Re-thinking Model Inversion Attacks Against Deep Neural Networks

Ngoc-Bao Nguyen* Keshigeyan Chandrasegaran* Milad Abdollahzadeh Ngai-Man Cheung†
Singapore University of Technology and Design (SUTD)
thibaongoc_nguyen@mymail.sutd.edu.sg,{keshigeyan,milad_abdollahzadeh,ngaiman_cheung}@sutd.edu.sg

Abstract

Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model. MI attacks have raised concerns about the leaking of sensitive information (e.g. private face images used in training a face recognition system). Recently, several algorithms for MI have been proposed to improve the attack performance. In this work, we revisit MI, study two fundamental issues ***pertaining to all state-of-the-art (SOTA) MI algorithms****, and propose solutions to these issues which lead to a significant boost in attack performance for all SOTA MI. In particular, our contributions are two-fold: 1) We analyze the optimization objective of SOTA MI algorithms, argue that the objective is sub-optimal for achieving MI, and propose an improved optimization objective that boosts attack performance significantly. 2) We analyze "MI overfitting", show that it would prevent reconstructed images from learning semantics of training data, and propose a novel "model augmentation" idea to overcome this issue. Our proposed solutions are simple and improve all SOTA MI attack accuracy significantly. E.g., in the standard CelebA benchmark, our solutions improve accuracy by* ***11.8%*** *and achieve for the first time over 90% attack accuracy.* ***Our findings demonstrate that there is a clear risk of leaking sensitive information from deep learning models.*** *We urge serious consideration to be given to the privacy implications. Our code, demo, and models are available at https://ngoc-nguyen-0.github.io/re-thinking_model_inversion_attacks/.*

1. Introduction

Privacy of deep neural networks (DNNs) has attracted considerable attention recently [2, 3, 23, 31, 32]. Today, DNNs are being applied in many domains involving private and sensitive datasets, e.g., healthcare, and security. There is a growing concern of privacy attacks to gain knowledge of confidential datasets used in training DNNs. One important category of privacy attacks is Model Inversion (MI) [5, 8, 11, 12, 16, 36, 37, 39, 40] (Fig. 1). Given access to a model, MI attacks aim to infer and reconstruct features of the private dataset used in the training of the model. For example, a malicious user may attack a face recognition system to reconstruct sensitive face images used in training. Similar to previous work [5,36,39], we will use face recognition models as the running example.

Related Work. MI attacks were first introduced in [12], where simple linear regression is the target of attack. Recently, there is a fair amount of interest to extend MI to complex DNNs. Most of these attacks [5, 36, 39] focus on the *whitebox* setting and the attacker is assumed to have complete knowledge of the model subject to attack. As many platforms provide downloading of entire trained DNNs for users [5, 39], whitebox attacks are important. [39] proposes Generative Model Inversion (GMI) attack, where generic public information is leveraged to learn a distributional prior via generative adversarial networks (GANs) [13, 35], and this prior is used to guide reconstruction of private training samples. [5] proposes Knowledge-Enriched Distributional Model Inversion (KEDMI), where an inversion-specific GAN is trained by leveraging knowledge provided by the target model. [36] proposes Variational Model Inversion (VMI), where a probabilistic interpretation of MI leads to a variational objective for the attack. KEDMI and VMI achieve SOTA attack performance (See Supplementary for further discussion of related work).

In this paper, we revisit SOTA MI, study two issues pertaining to all SOTA MI and propose solutions to these issues that are complementary and applicable to all SOTA MI (Fig. 1). In particular, despite the range of approaches proposed in recent works, common and central to all these approaches is an *inversion step* which formulates reconstruction of training samples as an optimization. The optimization objective in the inversion step involves the *identity loss*, which is the *same* for all SOTA MI and is formulated as the negative log-likelihood for the reconstructed samples under the model being attacked. While ideas have been proposed to advance other aspects of MI, *effective design of the identity loss has not been studied*.

