Computers & Security 144 (2024) 103967

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Targeted Model Inversion: Distilling style encoded in predictions

Hoyong Jeong [a,b], Kiwon Chung [a], Sung Ju Hwang [a], Sooel Son [a,*]

[a] School of Computing, KAIST, Daejeon, South Korea
[b] Deeping Source Inc., Seoul, South Korea

ARTICLE INFO

Keywords:
Model inversion attack
Machine learning
Privacy

ABSTRACT

Previous model inversion (MI) research has demonstrated the feasibility of reconstructing images representative of specific classes, inadvertently revealing additional feature information. However, there are two remaining challenges for practical black-box MI: (1) minimizing the number of queries to the target model, and (2) reconstructing a high-quality input image tailored to an observed prediction vector. We introduce **T**argeted **M**odel **I**nversion (TMI), a practical black-box MI attack. Our approach involves altering the mapping network in StyleGAN, which projects an observed prediction vector into a StyleGAN latent representation. Later, TMI leverages a surrogate model that is also derived from StyleGAN to guide instance-specific MI by optimizing the latent representation. These mapping and surrogate networks work together to conduct high-fidelity MI while significantly decreasing the number of necessary queries. Our experiments demonstrate that TMI outperforms state-of-the-art MI methods, demonstrating a new upper bound on the susceptibility to black-box MI attacks.

1. Introduction

Model inversion (MI) refers to an adversarial attack that reconstructs training data or class-representative instances based on the output from a target machine learning (ML) model. Assuming an adversary who is able to eavesdrop or obtain an output prediction from a target model, successful MI attacks either reconstruct an input image corresponding to that specific output or generate a representative image of the predicted class. Consequently, these reconstructed images expose privacy-sensitive features that the model owners or its users did not anticipate revealing through the output predictions.

Prior studies have vastly investigated the feasibility and efficacy of MI against deep neural networks (DNNs) (Fredrikson et al., 2015; He et al., 2019). Recently, Yang et al. (2019) proposed a training-based attack that utilizes a DNN-based inversion model, enabling it to reconstruct an image based on a given prediction vector. Subsequent works focused on improving the fidelity of reconstructed images by adopting generative adversarial networks (GANs) (Zhang et al., 2020b; Chen et al., 2021; Kahla et al., 2022; Yuan et al., 2023; Han et al., 2023) or StyleGANs (Wang et al., 2021; An et al., 2022; Struppek et al., 2022).

We posit that there still remains large room for improvement in conducting practical black-box MI attacks. Specifically, we propose two key challenges to overcome: (1) minimizing the number of necessary queries to a target model and (2) enabling instance-specific reconstruction. Numerous studies have assumed a strong white-box adversary who is able to access target model parameters, thereby leveraging gradients in performing MI (Fredrikson et al., 2015; Zhang et al., 2020b; Wang et al., 2021; An et al., 2022; Struppek et al., 2022). Moreover, existing black-box MI attacks (Yang et al., 2019; Kahla et al., 2022; Han et al., 2023) require an excessive number of queries to a target model, rendering them impractical. For example, attack by An et al. (2022) required 160k queries to reconstruct a single image. Furthermore, previous researchers have focused on reconstructing class-representative images rather than the original input images specific to the corresponding prediction vector. Class-representative images often omit intra-class differences within their class, which undermines the chances of reconstructing privacy-sensitive features. For instance, when a target task for MI is gender classification, class-representative images display a generic female face, not a specific woman involved in training (Melis et al., 2019). The difficulty is even exacerbated when a target task involves large variances in each class. For example. we observed that, with the NIH Chest X-ray dataset (Wang et al., 2017), previous methods are unable to reconstruct task-agnostic features such as gender or age.

To tackle the aforementioned challenges, we introduce Targeted Model Inversion (TMI), a novel MI framework that performs instance-specific reconstruction while leveraging only a restricted set of black-box queries to a target model. TMI consists of two steps: preparation and inversion. In the preparation step, TMI employs a StyleGAN (Karras et al., 2020) network trained on a dataset in which the underlying

* Corresponding author.
E-mail addresses: hoyong.jeong@deepingsource.io (H. Jeong), greenare@kaist.ac.kr (K. Chung), sjhwang@kaist.ac.kr (S.J. Hwang), sl.son@kaist.ac.kr (S. Son).

https://doi.org/10.1016/j.cose.2024.103967
Received 17 March 2024; Received in revised form 11 June 2024; Accepted 20 June 2024
Available online 25 June 2024

distribution is similar, yet different from the training set of the target model. Then, its mapping and discriminator networks are modified to project the prediction vector to the StyleGAN latent space and to act as a surrogate model, respectively. These networks are trained using StyleGAN-generated images and their corresponding predictions from the target model, eliminating the need for an additional dataset. In the inversion step, TMI locates the initial StyleGAN latent corresponding to a target prediction vector observed from the target model by using the modified mapping network. It then optimizes this latent to generate an image that prompts the surrogate model to emit a prediction vector similar to the target prediction vector.

The core idea of TMI is to construct a new mapping network that approximates a style latent corresponding to the target prediction vector, which is then further optimized through signals from the surrogate model. Both distilled from a benign StyleGAN network, these two modified components help significantly decrease the number of required queries for successful MI while conducting instance-specific inversion for a given prediction vector.

We evaluate TMI by comparing it with other state-of-the-art white-box and black-box MI attacks. We demonstrate the superiority of TMI even with a much smaller query budget; Against a facial recognition model, TMI achieves 43.5% higher coverage compared to the next best-performing white-box MI attack and shows even greater improvements when compared to the existing black-box MI methods. These experiment results demonstrate that the TMI attack enables practical black-box MI with high fidelity without requiring white-box access to a target model.

2. Background

2.1. Model inversion attack

Prior research has extensively explored the privacy implication that MI attacks pose, which contributes to leaking sensitive information across diverse domains such as healthcare, language modeling, and speech data (Dibbo, 2023). Fredrikson et al. (2014) introduced the first MI attack that enable a target ML model trained on medical data to inadvertently disclose private patient information. Moreover, recent studies Huang et al. (2022) and Zhang et al. (2022) have demonstrated the susceptibility of neural language models to MI attacks, thus leaking sensitive training data, including precise text inputs and personally identifiable information like email addresses and phone numbers through unintended memorization. Similarly, in the domain of speech recognition, MI attacks have proven effective in reconstructing spoken phrases from ML model outputs. A prior study by Pizzi et al. (2023) highlights the vulnerability of speech recognition systems, wherein the adversary is able to recover audio samples and voice features directly linked to the speaker's biometrics.

Formally, MI attack refers to an adversarial attempt to reconstruct an input image $x \in \mathcal{X}$ based on the target output prediction $\hat{y}_t \in \mathcal{Y}$ obtained from a target classifier $f : \mathcal{X} \mapsto \mathcal{Y}$. The reconstructed image x' may inadvertently leak privacy-sensitive features that were never expected by the model owner or its users. Formally, the adversary's objective is to derive an inversion image x' satisfying the following equation:

$$x' = \arg\min_{x \in \mathcal{X}} \mathcal{L}_{pred}(f(x), \hat{y}_t) \tag{1}$$

with a loss function $\mathcal{L}_{pred}$ (e.g., cross-entropy loss or ℓ_2 loss) that quantifies the dissimilarity between the observed prediction vector $\hat{y}_t$ and the target model output $f(x)$.

To overcome the challenge of reconstructing high-fidelity images in an input space (i.e., $\mathbb{R}^{3\times 224^2}$) based on a prediction vector in a limited model output space (i.e., $\mathbb{R}^{\mathcal{K}}$, where $\mathcal{K}$ refers to the number of classes in f), previous researches have explored different attack methods. Early MI studies focused on reconstructing low-resolution grayscale facial images or simple datasets like MNIST. For instance, Fredrikson et al. (2015) applied an analytic method of finding x' in Eq. (1). Later, Yang et al. (2019) proposed using a dedicated neural network consisting of multiple transposed convolution layers that directly map the observed prediction vectors onto the input space, expanding the attack vector of MI to relatively complicated neural networks. However, their attack was still limited to reconstructing low-resolution grayscale input images.

To further improve MI, subsequent studies have proposed leveraging the GAN generators (Zhang et al., 2020b; Chen et al., 2021; Wang et al., 2021; An et al., 2022; Struppek et al., 2022; Kahla et al., 2022; Yuan et al., 2023; Han et al., 2023). The generator $g : \mathcal{Z} \mapsto \mathcal{X}$ operates as an image prior, generating input images in $\mathcal{X}$ from Gaussian latent vectors in $\mathcal{Z}$. Instead of directly optimizing in the input image space $\mathcal{X}$ as in Eq. (1), GAN-based approaches perform optimization within a more constrained space $\mathcal{Z}$. Recent MI researchers have adopted StyleGANs (Karras et al., 2019, 2020) to attain higher-fidelity reconstruction (An et al., 2022; Struppek et al., 2022); they perform optimization in a newly introduced intermediate latent space $\mathcal{W}$.

