Label-Only Model Inversion Attacks via Boundary Repulsion

Mostafa Kahla
Virginia Tech
kahla@vt.edu

Si Chen
Virginia Tech
chensi@vt.edu

Hoang Anh Just
Virginia Tech
just@vt.edu

Ruoxi Jia
Virginia Tech
ruoxijia@vt.edu

Abstract

*Recent studies show that the state-of-the-art deep neural networks are vulnerable to model inversion attacks, in which access to a model is abused to reconstruct private training data of any given target class. Existing attacks rely on having access to either the complete target model (whitebox) or the model's soft-labels (blackbox). However, no prior work has been done in the harder but more practical scenario, in which the attacker only has access to the model's predicted label, without a confidence measure. In this paper, we introduce an algorithm, Boundary-Repelling Model Inversion (*BREP-MI*), to invert private training data using only the target model's predicted labels. The key idea of our algorithm is to evaluate the model's predicted labels over a sphere and then estimate the direction to reach the target class's centroid. Using the example of face recognition, we show that the images reconstructed by* BREP-MI *successfully reproduce the semantics of the private training data for various datasets and target model architectures. We compare* BREP-MI *with the state-of-the-art whitebox and blackbox model inversion attacks, and the results show that despite assuming less knowledge about the target model,* BREP-MI *outperforms the blackbox attack and achieves comparable results to the whitebox attack. Our code is available online.* [1]

1. Introduction

Machine learning (ML) algorithms are often trained on private or sensitive data, such as face images, medical records, and financial information. Unfortunately, since ML models tend to memorize information about training data, even when stored and processed securely, privacy information can still be exposed through the access to the models [21]. Indeed, the prior study of privacy attacks has demonstrated the possibility of exposing training data at different granularities, ranging from "coarse-grained" information, such as determining whether a certain point participates in training [11, 15, 17, 22] or whether a training dataset satisfies certain properties [10, 16], to more "fine-grained" information, such as reconstructing the raw data [2, 4, 8, 26].

We focus on model inversion (MI) attacks, which recreates training data or sensitive attributes given the access to the trained model. MI attacks cause tremendous harm due to the "fine-grained" information revealed by the attacks. For instance, MI attacks applied to personalized medicine prediction models result in the leakage of individuals' genomic attributes [9]. Recent works show that MI attacks could even successfully reconstruct high-dimensional data, such as images. For instance, [4, 8, 25, 26] demonstrated the possibility of recovering an image of a person from a face recognition model given just their name.

Existing MI attacks have either assumed that the attacker has the complete knowledge of the target model or assumed that the attack can query the model and receive model's output as confident scores. The former and the latter are often referred to as the whitebox and the blackbox threat model, respectively. The idea underlying existing whitebox MI attacks [4, 26] is to synthesize the sensitive feature that achieves the maximum likelihood under the target model. The synthesis is implemented as a gradient ascent algorithm. By contrast, existing blackbox attacks [2, 20] are based on training an attack network that predicts the sensitive feature from the input confidence scores. Despite the exclusive focus on these two threat models, in practice, ML models are often packed into a blackbox that only produces hard-labels when being queried. This label-only threat model is more realistic as ML models deployed in user-facing services need not expose raw confidence scores. However, the design of label-only MI attacks is much more challenging than the whitebox or blackbox attacks given the limited information accessible to the attacker.

In this paper, we introduce, BREP-MI, a general algorithm for MI attack in the label-only setting, where the attacker can make queries to the target model and obtain hard labels, instead of confidence scores. Similar to the main idea of whitebox attacks, we still try to synthesize the most likelihood input for the target class under the target model. However, in the label-only setting, we cannot directly cal-

[1] https://github.com/m-kahla/Label-Only-Model-Inversion-Attacks-via-Boundary-Repulsion

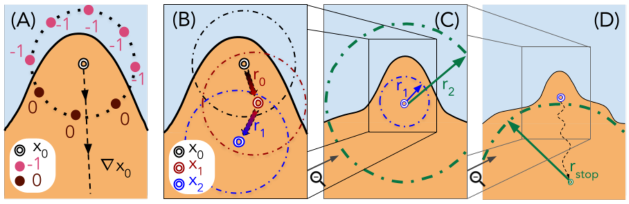

Figure 1. Intuitive explanation of BREP-MI. (A) Query the labels over a sphere and estimate the direction on the sphere that can potentially lead to the target label class. (B) Update the synthesized image according to the estimated direction. Alternate between the estimation and update until the sphere fits into the target class. (C) Increase the radius of the sphere. (D) Repeat the steps above until the attack hits some query budget.

culate the gradient information and leverage it to guide the data synthesis. Our key insight to resolve this challenge is that a high-likelihood region for a given class often lies at the center of the class and is far away from any decision boundaries. Hence, we design an algorithm that allows the synthesized image to iteratively move away from the decision boundary, as illustrated in Figure 1. Specifically, we first query the labels over a sphere and estimate the direction on the sphere that can potentially lead to the target label class (A). We progressively move according to estimated directions until the sphere fits into the target class (B). We then increase the radius of the sphere (C) and repeat the steps above until the attack hits some query budget (D). We theoretically prove that for linear target models, the direction estimated from hard labels queried on the spheres aligns with the gradient direction. We empirically show that BREP-MI can also lead to successful attacks against deep neural network-based target models. In particular, the efficacy of the attack is even higher than the existing blackbox attacks and comparable to the existing whitebox attacks.

Our contributions can be summarized as follows: **(1)** We propose the first algorithm for label-only model inversion attacks. **(2)** We provide theoretical justification for the algorithm in the linear target model case by proving the updates used in our algorithm align with the gradient and also analyze the error of alignment for nonlinear models. **(3)** We evaluate the attack on a range of model architectures and datasets, then show that despite exploiting less information about the target model, our attack outperforms the confidence-based blackbox attack by a large margin and achieves comparable performance to the state-of-the-art whitebox attack. Besides, we will release data, code, and models to facilitate future research.

2. Related Work

Model Inversion Attacks. Model inversion attempts to reconstruct from partial up to full training sample. Typically, MI attacks can be formalized as an optimization problem, which goal is to find the sensitive feature value that achieves the highest likelihood under the model being attacked. However, when the target model is a deep neural network (DNN) or the private data lie in high-dimensional space, such optimization problem becomes non-convex and directly solving it via gradient descent may result in poor attack performance [8]; for example, when attacking a face recognition model, the recovered images are blurry and do not contain much private information. Recent work [26] proposes a GAN-based MI attack method which is effective on DNNs. In particular, they learn a generic prior from public data via GAN and solve the optimization problem over the latent space rather than the unconstrained ambient space. However, their attack method does not fully exploit private information contained in the target model at the stage of training GAN. [4] significantly improves the attack performance through a special design of GAN which can distill knowledge from the target model; as a result, the generated images align better with the private distribution. They further improve the performance by ensuring that both the recovered image and its neighboring images have high likelihood. While [4, 26] achieve success on attacking various models and datasets, their attacks rely on whitebox access to the model. In many cases, the attacker can only make prediction queries against a model, but not actually download the model, which motivates the study of blackbox MI attacks. [25] analyzes the blackbox setting and proposes an attack model which swaps the input and prediction vector of the target model to perform model inversion. [2] proposes to train a GAN and a surrogate model simultaneously, with the GAN generating inputs that resemble private training data and the surrogate model mimicking the target model's behavior. All of the blackbox attacks make an assumption that prediction confidences of the target model are revealed to the attacker. However, it is more practical in real-world setting that an adversary, who only makes queries to the model, can only obtain the hard labels, without confidence scores. From this aspect, we aim to provide an effective MI attack method that only requires access to the hard label, which we refer to as label-only MI attacks.

