Breaking the Black-Box: Confidence-Guided Model Inversion Attack for Distribution Shift

Anonymous Author(s)

ABSTRACT

Model inversion attacks (MIAs) seek to infer the private training data of a target classifier by generating synthetic images that reflect the characteristics of the target class through querying the model. However, prior studies have relied on full access to the target model, which is not practical in real-world scenarios. Additionally, existing black-box MIAs assume that the image prior and target model follow the same distribution. However, when confronted with diverse data distribution settings, these methods may result in suboptimal performance in conducting attacks. To address these limitations, this paper proposes a **C**onfidence-**G**uided **M**odel **I**nversion attack method called CG-MI, which utilizes the latent space of a pre-trained publicly available generative adversarial network (GAN) as prior information and gradient-free optimizer, enabling high-resolution MIAs across different data distributions in a black-box setting. Our experiments demonstrate that our method significantly **outperforms the SOTA black-box MIA by more than 49% for Celeba and 58% for Facescrub in different distribution settings**. Furthermore, our method exhibits the ability to generate high-quality images **comparable to those produced by white-box attacks**. Our method provides a practical and effective solution for black-box model inversion attacks.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problem**.

KEYWORDS

Model Inversion Attack, Gradient-free Optimization, Generative Adversarial Network, Distribution Shift

ACM Reference Format:

Anonymous Author(s). 2024. Breaking the Black-Box: Confidence-Guided Model Inversion Attack for Distribution Shift. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Privacy protection and attacks have been extensively studied, attracting significant attention within the scientific community[24, 33, 34, 41]. Model inversion attacks (MIAs) represent a class of attacks aimed at compromising the privacy protection of models[36].

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

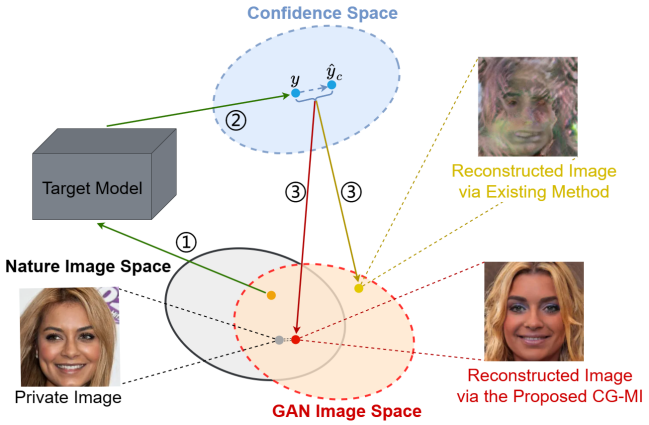

Figure 1: Illustration of private training data leakage for a specific target class c via the target model output confidence: ① Adversary inputs the initially generated image into the target model. ② Adversary obtains the model output confidence y. ③ Adversary attempts to reconstruct the private training data from the confidence y to $\hat{y}_c$ in a black-box setting, where $\hat{y}_c$ is the one-hot vector for class c. The existing method is the optimization result in W space with gradient-free optimizer.

MIAs target the retrieval of sensitive information about the model's training data by leveraging known model outputs, thus putting user privacy at risk. For instance, an attacker may query the output of a facial recognition model and, upon successful exploitation, generate synthetic images that reflect the user's facial features, thereby violating user privacy.

Model inversion attacks can currently be categorized into three types: white-box attacks[4, 29, 37, 39, 43, 44], black-box attacks[1, 2, 12, 32, 42], and label-only attacks[18], based on the level of access the attacker has to the target model. In the context of white-box attacks, the attacker has complete access to the target model, including its knowledge such as weights and output confidence scores. In the case of black-box attacks, the attacker has limited access and can only utilize the confidence scores provided by the target model without any internal knowledge. In the label-only setting, the attacker can only use the output labels provided by the model.