Note that the optimization process in MI typically requires computing gradients using the target model, thereby assuming the presence of a white-box adversary who is able to access the target model parameters. Follow-up studies have proposed attack methods to simulate the optimization using only black-box queries. These attack techniques include genetic algorithms (An et al., 2022), decision boundary estimation (Kahla et al., 2022), and reinforcement learning (Han et al., 2023) as proxies for the optimization process. Although achieving state-of-the-art performance compared to traditional black-box approaches, we argue that all the existing methods still fail to address two following challenges: practicality and instance-specific inversion.

Practicality. Existing black-box MI methods still demand a prohibitively large number of queries to the target model. This poses practical challenges, particularly when considering the limitations imposed by Machine Learning as a Service (MLaaS) providers. These providers often enforce rate limits on API calls, restricting the number of queries (e.g., Clarifai[1] - 5000/day, DatumBox[2] - 1000/day). The requirement for a high volume of queries not only tampers with the practicality of MI but also raises the risk of attack detection because an abnormal number of queries can potentially be flagged.

Instance-specific inversion. Test-time MI attacks can be categorized into two groups (Yang et al., 2019): instance-specific MI and class-representative MI. The former refers to a scenario in which the attacker infers a victim's input instance for an observed prediction output. On the other hand, class-representative MI focuses on reconstructing generic images for a single output class in a target model. Whereas a larger volume of previous research focused on conducting class-generic MI that reveals class-bound features (Fredrikson et al., 2015; Yang et al., 2019; Zhang et al., 2020b; Chen et al., 2021; Wang et al., 2021; An et al., 2022; Struppek et al., 2022; Kahla et al., 2022; Yuan et al., 2023; Han et al., 2023), instance-specific MI has been largely understudied. Due to the difficulty of instance-specific MI that requires the reconstruction of subtle and instance-specific image features, it was deemed possible under specific conditions, such as the collaborative inference setting, where intermediate representations and gradients are accessible to the adversary (Melis et al., 2019; He et al., 2019, 2021). For instance, class-generic MI reveals only class-bound features, including race, gender, and age, in facial recognition tasks. On the other hand, the instance-specific MI seek reconstruction of additional instance-specific features, such as accessories, facial expressions, or posture, as well as the class-bound features.

[1] https://clarifai.com
[2] https://www.datumbox.com

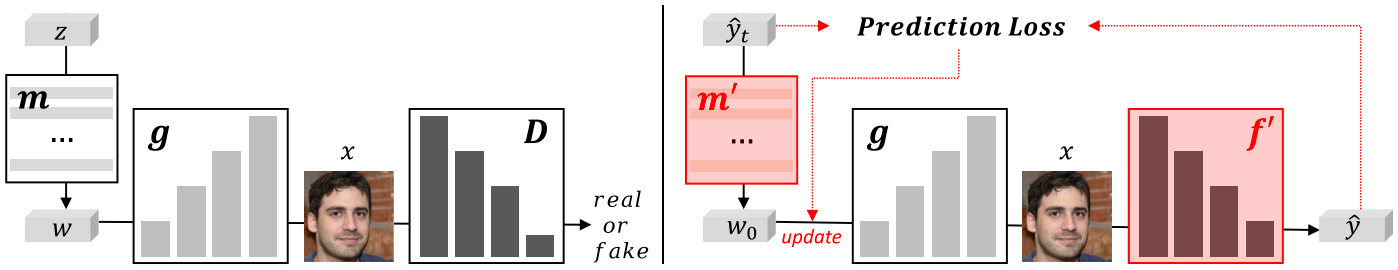

Fig. 1. An overview of the TMI attack workflow (right), compared with the original StyleGAN network (left). The StyleGAN components modified for TMI (m' and f') are highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Attribute inference attack

An attribute inference attack refers to an attempt to infer a private attribute of a target individual by leveraging benign outputs from a target system along with other known attributes of that individual. For example, such an attack may deduce a user's location, gender, or political view based on their public information or past behaviors. This attack involves exploiting the statistical correlations between a user's publicly available data points and their private attributes.

This emerging threat spans various application domains, including social media (Abdelberi et al., 2012; Jia et al., 2017), recommender systems (Otterbacher, 2010; Weinsberg et al., 2012), and mobile platforms (Michalevsky et al., 2015; Narain et al., 2016). In particular, on social media platforms, an attacker can examine a user's page likes to deduce personal information such as gender and political views with notable accuracy. Similarly, against a recommendation system, an attacker can utilize publicly accessible rating scores of items like movies or apps to infer the user's gender or age. The efficacy of these attribute inference attacks highlights the substantial privacy risks associated with the public availability of user data.

Attribute inference attacks and MI attacks are closely related, sharing the objective of extracting sensitive information from the outputs of a targeted ML system. Initially, early studies on MI (Fredrikson et al., 2014, 2015) demonstrated limited reconstruction capabilities, often restricted to inferring specific attributes rather than fully reconstructing the input data or its representative image class. However, follow-up MI studies have strengthened their capabilities, enabling them to reconstruct images with high fidelity (An et al., 2022; Struppek et al., 2022).

3. Threat model

We assume a target classifier f, providing black-box access where the adversary is able to query an input image x to obtain $\hat{y}$, where $\hat{y}$ is the corresponding output prediction in the form of a confidence vector. The designed goal of TMI is to reconstruct the specific input x that produced the target prediction $\hat{y}_t$. In general, one does not anticipate an output prediction to convey subtle details of the corresponding input, so they are regarded less confidential compared to the input data itself. This trend is evident in regulations like HIPAA, where the guidelines for storage and transmission of medical images are more strict than the rules regarding diagnostic predictions (Moore and Frye, 2019). Furthermore, in the field of confidential computing that protects the privacy of user input to a cloud-provided ML service, the prediction outputs are excluded from encryption, allowing direct access from cloud providers with malicious intents (Gu et al., 2018; Narra et al., 2019). Accordingly, $\hat{y}_t$ is often leaked, eavesdropped, forged, or carelessly exposed to cloud service providers or man-in-the-middle adversaries in real-world scenarios. These exemplary scenarios also include users posting their prediction results on social media (Yang et al., 2019) (e.g., celebrity look-alike apps that show the look-alike percentage to celebrities given facial images as input), split inference settings where the inference result is sent to different parties (He et al., 2019), or medical professionals sharing diagnosis predictions for educational or consultative purposes.

The adversary leverages an auxiliary dataset $\mathcal{D}_{aux}$ of which the underlying distribution is similar to those of the original dataset $\mathcal{D}$ upon which f is trained. It then uses $\mathcal{D}_{aux}$ to train a StyleGAN network. Alternatively, the adversary can leverage a pretrained StyleGAN network available on the Internet, which removes the need for $\mathcal{D}_{aux}$. We also evaluate TMI on using $\mathcal{D}_{aux}$ with a significant deviation from the input distribution of f in our ablation study. Lastly, the adversary has black-box access; it cannot access the model parameters, gradients, or intermediate results while performing MI. It is only permitted to send a limited number of benign input queries to f and use their output predictions. We emphasize that the advversary is bound to a predefined query budget.

We note that the adversary is even capable of populating an arbitrary prediction vector $\hat{y}$ or using only labels for conducting TMI. Under this scenario, it can apply label smoothing (Müller et al., 2019) to hard-coded prediction outputs, each of which represents a corresponding class, then conduct the TMI attack.

3.1. Adversarial capabilities

The adversary is capable of leveraging an auxiliary dataset $\mathcal{D}_{aux}$ of which the underlying distribution is similar to those of the original dataset $\mathcal{D}$ upon which f is trained. The adversary uses $\mathcal{D}_{aux}$ to train their own StyleGAN network. When the adversary leverages pretrained StyleGAN networks available on the Internet, it removes the need for $\mathcal{D}_{aux}$.

Unlike previous MI studies assuming white-box access (Fredrikson et al., 2015; Zhang et al., 2020b; Chen et al., 2021; Wang et al., 2021; An et al., 2022; Struppek et al., 2022; Yuan et al., 2023), we assume a black-box adversary who is unable to access the model parameters, gradients, and any intermediate results while performing MI. The adversary is only permitted to send a limited number of input queries to f and obtain their corresponding prediction vectors. Lastly, we assume an adversary is able to obtain or forge a prediction vector $\hat{y}_t$ for targeted MI under the scenarios aforementioned.

4. Design

The TMI attack consists of two distinct phases: *preparation* and *inversion*. Given query access to a target model f, the adversary leverages a pretrained StyleGAN network and alters its two components, the mapping network m and the discriminator D, during the preparation phase. Once preparation is complete, any observed prediction output can be fed into the modified network to perform a offline MI attack to reconstruct its corresponding input image. The original and modified StyleGANs are illustrated in Fig. 1.