*Equal Contribution †Corresponding Author

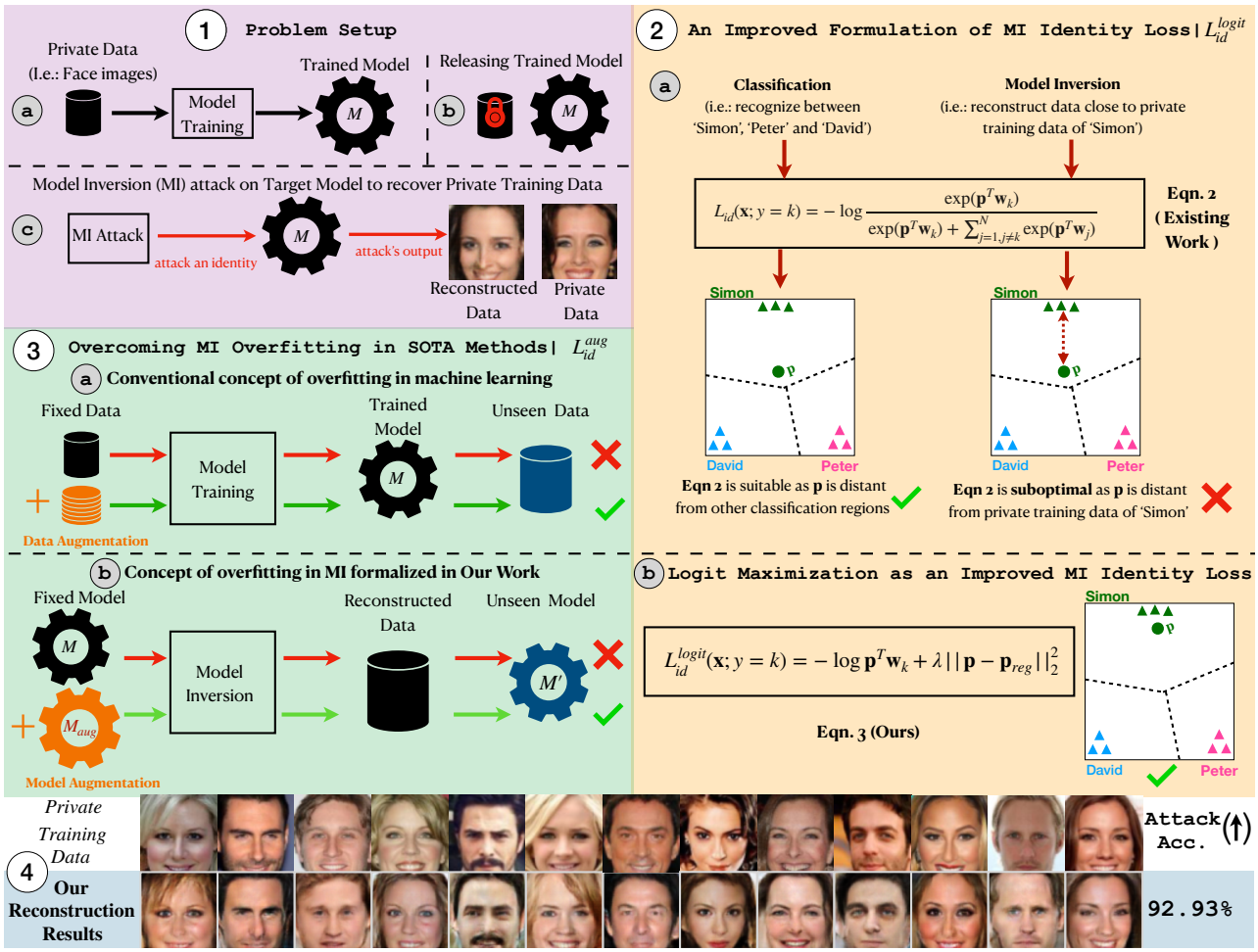

Figure 1. *Overview and our contributions.* ① We consider the problem of the Model Inversion (MI) attack to reconstruct private training data based on model parameters. Our work makes two foundational contributions to MI attacks. ② First, we analyse the optimization objective of existing SOTA MI algorithms and show that they are sub-optimal. Further, we propose an improved optimization objective that boosts MI attack performance significantly (Sec 3.1). ③ Second, we formalize the concept of "MI overfitting" showing that it prevents reconstructed images from learning identity semantics of training data. Further, we propose a novel "model augmentation" idea to overcome this issue (Sec 3.2). ④ Our proposed method significantly boosts MI attack accuracy. *E.g.* In the standard CelebA benchmark, our method boosts attack accuracy by *11.8%, achieving above 90% attack accuracy for the first time in contemporary MI literature*.

To address this research gap, our work studies subtleties of identity loss in all SOTA MI, analyzes the issues and proposes improvements that boost the performance of all SOTA significantly. In summary, our contributions are as follows:

- We analyze existing identity loss, argue that it could be sub-optimal for MI, and propose an improved identity loss that aligns better with the goal of MI (Fig. 1 ②).
- We formalize the concept of *MI overfitting*, analyze its impact on MI and propose a novel solution based on *model augmentation*. Our idea is inspired by the conventional issue of overfitting in model training and data augmentation as a solution to alleviate the issue (Fig. 1 ③).
- We conduct extensive experiments to demonstrate that our solutions can improve SOTA MI algorithms (GMI [39], KEDMI [5], VMI [36]) significantly. Our solutions achieve for the first time over 90% attack accuracy under standard CelebA benchmark (Fig. 1 ④).

Our work sounds alarm over the rising threats of MI attacks, and urges more attention on measures against the leaking of private information from DNNs.

2. General Framework of SOTA MI Attacks

Problem Setup. In MI, an attacker abuses access to a model M trained on a private dataset $\mathcal{D}_{priv}$. The attacker can access M, but $\mathcal{D}_{priv}$ is not intended to be shared. The goal of MI is to infer information about private samples in $\mathcal{D}_{priv}$. In existing work, for the desired class (label) y,

Table 1. Categorizing SOTA MI attacks based on their difference in latent code distribution and prior loss. $p_{\texttt{GAN}}(\mathbf{z})$ is a GAN prior. G and D are generator and discriminator of a GAN.

Method	**Latent distribution** $q(\mathbf{z})$	**Prior loss** L_{prior}
GMI [39]	Point estimate $\delta(\mathbf{z}-\mathbf{z}_0)$	$-D(G(\mathbf{z}))$
KEDMI [5]	Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	$-\log D(G(\mathbf{z}))$
VMI [36]	Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *or* Normalizing Flow [19]	Distance *w.r.t.* GAN prior $D_{\texttt{KL}}(q(\mathbf{z})\|\|p_{\texttt{GAN}}(\mathbf{z}))$

MI is formulated as the reconstruction of an input $\mathbf{x}$ which is most likely classified into y by the model M. For instance, if the problem involves inverting a facial recognition model, given the desired identity, MI is formulated as the reconstruction of facial images that are most likely to be recognized as the desired identity. The model subject to MI attacks is called *target model*. Following previous works [5, 36, 39], we focus on *whitebox* MI attack, where the attacker is assumed to have complete access to the target model. For high-dimensional data such as facial images, this reconstruction problem is ill-posed. Consequently, various SOTA MI methods have been proposed recently to constrain the search space to the manifold of meaningful and relevant images using a GAN: using a GAN trained on some public dataset $\mathcal{D}_{pub}$ [39], using an inversion-specific GAN [5], and defining variational inference in latent space of GAN [36].