Other Privacy Attacks. Asides from MI, there are two other categories of privacy attacks that allow adversaries to gain unauthorized information from the target model and its data. In a membership inference attack, the attacker attempts to evaluate whether a certain point is used in the target model's training. This attack technique was introduced by [22] who created multiple shadow models to estimate the target model. [11, 15, 17] pointed out that the membership inference attack exploits the overfitting of specific data points. Interestingly, [6] performs a membership inference attack under same setting as our BREP-MI attack and notes that the viable defense against such an attack is via differential privacy (DP). DP [1, 7] ensures that the trained model is stable to the change of any single record in the training set. However, with differential privacy, the target model's test accuracy will significantly degrade. Additionally, property inference attacks aim to infer from the properties about the training dataset [10]. Compared to these attacks, MI is arguably more challenging as the information it attempts to recover is higher in resolution.

3. Threat Model

Attack goal. In MI attacks, given the access to a target model $f : [0,1]^d \to \mathbb{R}^{|C|}$ and any target class $c^* \in C$, the attacker attempts to recover a representative point x^* of the training data from the class c^*; d represents the dimension of the model input; C denotes the set of all class labels and $|C|$ is the size of the label set. For example, an attack on the face recognition classifier would try to recover the face image for a given identity based on the access to the classifier.

Model knowledge. The attacker's knowledge about a target model can take different forms: (i) Whitebox: complete access to all target model parameters; ii) Blackbox: access to the confidence scores output by the target model; and iii) Label-only: access to only the hard labels output by the model without the confidence scores. Our paper will focus on the label-only setting. Specifically, given the target network f, the attacker can query the target network at any input x and obtain the corresponding predicted label $\hat{y}(x) = \arg\max_{c\in C} f_c(x)$.

Task Knowledge. For the rest of the paper, we assume that the attacker has knowledge about the task that the target model performs. This is a reasonable assumption, since this information is available for existing online models, or can be inferred from output labels.

Data Knowledge. Since we assume that attackers know the task of the attacked model, it is reasonable to assume that they can gain access to a public dataset from a related distribution. For example, if attackers know that the target model is trained to perform facial recognition, they can easily gather a public dataset by leveraging the existing open-sourced datasets or crawling data from the web. Throughout the paper, we assume that the public data and the private data do not share any classes (e.g., identities) in common.

Target models. Our approach neither makes assumption on the target model architecture, nor requires the attacker to have any information about it. In other words, our approach is model-agnostic. We will empirically show in Section 5 that our BREP-MI attack generalizes to a variety of models with different architectures and sizes.

Target labels. The attack can be *targeted*, when the goal is to find n input images that maximize a set of n *predefined* labels, or *untargeted*, when the goal is to find n input images that maximize a set of *any* n labels. The proposed algorithm can apply to *both* scenarios. In our evaluation, we will focus on the more challenging scenario, where the attack is targeted for n specific labels.

4. Algorithm Design

In this section, we will present the design of our proposed algorithm BREP-MI. We will start by formulating the MI attack as an optimization problem. Then, we describe an algorithm to estimate the gradient of the MI optimization objective based only on predicted labels. We will rigorously characterize the alignment between the estimate and the true gradient for the special case of linear models and provide insights into the attack efficacy for deep, nonlinear models.

4.1. Problem Formulation

Without loss of generality, we state the attack problem formulation for a single target label and define $M_{c^*} : \mathbb{R}^d \to \mathbb{R}$ such that

$$M_{c^*}(x) = f_{c^*}(x) - \max_{c\neq c^*} f_c(x), \tag{1}$$

where c^* is the target label. $M_{c^*}(x)$ represents the logit (or confidence score) difference between the target class c^* and the most likely label in the rest of the classes. Note that when x is predicted into the target class (i.e., $c^* = \arg\max_{c\in C} f_c(x)$), $M_{c^*}(x) > 0$. Clearly, the most representative input for the target class c^* should be most distinguishable from all the other classes. Hence, we cast the MI problem into an optimization problem that seeks for the input that achieves maximum difference between the confidence for the target class and the highest confidence for the other classes:

$$\arg\max_{x\in[0,1]^d} M_{c^*}(x). \tag{2}$$

However, for images, x usually lies in a high-dimensional continuous data space and optimizing over this space can easily get stuck in local minima that do not correspond to any meaningful images. To resolve this issue, we leverage the idea in [4, 25, 26] and optimize over a more semantically meaningful latent space. This is done by using a public dataset to train GAN models and then optimizing over the input to the GAN generator. Denote the publicly trained generator by $G(z)$, where $z \in \mathbb{R}^{d'}$ and $d' < d$. Now, the MI optimization problem can be updated to reflect the change of optimizing z rather than x as follows:

$$\arg\max_{z\in\mathbb{R}^d} M_{c^*}(G(z)). \tag{3}$$

Unlike the whitebox setting, we cannot directly optimize $M_{c^*}(G(z))$ using gradients as we do not have access to the model parameters f. Moreover, it is also not possible to apply zero-order optimization algorithms, as they require access to the confidence scores output by the model.

4.2. BREP-MI Algorithm

The intuition behind our algorithm is that the farther a point is from the decision boundary of a class, the more representative this point becomes to the class. Thus, the centroid of any class should be its good representative. Inspired by this, we design an algorithm which tries to gradually move away from the decision boundary. In a high level, our algorithm proceeds by first sampling points over a sphere and then querying their labels. Intuitively, the points that are not predicted into the target class represent the directions that we want to move away from. Hence, we take an average over those points and move in the direction opposite to the average. If all the points are predicted into the target class, then we will increase the radius.

Let sign$(\cdot)$ be a function that returns 1 if the input is positive and -1, otherwise. We define $\Phi_{c^*} : \mathbb{R}^d \to \{-1, 0\}$:

$$\Phi_{c^*}(z) = \frac{\text{sign}(M_{c^*}(z)) - 1}{2} \tag{4}$$

$$= \begin{cases} 0, \text{ if } c^* = \arg\max_{c \in C} f_c(G(z)) \\ -1, \text{ otherwise.} \end{cases} \tag{5}$$

Essentially, $\Phi_{c^*}(z)$ marks points that are not predicted into the target class. Then, we define our gradient estimator as

$$\widehat{M_{c^*}}(z, R) = \frac{1}{N} \sum_{n=1}^{N} \Phi_{c^*}(z + R u_n) u_n, \tag{6}$$

where u_n is a uniformly random point sampled over a d'-dimensional sphere with radius R and N is the number of points sampled on the sphere. Note that $\widehat{M_{c^*}}(z, R)$ can be calculated in the label-only setting as it only requires the knowledge of predicted label of the sampled points. We will then use $\widehat{M_{c'}}(z, R)$ to update z:

$$z \leftarrow z + \alpha \widehat{M_{c^*}}(z, R), \tag{7}$$

where α is the update step size. It can be either a fixed value or a function of the current radius R. When all points sampled from the sphere of the current radius are predicted into the target class, i.e., $\Phi_{c^*}(z + R u_n) = 0$ for all $n = 1, \ldots, N$, then we increase the radius and alternate between estimating $\widehat{M_{c'}}(z, R)$ using Eq. (6) and performing update with Eq. (7) at the new radius.