Algorithm 1 TMI Attack

1: **procedure** ATTACK(target $\hat{y}_t$, iteration n, step size η, exploration e, mix probability δ)
2: $w_0 \leftarrow m'(\hat{y}_t)$
3: $\ell_{best} \leftarrow \infty$
4: **for** $i \in \{0, \ldots, n\}$ **do**
5: $x_i \leftarrow g(w_i)$
6: $y'_i \leftarrow f'(x_i)$
7: $\ell_i \leftarrow Distance(y'_i, \hat{y}_t)$ ▷ ℓ_2 Loss
8: **if** $\ell_i < \ell_{best}$ **then**
9: $x_{best}, \ell_{best} \leftarrow x_i, \ell_i$
10: **end if**
11: $w_{i+1} \leftarrow w_i - \eta \nabla_w \ell_i$
12: **if** i is multiple of e **then**
13: $w_{i+1} \leftarrow RandomMix(w_{i+1}, \delta)$
14: **end if**
15: **end for**
16: **return** x_{best}
17: **end procedure**

4.1. Preparation phase: Tweaking StyleGAN

StyleGAN. In a StyleGAN network, the generator is composed of two parts: a mapping network $m : \mathcal{Z} \mapsto \mathcal{W}$ and a synthesis network $g : \mathcal{W} \mapsto \mathcal{X}$. Unlike traditional GAN generators that take a random Gaussian vector z from the latent space $\mathcal{Z}$ and pass it through the synthesis network, m first projects z to an intermediate latent space $w \in \mathcal{W}$. The synthesis network g, consisting of consecutive style blocks, takes w as the input for each style block. The style blocks synthesize images in a progressive manner, starting from low-resolution images and progressively refining them to higher resolutions. The discriminator D in the StyleGAN network receives the final result from g and determines whether this image is a real image or a synthesized sample. The minimax game between D and g gradually makes D to better distinguish fake samples.

An important characteristic of the intermediate latent space $\mathcal{W}$ in StyleGAN is that image features are disentangled, meaning that different features are represented by separate dimensions in $\mathcal{W}$. This characteristic is encouraged during the StyleGAN training phase because (1) generating realistic images is easier when the representation is disentangled, which allows independent control over different image attributes, and (2) the separation of style blocks causes different subsets of $\mathcal{W}$ to contribute to different levels of styles, enabling fine-grained control over the generated images. This characteristic of disentangled image features in $\mathcal{W}$ has been found to be beneficial not only for style mixing (Karras et al., 2019) or editing (Abdal et al., 2019), but also as an effective basis for MI attacks. By manipulating specific dimensions in $\mathcal{W}$, an adversary is able to exert control over certain features of the generated image, enabling the reconstruction of input images corresponding to specific prediction vectors.

Tailoring StyleGAN. In TMI, the adversary leverages a publicly available StyleGAN network, or trains their own using D_{aux}. In the evaluation section, we consider scenarios, including the use of pretrained StyleGAN networks, to assess the performance and effectiveness of the TMI attack.

The original m is trained to convert a Gaussian vector z into w. Therefore, the adversary trains a new mapping network $m' : \hat{\mathcal{Y}} \mapsto \mathcal{W}$ to emit w directly from an observed prediction vector $\hat{y}$. For this, the adversary exploits the StyleGAN network to generate triplets, each of which consists of an image x generated via the generator using a random vector z, the intermediate latent w used to generate x, and a prediction vector $\hat{y}$ obtained from f by querying x. That is, the adversary generates $D_{gen} = \{(w, x, \hat{y}) \mid z \in \mathcal{Z},\ w = m(z),\ x = g(w),\ \hat{y} = f(x)\}$. After the dataset is complete, m' is optimized until convergence:

$$\arg\min_{\theta} \mathbb{E}_{(w,x,\hat{y}) \in D_{gen}} \left[\left(w - m'_\theta(\hat{y}) \right)^2 \right]. \tag{2}$$

Although this training procedure does not require white-box access to f, it still requires sending a number of queries (i.e., $|D_{gen}|$) to f during D_{gen} construction. We set this number to be 100k throughout our main evaluation, which is significantly smaller than the number of required queries in prior works summarized in Table 1.

The new mapping network m' plays a key role in locating an initial latent point for each observed prediction $\hat{y}$. We expect m' to learn a way of distilling a style given a prediction vector during its training procedure, which the adversary exploits in the later inversion phase (Section 4.2). We exemplify the efficacy of m' in selecting a reliable initial latent point w_0 with high fidelity and the superiority of this approach compared to the prior MI methods.

Lastly, the adversary conducts transfer learning, to make a surrogate model $f' : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ from the original discriminator D in the StyleGAN network with its last layer changed to match the number of classes in f. We update f' by optimizing the following loss:

$$\arg\min_{\theta} \mathbb{E}_{(w,x,\hat{y}) \in D_{gen}} \left[(\hat{y} - f'_\theta(x))^2 \right]. \tag{3}$$

The goal of f' is to emit a prediction vector similar to the one produced by f for each $x \in D_{gen}$. This process does not send additional queries to f as it only leverages D_{gen} which is already obtained from the previous step of training m'.

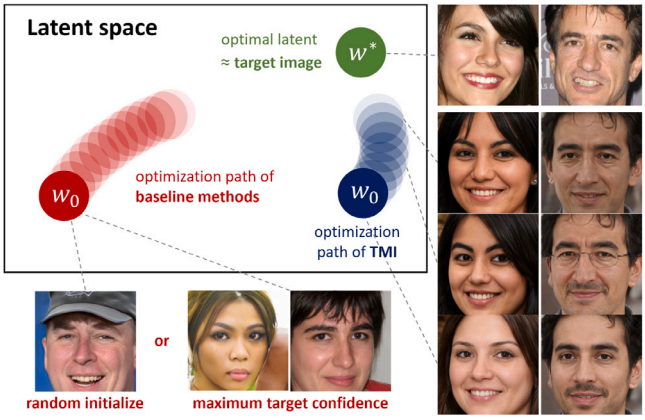

Fig. 2. Illustration of latent space exploration in TMI and baseline methods.

4.2. Inversion phase: Reconstructing input images

Once the preparation phase is complete, the adversary can launch the inversion phase on any observed target prediction $\hat{y}_t$ to reconstruct its input image. In TMI, white-box optimization using f's gradients is replaced with repetitive approximated optimization, starting from w_0 derived from the renewed mapping network m':

$$w_{i+1} := w_i - \eta \nabla_w \left[f'\left(g(w_i) \right) - \hat{y}_t \right]^2. \tag{4}$$

Algorithm 1 describes the overall process of the inversion phase. The adversary starts by obtaining an initial latent representation $w_0 = m'(\hat{y}_t)$ using m' obtained from the previous phase (Line 2). w_0 is fed into g to generate an image x (Line 5). In Lines 6–7, this synthesized image is fed into f' to produce a prediction result y', and it then computes the distance between y' and $\hat{y}$. In Line 11, w is optimized via gradient descent so that the generated image produces an approximated prediction y' that is closer to $\hat{y}$.

Table 1
Comparison of required queries. Number under `Attack` are the required queries for each attack attempt, whereas numbers under `Prep` are required once in the preparation phase.

Method	Query count		Image prior
	Prep	Attack	
AMI	$\lvert D_{aux}\rvert$	0	None
TMI (ours)	$\lvert D_{gen}\rvert$[a]	0	Style-GAN
MIRROR-w	100k	160k	
P&P	5k	34k	
MIRROR-b	100k	10k	
RLB-MI	0	80k	GAN
LO-MI	100k[b]	25k	

[a] We use 100k as default for main evaluation.

[b] LO-MI does not have an explicit query limit on `Prep`, however we observed it to have the highest number of queries in practice. Hence, we regard it as the upper-bound among baselines.

For every e steps, TMI performs $RandomMix$, where subsets of w are reset to w_0 with probability δ. This is to avoid w from overfitting only to a specific style, which leads to unnatural images. $RandomMix$ prevents the optimization routine from getting trapped in a local minimum and allows it to explore different styles and combinations. In addition, to bound the reconstructed images to the natural image domain, we applied a clipping technique after $RandomMix$. Formally, we computed the dimension-wise mean μ and deviation σ of $\mathcal{W}$ from $\mathcal{D}_{gen}$, then:

$$w^{(i)} = \max(\min(w^{(i)},\ \mu^{(i)} + \sigma^{(i)}),\ \mu^{(i)} - \sigma^{(i)}), \tag{5}$$

where $\cdot^{(i)}$ denotes the ith dimension.[3] Finally, the algorithm returns x that recorded the closest y' to $\hat{y}_t$.

Note that in TMI, the adversary exploits f' in an offline manner to refine the initial latent vector w_0, which is also obtained by an offline single-pass to m'. Therefore, TMI does not generate any queries or require white-box access to f throughout the inversion phase. This makes the attack completely passive once the preparation phase is complete.