Despite the differences in various SOTA MI, common and central to all these methods is an *inversion step* –called *secret revelation* in [39]–, which performs the following optimization:

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})} \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\{L_{id}(\mathbf{z}; y, M) + \lambda L_{prior}(\mathbf{z})\} \tag{1}$$

Here $L_{id}(\mathbf{z}; y, M) = -\log \mathbb{P}_M(y|G(\mathbf{z}))$ is referred to as *identity loss* in MI [39], which guides the reconstruction of $\mathbf{x} = G(\mathbf{z})$ that is most likely to be recognized by model M as identity y, and L_{prior} is some prior loss, and $q^*(\mathbf{z})$ is the optimal distribution of latent code used to generate inverted samples by GAN ($\mathbf{x} = G(\mathbf{z})$; $\mathbf{z} \sim q^*(\mathbf{z})$). Importantly, all SOTA MI methods use the *same* identity loss $L_{id}(\mathbf{z}; y, M)$, although they have different assumption about $q(\mathbf{z})$ and the prior loss L_{prior} (see Table 1 and Supplementary for more details on each algorithm). While advances observed by improving $q(\mathbf{z})$ and L_{prior}, ***the design of more effective L_{id} has been left unnoticed*** in all SOTA MI algorithms. Therefore, our work instead focuses on L_{id}, analyzes issues, and proposes improvement for L_{id} that can lead to a performance boost in all SOTA MI. To simplify notations, we denote $L_{id}(\mathbf{z}; y, M)$ by $L_{id}(\mathbf{x}; y)$ when appropriate, where $\mathbf{x} = G(\mathbf{z})$ is the reconstructed image.

3. A Closer Look at Model Inversion Attacks

3.1. An Improved Formulation of MI Identity Loss

In this section, we discuss our first contribution and take a closer look at the optimization objective of *identity loss*, $L_{id}(\mathbf{x}; y)$. Existing SOTA MI methods, namely GMI [39], KEDMI [5] and VMI [36] formulate the identity loss as an optimization to minimize the negative log likelihood of an identity under model parameters (*i.e*. cross-entropy loss). Particularly, the $L_{id}(\mathbf{x}; y)$ introduced in Eqn. 1 for an inversion targeting class k can be re-written as follows:

$$L_{id}(\mathbf{x}; y = k) = -\log \frac{\exp(\mathbf{p}^T\mathbf{w}_k)}{\exp(\mathbf{p}^T\mathbf{w}_k) + \sum_{j=1, j\neq k}^{N} \exp(\mathbf{p}^T\mathbf{w}_j)} \tag{2}$$

where $\mathbf{p}$ refers to penultimate layer activations [4, 26] for sample $\mathbf{x}$ and $\mathbf{w}_i$ refers to the last layer weights for the i^{th} class [1]in target model M.

Existing identity loss (Eqn. 2) used in SOTA MI methods [5,36,39] is sub-optimal for MI (Fig. 1 ②). Although the optimization in Eqn. 2 accurately captures the essence of a classification problem (*e.g*. face recognition), we postulate that such formulation is sub-optimal for MI. We provide our intuition through the lens of penultimate layer activations, $\mathbf{p}$ (Fig. 1 ②). In a classification setting, the main expectation for $\mathbf{p}$ is to be sufficiently discriminative for class k (*e.g*. recognize between 'Peter', 'Simon' and 'David'). This objective can be achieved by both maximizing $\exp(\mathbf{p}^T\mathbf{w}_k)$ and/or minimizing $\sum_{j=1, j\neq k}^{N} \exp(\mathbf{p}^T\mathbf{w}_j)$ in Eqn. 2. On the contrary, *the goal of MI is to reconstruct training data*. That is, in addition to $\mathbf{p}$ being sufficiently discriminative for class k, successful inversion also requires $\mathbf{p}$ to be close to the training data representations for class k represented by $\mathbf{w}_k$ (*i.e*. an inversion targeting 'Simon' needs to reconstruct a sample close to the private training data of 'Simon'; Fig. 1 ②). Specifically, we argue that MI requires a lot more attention on maximizing $\exp(\mathbf{p}^T\mathbf{w}_k)$ compared to minimizing $\sum_{j=1, j\neq k}^{N} \exp(\mathbf{p}^T\mathbf{w}_j)$ in Eqn. 2.

Motivated by this hypothesis, we conduct an analysis to investigate the proximity between private training data and reconstructed data in SOTA MI methods using penultimate layer representations [4, 25, 26, 29]. Particularly, our analysis using KEDMI [5] (SOTA) shows several instances where using Eqn. 2 for identity loss is unable to reconstruct data close to the private training data. We show this in Fig. 2 (top row). Consequently, our analysis motivates the search for an improved identity loss focusing on maximizing $\exp(\mathbf{p}^T\mathbf{w}_k)$ for MI.

[1] $\mathbf{p}$ is concatenated with 1 at the end to include bias as $\mathbf{w}_i$ includes biases at the end.