The pseudo-code of BREP-MI is provided in Algorithm 1. BREP-MI starts with the initial point correctly classified as the target class. To ensure this, images are sampled from the GAN until a point belonging to the target class is generated. Note that the initial point, although classified into the target class, is almost *never* a representative point for the target class (see more examples in Fig. 3). The radius of the sphere is initialized to a reasonably small value. Then, the algorithm will try to move away from the decision boundary, iteratively. At each iteration, we sample N points on the sphere with radius R centered at the current point and query their labels from the target model. If all the points are classified into the target class, the radius will be enlarged; otherwise, we estimate $\widehat{M_{c'}}(z, R)$ using Eq. (6) and update z according to Eq. (7). Note that the update is reverted if the new point z lies outside the target class. In that case, we will resample the points on the sphere and compute a new update. The algorithm will be halted when it is not possible to find a larger sphere such that all the samples on that sphere fall into the target class. The output of the algorithm is a point (z^*) with the largest sphere that can fit into the target class. This indicates that the point is the farthest from the boundary. We will use this point to evaluate the attack.

Algorithm 1: BREP-MI Decision-Based Zero Order Optimization Algorithm.

```
input : Target model's hard-label prediction $\hat{y}$ ;
        target class $c^*$, number of samples $N$;
        number of maximum iterations $maxIters$;
        initial sphere sampling radius $R_0$; radius
        multiplier $\gamma$; data point learning rate $\alpha$
output: Representative sample $z^*$ for $c^*$.
ensure: A sample $z$ in the target class $c^*$ by
        repeatedly sampling from the GAN's latent
        space.
Set $R \leftarrow R_0$.
Set $iters \leftarrow 0$.
Set $points \leftarrow vector(N)$
while $iters < maxIters$ do
    $points \leftarrow$ random N points on a sphere r=R
    // Check if all sampled points are
       in target class.
    if $points$ in $c^*$ then
        // Update radius and current best
           point
        $R \leftarrow R \times \gamma$ .
        $z^* \leftarrow z$ .
        $iters \leftarrow 0$.
    else
        Compute $\widehat{M_{c^*}}(z, R)$ via Eq. (6)
        $z_{\text{new}} \leftarrow$ the RHS of Eq. (7)
        if if $\hat{y}(z_{new}) = c^*$ then
            $z \leftarrow z_{\text{new}}$
        end
    end
end
```

4.3. Attack Justification

As our gradient estimator $\widehat{M_{c^*}}(z, R)$ repels non-target-class points, intuitively, it points towards the direction that increases the target class' likelihood. We provide a theorem that characterizes the alignment between the proposed estimator and the true gradient $\nabla M_{c^*}(z)$ for special cases of linear classification models (e.g., logistic regression).

Theorem 1. *Assume f has a linear classification model. Let z be an arbitrary point within the target class, i.e. $M_{c^*}(z) > 0$. Then, the cosine of the angle between $\mathbb{E}[\widehat{M_{c^*}}(z,R)]$ and $\nabla M_{c^*}(z)$ is bounded by*

$$\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) \tag{8}$$

$$\geq 1 - \mathcal{O}\left(\frac{M_{c^*}(z)^2 (d-1)^2}{\delta^2 R^2 \|\nabla M_{c^*}(z)\|_2^2}\right). \tag{9}$$

Therefore, with increasing radius R,

$$\lim_{R\to\infty} \cos\angle\left(\mathbb{E}\left[\widehat{M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) = 1,$$

which tells that the estimator is asymptotically unbiased for gradient estimation.

The proof is provided in the Appendix 1.1. Theorem 1 shows that as long as R is large enough, the gradient estimator aligns well with the actual gradient.

For the deep learning model with bounded nonlinearity, we can also derive the bound for the cosine of the angle between the estimate and the true gradient: $\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) \geq 1 - \mathcal{O}\left(\frac{\left[M_{c^*}(z)^2 + L^2\delta^4 R^4 + 4M_c(z)\delta^2 L R^2\right](d-1)^2}{\delta^2 R^2 \|\nabla M_{c^*}(z)\|_2^2}\right)$, where L characterizes the level of nonlinearity. A more formal statement of the result is provided in Appendix 1.2. The result shows that with increasing R the estimated gradient will align with the true gradient. However, after a certain point of inflection, increasing radius will only decrease the accuracy of the estimate. In all experiments, the maximum R we reach is small (usually less than 20) and the loss was always decreasing when R increases. This means that our gradient estimation was correct and we are not reaching this point of inflection in practice.

5. Evaluation

Our evaluation aims to answer the following questions: (1) Can BREP-MI successfully attack deep nets with different architectures and trained on different datasets? (2) How many queries does BREP-MI require to perform a successful attack? (3) How does the distributional shift between private data and public data affect the attack performance? (4) How sensitive is BREP-MI to the initialization and the sphere radius? In the main text, we will focus on a canonical application–face recognition–as our attack target. We will leave experiments on other applications to the Appendix 2.

5.1. Experimental Setup

Datasets. We experiment on three different face recognition datasets: CelebA [14], Facescrub [18], and Pubfig83 [19]. Similar to [4,25,26], we crop the images of all datasets to the center and resize them to 64×64. We split the identities into the public domain (which we train our GAN on), and the private domain (which we will train target models on). *There are no overlapping identities between the public and the private domain.* This means that the attacker has zero knowledge about the identities in the private domain. We then perform the attack on the classifier that is trained on the private domain. The details about each dataset are shown in Table 1. To study the impact of a large distributional shift between private and public domain on the attack performance, we use the FFHQ dataset [13] as our public domain to train the GAN and the aforementioned three datasets as the private domains.

Dataset	#Images	#Total id	#Public id	#Private id	# Target id
CelebA	202,599	10,177	9,177	1,000	300
Pubfig83	13,600	83	33	50	50
Facescrub	106,863	530	330	200	200

Table 1. Details on how we split datasets in evaluations into the public and the private domains.

Target Models. We also evaluate our attack on different models with a variety of architectures. To provide consistent results with the previous work, we use the same model architectures used in the state-of-the-art MI attack [4]: (1) face.evoLve adpated from [5]; (2) ResNet-152 adapted from [12]; and (3) VGG16 adapted from [23].

Evaluation Protocol. We perform a targeted attack as it is a more challenging setting compared to untargeted attack. Following [4, 26], we use the *attack accuracy* to measure the attack performance. The attack accuracy is based on an evaluation classifier, which predicts the identities for reconstructed face images and is a proxy for human judge. Specifically, the attack accuracy is calculated by the ratio of the number of reconstructed images that are correctly classified into the corresponding target classes over the total number of reconstructed images. As the evaluation classifier reflects a human judge, it should have high performance. At the same time, it should be different from the attacked target models to avoid some semantically meaningless reconstructed images that overfits to the target models to be considered as good reconstructions.

Hyperparameters. We manually fine-tuned the hyperparameters of BREP-MI in our evaluations. We found empirically that the best initial radius R_0 is 2, the radius expansion coefficient γ is 1.3, and the step size $\alpha_t = \min(R/3, 3)$. We choose N, the number of sampled points on a sphere, to be 32 unless otherwise specified and maxIters is chosen to be 1000, i.e., BREP-MI terminates when more than 1000 iterations have passed for a certain R without having all points on sphere to be classified as the target class.

Baselines. Since this is the first work that provides a solution to label-only MI attacks, we have no baselines to evaluate against. We opt to evaluate against whitebox and blackbox attacks in which the attacker has a *greater* advantage in terms of additional knowledge about target models. To ensure fair comparison, we apply all baselines over the same set of target identities for each dataset and the same target models. Then we evaluate attack accuracy against the same evaluation classifiers. Two of our baselines are white box attacks, including Generative Model Inversion (**GMI**) [26], which is the first MI attack algorithm against deep nets, and Knowledge-Enriched Distributional Model Inversion attack (**KED-MI**) [4], which provides the currently state-of-the-art performance for whitebox MI. The GAN models in GMI is set to be the same as the GAN in our attack. KED-MI relies on the access to target model parameters in training the GAN models. However, we cannot access such information and train the same GAN in our setting. We also employ a blackbox attack [25], referred to as the learning-based model inversion (**LB-MI**) as one of our baselines. LB-MI builds an inversion model that learns to reconstruct images from soft-labels produced by the target model. To reconstruct the most representative image for a given identity, we feed a one-hot encoding for that identity at the input of the inversion model and receive the output.