4.3. Summary and differences to prior MI attacks

Previous studies have focused on reconstructing class-generic images, overlooking the reconstruction of instance-specific features. This trend comes from the fact that GAN- and StyleGAN-based MI performs optimization from either a random initial latent (Kahla et al., 2022; Han et al., 2023) or the latent that yields the maximum target confidence (An et al., 2022; Struppek et al., 2022; Yuan et al., 2023). Such initial points tend to be biased toward singularities and become far from the optimal point as shown in Fig. 2. Prior works overcome this issue by performing a large number of optimization steps, which naturally demand an excessive number of input queries, thus undermining the practicality of MI. Table 1 shows the number of queries for each MI approach. Furthermore, using a large number of optimization steps frequently leads to convergence to local optima due to the nature of greedy updates. In contrast, we employ the customized mapping network m' to directly project the predictions to the latent space. By pinpointing a reliable starting point via m', TMI bypasses most of the early optimizations present in existing MI attack methods.

The choice of constructing f' also brings benefits compared with a naive black-box migration of white-box methods, which would be to train a surrogate model from scratch as a substitute for f in the optimization routine. Transfer learning from D offers a more reliable and generalized surrogate model due to the fact that D has already been exposed to numerous styles of images during the training of StyleGAN. We show that a surrogate model trained from scratch is insufficient to imitate the optimization path of the target model f, in our ablation study.

[3] Removing the clipping logic results in similar metric scores, however, the reconstructed images often appear visually unnatural.

Table 2
Inversion performance of TMI and SOTA methods in three different application domains (facial recognition, chest X-ray diagnosis, and car classification). The top five MI methods are black-box attacks, and the remaining ones are white-box attacks. The direction of the arrow following each metric name indicates the direction of better reconstruction performance. The best attack performances among black-box attacks are marked in bold.

(a) Facial recognition

Method	Acc@1 ↑	Acc@5 ↑	F-dist ↓	Cover ↑	MS-SSIM ↑
TMI	**.3408**$_{\pm.0061}$	**.6255**$_{\pm.0092}$	**.2950**$_{\pm.0009}$	**.2067**$_{\pm.0126}$	**.2407**$_{\pm.0107}$
AMI	.0443$_{\pm.0179}$	.0906$_{\pm.0278}$	.3860$_{\pm.0011}$	.0033$_{\pm.0009}$	.1436$_{\pm.0010}$
MIRROR-b	.2026$_{\pm.0267}$	.4533$_{\pm.0394}$	.3564$_{\pm.0058}$	.0613$_{\pm.0059}$	.2218$_{\pm.0315}$
RLB-MI	.2568$_{\pm.0172}$	.5044$_{\pm.0246}$	.3804$_{\pm.0049}$	.0514$_{\pm.0038}$	.2344$_{\pm.0238}$
LO-MI	.2611$_{\pm.0079}$	.5155$_{\pm.0115}$	.3921$_{\pm.0008}$	.0536$_{\pm.0038}$	.1909$_{\pm.0027}$
P&P	.7779$_{\pm.0308}$	.9476$_{\pm.0076}$	.2470$_{\pm.0029}$	.1440$_{\pm.0059}$	.2111$_{\pm.0407}$
MIRROR-w	.8129$_{\pm.0228}$	.9531$_{\pm.0085}$	.2491$_{\pm.0038}$	.1257$_{\pm.0048}$	.2356$_{\pm.0229}$

(b) Chest X-ray diagnosis

Method	Acc@1 ↑	Acc@5 ↑	F-dist ↓	Cover ↑	MS-SSIM ↑
TMI	.5158$_{\pm.0127}$	**.9999**$_{\pm.0004}$	**.0851**$_{\pm.0006}$	**.2415**$_{\pm.0396}$	.0766$_{\pm.0097}$
AMI	.0743$_{\pm.0080}$	.8717$_{\pm.0103}$	.1788$_{\pm.0008}$	.0002$_{\pm.0002}$	.0765$_{\pm.0019}$
MIRROR-b	**.7786**$_{\pm.0757}$	.9983$_{\pm.0049}$	.1094$_{\pm.0183}$	.0172$_{\pm.0154}$	.0709$_{\pm.0296}$
RLB-MI	.0790$_{\pm.0000}$	.8988$_{\pm.1026}$	.1827$_{\pm.0069}$	.0002$_{\pm.0000}$	**.0768**$_{\pm.0002}$
LO-MI	.0790$_{\pm.0000}$	.9860$_{\pm.0000}$	.1769$_{\pm.0006}$	.0002$_{\pm.0000}$	.0708$_{\pm.0000}$
P&P	.7250$_{\pm.2769}$	.9960$_{\pm.0057}$	.1155$_{\pm.0211}$	.0084$_{\pm.0074}$	.0739$_{\pm.0038}$
MIRROR-w	.8634$_{\pm.1339}$	.9983$_{\pm.0049}$	.1138$_{\pm.0264}$	.0137$_{\pm.0140}$	.0679$_{\pm.0189}$

(c) Car classification

Method	Acc@1 ↑	Acc@5 ↑	F-dist ↓	Cover ↑	MS-SSIM ↑
TMI	**.0625**$_{\pm.0075}$	**.1693**$_{\pm.0247}$	**.4901**$_{\pm.0153}$	**.1162**$_{\pm.0035}$	**.1844**$_{\pm.0201}$
AMI	.0000$_{\pm.0000}$	.0000$_{\pm.0000}$	.6020$_{\pm.0035}$	.0000$_{\pm.0000}$	.0930$_{\pm.0028}$
MIRROR-b	.0299$_{\pm.0038}$	.1156$_{\pm.0363}$	.5619$_{\pm.0047}$	.0911$_{\pm.0027}$	.1423$_{\pm.0185}$
RLB-MI	.0000$_{\pm.0000}$	.0020$_{\pm.0023}$	.5752$_{\pm.0028}$	.0000$_{\pm.0000}$	.1082$_{\pm.0170}$
LO-MI	.0000$_{\pm.0000}$	.0007$_{\pm.0003}$	.5845$_{\pm.0066}$	.0000$_{\pm.0000}$	.1091$_{\pm.0050}$
P&P	.0668$_{\pm.0083}$	.1836$_{\pm.0193}$	.5588$_{\pm.0808}$	.0496$_{\pm.0058}$	.1625$_{\pm.0254}$
MIRROR-w	.0533$_{\pm.0076}$	.1568$_{\pm.0592}$	.5578$_{\pm.0454}$	.0336$_{\pm.0035}$	.1521$_{\pm.0203}$

5. Experiments

We conducted a comprehensive comparison of the inversion capability of TMI with state-of-the-art MI attacks, including both black-box and white-box methods. The black-box methods include MIRROR-b, RLB-MI (Han et al., 2023), LO-MI (Kahla et al., 2022), and AMI (Yang et al., 2019), while the white-box methods include MIRROR-W and P&P (Struppek et al., 2022). MIRROR-b and MIRROR-w represent the black-box and white-box MI methods using MIRROR (An et al., 2022), respectively.

5.1. Experimental setup

We selected three tasks for MI: facial recognition, chest X-ray diagnosis, and car classification. For facial recognition, we prepared target networks trained on the FaceScrub (Ng and Winkler, 2014) and CelebA (Liu et al., 2015) datasets. We used the official bounding-box information for FaceScrub and use its 530 identities as each class for classification. For CelebA, we randomly selected 1000 identities from the entire dataset for classification. With the aligned version of CelebA images, we further applied 108 × 108 center-crop to align them with the FaceScrub images.

We used ResNeSt-101 (Zhang et al., 2020a) as the architecture for f in our main experiments, along with DenseNet-169 (Huang et al., 2017) and MobileNet-v3 (Howard et al., 2019) in the additional experiments. As for the adversary's image prior, we used a pretrained

Fig. 3. Comparison of the inversion results on facial recognition (left) chest X-ray diagnosis (middle), and car classification (right). Gender and age information for each X-ray target images are shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

StyleGAN2 (Karras et al., 2020) network available online[4] trained on the Flickr-Faces-HQ (FFHQ) (Karras et al., 2019) dataset. We note that the underlying distribution of the StyleGAN network is different from that of the target networks, reflecting a practical attack setting. In our ablation study, we have also verified that TMI remains effective under conditions of an even greater distribution shift. For chest X-ray diagnosis, we used a StyleGAN2 network trained on the NIH Chest X-ray dataset (Wang et al., 2017). The target network was trained using the PadChest (Bustos et al., 2019) dataset. We used only the front-facing X-ray images, and selected the seven most frequent findings (normal, pneumonia, tuberculosis sequelae, emphysema, heart insufficiency, pulmonary fibrosis, COPD signs) from the PadChest dataset for disease classification.

For the car classification task, we used a publicly available StyleGAN2 network[5] trained on the LSUN car dataset (Yu et al., 2015); we then conducted TMI against a target model trained upon the CompCars dataset (Yang et al., 2015).