Currently, there has been significant research focus on white-box MIAs [4, 44]. These attack methods require complete access to the target model for performing MIAs. Moreover, these methods assume that the private training data of the target model and the public data used to train the generative model follow the same distribution, which is not practical in real-world scenarios. To address this challenge, Plug & Play Attack(PPA) [37] proposed an independent white-box MIA that works under different data distribution settings. However, it still relies on complete access to the target

model. In the black-box domain, Reinforcement Learning-Based Model Inversion attack(RLB-MI) [12] focused on black-box MIA within the same data distribution using reinforcement learning techniques. However, their attack performance on different data distributions is not satisfactory, and the synthesized images have lower resolution. **Therefore, a crucial challenge in the black-box MIAs scenario is how to generate high-resolution and effective synthetic images solely based on the confidence scores provided by the target model across different data distributions.**

The challenges in this task can be summarized as follows: Firstly, generating high-resolution synthetic images in a high-dimensional latent space poses optimization difficulties. Secondly, the absence of gradient information in black-box model inversion attacks may lead optimization algorithms to explore GANs' latent space excessively, resulting in the generation of images without meaningful features and leading to the failure of the attack. These challenges hinder the direct application of existing white-box attack method[37] to black-box scenarios or limit the attack effectiveness in different data distribution scenarios[12]. The adversary's attack process in a black-box scenario is illustrated in 1.

To address the limitations described above, our paper proposes CG-MI, a novel approach that achieves MIAs in black-box settings with different data distributions. The main idea is to leverage a pre-trained, target-independent generative adversarial network (GAN)[11, 19] as image prior, then employ gradient-free optimization methods to minimize the confidence loss, which measures the matching between the GAN image manifold and the target model. To overcome dimensionality issues and avoid generating meaningless images during the optimization process, we propose a novel objective optimization function. The core idea is to incorporate the mapping network of StyleGAN2[19] into the gradient-free optimization process, thereby ensuring that the solution of the optimization problem remains within a meaningful exploration space. Extensive experiments demonstrate that our method significantly outperforms existing black-box MIAs and its ability to generate high-quality images comparable to white-box attacks. Our main contributions are as follows:

- We present a novel approach to black-box MIAs by utilizing gradient-free optimizer-based method. Our method enables MIAs in black-box scenarios, accommodating various data distributions and generate high-resolution synthesis images.
- We propose the concept of *synthesis image transferability in model inversion*, analysis its impact on MIAs, and address this issue by designing a novel objective optimization function.
- We demonstrate on different datasets and models with the proposed CG-MI. Compared to state-of-the-art black-box attack methods, our approach significantly improves attack performance. Furthermore, visual comparisons indicate comparable synthesis image quality to white-box approaches.

Our work shows that in more challenging scenarios, MIAs still can lead to the leakage of private information from DNNs.

2 RELATED WORK

MIAs can be viewed as an optimization problem, where the objective is to maximize the confidence scores of a given target class in order to generate images that reveal sensitive data features. MIAs were first introduced by [10] for attacking linear regression models. Subsequently, [9] proposed a gradient descent-based algorithm to attack shallow networks. In the following sections, we will introduce recent attack methods based on the types of MIAs.

White-Box MIAs. White-box MIAs leverage full access to the target model and utilize gradient-based optimization techniques. [44] was the first to propose a generative attack method for MIAs, to enable MIAs for deeper networks. They trained a DCGAN [31] on publicly available data that had no overlap with the private training data, and conducted the attack by optimizing the latent input vector z of the GAN to attack the target model trained on the private data. Subsequent work, such as [4], incorporated soft labels of the target model to guide the GAN training and allow the generator to learn the latent distribution, enabling specific GANs for MIAs. Furthermore, [43] introduced a Conditional GAN[40] as an image prior model for MIAs, addressing the issue of the generator in [4] not fully utilizing the target model's knowledge. Additionally, [29] proposed a method that directly maximizes confidence scores instead of minimizing negative log-likelihood scores, aiming to improve the attack performance of [44] and [4]. [39] introduced a variation-based MIAs using StyleGAN2 [22], capable of generating high-resolution images that reflect the target model's private training data. Moreover, [37] introduced a dataset-agnostic MIAs approach utilizing pre-trained StyleGAN2 [19] models. This method targets the vulnerability of prior works to dataset distribution shifts, aiming to address this concern. [30] proposed a novel Dynamic Memory Model Inversion Attack that employs within-class multi-center representation (IMR) and between-class discriminative representation (IDM) terms to model the representation of the target class through memory induction.