5.2. Results

Performance on Different Datasets. We compare BREP-MI to whitebox and blackbox methods on three different face datasets. We use FaceNet64 as the target model across all datasets. For each dataset, the GAN models are trained on its public identities, and target models are trained on the private identities. Table 2 shows that our approach considerably outperforms both the whitebox GMI attack and the blackbox attack on all datasets. Further, our method surpasses the state-of-the-art whitebox KED-MI attack on Pubfig83 and achieves a close attack accuracy on the CelebA dataset. On the other hand, we fall behind by 15% on the Facescrub dataset. It is worth noting that the outcome of this experiment implies that there is still a considerable potential for development in the other threat models in MI attacks, particularly blackbox attacks (which perform poorly with respect to the other threat models). The reason why GMI performs poorly even with the whitebox knowledge is that it optimizes the likelihood of only the synthesized data point without considering the neighborhood of the point. Hence, it is possible that optimization gets stuck in a sharp local maximum that does not represent the class. On the other hand, both BREP-MI and KED-MI explicitly finds a neighborhood with high likelihood, which turns out to be crucial to produce representative points and enhance attack performance. The blackbox attack, although leveraging more knowledge about target model than our attack, consistently achieves the worst performance. Compared to the other attacks, the blackbox attack utilizes a different idea for distilling knowledge from public datasets. It uses the public data to train the inversion model whereas all the other attacks train GANs on the public data. Results suggest that GANs are more effective in distilling public knowledge than an inversion model. So a potential way to improve blackbox attack is to regularize the synthesized images via GAN.

Dataset	[Whitebox]		[Blackbox]	[Label-only]
	GMI	KED-MI	LB-MI	BREP-MI
CelebA	32.00%	82.00%	1.67%	75.67%
Pubfig83	24.00%	62.00%	2.00%	66.00%
Facescrub	19.00%	48.00%	0.50%	35.68%

Table 2. Attack performance comparison various datasets.

Performance on Different Models. We also evaluate our attack on multiple different models trained on the same dataset (CelebA). This experiment is intended to test whether our approach can generalize to different model architectures. Table 3 shows that BREP-MI indeed continues to perform well on a variety of target model architectures. In particular, BREP-MI outperforms GMI and the blackbox attack by a substantial margin for all model architectures. As we can see, the attack accuracy is $2\times$ to $4\times$ that of GMI attack, while the blackbox attack continues to have $< 2\%$ accuracy. Additionally, our performance on all model architectures is comparable to that of the state-of-the-art whitebox. Similar to other attacks, our attack becomes more successful when the target model has higher predictive power.

Model Archt.	[Whitebox]		[Blackbox]	[Label-only]
	GMI	KED-MI	LB-MI	BREP-MI
FaceNet64	32.00%	82.00%	1.67%	75.67%
IR152	26.00%	83.00%	0.33%	72.00%
VGG16	15.00%	69.00%	1.33%	63.33%

Table 3. Attack performance comparison on different model architectures trained on the CelebA dataset.

Public→Private	[Whitebox]		[Blackbox]	[Label-only]
	GMI	KED-MI	LB-MI	BREP-MI
FFHQ→CelebA	9.00%	48.33%	0.67%	46.00%
FFHQ→Pubfig83	28.00%	88.00%	4.00%	80.00%
FFHQ→Facescrub	12.00%	60.00%	.015%	39.20%

Table 4. Performance comparison when there is a large distribution shift between public and private data.

Cross-Dataset Evaluation. In prior experiments, we assumed that the attacker had access to public data with low distributional shift with the private data. This is because both public and private domains are derived from the same dataset. It is important to consider a more pragmatic scenario, in which the attacker has access to only public data that have a larger distributional shift. To investigate this scenario, we perform an experiment in which we use the FFHQ dataset as public data.

As shown in Table 4, the accuracy indeed decreases significantly for the CelebA dataset when we utilize FFHQ as our public dataset. Interestingly, the attack accuracy for Pubfig83 and Facescrub datasets increased. The rationale for this performance boost is that Pubfig83 and Facescrub datasets have just 33 and 330 identities in their public distributions, respectively, as shown in Table 1. This means that the GAN models trained on these datasets would lack the ability to generalize and thus, produce bad results. Therefore, the ability of GAN models to generalize to the large number of identities in FFHQ compensate for the distributional shift and, consequently, the results improve. On the other hand, the CelebA dataset has a rather significant number of public identities (9177 identities). Thus, the GAN is already capable of generalizing across different identities, and the performance increase associated with generalizing on a more varied dataset is insufficient to compensate for the performance reduction associated with distributional shift. The takeaway from this experiment is that having a large, diverse public data for distilling a distributional prior is crucial to MI attack performance.

Limited Query Budget. We investigate the performance of our attack at various query budgets. In practice, some online models, such as Google's cloud vision API, limit the number of queries per minute, others may ban users if they identify an unusually high volume of queries. Due to the fact that some attack scenarios restrict the amount of queries that may be sent to the target model, it is important to investigate the impact of this restriction on the attack performance.

This restriction has not been addressed in prior works that conduct whitebox MI attacks in the literature. This is because the attacker, by definition, has complete access to the model parameters and can thus create an offline copy of the model, and then proceed with the attack offline with unlimited queries. However, for blackbox attacks in general (including label-only attack), the user cannot copy model parameters to an offline model. As a result, the query budget may become a constraint.

Fig.2 (a) demonstrates how BREP-MI performs under different query budgets. We see that the attack accuracy increases exponentially by increasing the query budget. This is true until we hit some query budget, then attack accuracy starts decreasing again. We will provide some insights in Section 5.2. For all datasets examined in this paper, recovering a representative image to a private class requires from $10k$ to $16k$ queries to the model, which is reasonable.

The attacker should also be concerned when choosing the hyperparameter N under limited query budgets. Choosing large N would increase the number of sampled points on sphere, and produce a better estimator for our update direction. On the other hand, for a fixed query budget, increasing N means decreasing the number of possible iterations in the attack. We conducted experiments to show this trade-off between spending queries to get better gradient estimator per iteration vs using queries to apply more iterations. Fig. 3 (b), (c), and (d) indicate that, for small query budgets, BREP-MI performs slightly better when spending query budget on increasing the number of iterations, instead of increasing N. However, for sufficiently large query budget, increasing N produces better results.

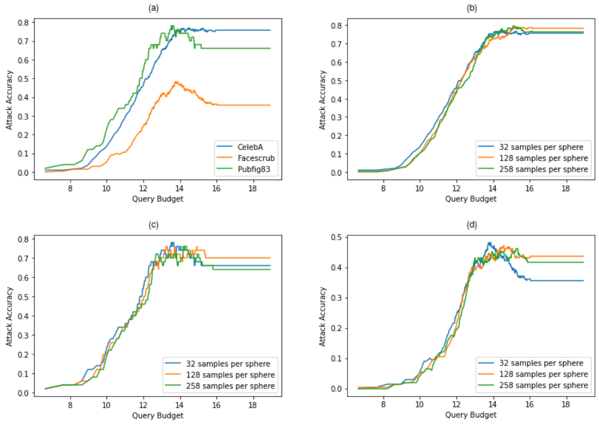

Figure 2. Attack accuracy of BREP-MI under different query budgets. (a) compares different datasets. (b), (c), and (d) compare different sampling strategies for CelebA, pubfig, and Facescrub respectively. Query Budget is shown in a base-2 log.