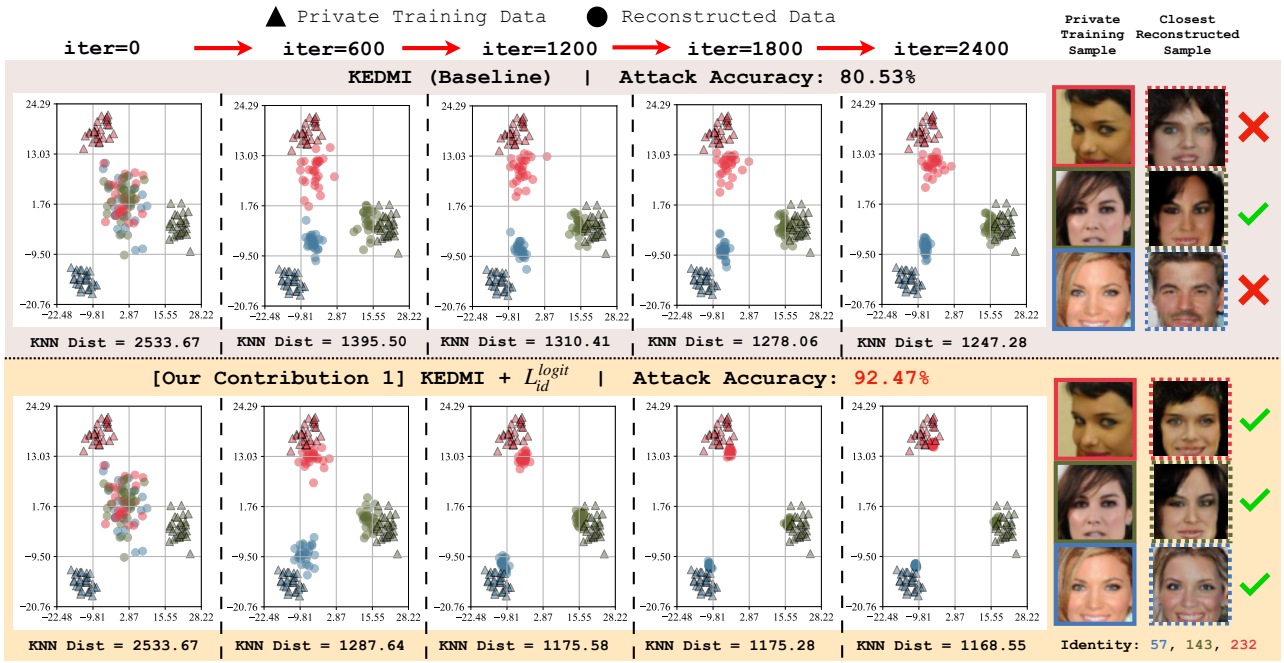

Figure 2. Visualization of the penultimate layer representations ($\mathcal{D}_{priv}$ = CelebA [24], $\mathcal{D}_{pub}$ = CelebA [24], Target Model = IR152 [5], Evaluation Model = face.evoLve [6], Inversion iterations = 2400) for private training data and reconstructed data using KEDMI [5]. Following the exact evaluation protocol in [5], we use face.evoLve [6] to extract representations. We show results for 3 randomly chosen identities. We include KNN distance (for different iterations) and final attack accuracy following the protocol in [5]. For each identity, we also include randomly selected private training data and the closest reconstructed sample at iteration=2400. ① **Identity loss in SOTA MI methods [5, 36, 39] (Eqn. 2) is sub-optimal for MI (Top).** Using penultimate representations during inversion, we observe 2 instances (*e.g.* target identity **57** and **143**) where KEDMI [5] (using Eqn. 2 for identity loss) is unable to reconstruct data close to private training data. Hence, private and reconstructed facial images are qualitatively different. ② **Our proposed identity loss, L_{id}^{logit} (Eqn. 3), can effectively guide the reconstruction of data close to private training data (Bottom)**. This can be clearly observed using both penultimate layer representations and KNN distances for all 3 target classes **57**, **143** and **252**. We show similar results using additional MI algorithms (GMI [39], VMI [36]) and target classifiers (face.evoLve, VGG16) in Supplementary. Best viewed in color.

Logit Maximization as an improved MI identity loss. In light of our analysis / observations above, we propose to directly maximize the logit, $\mathbf{p}^T\mathbf{w}_k$, instead of maximizing the log likelihood of class k for MI. Our proposed identity loss objective is shown below:

$$L_{id}^{logit}(\mathbf{x}; y=k) = -\log \mathbf{p}^T\mathbf{w}_k + \lambda||\mathbf{p}-\mathbf{p}_{reg}||_2^2 \quad (3)$$

where $\lambda(>0)$ is a hyper-parameter and $\mathbf{p}_{reg}$ is used for regularizing $\mathbf{p}$. Particularly, if the regularization in Eqn. 3 is omitted and hence $||\mathbf{p}||$ is unbounded, a crude simplified way to solve Eqn. 3 is to maximize $||\mathbf{p}||$. Hence, we use $\mathbf{p}_{reg}$ to regularize $\mathbf{p}$. Given that the attacker has no access to private training data, we estimate $\mathbf{p}_{reg}$ by a simple method using *public* data (See Supplementary). We remark that $\mathbf{p} = M^{\texttt{pen}}(\mathbf{x})$ where $\mathbf{x} = G(\mathbf{z})$ and $M^{\texttt{pen}}()$ operator returns the penultimate layer representations for a given input.

Our analysis shows that our proposed identity loss, L_{id}^{logit} (Eqn. 3), can significantly improve reconstruction of private training data compared to existing identity loss used in SOTA MI algorithms [5, 36, 39]. This can be clearly observed using both penultimate layer representations and KNN distances in Fig. 2 (bottom row). Here KNN Dist refers to the shortest Euclidean feature distance from a reconstructed image to private training images for a given identity [5, 39]. Our proposed L_{id}^{logit} can be easily plugged in to all existing SOTA MI algorithms by replacing L_{id} with our proposed L_{id}^{logit} in Eqn. 1 (in the inversion step) with minimal computational overhead.

3.2. Overcoming MI Overfitting in SOTA methods

In this section, we discuss our second contribution. In particular, we formalize a concept of *MI overfitting*, observe its considerable impacts even in SOTA MI methods [5, 36, 39], and propose a new, simple solution to overcome this issue (Fig. 1 ③). To better discuss our MI overfitting concept, we first review the conventional concept of overfitting in machine learning: Given the fixed training dataset and our goal of learning a model, conventionally, overfitting is defined as instances which during model learning (training stage), the model fits too closely to the training data and adapts to the random variation and noise of training data, failing to adequately learn the semantics of the train-

ing data [1, 28, 34, 38, 41]. As the model lacks semantics of training data, it could be observed that the model performs poorly under unseen data (Fig. 1 ③ ⓐ).