Black-Box MIAs. Black-box MIAs require access to the confidence scores of the target model. In black-box MIAs, [32] proposed a method that simultaneously trains a GAN and a surrogate model. The GAN is used to generate inputs similar to the private training data, while the surrogate model imitates the behavior of the target model for inversion attacks. Additionally, [2] attempted to recover faces from deep feature vectors of a face recognition model in a black-box setting without prior knowledge. Another attack model was introduced by [42], where they perform MIAs by swapping the input and prediction vectors of the target model. Furthermore, [1] proposed a black-box MIAs method based on StyleGAN, using a classical genetic algorithm for optimization. Recently, [12] presented a reinforcement learning-based approach for MIAs, where the confidence scores of the target model's outputs serve as rewards.

Label-Only MIAs. Label-only MIAs focus on querying the model to obtain hard labels without confidence scores. [18] introduces an algorithm called Boundary Repulsion Model Inversion (Brep-MI). The core idea of this algorithm is to evaluate the model's predicted labels on a spherical surface and then estimate the direction towards the center of the target class to generate the most representative image. [28] proposed a knowledge transfer-based

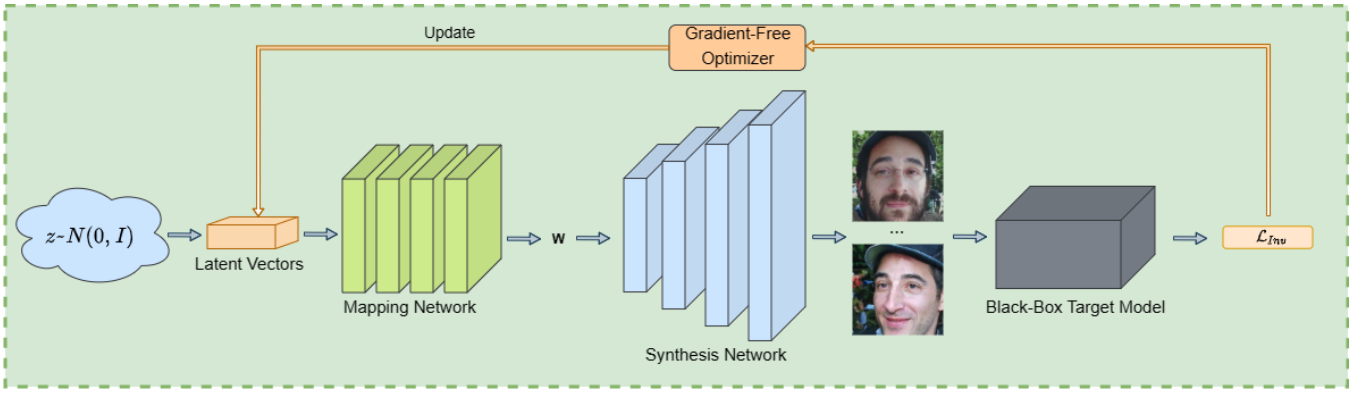

Figure 2: The overview of the proposed attack. Latent vectors z are sampled from a standard normal distribution $N(0, 1)$. These latent vectors are then passed through a mapping network to obtain style vectors w. The style vectors w are subsequently fed into a synthesis network to generate corresponding images. These generated images are further inputted into the target model, and the loss is calculated based on the objective function. The latent vectors z are updated using a gradient-free optimization method. This process continues until we obtain optimized synthesized images.

model inversion attack specifically designed for the label-only scenario.

3 THREAT MODEL

Attack goal. When the target model M is a facial recognition classifier, the aim of MIAs is to exploit the attacker's access to generate facial images that reflect the features of a specific class, represented as $c \in C$, C represents all classes.

Model Knowledge. In white-box MIAs, the attacker possesses the ability to download the model and exploit its weights and confidence information for launching the attack. In contrast, black-box MIAs restrict the attacker to using only the confidence scores provided by the target model. In label-only MIAs, the attacker is limited to utilizing the model's output labels. Our research focus on the black-box MIAs.

Data Knowledge. In the majority of existing white-box MIAs[4, 43, 44] and black-box MIAs [12], they assume that the attacker can launch attacks on data from the same distribution, i.e., $P(X_{prior}) = P(X_{target})$, $P(X_{prior})$ represents the distribution of image priors and $P(X_{target})$ represents the distribution of target model. In our work, we relax the assumption that the attacker is only aware of the targeted model's classification task, such as facial recognition, under the setting where $P(X_{prior}) \neq P(X_{target})$.