Analyzing BREP-MI. A qualitative analysis for our BREP-MI can be seen in Fig. 3. It is noticeable that the first generated image at the beginning of the attack is not a good representative for the target class. The progression of the image towards the groundtruth images is clearly seen with the increase of R .

Below, we provide some quantitative analysis. Table 5 analyzes the intermediate steps when attacking FaceNet64 model on CelebA dataset. We say the attack reached a radius R when it finds a center point, for which all points sampled on a sphere with radius R lie in the correct target class. We report for each reached radius during the attack the following measurements: (i) the percentage of the target identities that successfully reach it (column: labels %); (ii) the minimum, maximum, average number of iterations

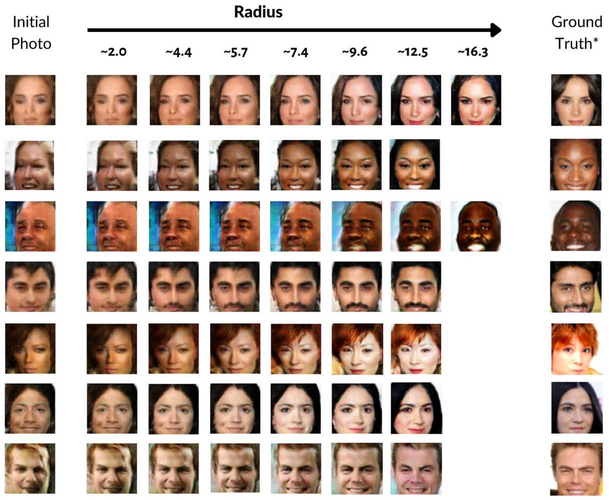

Figure 3. BREP-MI's progression along each radius from the first random initial point until the algorithm's termination.

required to reach it; and (iii) the attack success accuracy of the points that reached it (column: success %).

As we can see in the "labels %" column, BREP-MI is able to increase R multiple times for all target identities. In fact, all target identities had their R increased by BREP-MI at least 5 times. This shows the effectiveness of our algorithm to repel away from the boundary and get closer to the center of the class (which is our goal).

Another interesting observation is that the bigger the radius is, the higher the attack accuracy we get. This is true until reaching a certain radius size then the accuracy starts dropping (Similar behavior is observed in Fig.2). It seems at higher radii, we are overfitting to the target model and learning meaningless noisy features that are not actually related to the target identity. Consequently, the accuracy over the evaluator model decreases at those radii. This observation is consistent for all our conducted experiments regardless the model or dataset. Unfortunately, since the attacker does not have any ground truth images of the target class (or an evaluator classifier), it is not possible to decide what is the best radius that the algorithm should stop at. Nevertheless, as seen in the table 5, the number of identities that reached those radii is very low and their contribution to the final attack accuracy is small. Additionally, our stopping criterion for the algorithm empirically provides close results to the best radius. For this experiment particularly, we were able to get 75.67% which is close to stopping at the best radius.

Effect of Random Sampling As BREP-MI starts from an initial random point, it is important to show whether different initial points would affect the algorithm outcome (both qualitatively and quantitatively). We started three attacks with different random seeds on a FaceNet64 model trained on the CelebA dataset. The attack accuracy was 75.67%, 76.33%, and 75.67%, respectively. This shows that the random initial point has little effect quantitatively on our algorithm. Additionally, Fig.4 demonstrates our qualitative results. We can observe that even under different initial points, the output is close to the ground truth images.

Radius	labels %	min iters	max iters	avg iters	success %
2.00	100.00%	0	191	37.29	23.00%
2.60	100.00%	0	246	63.20	30.33%
3.38	100.00%	0	374	103.43	45.00%
4.39	100.00%	2	627	156.65	56.00%
5.71	100.00%	28	947	230.64	63.67%
7.43	100.00%	53	1721	336.38	71.67%
9.65	97.00%	89	1899	502.60	77.66%
12.55	71.00%	141	1909	746.60	80.28%
16.31	20.33%	298	1823	939.90	70.49%
21.21	1.67%	492	1875	1122.00	60.00%
27.57	0.67%	660	728	694.00	50.00%
35.84	0.33%	877	877	877.00	0.00%

Table 5. Analysis on the intermediate steps of our algorithm.

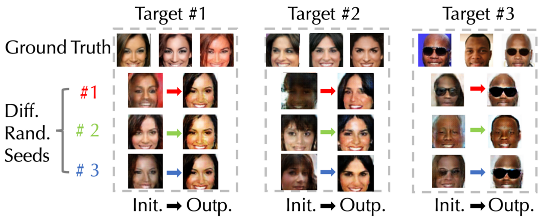

Figure 4. Ablation study of initialization of BREP-MI. "Init" is the initial sampled point that we start our attack from. "Outp" is the final output of the attack.

6. Conclusion

We presented a novel algorithm to perform the first label-only MI attack. Experiments showed the effectiveness of our approach on different datasets and model architectures. Interestingly, BREP-MI provides comparable results with the state-of-the-art whitebox attacks and outperforms all the other baselines despite the fact that they make stronger assumption about the attacker knowledge. This indicates that there is still room for improvement for whitebox attack. Similarly, the blackbox attack is outperformed by our label-only attack with a huge margin although it can access more fine-grained model output than our label-only attack.

Acknowledgements

This work was supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. For more information about CCI, please visit www.cyberinitiative.org. RJ and the ReDS Lab also gratefully acknowledge the support from Cisco.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 2

[2] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther. Gamin: An adversarial approach to black-box model inversion. *arXiv preprint arXiv:1909.11835*, 2019. 1, 2

[3] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020. 10

[4] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks, 2021. 1, 2, 3, 5, 6

[5] Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1924–1932, 2017. 5

[6] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning*, pages 1964–1974. PMLR, 2021. 2

[7] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 2

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015. 1, 2

[9] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32, 2014. 1

[10] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 619–633, 2018. 1, 3

[11] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, pages 133–152. De Gruyter, 2019. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

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

[15] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018. 1, 2

[16] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019. 1

[17] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019. 1, 2

[18] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014. 5

[19] Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *CVPR 2011 WORKSHOPS*, pages 35–42. IEEE, 2011. 5

[20] Anton Razzhigaev, Klim Kireev, Edgar Kaziakhmedov, Nurislam Tursynbek, and Aleksandr Petiushko. Black-box face recovery from identity features. In *European Conference on Computer Vision*, pages 462–475. Springer, 2020. 1

[21] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*, 2020. 1

[22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. 1, 2

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[24] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11666–11673, 2021. 13

[25] Ziqi Yang, Ee-Chien Chang, and Zhenkai Liang. Adversarial neural network inversion via auxiliary knowledge alignment. *arXiv preprint arXiv:1902.08552*, 2019. 1, 2, 3, 5, 6

[26] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020. 1, 2, 3, 5, 6, 13

Appendix

A. Theorem

The following proof technique was inspired by [3]. However, in our case the major difference is that our initial point z is an arbitrary point within the target class instead of being near the decision boundary, the fact which the authors [3] used to assume $M_{c^*}(z) = 0$. Furthermore, our goal is to move towards the centroid of a target class not to reach the boundary, which is completely opposite direction than ours. In summary, by exploiting the Taylor's theorem and the Minkowski inequality for expectations, we bounded the cosine angle between our gradient estimator and the true gradient, which we present for two different attack settings.