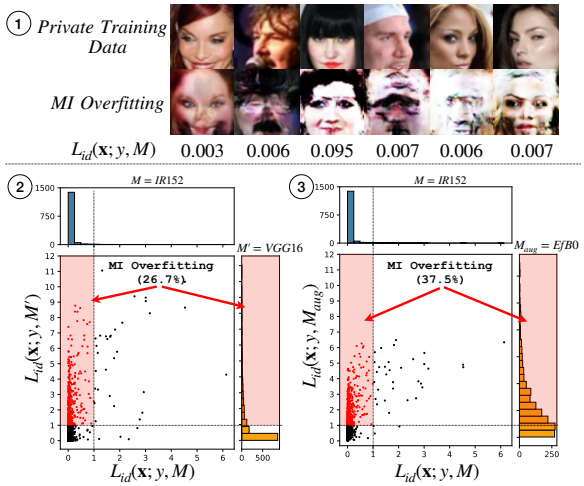

Figure 3. Qualitative / Quantitative studies to demonstrate MI overfitting in SOTA methods. We demonstrate this observation using KEDMI [5]. We use $\mathcal{D}_{priv}$ = CelebA [24], $\mathcal{D}_{pub}$ = CelebA [24] and M = IR152 [5]. ① **We show qualitative results to illustrate MI overfitting.** We show 6 identities, top: private data, bottom: reconstructed data from M. The reconstructed samples have fit too closely to M during inversion resulting in samples with lack of identity semantics. Particularly, we remark that these samples have very low identity loss under the target model M. ② **Quantitative results validating the prevalence of MI overfitting in SOTA MI methods.** We use an additional target classifier M' = VGG16 released by [5, 39] to quantitatively verify the presence of MI overfitting using identity loss. For 1,500 reconstructed samples from M, we visualize their identity loss w.r.t. M and M' in the scatter plot and respective histograms. Particularly, we find that there are 26.7% of samples with low identity loss under the target model M, but large identity loss under unseen VGG16 model M', hinting that these samples might lack identity semantics. This result shows that MI overfitting is a considerable issue even in SOTA MI methods. Note that VGG16 is used here only for analysis and is not part of our solution, as private data is not available. ③ **Model Augmentation to alleviate MI overfitting during inversion.** We repeat the above analysis, with M' = VGG16 replaced by M_{aug} = EfficientNet-B0. Importantly, M_{aug} is trained by *public data using knowledge distillation* [15]. We similarly observe samples with large identity loss under M_{aug}.

Overfitting in MI. We formalize the concept of overfitting in MI (Fig. 1 ③ ⓑ). Given the fixed (target) model and our goal of learning reconstructed samples, we define MI overfitting as instances which during model inversion, the reconstructed samples fit too closely to the target model and adapt to the random variation and noise of the target model parameters, failing to adequately learn semantics of the identity. As these reconstructed samples lack identity semantics, it could be observed that they perform poorly under another unseen model.

Analysis. In what follows, we discuss our analysis to demonstrate MI overfitting and understand its impact in SOTA. See Fig. 3 for analysis setups and results. In particular, in Fig. 3 ①, we show some reconstructed samples which achieve low identity loss under the target model M, yet they lack identity semantics. In Fig. 3 ②, we show that for a considerable percentage of reconstructed samples from target model M with low identity loss under M, their identity loss under another unseen model M' is large as shown in the scatter plot and histograms, hinting that these samples might have suffered from MI overfitting and lack identity semantics. We note that the identity loss under M' is obtained by feeding the reconstructed sample into M' in a forward pass. We also note that SOTA KEDMI [5] is used in this analysis but the issue persists in [36, 39].

Our proposed solution to MI overfitting. We propose a novel solution based on *model augmentation*. Our idea is inspired by the conventional issue of overfitting in model training and data augmentation as a solution to alleviate the issue. In particular, for conventional overfitting, augmenting the training dataset could alleviate the issue [21]. Therefore, we hypothesize that by augmenting the target model we can alleviate MI overfitting.

Specifically, we propose to apply knowledge distillation (KD) [15], with target model M_t as the teacher, to train augmented models $M_{aug}^{(i)}$. Importantly, as we do not have access to the private data, during KD, each $M_{aug}^{(i)}$ is trained on the *public dataset* to match its output to the output of M_t. We select different network architectures for $M_{aug}^{(i)}$ and they are different from M_t (Detailed discussion in the Supplementary). After performing KD, we apply $M_{aug}^{(i)}$ together with the target model M_t in the inversion step and compute the identity loss (with model augmentation):

$$L_{id}^{aug}(\mathbf{x}; y) = \gamma_t \cdot L_{id}(\mathbf{x}; y, M_t) + \gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} L_{id}(\mathbf{x}; y, M_{aug}^{(i)}) \tag{4}$$

Here, γ_t and γ_{aug} are two hyper-parameters. In particular, we use $\gamma_t = \gamma_{aug} = \frac{1}{N_{aug}+1}$, where N_{aug} is the number of augmented models. L_{id}^{aug} in Eqn. 4 is used to replace L_{id} in the inversion step in Eqn. 1. Furthermore, our proposed L_{id}^{logit} in Eqn. 3 can be used in Eqn. 4 to combine the improvements. See details in Supplementary.

In Fig. 3 ③, we analyze the performance of $M_{aug}^{(i)}$. Similar to using the unseen model M', we observe samples with large identity loss under $M_{aug}^{(i)}$, suggesting that samples with MI overfitting perform poorly under $M_{aug}^{(i)}$ as these samples lack identity semantic.

Table 2. We follow the exact the experiment setups in [5] for GMI [39] and KEDMI [5]. For VMI [36], we follow the exact experiment setups in [36]. In total, we conduct 72 experiments spanning 18 setups to demonstrate the effectiveness of our proposed method.

Method	Private Dataset	Public Dataset	Target model	Evaluation Model	Model Augmentation
GMI [39] / KEDMI [5]	CelebA [24]	CelebA / FFHQ [18]	VGG16 [30] / IR152 [14] / face.evoLve [7]	face.evoLve	EfficientNet-B0 [33], EfficientNet-B1 [33], EfficientNet-B2 [33]
	CIFAR-10 [20]	CIFAR-10	VGG16	ResNet-18 [14]	
	MNIST [22]	MNIST	CNN(Conv3)	CNN(Conv5)	CNN(Conv2), CNN(Conv4)
VMI [36]	CelebA	CelebA	ResNet-34 [14]	IR-SE50 [10]	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2
	MNIST	EMNIST [9]	ResNet-10	ResNet-10	CNN(Conv2), CNN(Conv4)

4. Experiments

In this section, we evaluate the performance of the proposed method in recovering a representative input from the target model, against current SOTA methods: GMI [39], VMI [36], and KEDMI [5]. More specifically, as our proposed method identifies two major limitations in current $L_{id}(\mathbf{x}; y)$ –used commonly in all SOTA MI approaches– we will evaluate the improvement brought by our improved identity loss L_{id}^{logit}, and model augmentation L_{id}^{aug} for all SOTA MI approaches.

