451	0.634	**26.811**	0.246	0.528	26.471
	SecretGen	✓	**0.551**	**0.950**	26.607	**0.326**	**0.913**	**26.609**

levels: Protocol 1 evaluates how much "detailed" information the recovered images contain, while Protocol 2 evaluates how much distributional information we can recover by training another model based on the reconstructed data. Clearly, Protocol 2 is relatively easier by recovering distributional level information and thus achieves higher scores. We believe such observations will inspire interesting future work and narrow down such a gap.

Blackbox Attacks. In the *blackbox* setting, the adversary is not capable of performing backpropagation with the target model. We make the following changes to our attack pipeline: (1) In Section 3.3, when training the generation backbone, we use a public feature extractor from [5] pre-trained on `MS-Celeb-1M` to substitute the target model for extracting the feature embeddings in computing the diversity loss ($\mathcal{L}_{\text{div}}$, Eqn. (2)); (2) In Section 3.6, when performing SecretGen optimization, we remove the identity loss $\mathcal{L}_{\text{id}}$ and optimize the selected latent vector only with discriminator loss $\mathcal{L}_{\text{disc}}$.

Table 3 compares our results with PII under the *blackbox* setting on CelebA. The only difference for PII under *blackbox* and *whitebox* scenarios is whether the target model is accessed when training the generation backbone. We can see that with the ground truth labels, SecretGen significantly outperforms PII. Without ground truth labels, which is the most general case, we still outperform PII by a large margin. As far as we are concerned, we are the first to propose an effective model inversion attack against deep classification models under the *blackbox* case without ground truth label.

We note that GMI (with ground truth label) performs better on `face.evoLVe` than SecretGen (without ground truth label), as shown in Table 2 and Table 3. Under this setting, attack performance is largely dependent on the pseudo label predictor. We demonstrate that the label prediction accuracy of `face.evoLVe` is significantly lower than that of `VGG16` and `ResNet152` in the appendix. We believe the reason is that `face.evoLVe` is less overfitted due to the difference in

Table 3. *Blackbox* attack performance on CelebA. We report results for both cases where the adversary has or does not have ground truth labels. (Note: GMI does not support blackbox attack, and PII in the blackbox setting does not use the target model.)

Methods	Target Model	Ground Truth Label	Center Mask			Face T Mask		
			Protocol 1	Protocol 2	PSNR	Protocol 1	Protocol 2	PSNR
PII	Any	✗	0.216	0.759	27.319	0.081	0.484	25.705
SecretGen	VGG16	✗	0.351	0.915	27.638	0.164	0.837	26.045
		✓	**0.380**	**0.955**	**27.737**	**0.377**	**0.927**	**26.821**
	ResNet152	✗	0.334	0.933	27.737	0.152	0.765	26.144
		✓	**0.347**	**0.959**	**27.840**	**0.172**	**0.886**	**26.284**
	face.evoLVe	✗	0.447	0.711	27.568	0.156	0.353	25.787
		✓	**0.603**	**0.894**	**27.694**	**0.305**	**0.586**	**26.002**
	ViT	✗	0.285	0.709	27.480	0.119	0.685	25.828
		✓	**0.335**	**0.924**	**27.665**	**0.160**	**0.902**	**26.123**

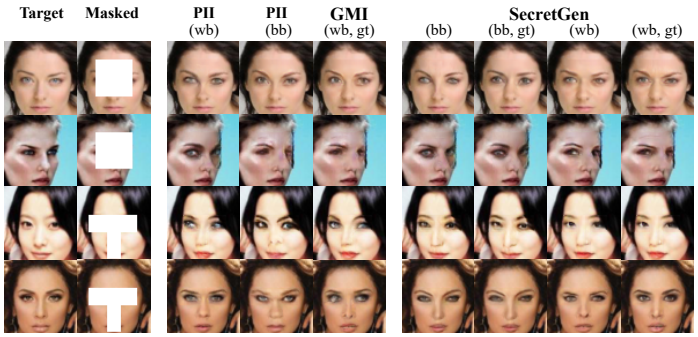

Fig. 3. Qualitative results of SecretGen on CelebA. "bb"/"wb" indicates the method requires *blackbox*/*whitebox* access to the model. "gt" indicates the method requires ground truth labels.

pre-training datasets. (`face.evoLVe` is pre-trained on `MS-Celeb-1M` while others are on `ImageNet`.)