A.1. Linear Case

Theorem 2. *Assume f has a linear classification model. Let z be an arbitrary point within the target class, i.e. $M_{c^*}(z) > 0$. Then, the cosine of the angle between $\mathbb{E}[\widehat{M_{c^*}}(z,R)]$ and $\nabla M_{c^*}(z)$ is bounded by*

$$\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) \geq 1 - \mathcal{O}\left(\frac{M_{c^*}(z)^2 (d-1)^2}{\delta^2 R^2 \left\|\nabla M_{c^*}(z)\right\|_2^2}\right). \quad (10)$$

Therefore, with increasing radius R,

$$\lim_{R\to\infty} \cos\angle\left(\mathbb{E}\left[\widehat{M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) = 1, \quad (11)$$

which tells that the estimator is asymptotically unbiased for gradient estimation.

Proof. Let Ru be a random vector uniformly distributed on the sphere, where $R > 0$ is a radius of the sphere. By Taylor's theorem, for any $\delta \in (0,1)$, we have that

$$M_{c^*}(z+\delta Ru) = M_{c^*}(z) + \delta \nabla M_{c^*}(z)^T Ru. \quad (12)$$

Recall that

$$M_{c^*}(z) > 0$$

and let $w := \frac{M_{c^*}(z)}{\delta R}$.
For the case $\nabla M_{c^*}(z)^T u > w$, using Taylor series expansion and the fact that

$$\delta \nabla M_{c^*}(z)^T Ru > w\delta R = M_{c^*}(z),$$

we derive that $M_{c^*}(z+\delta Ru) > M_{c^*}(z) + M_{c^*}(z) > 0$.
Similarly, for the case $\nabla M_{c^*}(z)^T u < -w$, using Taylor expansion and the fact that

$$\delta \nabla M_{c^*}(z)^T Ru < -w\delta R = -M_{c^*}(z),$$

we have that $M_{c^*}(z+\delta Ru) < M_{c^*}(z) - M_{c^*}(z) = 0$.
Therefore, from these two cases, we arrive at

$$\phi_{c^*}(z+\delta Ru) = \begin{cases} 0 \text{ if } \nabla M_{c^*}(z)^T Ru > w \\ -1 \text{ if } \nabla M_{c^*}(z)^T Ru < -w. \end{cases} \quad (13)$$

We define $v_1 = \nabla M(z) / \left\|\nabla M_{c^*}(z)\right\|_2, v_2, \ldots, v_d$ by expanding the vector $\nabla M_{c^*}(z)$ to orthogonal bases in $\mathbb{R}^d$. Then, we can write a random vector $Ru = \sum_{i=1}^d \beta_i v_i$, where β is uniformly distributed on the sphere of radius R. We construct an upper cap $E_1 := \left\{\nabla M_{c^*}(z)^T Ru > w\right\}$, the annulus $E_2 := \left\{\left|\nabla M_{c^*}(z)^T Ru\right| < w.\right\}$, and the lower cap $E_3 := \left\{\nabla M_{c^*}(z)^T Ru < -w\right\}$. Let $p := \mathbb{P}(E_2)$ be the probability of event E_2, then $\mathbb{P}(E_1) = \mathbb{P}(E_3) = (1-p)/2$.

For any $i \neq 1$ by symmetry:

$$\mathbb{E}[\beta_i \mid E_1] = \mathbb{E}[\beta_i \mid E_3] = 0. \quad (14)$$

Then, the expected value of the estimator becomes

$$\begin{aligned} \mathbb{E}[\phi_{c^*}(z+\delta Ru)u] &= p \cdot (\mathbb{E}[\phi_{c^*}(z+\delta Ru)u \mid E_2]) \\ &-\frac{1}{2}p \cdot (\mathbb{E}[\beta_1 v_1 \mid E_1] + \mathbb{E}[-\beta_1 v_1 \mid E_3]) \\ &+\frac{1}{2}\mathbb{E}[\beta_1 v_1 \mid E_1] + \frac{1}{2}\mathbb{E}[-\beta_1 v_1 \mid E_3]. \end{aligned} \quad (15)$$

Now, we can bound the difference between $\mathbb{E}[|\beta_1| v_1] = \frac{\mathbb{E}[|\beta_1|]}{\left\|\nabla M_{c^*}(z)\right\|_2} \nabla M_{c^*}(z)$ and $\mathbb{E}[\phi_{c^*}(z+\delta Ru)u]$ using Eq. 15:

$$\begin{aligned} &\left\|\mathbb{E}[\phi_{c^*}(z+\delta Ru)u] - \mathbb{E}[|\beta_1| v_1]\right\|_2 \\ &\leq \Big\| p \cdot (\mathbb{E}[\phi_{c^*}(z+\delta Ru)u \mid E_2]) \\ &- \frac{1}{2}p \cdot (\mathbb{E}[\beta_1 v_1 \mid E_1] + \mathbb{E}[-\beta_1 v_1 \mid E_3]) \Big\|_2 \\ &\leq p \cdot (R + \frac{1}{2}R + \frac{1}{2}R) = 2Rp. \end{aligned} \quad (16)$$

In Ineq. 16, we first substitute the LHS with $\mathbb{E}[|\beta_1| v_1] = \frac{\mathbb{E}[|\beta_1|]}{\left\|\nabla M_{c^*}(z)\right\|_2} \nabla M_{c^*}(z)$, then square both sides, and lastly divide by $\mathbb{E}[|\beta_1|]^2$ to derive the following inequality:

$$\left\|\frac{\mathbb{E}[\phi_{c^*}(z+\delta Ru)u]}{\mathbb{E}[|\beta_1|]} - \frac{\nabla M_{c^*}(z)}{\left\|\nabla M_{c^*}(z)\right\|_2}\right\|_2^2 \leq \left(\frac{2Rp}{\mathbb{E}[|\beta_1|]}\right)^2. \quad (17)$$

Using the property of the angle between two vectors, we obtain the cosine inequality:

$$\cos\angle\left(\mathbb{E}[\phi_{c^*}(z+\delta Ru)u], \nabla M_{c^*}(z)\right) \geq 1 - \frac{1}{2}\left(\frac{2Rp}{\mathbb{E}[|\beta_1|]}\right)^2.$$

(18)

The probability p can be bounded using the fact that $\left\langle \frac{\nabla M_{c^*}(z)}{\|\nabla M_{c^*}(z)\|_2}, u \right\rangle^2$ is a Beta distribution $\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)$:

$$\begin{aligned} p &= \mathbb{P}\left(\left\langle \frac{\nabla M_{c^*}(z)}{\|\nabla M_{c^*}(z)\|_2}, u \right\rangle^2 \leq \frac{w^2}{\|\nabla M_{c^*}(z)\|_2^2}\right) \\ &\leq \frac{2w}{\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)\|\nabla M_{c^*}(z)\|_2}. \end{aligned} \tag{19}$$

Substituting the Bound 19 into Eq. 18, we receive a bound for the cosine

$$\begin{aligned} &\cos\angle\left(\mathbb{E}\left[\phi_{c^*}(z+\delta Ru)\, u\right], \nabla M_{c^*}(z)\right) \\ &\geq 1 - \frac{8R^2w^2}{\left(\mathbb{E}\left[|\beta_1|\right]\right)^2 \mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)^2 \|\nabla M_{c^*}(z)\|_2^2} \\ &= 1 - \frac{2M_{c^*}(z)^2 R^2 (d-1)^2}{R^2\delta^2R^2\|\nabla M_{c^*}(z)\|_2^2} \\ &= 1 - \frac{2M_{c^*}(z)^2 (d-1)^2}{\delta^2R^2\|\nabla M_{c^*}(z)\|_2^2}. \end{aligned} \tag{20}$$