4.1. Experimental Setup

In order to have a fair comparison, when evaluating our method against each SOTA MI approach, we follow the exactly same experimental setup of that approach. In what follows, we discuss the details of these setups.

Dataset. Following previous works, we evaluate the proposed method on different tasks: face recognition and digit classification is used for comparison with all three SOTA approaches, and image classification is used for comparison with GMI [39], and KEDMI [5]. For the face recognition task, we use CelebA dataset [24] that includes celebrity images, and the FFHQ dataset [18] which contains images with larger variation in terms of background, ethnicity, and age. The MNIST handwritten digits dataset [22] is used for digit classification. We utilize the CIFAR-10 dataset [20] for image classification.

Data Preparation Protocol. Following previous SOTA approaches [5, 36, 39], we split each dataset into two disjoint parts: one part is used as private dataset $\mathcal{D}_{priv}$ for training target model, and another part is used as a public dataset $\mathcal{D}_{pub}$ to extract the prior information. Most importantly, *throughout all experiments, public dataset $\mathcal{D}_{pub}$ has no class intersection with private dataset $\mathcal{D}_{priv}$ used for training target model.* Note that this is essential to make sure that adversary uses $\mathcal{D}_{pub}$ only to gain prior knowledge about features that are general to that task (i.e., face recognition), and does not have access to information about class-specific and private information used for training target model.

Models. Following previous works, we implement several different models with varied complexities. As GMI [39] and KEDMI [5] use exactly similar model architecture in experiments, for comparison with these two algorithms, we use the same models. More specifically, for face recognition on CelebA and FFHQ, we use VGG16 [30], IR152 [14], and face.evoLve [7]–as SOTA face recognition model. For digit classification on MNIST, we use a CNN with 3 convolutional layers and 2 pooling layers. Finally, for image classification, following [5] we use VGG16 [30]. For a fair comparison with VMI, we follow its design in [36] and use ResNet-34 for face recognition CelebA, and ResNet-10 for digit classification on MNIST. The details of the target models, augmented models and datasets used in experiments are summarized in Table 2. We remark that when comparing our proposed method with each of the SOTA MI approaches, we use exactly the same target model and GAN for both SOTA and our approach.

Evaluation Metrics. To evaluate the performance of a MI attack, we need to assess whether the reconstructed image exposes private information about a target label/identity. In this work, following the literature, we conduct both qualitative evaluations by visual inspection, and quantitative evaluations using different metrics, including:

- **Attack Accuracy (Attack Acc).** Following [5, 36, 39], we use an *evaluation model* that predicts the label/identity of the reconstructed image. Similar to previous works, the evaluation model is different from the target model (different structure/ initialization seed), but it is trained on the same private dataset (see Table 2). Intuitively, considering a highly accurate evaluation model, it can be viewed as a proxy for human inspection [39]. Therefore, if the evaluation model infers high accuracy on reconstructed images, it means these images are exposing private information about the private dataset, i.e. high attack accuracy.
- **K-Nearest Neighbors Distance (KNN Dist).** KNN Dist indicates the distance between the reconstructed image for a specific label/id and corresponding images in the private training dataset. More specifically, it measures the

shortest feature distance from the reconstructed image to the real images in the private dataset, given a class/id. It is measured as l_2 distance between two images in the feature space, i.e., the penultimate layer of the evaluation model.

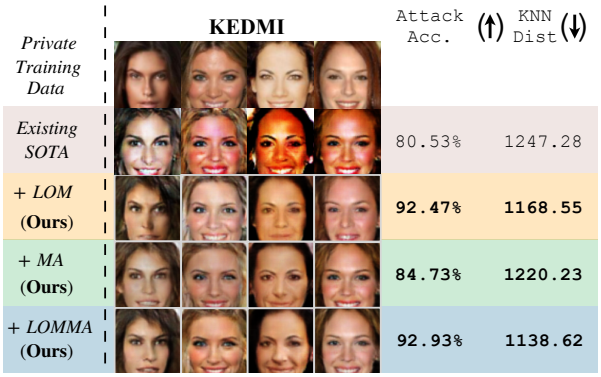

Figure 4. Qualitative / Quantitative (Top1 Attack Acc., KNN Dist) results to demonstrate the efficacy of our proposed method. We use KEDMI [5] (SOTA), $\mathcal{D}_{priv}$ = CelebA [24], $\mathcal{D}_{pub}$ = CelebA [24] and M = IR152 [14]. As one can observe, our proposed method achieves better reconstruction of private data both visually and quantitatively (validated by KNN results) resulting in a significant boost in attack performance.

4.2. Experimental Results

Comparison with previous state-of-the-art. We use GMI [39], KEDMI [5], and VMI [36] as SOTA MI baselines. We reproduce all baseline results using official public implementations. We report results for GMI and KEDMI for CelebA/ CelebA experiments in Table 3. We report VMI results for CelebA/ CelebA experiments in Table 4. For each baseline setup, we report results for 3 variants: • *LOM* (Logit Maximization, Sec. 3.1), • *MA* (Model Augmentation, Sec. 3.2), • *LOMMA* (Logit Maximization + Model Augmentation). The details are as follows:

1. + LOM (Ours): We replace existing identity loss, L_{id} with our improved identity loss L_{id}^{logit} (Sec. 3.1).
2. + MA (Ours): We replace existing identity loss, L_{id} with our proposed L_{id}^{aug} (Sec. 3.2).
3. + LOMMA (Ours): We combine both L_{id}^{logit} (Sec. 3.1) and L_{id}^{aug} (Sec. 3.2) for model inversion.