Qualitative Results. In Fig. 3 we exhibit the images recovered with SecretGen on CelebA to demonstrate that our recovered images are both identity-revealing and visually plausible. We also show qualitative results of PII and GMI for comparison. From the figure, we see that although all of the three methods generate realistic images, PII cannot effectively recover the original identity of private data, while SecretGen is more effective in identity revealing. More examples are shown in the appendix.

4.4 Robustness Evaluation

We evaluate the robustness of our proposed method against purification defense [27], which has been shown to effectively defend against model inversion attacks while inducing negligible utility loss. We use Purifier I in [27] which is specialized for model inversion attacks. We follow the default architectures and settings for training the purifier. See Table 4 for quantitative results on CelebA against VGG16 under the blackbox setting. We also assume the ground truth label is not provided. We do not evaluate the whitebox setting because the ad-

Table 4. Robustness evaluation for SecretGen against prediction purification on CelebA. Target model: `VGG16`. Blackbox setting.

Methods	Center Mask			Face T Mask		
	Protocol 1	Protocol 2	PSNR	Protocol 1	Protocol 2	PSNR
PII	0.216	0.759	27.319	0.081	0.484	25.705
SecretGen	0.351	0.915	27.638	0.164	0.837	26.045
SecretGen (purified)	0.328	0.913	27.590	0.151	0.747	26.007

versary can simply remove the purifier and directly attack the original private model. It can be seen that attack accuracy slightly decreases after the defense, but still outperforms the baseline by a large margin. Therefore, our method is robust against [27].

4.5 Ablation Studies

Discrimination Metrics. As discussed in Section 3.4, there may exist various choices for the discrimination metric. One intuitive choice may be derived by removing the transformations from our current discrimination metric $\mathcal{M}$ (Eqn. (4)), and the simplified discrimination metric is defined as follows:

$$\mathcal{M}'(c; n) \triangleq \frac{1}{n} \sum_{i=1}^{n} F_c(\hat{x}_i), \quad \forall c \in [1, C]. \tag{10}$$

We perform an *end-to-end* ablation study on `face.evoLVe` and CelebA. We remove the transformations in our pseudo label predictor and substitute $\mathcal{M}$ with $\mathcal{M}'$. Quantitative results on `face.evoLVe` are shown in Table 5. See the appendix for results regarding other model architectures. We conclude that incorporating transformations improves the performance of our framework for most model architectures that we used for evaluation.

Table 5. Attack accuracy of SecretGen with and without transformations on CelebA. Evaluated on 3,200 private instances under Protocol 1. Target model: `face.evoLVe`.

Metric	Center Mask		Face T Mask	
	Attack Acc	PSNR	Attack Acc	PSNR
w/o transformation	0.528	27.505	0.256	26.527
w/ transformation	**0.550**	**27.522**	**0.273**	26.527

To further understand why and how transformations help, we compare the performance of pseudo label predictor equipped with $\mathcal{M}$ and $\mathcal{M}'$. We evaluate the performance of pseudo label predictor using *label prediction accuracy*, which measures the percentage of the predicted labels matching the ground truth labels. We plot out the label prediction accuracy with $\mathcal{M}$ and $\mathcal{M}'$ on 3,200 recovered images for `face.evoLVe` with varying n in Fig. 4. We observe that our pseudo label predictor can predict the pseudo labels more accurately if transformations are incorporated. See the appendix for results of other model architectures.

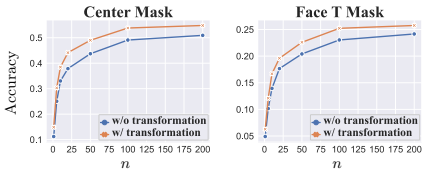

Fig. 4. Label prediction accuracy with and without transformations on CelebA. We plot the label prediction accuracy w.r.t. the number of random latent vectors n. Target model: `face.evoLVe`.