We used StyleGAN networks of an image size $\mathbb{R}^{3\times 256^2}$ for face recognition and chest X-ray diagnosis. For car classification, we used a StyleGAN network of an image size $\mathbb{R}^{3\times 512^2}$. We employed same StyleGAN2 networks as image priors in all baseline methods using StyleGAN. For other attacks that incorporate GANs (Kahla et al., 2022; Han et al., 2023), we trained their GANs using the same dataset used to train StyleGANs[6].

Throughout the experiments, the input to the target models and evaluation classifiers were unified to $\mathbb{R}^{3\times 224^2}$ and $\mathbb{R}^{3\times 299^2}$, respectively. Input images were resized to match the respective input dimensions using bilinear interpolation. StyleGAN generated images given as input to f during the preparation phase should also be cropped & resized appropriately. We applied 180×180 crop, then resized them to match the input size. We used pretrained ImageNet checkpoints provided by Torchvision or PyTorch Hub as initial weights for the target and evaluation classifiers, and replaced their final fully-connected layer to match the number of classes in respective datasets. 10% of each dataset were used as the test split. Note that the TMI adversary has no knowledge of the cropping or resizing logic of f, as it is processed inside the black-box service. Accordingly, $f' : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ receives the full images generated from StyleGAN, whereas the actual $\hat{\mathcal{Y}}$ is calculated inside f upon cropped & resized versions.

In constructing $\mathcal{D}_{gen}$, we applied the truncation trick, which is a generally used technique to promote natural image generation with StyleGAN (Karras et al., 2019). Specifically, we generated images with truncation $\psi = 0.7$ in order to avoid unnatural synthesis results. To increase the dispersion between confidence values for easier training of m' and f', we apply natural logarithm to the prediction vectors from f.

For the inversion phase of the TMI experiments, we used $n = 5000$, $\eta = 10^{-5}$, $e = 500$, $\delta = 0.05$ for Algorithm 1. Note that despite these hyperparameters can be fine-tuned to each attack scenario, we fixed them for simplicity and to demonstrate the robustness of TMI.

For each evaluation scenario, we attacked 1000 randomly selected samples from the test split of f's dataset. This simulates the real-world attack, where the target images correspond to one of f's classes, however not directly included in its training set. For TMI and AMI in Table 2, we repeated the experiment 8 times and reported the mean and standard deviation of each metrics. For the other baseline attacks, we selected the 8 best candidates from their final output and report their mean and standard deviation.

5.2. Evaluation metrics

To evaluate the effectiveness of TMI, we employed various metrics to assess the quality of the inversion results and the instance-specific MI capabilities. Accuracy and feature distance are widely used metrics in the MI literature (An et al., 2022; Struppek et al., 2022; Wang et al.,

[4] https://github.com/rosinality/stylegan2-pytorch.

[5] https://github.com/NVlabs/stylegan2.

[6] Pretrained networks uploaded by the authors were unusable since it used a much tighter crop compared to the FaceScrub official bounding-box. We also observed worse results when the image priors were replaced with StyleGAN.

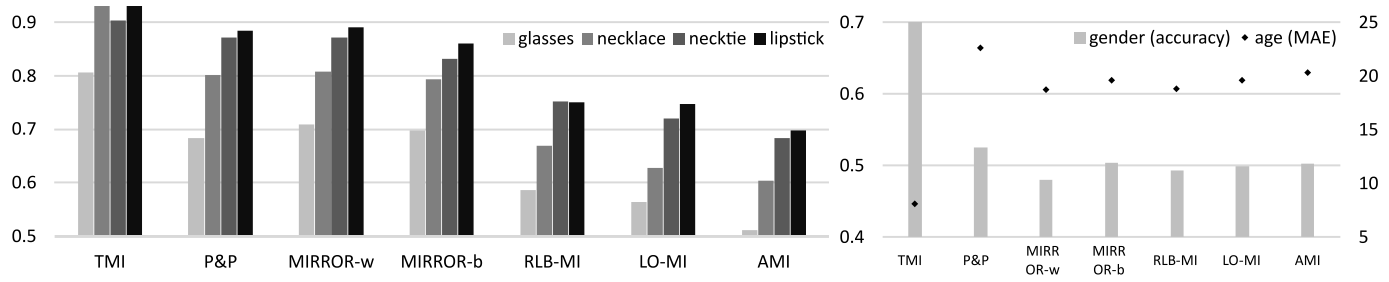

Fig. 4. Attribute accuracy on facial recognition (left) and chest X-ray diagnosis (right). For age, we reported the mean absolute error.

2021; Zhang et al., 2020b). We also used three additional metrics, MS-SSIM, class-wise coverage, and attribute accuracy to demonstrate the capabilities of instance-specific MI.

Accuracy (`Acc@1` and `Acc@5`). To assess the resemblance of the reconstructed images to the target image class, we computed the proportion of reconstructed images that were classified into the same class as the target image by f_E, an evaluation classifier using Inception-v3 (Szegedy et al., 2016) trained on the same dataset as f. This proportion represents the accuracy of the reconstruction process, indicating how well the reconstructed images capture the features of the target class in general.

Feature distance (`F-dist`). The feature distance metric captures the similarity between two images at an intermediate representation layer (Dosovitskiy and Brox, 2016) of f_E, which quantifies the perceptual similarity between the images. Specifically, we computed the average ℓ_2 distance between features extracted from the penultimate layer of a f_E, hence computing the similarity in high-level visual features perceived by the classifier (Zhao et al., 2021).

Multi-scale structural similarity (`MS-SSIM`). The multi-scale structural similarity index (MS-SSIM) (Wang et al., 2003) is an image quality metric that extends the single-scale SSIM by assessing the structural similarity between two images across multiple resolutions in order to simulate various viewing conditions. This metric evaluates image details across different scales by downsampling and integrating the SSIM values at each scale using a weighted geometric mean. MS-SSIM is widely used for evaluating image reconstruction due to its capability to capture image details at various scales. This allows it to discern subtle differences and prioritize important visual information.

Class-wise coverage (`Cover`). We adopted the class-wise coverage metric to assess whether the reconstructed samples successfully captured the intra-class diversity, which is crucial in the instance-specific MI task. We use a slightly modified version of the original notion introduced by Naeem et al. (2020). This metric evaluates the extent to which the reconstructed samples cover the range of variations within each target class. It measures the fraction of target images that have a reconstructed sample in close proximity, providing insight into how well the reconstruction process captures the intra-class diversity. The class-wise coverage is formally defined as follows:

$$\texttt{Cover} = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists\, j \text{ s.t. } Y_j \in B(X_i,\, NND_k(X_i))}. \tag{6}$$

where N and $1_{(\cdot)}$ are the number of samples and the indicator function, respectively. Whereas the original notation considered the intermediate representations of real and fake samples as X_i and Y_i, we replaced them with the intermediate representations of the target and reconstructed images, respectively. $B(x, r)$ indicates a sphere in the representation space around x with radius r, and $NND_k(X_i)$ denotes the distance from X_i to its kth nearest neighbor. We used $k = 1$ throughout our evaluations.

Attribute accuracy. To evaluate the success of feature reconstruction, we trained attribute classifiers using the Inception-v3 architecture (Szegedy et al., 2016), where the final layer of the classifier was adjusted to accommodate the number of categories for each attribute. Specifically, we trained attribute classifiers using the respective attribute labels available in CelebA and the gender and age information in PadChest.

5.3. Experimental results

5.3.1. Main experiment

Fig. 3 shows the inversion results obtained by other MI methods and TMI. It is evident that the samples reconstructed using TMI are visually more similar to their corresponding original images, making it easier to identify them as the same identity. We also note that TMI reconstructs facial expressions (columns 1, 2, 4), and instance-specific attributes (columns 3, 5) such as glasses. The other methods did succeed in reconstructing some general features of the original images. However, they failed to capture the fine and specific characteristics of the original image. Moreover, we observed that RLB-MI, LO-MI, and AMI methods are not suitable for real-world MI attacks on high-dimensional images. While the authors had demonstrated their success on $\mathbb{R}^{3\times 64^2}$ tightly-cropped images, we found that these black-box attacks experienced difficulties in reconstructing $\mathbb{R}^{3\times 224^2}$ input images. In contrast, the white-box attacks exhibited high accuracy since they explicitly took into account the classification loss on the target model during their optimization steps.

Table 2 provides the quantitative evaluation results of TMI, along with state-of-the-art MI methods. The experimental results clearly demonstrate that TMI outperformed all other black-box MI attacks according to the reported metrics. These results confirm the superiority of TMI in conducting practical black-box MI.

We emphasize the significant decrease in the number of required queries in performing MI. When assuming a scenario in which the adversary aims to generate 530 class facial images in the FaceScrub dataset, TMI requires the default query budget of only 100k queries to a target model in achieving the reported metrics in Table 2. In contrast, MIRROR-w and MIRROR-b, which performed the best beside TMI, required 84 900 k = 100 k+160 k×530 and 5400 k = 100 k+10 k×530 queries, respectively. Outside of AMI which failed to produce any meaningful results, LO-MI required the least number of queries among the baselines: 1160 k = 100 k + 2 k × 530, which is still significantly higher than the one that TMI required.