As one can clearly observe from Table 3 and Table 4, our proposed methods yield significant improvement in MI attack accuracy in *all experiment setups* showing the efficacy of our proposed methods. Further, by combining both our proposed methods, we significantly boost attack accuracy. The KNN results also clearly show that our proposed methods are able to reconstruct data close to the private training data compared to existing SOTA MI algorithms. Particularly, we improve the KEDMI baseline [5] attack accuracy by 12.4% under IR152 target classifier. We show private

Table 3. We report the results for KEDMI and GMI for IR152, face.evoLve and VGG16 target model. Following exact experiment setups in [5], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve. We report top 1 accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance. Top 5 attack accuracies are included in the Supplementary. The best results are in **bold**. By alleviating both these major problems in MI algorithms, we achieve new SOTA MI performance (face.evoLve: 81.40% → **93.20%**).

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/CelebA/IR152			
KEDMI	80.53 ± 3.86	-	1247.28
+ LOM (Ours)	92.47 ± 1.41	11.94	1168.55
+ MA (Ours)	84.73 ± 3.76	4.20	1220.23
+ LOMMA (Ours)	**92.93 ± 1.15**	**12.40**	**1138.62**
GMI	30.60 ± 6.54	-	1609.29
+ LOM (Ours)	78.53 ± 3.41	47.93	1289.62
+ MA (Ours)	61.20 ± 4.34	30.60	1389.99
+ LOMMA (Ours)	**82.40± 4.37**	**51.80**	**1254.32**
CelebA/CelebA/face.evoLve			
KEDMI	81.40 ± 3.25	-	1248.32
+ LOM (Ours)	92.53 ± 1.51	11.13	1183.76
+ MA (Ours)	85.07 ± 2.71	3.67	1222.02
+ LOMMA (Ours)	**93.20 ± 0.85**	**11.80**	**1154.32**
GMI	27.07 ± 6.72	-	1635.87
+ LOM (Ours)	61.67 ± 4.92	34.60	1405.35
+ MA (Ours)	74.13 ± 4.32	47.06	1352.25
+ LOMMA (Ours)	**82.33 ± 3.51**	**55.26**	**1257.50**
CelebA/CelebA/VGG16			
KEDMI	74.00 ± 3.10	-	1289.88
+ LOM (Ours)	89.07 ± 1.46	15.07	1218.46
+ MA (Ours)	82.00 ± 3.85	8.00	1248.33
+ LOMMA (Ours)	**90.27 ± 1.36**	**16.27**	**1147.41**
GMI	19.07 ± 4.47	-	1715.60
+ LOM (Ours)	69.67 ± 4.80	50.60	1363.81
+ MA (Ours)	51.73 ± 6.03	32.66	1467.68
+ LOMMA (Ours)	**77.60 ± 4.64**	**58.53**	**1296.26**

training data and reconstructed samples for KEDMI [5] under IR152 target model including all 3 variants in Fig. 4. We remark that in the standard CelebA benchmark, our method boosts attack accuracy significantly thereby achieving more than 90% attack accuracy (Table 3) for the first time in contemporary MI literature. We also include CIFAR-10, MNIST and additional results in Supplementary.

Cross-dataset. Following [5], we conduct a series of experiments to study the effect of distribution shift between public and private data on attack performance and KNN distance. We use FFHQ [18] as the public dataset. In particular, we use FFHQ as public data for CelebA experiments. We train GAN models and three model augmentations using the public data. We remark that such setups closely replicate real-world MI attack scenario. We report top 1 accuracy and KNN distance for IR152, face.evoLve, and VGG16 target classifiers in Table 6. It is well known that baseline

Table 4. We follow exact the experiment setup of [36] for VMI experiments. Specifically, we use StyleGAN [17] and Flow model [19] to learn the distribution of z. The best results are in **bold**. Following exact experiment setups in [36], here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, target model = ResNet-34, evaluation model = IR-SE50. We report top 1 attack accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance (KNN Dist). The top 5 attack accuracies are included in the Supplementary. The best results are in **bold**. By alleviating both these major problems in MI algorithms, we improve the attack accuracy by 14.94% (59.96% $\rightarrow$ **74.90%**).

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/CelebA/ResNet-34			
VMI	59.96 ± 0.27	-	1.144
+ LOM (Ours)	68.34 ± 0.36	8.38	1.131
+ MA (Ours)	64.16 ± 0.27	4.20	1.140
+ LOMMA (Ours)	**74.90** ± 0.34	**14.94**	**1.109**

Table 5. Results for SOTA defense model BiDO-HSIC [27]: Following exact experiment setups in BiDO-HSIC, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA, evaluation model = face.evoLve, target model = BiDO-HSIC. We report top 1 attack accuracies (Attack Acc.), and KNN distance (KNN Dist).

Method	GMI		KEDMI	
	Attack Acc ↑	KNN Dist ↓	Attack Acc ↑	KNN Dist ↓
No Def.	19.07 ± 4.47	1715.60	74.00 ± 3.10	1289.88
Def. Model	5.20 ± 2.75	1962.58	42.80 ± 5.02	1469.75
+ LOM (Ours)	55.80 ± 3.64	1397.05	64.33 ± 1.82	1360.57
+ MA (Ours)	23.93 ± 5.50	1634.84	49.27 ± 4.02	1413.81
+ LOMMA (Ours)	**62.13** ± 4.04	**1358.54**	**70.47** ± 2.36	**1293.25**

attack performances will degrade due to distribution shift between public and private data [5]. But we remark that our proposed methods consistently improves the baseline SOTA attack performances. *i.e*. Our method boosts the attack accuracy of IR152 target model from 52.87% $\rightarrow$ 77.27%.