4 METHODOLOGY

4.1 Background

Problem Formulation. We define the target classification model as M, with x representing the input image to the target model and c denoting the target class for the attack. To obtain a synthesized image x^* that reflects the private features of the target class c, we optimize the following loss function:

$$x^* = arg\min_x \mathcal{L}\left(M\left(x\right), c\right) \tag{1}$$

Here, $\mathcal{L}$ can be the cross-entropy loss or other suitable loss functions. The purpose of this loss is to directly optimize the image x in order to leak the private training data of the model. As directly optimizing the high-dimensional vector x is not efficient. The following section will delve into generative MIAs.

Generative Model Inversion Attacks. The idea of training a generative model as an image prior to optimize the latent vector z in the GAN for image synthesis was first introduced by [44]. This approach addresses the problem discussed in [9], where directly optimizing x in the high-dimensional, nonlinear, and non-convex solution space when attacking deep neural networks can lead to the generation of meaningless results. Their method involves training a DCGAN [31] on publicly available data that does not overlap with the private training data, and then optimizing the latent input vector of the GAN to attack a target model trained on private data. By introducing the image prior GAN, the optimization problem in equation 1 can be expressed as:

$$z^* = arg\min_z \mathcal{L}\left(M\left(G\left(z\right)\right), c\right) \tag{2}$$

After optimizing equation 2 to obtain z^*, then input z^* into the GAN[31] to generate the synthesized image $x^* = G(z^*)$. This approach helps mitigate the issue of generating meaningless images to a certain extent.

4.2 Breaking the Black-Box

In this section, we will present our approach, Confidence-Guided Model Inversion (CG-MI), for attacking different data distribution models in a black-box scenario. An overview of CG-MI is illustrated in figure 2.

Pre-Trained Publicly Available Image Prior. In the architecture of a GAN[3, 19, 31], the generator model learns to map latent vectors sampled from a simple distribution (e.g., Gaussian or uniform distribution) denoted as z, to the generated image x. However, StyleGAN2[19] consists of two main components: $G_{mapping}$ and $G_{synthesis}$. In StyleGAN2, the latent vector z is first transformed into a style vector w using a non-linear mapping network f, implemented as an 8-layer Multi-Layer Perceptron (MLP). The style

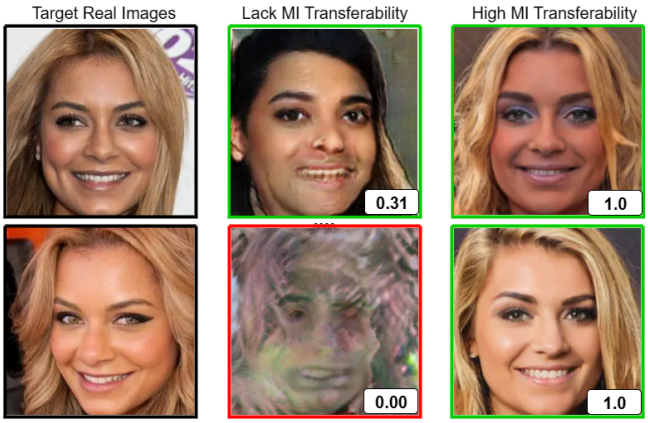

Figure 3: In the comparison between lack synthesis image transferability and high synthesis image transferability in MIAs targeting the same identity. The first image in the second column represents an attack[12] generated using DCGAN, while the second image in the second column represents an attack result achieved by combining a objective function proposed by PPA[37] with gradient-free optimization. The third column displays attack results generated by combining our proposed objective function with gradient-free optimization algorithm. The scores inside the pictures represent the confidence scores provided by the evaluation model.

vector w is then transformed into a synthesized image x. Specifically, a mapping network $G_{mapping} : Z \rightarrow W$, where $z \in Z$ and $z \sim \mathcal{N}(0, 1)$, a synthesis network $G_{synthesis} : W \rightarrow X$, generating the corresponding image x based on the input style vector $w \in W$. Previous works on PPA[37] have demonstrated the tremendous potential of StyleGAN2 across diverse data distributions. In our paper, while ensuring the independence between the generator model and the target model, we leverage a pre-trained publicly available StyleGAN2 model as our image prior to perform attacks between different data distributions.