We also investigated the change in reconstruction performance with different query budgets: 100, 500, 1k, 5k, 10k, 50k, and 100k (default). The query budgets denote the total number of queries allowed to attack every label (i.e., 530 attacks in case of FaceScrub). In TMI, this directly translates to the size of D_{gen} that is used during the preparation phase to train m' and f'. Note that we were unable to evaluate the performance of the baseline methods using 100, 500, and 1k query budgets since

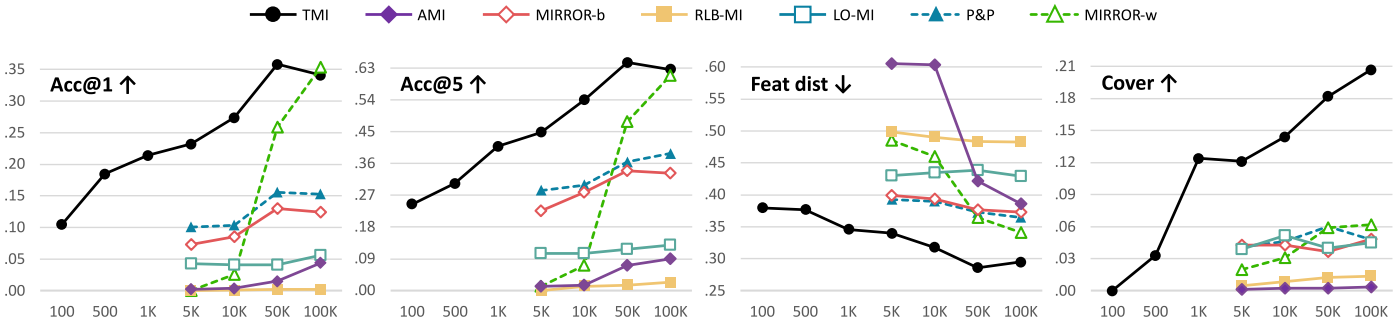

Fig. 5. Changes in attack performance while varying the query budget. White-box methods are represented by dashed lines.

Table 3
MI performance across scenarios involving different target domains and f. A→B indicates that the attacker utilizes an image prior (i.e., StyleGAN) obtained from A to attack f that was trained to classify B images.

Scenario	Arch	Acc@1	Acc@5	F-dist	Cover
FFHQ	MobileNet-v3	.3380	.6390	.2878	.1813
↓	ResNeSt-101	.3408	.6255	.2950	.2067
FaceScrub	DenseNet-169	.3265	.6035	.3027	.1967
FFHQ	MobileNet-v3	.2080	.4345	.9215	.2343
↓	ResNeSt-101	.2880	.5520	.8503	.2738
CelebA	DenseNet-169	.2015	.4335	.5751	.2965
CXR14	MobileNet-v3	.3540	.9880	.1161	.2316
↓	ResNeSt-101	.5158	.9999	.0851	.2415
PadChest	DenseNet-169	.4840	.9940	.1188	.1972

each attack per class requires a certain number of queries. For example, attacking a 530-class facial recognition model with a 1k query budget allows only two queries for each class.

As shown in Fig. 5, the overall MI performance decreases as the query budget becomes smaller. However, TMI consistently outperforms the baseline attacks across different query budgets. We also note that TMI holds a distinctive advantage of requiring zero queries to a target model after the preparation phase, which significantly facilitates the reconstruction of inputs for a large number of prediction vectors. Observe that with only 5k queries, TMI attained similar performance to MIRROR-b *without* a query budget; which required a total of 5400 K queries, resulting in a 1080 times decrease in the query budget.

For chest X-ray diagnosis, TMI significantly outperformed all other methods by a large margin. The performance gains can be attributed to the unique characteristics of the chest X-ray classifier. Unlike facial recognition, where each class corresponds to a single identity, the classes in the chest X-ray classifier encompass a diverse range of identities and features. For example, the *pneumonia* class includes samples from both male and female individuals, and the age of the subjects spans across various age groups. In this particular context, we argue that traditional metrics such as Acc@1 and Acc@5 are not indicators of instance-specific privacy leakage. Thus, we additionally used intra-class metrics (i.e., F-dist, Cover, and attribute accuracy) to evaluate TMI.

Fig. 3 presents a comparison of the inversion results on chest X-ray images. Notably, TMI successfully captures private information such as gender and body shape. For example, the highlighted TMI inversion results in the second row clearly shows a female X-ray image, even to the human eye, due to the accurate reconstructions of the chest shape when comparing its reconstruction quality with the other baseline results.

A similar trend is observed in the car classification task, where TMI is the only method capable of reconstructing the orientation and color of the input image. We note that each class in this classification task corresponds to a specific car model, which varies in color and orientation. Consequently, class-representative inversion tends to produce seemingly random results, as shown in Fig. 3.

Fig. 6. Qualitative demonstration of attribute reconstruction. Each two target images belong to the same class.

Table 4
MI performance of different initialization methods. Best cases are marked in bold.

Method	Acc@1	Acc@5	F-dist	Cover
m' only	.0840	**.2620**	**.1523**	**.0817**
Maximum target confidence	**.0900**	.2200	.4135	0
Random initialize	.0010	.0100	.4941	.0309

In addition to ResNeSt-101, we further investigated how MI performance varies across different architectures for f; DenseNet-169 and MobileNet-v3. The tendency of the results were constant to the main evaluation throughout all configurations (Table 3). For example, in the `FFHQ→FaceScrub` scenario, TMI surpassed all black-box baselines in every metric. This suggests that TMI is generally applicable across various target systems under a truly black-box setting, i.e., in a target model-agnostic manner.

To emphasize the capability of targeted MI, we further conducted comparison evaluations that measure the attribute accuracy across instance-specific attributes in Fig. 4. As the figure shows, TMI strictly outperformed all other methods in capturing subtle and intra-class features. For example, the attribute classifier for checking glasses on the TMI reconstructed facial images reported an accuracy of 80.7% while MIRROR-b and P&P reported 69.8% and 68.3%, respectively. Attribute reconstructions are better observed qualitatively in Fig. 6; sample-specific attributes such as glasses, skin tone, hair color and gaze are accurately reconstructed, even where the attribute is not correlated to the class itself. Also, when inferring the ages of the reconstructed chest X-ray images, TMI-generated images contribute to reporting a mean absolute error (MAE) of 8.1198, significantly outperforming all other methods. These results highlight the capability of TMI to capture instance-specific private information compared to the baselines.

5.3.2. Ablation study

Effect of m'. In order to assess the efficacy of the new mapping network m' in locating a reliable initial latent point (w_0), we evaluated the synthesis result directly from w_0 without the optimization steps (i.e., $g(w_0)$), and compared it with initialization techniques of existing methods. The results in Table 4 suggest that w_0 is closer to the target image in terms of F-dist and Cover compared to random initialization or maximum target confidence. Note that *maximum target confidence*

Table 5
Change in attack performance when replacing f' with f.

Performance gain	
Acc@1 ↑	0.3408→0.8264 (142.49% ↑)
Acc@5 ↑	0.6255→0.9534 (52.42% ↑)
F-dist ↓	0.2950→0.1975 (33.05% ↓)
Cover ↑	0.2067→0.2619 (26.71% ↑)

(a)

Performance drop	
Acc@1 ↑	0.3408→0.1100 (67.72% ↓)
Acc@5 ↑	0.6255→0.2660 (57.47% ↓)
F-dist ↓	0.2950→0.3816 (29.35% ↑)
Cover ↑	0.2067→0.0982 (52.49% ↓)

(b)

Fig. 7. Qualitative (a) and quantitative (b) comparison of TMI and f''.

achieves high Acc@1 due to its selection policies, however, falls behind in every other metrics. This indicates that the latent point drawn from such initialization technique is merely an overfit to the target class, which fails to capture all intra-class features.

In addition, we investigated the attack scenario of using the original model instead of the surrogate model. This setting demonstrates the effect of m' in an undisturbed setting, where the following optimization steps no longer have to use an imperfect substitution of the target model. As listed in Table 5, attack performance is improved by an average of 63.7%, outperforming even the white-box attacks across all four metrics. This also indicates that the strategy of retraining a mapping network can be applied to white-box TMI attacks for better reconstruction performance.

Effect of f'. A naive black-box migration of the white-box approaches would be to train a surrogate model from scratch, instead of using f' (which is transfer-learned from D of the StyleGAN network). Similar to f', this new surrogate model works as a substitute for f, removing the white-box requirement. We refer to such MI attack scenario as f''. Here, we empirically demonstrate that f'' is insufficient to provide reliable optimization, thus justifying the use of f'. Reconstructions from f'' only succeed in capturing some general coarse-grained features, failing to reconstruct the instance-specific details or even the identity, as can be observed from Fig. 7, both qualitatively and quantitatively. We argue that it is necessary to use f' and take advantage of its pre-exposure to various styles during the StyleGAN training.