In addition, we have that

$$\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right] = \mathbb{E}\left[\phi_{c^*}(z+\delta Ru)\, u\right], \tag{21}$$

so plugging it to Ineq. 20, we have the final bound to be

$$\begin{aligned} &\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) \\ &\geq 1 - \frac{9M_{c^*}(z)^2 (d-1)^2}{2\delta^2R^4\|\nabla M_{c^*}(z)\|_2^2}. \end{aligned} \tag{22}$$

Therefore, we can observe that with increasing radius R, we will achieve

$$\lim_{R\to\infty} \cos\angle\left(\mathbb{E}\left[\widehat{M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) = 1. \tag{23}$$

□

A.2. Nonlinear Case

Theorem 3. *Let z be an arbitrary point within the target class, i.e. $M_{c^*}(z) > 0$. Assume f be a nonlinear classification model with a Lipschitz continuous gradient in a neighborhood of z. Then, the cosine of the angle between $\mathbb{E}[\widehat{M_{c^*}}(z,R)]$ and $\nabla M_{c^*}(z)$ is bounded by*

$$\begin{aligned} &\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) \\ &\geq 1 - \frac{4M_{c^*}(z)^2 + L^2\delta^4R^4 + 4M_{c^*}(z)^2 L\delta^2R^2}{2\delta^2R^2(d-1)^{-2}\|\nabla M_{c^*}(z)\|_2^2}. \end{aligned} \tag{24}$$

Therefore, with increasing radius R up to the point of inflection K,

$$\lim_{R\to K} \cos\angle\left(\mathbb{E}\left[\widehat{M_{c^*}}(z,R)\right], \nabla M_{c^*}(z)\right) = 1, \tag{25}$$

which tells that the alignment between the proposed gradient estimator and the true gradient $\nabla M_{c^*}(z)$ is increased with greater radius R until reaching a certain point K. After passing that point of inflection, the alignment between them branches off.

Proof. Let Ru be a random vector uniformly distributed on the sphere, where $R > 0$ is a radius of the sphere. By Taylor's theorem, for any $\delta \in (0,1)$, we have that

$$\begin{aligned} &M_{c^*}(z+\delta Ru) \\ &= M_{c^*}(z) + \delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}\delta^2\nabla^2 M_{c^*}(z)^T R^2 u. \end{aligned} \tag{26}$$

For any function f that has a Lipschitz continuous gradient with Lipschitz constant $L > 0$, the following inequality holds

$$\begin{gathered} \left\|\nabla^2 M_{c^*}(z)\, u\right\|_2 \leq L \\ \Longleftrightarrow \\ \left\|\frac{1}{2}\delta^2\nabla^2 M_{c^*}(z)\, R^2 u\right\|_2 \leq \frac{1}{2}L\delta^2R^2. \end{gathered} \tag{27}$$

Recall that

$$M_{c^*}(z) > 0$$

and let $w := \frac{M_{c^*}(z)}{\delta R} + \frac{1}{2}L\delta R$.

For the case $\nabla M_{c^*}(z)^T u > w$, using Taylor series expansion with the negative Lipschitz bound and the fact that

$$\begin{aligned} &\delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}\delta^2\nabla^2 M_{c^*}(z)^T R^2 u \\ &\geq \delta\nabla M_{c^*}(z)^T Ru - \frac{1}{2}L\delta^2R^2 \\ &> \delta\nabla M_{c^*}(z)^T Ru - \frac{1}{2}L\delta^2R^2 - M_{c^*}(z) \\ &= \delta R\left(\nabla M_{c^*}(z)^T u - w\right) > 0, \end{aligned} \tag{28}$$

we derive that $M_{c^*}(z+\delta Ru) > M_{c^*}(z) > 0$.

Similarly, for the case $\nabla M_{c^*}(z)^T u < -w$, which can be rewritten as

$$\delta R\nabla M_{c^*}(z)^T u + \delta Rw < 0, \tag{29}$$

using Taylor expansion with the positive Lipschitz bound and the fact that

$$\begin{aligned} &\delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}\delta^2\nabla^2 M_{c^*}(z)^T R^2 u \\ &\leq \delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}L\delta^2R^2 \\ &= \left(\delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}L\delta^2R^2 + M_{c^*}(z)\right) - M_{c^*}(z) \\ &= \left(\delta R\nabla M_{c^*}(z)^T u + \delta Rw\right) - M_{c^*}(z) \\ &< -M_{c^*}(z) \qquad \text{(by Inequality 29)} \end{aligned} \tag{30}$$

we obtain that $M_{c^*}(z+\delta Ru) = M_{c^*}(z) + \delta\nabla M_{c^*}(z)^T Ru + \frac{1}{2}\delta^2\nabla^2 M_{c^*}(z)^T R^2 u < M_{c^*}(z) - M_{c^*}(z) = 0$.
Therefore, from these two cases, we arrive at

$$\phi_{c^*}(z+\delta Ru) = \begin{cases} 0 \text{ if } \nabla M_{c^*}(z)^T Ru > w \\ -1 \text{ if } \nabla M_{c^*}(z)^T Ru < -w. \end{cases} \tag{31}$$

We define $v_1 = \nabla S(z) / \|\nabla M_{c^*}(z)\|_2, v_2, \ldots, v_d$ by expanding the vector $\nabla M_{c^*}(z)$ to orthogonal bases in $\mathbb{R}^d$. Then, we can write a random vector $Ru = \sum_{i=1}^{d}\beta_i v_i$, where β is uniformly distributed on the sphere of radius R. We construct an upper cap $E_1 := \left\{\nabla M_{c^*}(z)^T Ru > w\right\}$, the annulus $E_2 := \left\{\left|\nabla M_{c^*}(z)^T Ru\right| < w.\right\}$, and the lower cap $E_3 := \left\{\nabla M_{c^*}(z)^T Ru < -w\right\}$. Let $p := \mathbb{P}(E_2)$ be the probability of event E_2, then $\mathbb{P}(E_1) = \mathbb{P}(E_3) = (1-p)/2$. For any $i \neq 1$ by symmetry:

$$\mathbb{E}[\beta_i \mid E_1] = \mathbb{E}[\beta_i \mid E_3] = 0. \tag{32}$$

Then, the expected value of the estimator becomes

$$\begin{aligned} \mathbb{E}[\phi_{c^*}(z+\delta Ru)u] = p\cdot(\mathbb{E}[\phi_{c^*}(z+\delta Ru)u \mid E_2]) \\ -\frac{1}{2}p\cdot(\mathbb{E}[\beta_1 v_1 \mid E_1] + \mathbb{E}[-\beta_1 v_1 \mid E_3]) \\ +\frac{1}{2}\mathbb{E}[\beta_1 v_1 \mid E_1] + \frac{1}{2}\mathbb{E}[-\beta_1 v_1 \mid E_3]. \end{aligned} \tag{33}$$

Now, we can bound the difference between $\mathbb{E}[|\beta_1| v_1] = \frac{\mathbb{E}[|\beta_1|]}{\|\nabla M_{c^*}(z)\|_2}\nabla M_{c^*}(z)$ and $\mathbb{E}[\phi_{c^*}(z+\delta Ru)u]$ using Eq. 33:

$$\begin{aligned} &\|\mathbb{E}[\phi_{c^*}(z+\delta Ru)u] - \mathbb{E}[|\beta_1| v_1]\|_2 \\ &\leq \Big\|p\cdot(\mathbb{E}[\phi_{c^*}(z+\delta Ru)u \mid E_2]) \\ &- \frac{1}{2}p\cdot(\mathbb{E}[\beta_1 v_1 \mid E_1] + \mathbb{E}[-\beta_1 v_1 \mid E_3])\Big\|_2 \\ &\leq p\cdot(R + \frac{1}{2}R + \frac{1}{2}R) = 2Rp. \end{aligned} \tag{34}$$