Synthesis Image Transferability in MIAs. Consider two well-trained face recognition models, denoted as M_1 and M_2, trained on the same dataset $P(X_{target})$. Let x^* be a synthetic image generated through an attack on M_1, classification result as $max(M_1(x^*)) = c$. In the ideal scenario of transferability, the generated image x^* satisfies both $max(M_1(x^*)) = max(M_2(x^*)) = c$. This indicates that both M_1 and M_2 are able to recognize x^* and classify it into the target class c. Conversely, if x^* lacks transferability, it may result in $max(M_2(x^*)) \neq c$, leading to attack failure.

Enhancing Synthesized Images Transferability with Meaningful Exploration. In previous works, such as Brep-MI [18] and RLB-MI [12], the chosen image prior model was DCGAN[31]. However, the overall quality of the synthetic images generated by DCGAN is not satisfactory, as illustrated in figure 3, thereby undermining the transferability of the attack. In the white-box PPA[37] work, StyleGAN2 [19] was selected as a replacement for DCGAN. They achieved success by optimizing the $w[bs, 14, 512]$ vectors, where bs represents the batch size, by leveraging the target model's weights and utilizing a gradient descent algorithm. The objective function for their method is shown in equation 3.

$$w^* = \arg\min_{w} \mathcal{L}(M(G_{synthesis}(w), c)) \tag{3}$$

where $G_{synthesis}$ is the synthesis network of StyleGAN2, c is the target class for the attack, M denotes the target model, w is the style vector of the generative model. However, in a black-box scenario, the use of gradient-free optimization algorithms for the direct optimization of the equation above encounters several model inversion issues, leading to the generation of images lacking meaningful features. The main issue arises because gradient-free optimization algorithms focus solely on minimizing the loss function without considering the preservation of the underlying structure of the GAN latent space. Without constraints, they can deviate from the natural image space of the GAN, leading to synthesized images that receive high confidence scores from the target model while the evaluation model assigns them low confidence scores. Moreover, the high dimensionality of w hampers the efficiency of gradient-free optimization algorithms.

To address the issue of generating images lack meaningful features and to reduce the dimensionality of the optimization variables, we focus on the z vectors before they enter the input mapping network. Regardless of how z is changed during the optimization process, the mapping network consistently maps z to a meaningful latent space, thereby avoiding the generation of meaningless images. In comparision to the high dimensionality of w, z has dimensions of $[bs, 512]$. Specifically, with a pre-trained StyleGAN2, we aim to solve the following optimization problem:

$$z^* = \arg\min_{z} \mathcal{L}_{Inv}(M(G_{synthesis}(G_{mapping}(z)), c)) \tag{4}$$

where $z \in \mathbb{R}^k$, $\mathcal{L}_{Inv}$ is confidence matching loss, $G_{mapping}$ is the mapping network of stylegan2. By incorporating a mapping network into the optimization process and utilizing a gradient-free optimization algorithm, we have successfully transitioned the problem of solving the black-box MIA from the unrestricted latent space to a space characterized by meaningful facial features.

Confidence Matching Loss. The Confidence Matching Loss encourages the solver to find images that can reflect the characteristics of the private training data in the image prior's latent space by minimizing the loss between the target model output confidence $M(x)_c$ and the label c. We explore the following confidence matching loss: (1) Cross-Entropy Loss[4]; (2) Max-Margin Loss[43]; and (3) Poincaré Loss[37]. We performed several comparisons in table 4 and we final chose poincaré loss. For specific details of the loss functions, please refer to Appendix A.

Gradient-free Optimizer. The optimization problem in equation 4 is a non-linear and non-convex problem. Choosing a suitable optimization algorithm is crucial for achieving good performance. In this study, we consider the Covariance Matrix Adaptation Evolution Strategy (CMA-ES)[13], a gradient-free optimization algorithm that is particularly well-suited for high-dimensional problems. CMA-ES is a variant of evolutionary strategies [6] and utilizes an adaptive covariance matrix to optimize the probability distribution. We initiate the optimization process by inputting an initial latent vector z. After obtaining the optimized z^*, we can generate an image x^* that reflects the private training data features with a target model class label of c using $G_{synthesis}(G_{mapping}(z^*))$. Please refer

Algorithm 1: CMA-ES Algorithm for MIA

Input: Population size λ, Learning rate for covariance matrix update c_1, Learning rate for step-size adaptation c_σ, d_σ, Initial mean vector $\mathbf{m}$, Initial step size σ, Covariance matrix $\mathbf{C}$, Fitness based recombination weight w, Target model M, Target class c, Synthesis network $G_{\text{synthesis}}$, Mapping network G_{mapping}, Maximum iterations T