Using a Different GAN Architecture. We investigated TMI's efficacy when using a GAN architecture other than StyleGAN. We conducted an additional experiment with UNet-GAN (Schönfeld et al., 2020) pretrained on FFHQ and compared the results to the main experiment. For the surrogate model f', we used the left-half of the UNet-GAN's discriminator. Since UNet-GAN does not incorporate a mapping network, a custom mapping layer was trained from scratch. Observe from Table 6 that while the overall performance is slightly decreased compared to TMI with StyleGAN, it is still more effective than MIRROR-b. The slight drop in performance is largely due to the entangled latent space of UNet-GAN, where the latent space $\mathcal{Z}$ is directly used without mapping it to an intermediate disentangled latent space $\mathcal{W}$ in advance. This result indicates that while StyleGAN is still the most effective image prior to be utilized, other GAN-based methods can generally benefit from the suggested approach of leveraging a new mapping layer and distilling the discriminator for a surrogate model.

Table 6
Attack performance of TMI with UNet-GAN, compared to the original TMI attack and MIRROR-b.

	Acc@1	Acc@5	F-dist	Cover
TMI	.3804	.6255	.2950	.2067
TMI with UNet-GAN	.2201	.4811	.3254	.2063
MIRROR-b	.2026	.4533	.3564	.0613

Fig. 8. Target FaceScrub images (top) and TMI-reconstructed using StyleGAN trained on art-style face portraits (bottom).

Using D_{aux} with Significant Deviation from D. Sometimes the TMI adversary may not be able to obtain D_{aux} with a distribution similar to D. In this experiment, we show that TMI is robust enough to capture features of the target images $x \in D$ that are embedded in the distribution of D_{aux}, even when D and D_{aux} as a whole have distinct distributions. We used a StyleGAN trained on art portraits[7] as the image prior to attack a ResNeSt-101 network trained to classify FaceScrub identities. Fig. 8 clearly demonstrates that TMI can successfully reconstruct input images with high fidelity, including facial features, posture, and rough color.

6. Mitigation

Label-only. To prevent TMI attackers from obtaining additional information from the model output, one could modify the system to return only the final decision (i.e., the predicted label) instead of the full confidence vector. However, we found that a slight modification to the TMI workflow extends the attack to invalidate such defense. Specifically, we implement label smoothing on each prediction output received from the target model, converting each label prediction into a pseudo confidence vector. Given the pseudo confidence vector, TMI can operate in the same way as the original version.

Table 7 demonstrates that the attack is still effective in label-only situations. While there is a slight drop in performance compared to the original use-case, notice that the attack continues to surpass the capabilities of MIRROR-b in all metrics outside of F-dist.

Random Noise. TMI utilizes subtle changes in the confidence vector that respond to certain styles in the input images. Therefore, to reduce TMI's attack performance, one could inject subtle noise to the predicted output while preserving the prediction label. Table 7 demonstrates the mitigation effect showing that its performance dropped below MIRROR-b, in particular, F-dist and Cover metrics are greatly degraded from the original TMI experiment (21.1% and 81.6%). This suggests that random noise successfully distracts TMI from reconstructing subtle features.

7. Discussion and limitations

We demonstrated that TMI is effective in reconstructing inputs even when the adversary is able to only obtain labels, not their prediction

[7] https://github.com/ak9250/stylegan-art.

Table 7
MI performance on label-only setting and random noise defense. The original TMI and MIRROR-b experiment results are also included for comparison.

Method	Acc@1	Acc@5	F-dist	Cover
TMI (original)	.3804	.6255	.2950	.2067
TMI against label-only	.2399	.4800	.3637	.1167
TMI against random noise	.1792	.4297	.3739	.0381
MIRROR-b	.2026	.4533	.3564	.0613

vectors (Section 6). One way to mitigate this threat is to inject Gaussian noise into prediction vectors, effectively deterring the adversary's attempts using TMI. However, the adversary can still conduct label-only attacks, avoiding to use the injected noise. We leave developing more robust defenses against TMI for future work.

One limitation of TMI is its dependence on the pretrained StyleGAN model. When this pretrained model is unavailable or trained with data instances whose underlying distribution differs from the target dataset (e.g., performing MI attacks against a facial recognition classifier with a chest X-ray dataset), the performance of TMI diminishes in reconstructing input images.

TMI depends on the generative capability of a StyleGAN network, making its success heavily dependent on the expressiveness of the underlying image prior. Unfortunately, several data domains are challenging to model through a generative prior due to either a lack of data or the generator's limited expressiveness. Training a generative model in such domains often results in model collapse and non-convergence. Although prior studies have proposed methods to mitigate these issues (Salimans et al., 2016), domains with extremely high variability still remain difficult to be captured properly. For instance, complex natural scenes, which feature a vast diversity of objects and intricate spatial relationships, are notoriously difficult to model. Furthermore, even in cases where a generative model is present, data points not fully represented in the manifold of this generative model will show poor reconstruction, thus impeding the success of MI attacks using TMI.

Lastly, the success of TMI depends on the size of D_{gen}, as shown in Fig. 5. TMI requires the attacker to send queries comprised solely of synthetic images to the target model. However, we emphasize that TMI demands a significantly smaller number of queries compared to the other state-of-the-art MI attacks, advancing the current lower bound in conducting effective MI attacks.

8. Conclusion

We have proposed TMI, a novel black-box MI attack that achieves instance-specific MI using a limited query budget. TMI alters the mapping network of a benign StyleGAN network to find a reliable initial latent point corresponding to a target prediction output, then performs further optimization by leveraging a surrogate model distilled from the StyleGAN discriminator. TMI significantly decreases the number of required queries while improving the reconstruction quality over state-of-the-art black-box MI methods.

CRediT authorship contribution statement

Hoyong Jeong: Conceptualization. **Kiwon Chung:** Conceptualization. **Sung Ju Hwang:** Supervision. **Sooel Son:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sooel Son reports financial support was provided by Institute of Information & communications Technology Planning & Evaluation (IITP). Hoyong Jeong reports financial support was provided by Institute of Information & communications Technology Planning & Evaluation (IITP). Kiwon Chung reports financial support was provided by Institute of Information & communications Technology Planning & Evaluation (IITP). Sung Ju Hwang reports financial support was provided by Institute of Information & communications Technology Planning & Evaluation (IITP). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00153, Penetration Security Testing of ML Model Vulnerabilities and Defense).

References

Abdal, R., Qin, Y., Wonka, P., 2019. Image2StyleGAN: How to embed images into the stylegan latent space? In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, pp. 4431–4440, URL https://doi.org/10.1109/ICCV.2019.00453.

Abdelberi, C., Ács, G., Kâafar, M.A., 2012. You are what you like! information leakage through users' interests. In: 19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5-8, 2012. The Internet Society, URL https://www.ndss-symposium.org/ndss2012/you-are-what-you-information-leakage-through-users-interests.

An, S., Tao, G., Xu, Q., Liu, Y., Shen, G., Yao, Y., Xu, J., Zhang, X., 2022. Mirror: Model inversion for deep learning network with high fidelity. In: Proceedings of the 29th Network and Distributed System Security Symposium. http://dx.doi.org/10.14722/ndss.2022.24335.

Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M., 2019. PadChest: A large chest x-ray image dataset with multi-label annotated reports. CoRR arXiv:1901.07441. arXiv:1901.07441. URL http://arxiv.org/abs/1901.07441.

Chen, S., Kahla, M., Jia, R., Qi, G., 2021. Knowledge-enriched distributional model inversion attacks. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, pp. 16158–16167, URL https://doi.org/10.1109/ICCV48922.2021.01587.

Dibbo, S.V., 2023. SoK: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In: 36th IEEE Computer Security Foundations Symposium, CSF 2023, Dubrovnik, Croatia, July 10-14, 2023. IEEE, pp. 439–456, URL https://doi.org/10.1109/CSF57540.2023.00027.

Dosovitskiy, A., Brox, T., 2016. Generating images with perceptual similarity metrics based on deep networks. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 658–666, URL https://proceedings.neurips.cc/paper/2016/hash/371bce7dc83817b7893bcdeed13799b5-Abstract.html.

Fredrikson, M., Jha, S., Ristenpart, T., 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In: Ray, I., Li, N., Kruegel, C. (Eds.), Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015. ACM, pp. 1322–1333, URL https://doi.org/10.1145/2810103.2813677.

Fredrikson, M., Lantz, E., Jha, S., Lin, S.M., Page, D., Ristenpart, T., 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Fu, K., Jung, J. (Eds.), Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014. USENIX Association, pp. 17–32, URL https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.

Gu, Z., Huang, H., Zhang, J., Su, D., Lamba, A., Pendarakis, D., Molloy, I.M., 2018. Securing input data of deep learning inference systems via partitioned enclave execution. CoRR arXiv:1807.00969. arXiv:1807.00969. URL http://arxiv.org/abs/1807.00969.