In Inequality 34, we first substitute the LHS with $\mathbb{E}[|\beta_1| v_1] = \frac{\mathbb{E}[|\beta_1|]}{\|\nabla M_{c^*}(z)\|_2}\nabla M_{c^*}(z)$, then square both sides, and lastly divide by $\mathbb{E}[|\beta_1|]^2$ to derive the following

$$\left\|\frac{\mathbb{E}[\phi_{c^*}(z+\delta Ru)u]}{\mathbb{E}[|\beta_1|]} - \frac{\nabla M_{c^*}(z)}{\|\nabla M_{c^*}(z)\|_2}\right\|_2^2 \leq \left(\frac{2Rp}{\mathbb{E}[|\beta_1|]}\right)^2. \tag{35}$$

Using the property of the angle between two vectors, we obtain the cosine inequality:

$$\begin{aligned} &\cos\angle(\mathbb{E}[\phi_{c^*}(z+\delta Ru)u], \nabla M_{c^*}(z)) \\ &\geq 1 - \frac{1}{2}\left(\frac{2Rp}{\mathbb{E}[|\beta_1|]}\right)^2. \end{aligned} \tag{36}$$

The probability p can be bounded using the fact that $\left\langle\frac{\nabla M_{c^*}(z)}{\|\nabla M_{c^*}(z)\|_2}, u\right\rangle^2$ is a Beta distribution $\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)$:

$$\begin{aligned} p &= \mathbb{P}\left(\left\langle\frac{\nabla M_{c^*}(z)}{\|\nabla M_{c^*}(z)\|_2}, u\right\rangle^2 \leq \frac{w^2}{\|\nabla M_{c^*}(z)\|_2^2}\right) \\ &\leq \frac{2w}{\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)\|\nabla M_{c^*}(z)\|_2}. \end{aligned} \tag{37}$$

Substituting the Bound 37 into equality 36, we receive a bound for the cosine

$$\begin{aligned} &\cos\angle(\mathbb{E}[\phi_{c^*}(z+\delta Ru)u], \nabla M_{c^*}(z)) \\ &\geq 1 - \frac{8R^2w^2}{(\mathbb{E}[|\beta_1|])^2\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)^2\|\nabla M_{c^*}(z)\|_2^2} \\ &\geq 1 - \frac{8M_{c^*}(z)^2 + 2L^2\delta^4R^4 + 8M_{c^*}(z)^2 L\delta^2R^2}{\delta^2(\mathbb{E}[|\beta_1|])^2\mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)^2\|\nabla M_{c^*}(z)\|_2^2} \\ &\geq 1 - \frac{4M_{c^*}(z)^2 + L^2\delta^4R^4 + 4M_{c^*}(z)^2 L\delta^2R^2}{2\delta^2R^2(d-1)^{-2}\|\nabla M_{c^*}(z)\|_2^2}. \end{aligned} \tag{38}$$

In addition, we have that

$$\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z, R)\right] = \mathbb{E}[\phi_{c^*}(z+\delta Ru)u], \tag{39}$$

so plugging it to inequality 38, we have the final bound to be

$$\begin{aligned} &\cos\angle\left(\mathbb{E}\left[\widehat{\nabla M_{c^*}}(z, R)\right], \nabla M_{c^*}(z)\right) \\ &\geq 1 - \frac{4M_{c^*}(z)^2 + L^2\delta^4R^4 + 4M_{c^*}(z)^2 L\delta^2R^2}{2\delta^2R^2(d-1)^{-2}\|\nabla M_{c^*}(z)\|_2^2}. \end{aligned} \tag{40}$$

Therefore, we can observe that with increasing radius R, we will achieve

$$\lim_{R\to K}\cos\angle\left(\mathbb{E}\left[\widehat{M_{c^*}}(z, R)\right], \nabla M_{c^*}(z)\right) = 1, \tag{41}$$

□

B. Experiments

B.1. Performance on MNIST

We extend our experiments to include tasks other than facial recognition. Particularly, we perform our attack on MNIST dataset. We use identical attack model to the main experiments, in which the attacker only gets access to the target model's decision. We also assume that that the attacker has access to a prior knowledge (public dataset). In our case, these are the digits from “5” to “9”. The attacker’s goal is to infer information on the private dataset (i.e., the digits from “0” to “4”). This is not trivial since the public knowledge available to the attacker contains only 5 classes which makes it harder for the attacker to generalize to other classes. Nevertheless, BREP-MI was able to successfully attack 4 out of the 5 private classes.

Similar to the main experiments, we train two models. The target model is a network with 2 CNN layers followed by two linear layers and the evaluation classifier has 3 CNN layers followed by two linear layers. As shown in Fig.5, the initial image generated by GAN for class “0” resembles class “6”. This is expected, since the attacker has only knowledge of classes from “5” to “9”. The final output however, looks more similar to “0”. Similar behavior can be seen when attacking class “4”. For classes “2” and “3”, the initial images were noises, yet the we were still able to successfully attack class “3”.

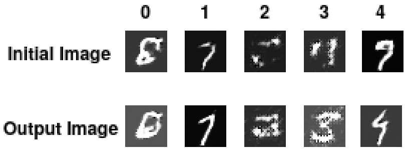

Figure 5. Performing BREP-MI on MNIST dataset. The output is shows for each private label (digit).

B.2. Performance Against Defenses

Unfortunately, there is currently no viable defense against Model Inversion. Particularly, [26] showed that differential privacy –a gold standard for protecting privacy– is not protective against Model Inversion attacks.

We conducted experiments to attack the most recent and effective defensive mechanism [24]. Our results, shown in table 6, indicate that there is a direct tradeoff between privacy and model performance. For a high value of λ (a regularization factor that controls robustness against MI attacks), the attack accuracy decreases significantly, but so does the performance of the target model. The discussion about this tradeoff is also reported by [24] in which they emphasize that this defense does not seek to prevent the attack, but rather to optimize the tradeoff between dataset privacy and model performance.

The absence of an effective defense in Model Inversion setting is concerning for the privacy of datasets. Indeed, it is a crucial future direction.

Defense	**Target Model Acc**	**Attack Acc**		
		GMI	KED	BREP
No defense	89	15	69	63
$\lambda = 0.001$	82	17	64	62
$\lambda = 0.01$	66	15	39	34
$\lambda = 0.1$	No train	-	-	-

Table 6. Several values for λ and their corresponding attack accuracy and accuracy of the target model (VGG16 that is trained on CelebA dataset). For $\lambda = 0.1$, the model was not able to train.

B.3. Implicit Attributes Recovery

Another important metric for evaluating MI attacks is to show their ability to recover implicit attributes from private datasets. Table 7 shows the accuracy of recovering certain attributes on CelebA dataset. We notice that we still provide accuracy that is close to the SOTA whitebox attack.

Attributes	**Attack Acc**		
	GMI	KED	BREP
Blond Hair	69	**81**	74
Straight Hair	**76**	72	75
Wavy Hair	70	**71**	70
Gray Hair	84	**94**	88
Bald	**96**	**96**	**96**
Bushy Eyebrows	83	**85**	83
Arched Eyebrows	63	**73**	70
Glasses	90	**91**	**91**
Mustache	91	**95**	**95**
Young	**58**	**58**	**58**
Big Nose	72	**75**	70
Double Chin	94	**96**	95

Table 7. Implicit attributes recovery of BREP-MI against baselines. Attack accuracy is measured by an attributes classifier that is trained on CelebA.