Output: Optimized latent vector z^*

while *current step* $t < T$ **do**
- Generate λ latent vectors $\{z_0, ..., z_{\lambda-1}\}$ from multivariate normal distribution $\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{C})$;
- Evaluate fitness based on $\mathcal{L}(M(G_{\text{synthesis}}(G_{\text{mapping}}(z_i)), c))$, where $z_i \in \{z_0, ..., z_{\lambda-1}\}$;
- Select the top μ latent vectors with the highest fitness;
- Update mean vector:
  - $\mathbf{m} = \mathbf{m} + (1/\mu) \sum_{i=1}^{\mu} w_i \cdot \mathbf{z}_i$;
- Update covariance matrix:
  - $\mathbf{u}_i = \mathbf{C}^{-1} \cdot (\mathbf{z}_i - \mathbf{m})$;
  - $\mathbf{C} = (1 - c_1) \cdot \mathbf{C} + c_1 \sum_{i=1}^{\mu} w_i \cdot \mathbf{u}_i \cdot \mathbf{u}_i^T$;
- Update step size:
  - $\sigma = \sigma \cdot \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\mathbf{C}\|}{E[\|\mathcal{N}(0,\mathbf{I})\|]} - 1\right)\right)$;
- current step t ++;

to Algorithm 1 for details on the working principle of CMA-ES in MIA.

Transformation-based Selection. Following [37], we utilize transformation-based selection to choose the images from the optimized population that best reflect the private training data. Let $x = \{x_1, x_2\}$ denote the set of optimized synthetic images. We define a transformation operation T, which includes scaling, modifying aspect ratio, and random horizontal flipping. By applying T to x_1 and x_2, we obtain transformed images, $T(x_1)$ and $T(x_2)$, which are then input to the target model, yielding new confidence scores, l_1 and l_2. Typically, if x_1 captures the target class features more effectively than x_2, the confidence scores after transformations satisfy $l_1 > l_2$.

$$E\left[M\left(T\left(x\right)\right)_c\right] \approx \frac{1}{N} \sum_{i=1}^{N} M\left(T\left(x\right)\right)_c \tag{5}$$

Hence, the image with the highest logit after transformations, denoted by x_1, is selected as the final image. In equation 5, the Monte Carlo estimation method are employed, where N represents the number of applied transformations.

5 EXPERIMENTS

This section begins with a comprehensive explanation of our experimental setup. Subsequently, we assess the effectiveness of our CG-MI attack by considering different factors such as the performance of various MIA methods in different scenarios, different datasets and different target models.

5.1 Experimental Settings

Datasets. In our experiments, we focus on five datasets: CelebA[25], FFHQ[21], FaceScrub[27], AFHQ Dogs[5], Metfaces[20] and Stanford Dogs[23]. For the purpose of conducting attack evaluations, we trained our target models on CelebA, FaceScrub, and Stanford Dogs datasets. As a prior for image synthesis, we utilized pre-trained stylegan2 models on FFHQ, Metfaces and AFHQ Dogs datasets, enabling us to perform attacks on different data distributions.

Models. Our models are divided into target models and evaluation models. To facilitate a fair comparison, we conducted attack experiments on several popular network architectures, including Resnet[15] and DenseNet[17], while selecting Inception-V3[38] as the evaluation model. For the facial recognition task, we trained Resnet18, Resnet152, Densenet169, and InceptionV3 models on the CelebA and FaceScrub datasets, respectively. Similarly, for the dog breed classification task, we trained Resnet18, Resnet152, Densenet169, and InceptionV3 models on the Stanford dogs dataset. To facilitate comparison with prior work, we attacked the Resnet18 model trained on the CelebA and FaceScrub datasets for comparative experiments. The details of the data partition used for training the target model and the parameters for model training, the attack process, and comparative experiments can be referenced in section 5.1.

Attack Implementation. All models were trained using the Adam optimizer with a learning rate of 0.001 and β values of $[0.9, 0.999]$ for a total of 100 epochs with a batch size of 128. The training data was normalized with mean μ and standard deviation σ set to 0.5. The input images of the target model were resized to 224×224, and the evaluation model InceptionV3 was resized to 299×299. Additionally, data augmentation techniques were applied, including 50% horizontal flipping and adjustments in brightness and contrast within the range $[0.8, 1.2]$, saturation within $[0.9, 1.1]$, and hue within $[-0.1, 0.1]$.