Han, G., Choi, J., Lee, H., Kim, J., 2023. Reinforcement learning-based black-box model inversion attacks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. IEEE, pp. 20504–20513, URL https://doi.org/10.1109/CVPR52729.2023.01964.

He, Z., Zhang, T., Lee, R.B., 2019. Model inversion attacks against collaborative inference. In: Balenson, D.M. (Ed.), Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019. ACM, pp. 148–162, URL https://doi.org/10.1145/3359789.3359824.

He, Z., Zhang, T., Lee, R.B., 2021. Attacking and protecting data privacy in edge-cloud collaborative inference systems. IEEE Internet Things J. 8 (12), 9706–9716, URL https://doi.org/10.1109/JIOT.2020.3022358.

Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., 2019. Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, pp. 1314–1324, URL https://doi.org/10.1109/ICCV.2019.00140.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, pp. 2261–2269, URL https://doi.org/10.1109/CVPR.2017.243.

Huang, J., Shao, H., Chang, K.C., 2022. Are large pre-trained language models leaking your personal information? pp. 2038–2047, URL https://doi.org/10.18653/v1/2022.findings-emnlp.148.

Jia, J., Wang, B., Zhang, L., Gong, N.Z., 2017. AttriInfer: Inferring user attributes in online social networks using Markov random fields. In: Barrett, R., Cummings, R., Agichtein, E., Gabrilovich, E. (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017. ACM, pp. 1561–1569, URL https://doi.org/10.1145/3038912.3052695.

Kahla, M., Chen, S., Just, H.A., Jia, R., 2022. Label-only model inversion attacks via boundary repulsion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, la, USA, June 18-24, 2022. IEEE, pp. 15025–15033, URL https://doi.org/10.1109/CVPR52688.2022.01462.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp. 4401–4410. http://dx.doi.org/10.1109/CVPR.2019.00453, URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020. Analyzing and improving the image quality of stylegan. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, pp. 8107–8116. http://dx.doi.org/10.1109/CVPR42600.2020.00813, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, pp. 3730–3738, URL https://doi.org/10.1109/ICCV.2015.425.

Melis, L., Song, C., Cristofaro, E.D., Shmatikov, V., 2019. Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019. IEEE, pp. 691–706, URL https://doi.org/10.1109/SP.2019.00029.

Michalevsky, Y., Schulman, A., Veerapandian, G.A., Boneh, D., Nakibly, G., 2015. PowerSpy: Location tracking using mobile device power analysis. In: Jung, J., Holz, T. (Eds.), 24th USENIX Security Symposium, USENIX Security 15, Washington, D.C., USA, August 12-14, 2015. USENIX Association, pp. 785–800, URL https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/michalevsky.

Moore, W., Frye, S., 2019. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. J. Nucl. Med. Technol. 47 (4), 269–272.

Müller, R., Kornblith, S., Hinton, G.E., 2019. When does label smoothing help? URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J., 2020. Reliable fidelity and diversity metrics for generative models. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. In: Proceedings of Machine Learning Research, vol. 119, PMLR, pp. 7176–7185, URL http://proceedings.mlr.press/v119/naeem20a.html.

Narain, S., Vo-Huu, T.D., Block, K., Noubir, G., 2016. Inferring user routes and locations using zero-permission mobile sensors. In: IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. IEEE Computer Society, pp. 397–413, URL https://doi.org/10.1109/SP.2016.31.

Narra, K.G., Lin, Z., Wang, Y., Balasubramaniam, K., Annavaram, M., 2019. Privacy-preserving inference in machine learning services using trusted execution environments. CoRR arXiv:1912.03485. arXiv:1912.03485. URL http://arxiv.org/abs/1912.03485.

Ng, H., Winkler, S., 2014. A data-driven approach to cleaning large face datasets. In: 2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014. IEEE, pp. 343–347, URL https://doi.org/10.1109/ICIP.2014.7025068.

Otterbacher, J., 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In: Huang, J.X., Koudas, N., Jones, G.J.F., Wu, X., Collins-Thompson, K., An, A. (Eds.), Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010. ACM, pp. 369–378, URL https://doi.org/10.1145/1871437.1871487.

Pizzi, K., Boenisch, F., Sahin, U., Böttinger, K., 2023. Introducing model inversion attacks on automatic speaker recognition. CoRR arXiv:2301.03206. arXiv:2301.03206. URL https://doi.org/10.48550/arXiv.2301.03206.

Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. pp. 2226–2234, URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Schönfeld, E., Schiele, B., Khoreva, A., 2020. A U-net based discriminator for generative adversarial networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, pp. 8204–8213. http://dx.doi.org/10.1109/CVPR42600.2020.00823, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Schonfeld_A_U-Net_Based_Discriminator_for_Generative_Adversarial_Networks_CVPR_2020_paper.html.

Struppek, L., Hintersdorf, D., Correia, A.D.A., Adler, A., Kersting, K., 2022. Plug & play attacks: Towards robust and flexible model inversion attacks. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. In: Proceedings of Machine Learning Research, vol. 162, PMLR, pp. 20522–20545, URL https://proceedings.mlr.press/v162/struppek22a.html.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, pp. 2818–2826, URL https://doi.org/10.1109/CVPR.2016.308.

Wang, K., Fu, Y., Li, K., Khisti, A., Zemel, R.S., Makhzani, A., 2021. Variational model inversion attacks. URL https://proceedings.neurips.cc/paper/2021/hash/50a074e6a8da4662ae0a29edde722179-Abstract.html.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-Ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, pp. 3462–3471, URL https://doi.org/10.1109/CVPR.2017.369.

Wang, Z., Simoncelli, E., Bovik, A., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Vol. 2, pp. 1398–1402 Vol.2. http://dx.doi.org/10.1109/ACSSC.2003.1292216.

Weinsberg, U., Bhagat, S., Ioannidis, S., Taft, N., 2012. Blurme: inferring and obfuscating user gender based on ratings. In: Cunningham, P., Hurley, N.J., Guy, I., Anand, S.S. (Eds.), Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012. ACM, pp. 195–202, URL https://doi.org/10.1145/2365952.2365989.

Yang, L., Luo, P., Loy, C.C., Tang, X., 2015. A large-scale car dataset for fine-grained categorization and verification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, pp. 3973–3981, URL https://doi.org/10.1109/CVPR.2015.7299023.

Yang, Z., Zhang, J., Chang, E., Liang, Z., 2019. Neural network inversion in adversarial setting via background knowledge alignment. In: Cavallaro, L., Kinder, J., Wang, X., Katz, J. (Eds.), Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019. ACM, pp. 225–240, URL https://doi.org/10.1145/3319535.3354261.

Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J., 2015. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. CoRR arXiv:1506.03365. arXiv:1506.03365. URL http://arxiv.org/abs/1506.03365.

Yuan, X., Chen, K., Zhang, J., Zhang, W., Yu, N., Zhang, Y., 2023. Pseudo label-guided model inversion attack via conditional generative adversarial network. CoRR arXiv:2302.09814. arXiv:2302.09814. URL https://doi.org/10.48550/arXiv.2302.09814.

Zhang, R., Hidano, S., Koushanfar, F., 2022. Text revealer: Private text reconstruction via model inversion attacks against transformers. CoRR arXiv:2209.10505. arXiv:2209.10505. URL https://doi.org/10.48550/arXiv.2209.10505.

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D., 2020b. The secret revealer: Generative model-inversion attacks against deep neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, pp. 250–258. http://dx.doi.org/10.1109/CVPR42600.2020.00033, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhang_The_Secret_Revealer_Generative_Model-Inversion_Attacks_Against_Deep_Neural_Networks_CVPR_2020_paper.html.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A.J., 2020a. ResNeSt: Split-attention networks. CoRR arXiv:2004.08955. arXiv:2004.08955. URL https://arxiv.org/abs/2004.08955.

Zhao, X., Zhang, W., Xiao, X., Lim, B.Y., 2021. Exploiting explanations for model inversion attacks. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, pp. 662–672, URL https://doi.org/10.1109/ICCV48922.2021.00072.

Hoyong Jeong is a machine learning researcher at Deeping Source. He received his master's degree from KAIST, under the direction of professor Sooel Son. His research explores various topics in AI privacy, spanning from privacy-preserving algorithms to strategies for privacy attacks.

Kiwon Chung is a Master's student from the Graduate School of Information Security at KAIST under the direction of professor Sooel Son. His research interests include security of Machine Learning, Computer vision, and Natural Language Processing.

Sung Ju Hwang is an endowed chair professor in the Kim Jaechul School of AI and School of Computing at KAIST. He received his Ph.D. in computer science at University of Texas at Austin. His research mainly focuses on developing novel models and algorithms for tackling practical challenges in deploying AI systems to various real-world application domains.

Sooel Son is an associate professor of School of Computing at KAIST. He received his Ph.D. in the department of Computer Science at the University of Texas at Austin. He is working on various topics regarding web security and privacy.