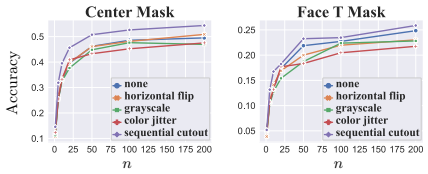

Fig. 5. Label prediction accuracy with different transformations on CelebA. We plot the label prediction accuracy w.r.t. the number of randomly sampled latent vectors n. Target model: `face.evoLVe`.

Data Transformations. Next, we discuss the performance of various data transformations on CelebA. We plot out label prediction accuracy w.r.t. n for various transformations including the proposed sequential cutout, horizontal flipping, gray-scale, and color jittering in Fig. 5. We also plot the results without transformations. We can see that sequential cutout performs better than other transformations in terms of label prediction accuracy. Although it is also possible to adopt other transformations within our pipeline, it is non-trivial to select the best hyper-parameters for other transformations (*e.g.*, cropping and color jittering). We leave the analysis of how different transformations impact attack performance as future work.

Overfitting Levels. We also evaluate the impact of *higher overfitting levels* of the target model on the performance of SecretGen, since the overfitting phenomenon is key to model inversion attacks. Note that results reported in Table 2 and Table 3 are based on standard well-trained models. We demonstrate that highly overfitted models are more vulnerable to our proposed attack. We describe the relevant experiment setup and quantitative results in the appendix.

5 Conclusion

In this paper, we propose an effective private data recovery framework SecretGen, which can effectively recover private information under a wide range of scenarios. To our full knowledge, we are the first to propose an effective model inversion attack without prior knowledge of ground truth labels, which can achieve comparable results with previous methods that require ground truth labels. If we are given such prior knowledge, we significantly outperform previous methods. Our attack can also be applied under the *blackbox* setting where the target model is provided as a service and not locally available. We perform a comprehensive analysis of the performance of SecretGen and our design choices. We also demonstrate that our attack is robust against the purification defense. We hope to raise people's concerns about possible negative effects of releasing pre-trained models online. For future work, we are interested in whether we can perform privacy recovery simply with the target model and develop defenses against our attack.

Acknowledgements. This work is partially supported by NSF grant No.1910100, NSF CNS No.2046726, C3 AI, and the Alfred P. Sloan Foundation.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16). pp. 265–283 (2016)
2. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML workshop on unsupervised and transfer learning. pp. 17–36. JMLR Workshop and Conference Proceedings (2012)
3. Chen, S., Kahla, M., Jia, R., Qi, G.J.: Knowledge-enriched distributional model inversion attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16178–16187 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Cheng, Y., Zhao, J., Wang, Z., Xu, Y., Jayashree, K., Shen, S., Feng, J.: Know you at one glance: A compact vector representation for low-shot learning. In: ICCVW. pp. 1924–1932 (2017)
6. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-only membership inference attacks. In: International Conference on Machine Learning. pp. 1964–1974. PMLR (2021)
7. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979 (2020)
8. 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)
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. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322–1333 (2015)
11. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd {USENIX} Security Symposium ({USENIX} Security 14). pp. 17–32 (2014)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
14. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
15. 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)

16. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Leino, K., Fredrikson, M.: Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In: 29th {USENIX} Security Symposium ({USENIX} Security 20). pp. 1605–1622 (2020)
19. Liu, H., Jia, J., Qu, W., Gong, N.Z.: Encodermi: Membership inference against pre-trained encoders in contrastive learning. arXiv preprint arXiv:2108.11023 (2021)
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
21. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). pp. 343–347. IEEE (2014)
22. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Wang, X., Xiong, W., Wang, H., Wang, W.Y.: Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 37–53 (2018)
25. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. arXiv preprint arXiv:1901.09024 (2019)
26. Yang, Z., Chang, E.C., Liang, Z.: Adversarial neural network inversion via auxiliary knowledge alignment. arXiv preprint arXiv:1902.08552 (2019)
27. Yang, Z., Shao, B., Xuan, B., Chang, E.C., Zhang, F.: Defending model inversion and membership inference attacks via prediction purification. arXiv preprint arXiv:2005.03915 (2020)
28. 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)