During the attack process, the StyleGAN2 synthesis network was configured with a truncation parameter ψ of 0.5 and a truncation cutoff value of 8. The CMA-ES algorithm was employed with 8 rounds, 300 maximum iterations, and a population size of 25. For other parameters, we maintain the settings as specified in [14]. The rotation transformation selection strategy involved 100 transformations. It included center cropping of images generated by the generative model, resizing them to 224, and applying random adjustments to cropping parameters within the ranges of size $[224, 224]$, scale $[0.5, 0.9]$, and ratio $[0.8, 1.2]$. To accelerate the multiple CMA algorithm optimizations, 8 parallel processes were utilized. In the extended experiments, we performed central cropping to 800 on the images generated by Metfaces for Celeba and Facescrub datasets, followed by resizing to 224. For the remaining different architectures of target models, we maintained consistent attack parameters.

In the comparative experiments to perform RLB-MI and Brep-MI, the latent vector size in the GAN was set to 100. For GAN training, the ADAM optimizer was used with a batch size of 64, a learning rate of 0.0002, and β values of $[0.5, 0.999]$. The training was conducted for 280 epochs.

Evaluation Metrics. To align with prior work, we followed PPA[37] to calculate various evaluation metrics. Firstly, we trained

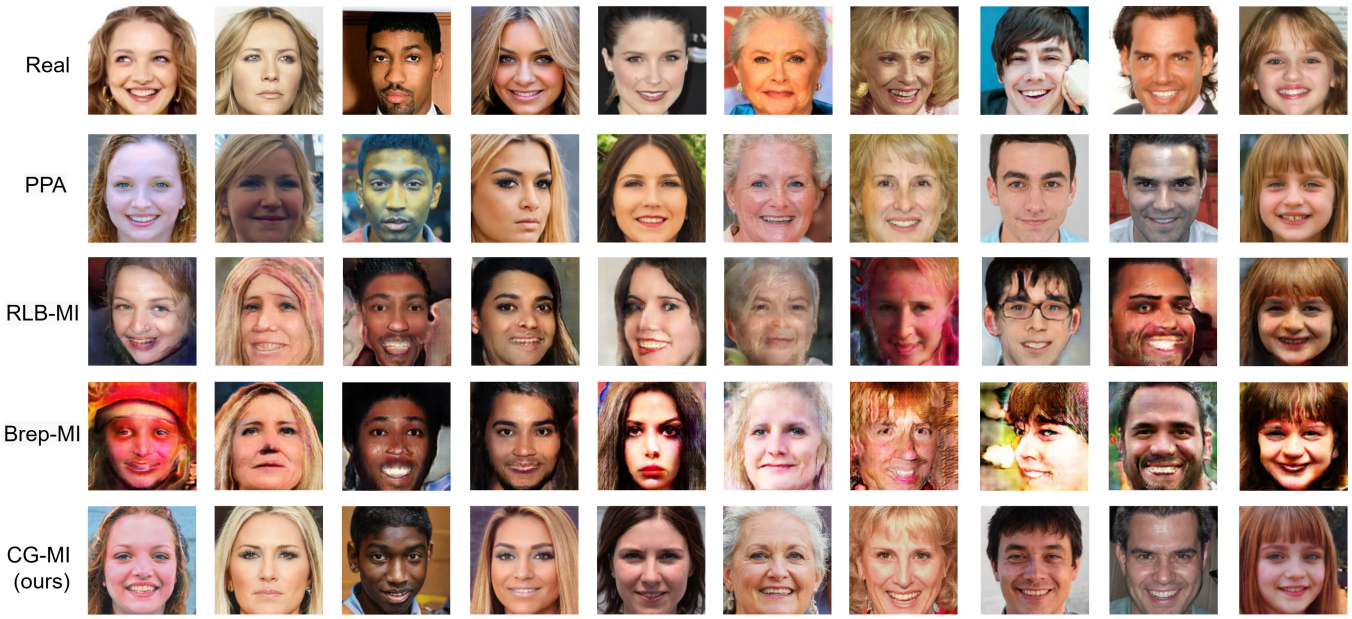

Figure 4: We present a visual comparison of the attack results for different methods in the scenario where the $P(X_{prior})$ = FFHQ, $P(X_{target})$ = CelebA and the target model architecture is Resnet18. The first row shows ground truth images of target class. The second row represents PPA[37], the third row represents RLB-MI[12], and the fourth row represents Brep-MI[18]. The last row introduces our proposed method, CG-MI.

an independent Inception-v3 evaluation model on the training data of the target model. Then, we used the evaluation model to predict labels on the generated attack results and computed the TOP-1 and TOP-5 accuracy for the target class.