MI under SOTA defense models. We further evaluate our method on SOTA MI defense models provided by BiDO-HSIC [27]. Specifically, we use the exact GAN and defense models provided by BiDO-HSIC which are trained on CelebA dataset. We then transfer knowledge from the defense model to M_{aug} = {Efficientnet-B0, Efficientnet-B1, Efficientnet-B2} using D_{pub}. Results using GMI and KEDMI are shown in Table 5. We observe that SOTA defense BiDO-HSIC is rather ineffective for our proposed MI.

5. Discussion

Conclusion. We revisit SOTA MI and study two issues pertaining to all SOTA MI approaches. First, we analyze existing identity loss in SOTA and argue that it is sub-optimal for MI. We propose a new logit based identity loss that aligns better with the goal of MI. Second, we formalize the concept of MI overfitting and show that it has a considerable impact even in SOTA. Inspired by conventional data augmentation, we propose model augmentation to alleviate MI overfitting. Extensive experiments demonstrate that our solutions can improve SOTA significantly, achieving for the first time over 90% attack accuracy under the standard benchmark. Our findings highlight rising threats based on MI and prompt serious consideration on privacy of machine learning.

Table 6. We report the results for KEDMI and GMI for IR152, face.evoLve and VGG16 target model. Here $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ, evaluation model = face.evoLve. We report top 1 accuracies, the improvement compared to the SOTA MI (Imp.), and KNN distance. Top 5 attack accuracies are included in the Supplementary. The best results are in **bold**. By alleviating both these major problems in MI algorithms, we improve the attack accuracy 24.40% (IR152: 52.87% $\rightarrow$ **77.27%**).

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/FFHQ/IR152			
KEDMI	52.87 ± 4.96	-	1418.83
+ LOM (Ours)	67.73 ± 2.29	14.86	1325.28
+ MA (Ours)	64.13 ± 4.49	11.26	1373.42
+ LOMMA (Ours)	**77.27** ± **2.01**	**24.40**	**1292.80**
GMI	17.20 ± 5.31	-	1701.76
+ LOM (Ours)	56.00 ± 5.20	38.80	1427.59
+ MA (Ours)	50.80 ± 6.89	33.60	1462.92
+ LOMMA (Ours)	**72.00** ± **6.62**	**54.80**	**1338.35**
CelebA/FFHQ/face.evoLve			
KEDMI	51.87 ± 3.88	-	1440.19
+ LOM (Ours)	69.73 ± 2.47	17.86	1379.73
+ MA (Ours)	65.73 ± 3.51	13.86	1379.09
+ LOMMA (Ours)	**73.20** ± **2.24**	**21.33**	**1321.00**
GMI	14.27 ± 4.42	-	1744.47
+ LOM (Ours)	47.93 ± 4.87	33.66	1498.19
+ MA (Ours)	46.07 ± 4.88	31.80	1500.10
+ LOMMA (Ours)	**64.33** ± **4.69**	**50.06**	**1386.33**
CelebA/FFHQ/VGG16			
KEDMI	41.27 ± 3.50	-	1490.09
+ LOM (Ours)	55.07 ± 1.88	13.80	1438.72
+ MA (Ours)	52.07 ± 2.92	10.80	1428.77
+ LOMMA (Ours)	**62.67** ± **2.29**	**21.40**	**1366.94**
GMI	10.93 ± 3.47	-	1766.27
+ LOM (Ours)	44.40 ± 5.96	33.47	1508.84
+ MA (Ours)	34.93 ± 4.52	24.00	1547.93
+ LOMMA (Ours)	**58.73** ± **6.18**	**47.80**	**1415.06**

Limitations and Ethical Concerns. We follow previous work in experimental setups. The scale of our experiments is comparable to previous works. Furthermore, extension of our methods for blackbox/ label-only attacks can be considered in future. While our improved MI methods could have negative societal impacts if it is used by malicious users, our work contributes to increased awareness about privacy attacks on DNNs.

Acknowledgements. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-RP-2021-021; AISG Award No.: AISG2-TC-2022-007). This project is also supported by SUTD project PIE-SGP-AI-2018-01. We thank reviewers for their valuable comments. We also thank Loo Yi and Kelly Kuo for helpful discussion.

References

[1] Milad Abdollahzadeh, Touba Malekzadeh, and Ngai-Man Cheung. Revisit multimodal meta-learning through the lens of multi-task learning. *Advances in Neural Information Processing Systems*, 34:14632–14644, 2021.

[2] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[3] Hervé Chabanne, Amaury De Wargny, Jonathan Milgram, Constance Morel, and Emmanuel Prouff. Privacy-preserving classification on deep neural network. *Cryptology ePrint Archive*, 2017.

[4] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2890–2916. PMLR, 17-23 Jul 2022.

[5] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16178–16187, 2021.

[6] 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 *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1924–1932, 2017.

[7] 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.

[8] 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.

[9] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

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

[11] 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.

[12] 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.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[14] 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.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[16] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15045–15053, 2022.

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[18] 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.

[19] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[20] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

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

[23] Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, et al. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *IEEE Access*, 10:30039–30054, 2022.

[24] 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.

[25] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458. PMLR, 13–18 Jul 2020.

[26] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[27] Xiong Peng, Feng Liu, Jingfen Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. *KDD*, 2022.

[28] Claudio Filipi Gonçalves Dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR)*, 54(10s):1–25, 2022.

[29] Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. In *International Conference on Learning Representations*, 2021.

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

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

[32] Nagesh Subbanna, Matthias Wilms, Anup Tuladhar, and Nils D Forkert. An analysis of the vulnerability of two common deep learning-based medical image segmentation techniques to model inversion attacks. *Sensors*, 21(11):3874, 2021.

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[34] Christopher TH Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. Fair generative models via transfer learning. *arXiv preprint arXiv:2212.00926*, 2022.

[35] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021.

[36] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34:9706–9719, 2021.

[37] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

[38] Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pages 25595–25610. PMLR, 2022.

[39] 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.

[40] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–692, 2021.

[41] Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. In *Advances in Neural Information Processing Systems*, 2022.