Next, we computed the shortest feature distance from each generated image to any training sample in the target class and denoted the average distance as δ_{eval}. The distance was measured using the squared L2 distance between activation layers in the penultimate layer of the evaluation model. For facial images, we utilized a pre-trained FaceNet model [35] to measure the feature distance δ_{Face}. Lower values indicate that the attack results are visually closer to the training data.

The third metric is the FID (Fréchet Inception Distance) score [16]. FID calculates the distance between the feature vectors of the generated attack results and the training data of the target. The feature vectors are extracted using an Inception-v3 model trained on ImageNet [7]. A lower FID score indicates a higher similarity between the two datasets.

5.2 Experimental Results

Comparision with Previous MIAs Approachs. We compared CG-MI with various MIAs methods in different scenarios, including white-box, black-box, and MIAs in the label-only setting. For white-box MIAs, we used PPA[37] as the baseline method. In contrast to previous white-box attack methods (such as [4, 43, 44]), PPA focuses on scenarios with different data distributions, image priors, and independence from the target model, making it more meaningful for comparison. For the black-box attack scenario, we selected RLB-MI[12] and Brep-MI[18] as baseline methods, representing state-of-the-art MIA methods in black-box attacks and the label-only setting, respectively.

To ensure fair comparison, we trained the Resnet18 model on the CelebA and Facescrub datasets for conducting comparative experiments using different MIAs methods. In our CG-MI method, we first ran the CMA-ES optimization algorithm multiple times for each class, generating a batch of 200 synthetic images. This approach ensured the stability and reliability of the results and reduced the impact of randomness. Next, we adopted a transformation selection strategy to choose the most representative 50 images from the optimized batch of 200 synthetic images for evaluation. For the other MIAs methods, we combined the characteristics of each attack method and ran it multiple times to generate a total of 200 synthetic images. Additionally, these methods also incorporate a transformation-based selection strategy to choose 50 images. We then used the same evaluation models and metrics to assess all the methods.

It is worth mentioning that RLB-MI and Brep-MI both employ the same GAN[31] structure, which is designed specifically for generating low-resolution 64x64 pixel images. To ensure a fair comparison, followed PPA[37], we made adjustments to the aforementioned methods by using a deeper GAN architecture capable of generating higher-resolution images[37]. For the training of DCGAN in the baseline, we follow the standard experiment setup for GMI as outlined in the PPA. This decision is based on the fact that GMI, along with RLB-MI and Brep-MI, employs the DCGAN architecture. PPA has also deepened the DCGAN structure for GMI, enabling

	Type	Method	↑acc@1	↑acc@5	↓δ_{face}	↓δ_{eval}	↓FID
CelebA	White-box	PPA	**88.28%**	**97.34%**	0.6992	283.89	40.43
	Black-box	RLB-MI	29.25%	53.77%	1.0740	358.18	101.86
		CG-MI(Ours)	**77.86%**	**94.16%**	0.7465	292.14	46.66
	Label-only	Brep-MI	38.50%	61.25%	0.9700	356.83	93.05
Facescrub	White-box	PPA	**98.32%**	**99.84%**	0.6735	107.35	45.73
	Black-box	RLB-MI	33.28%	64.52%	1.1097	135.16	111.06
		CG-MI(Ours)	**90.92%**	**99.34%**	0.7570	111.75	62.24
	Label-only	Brep-MI	51.33%	73.82%	1.0664	132.59	102.94

Table 1: Different MIA methods were applied to attack a Resnet18 model trained on CelebA and Facescrub datasets with $P(X_{prior})$ = FFHQ. In the black-box scenario, CG-MI significantly outperforms existing methods.

inversion at the 256x256 resolution. Firstly, we added two additional upsampling blocks (consisting of a transpose convolution layer and a batch normalization layer) to the generator. We also expanded the discriminator by adding two convolution blocks, with each block consisting of a convolution layer and an instance normalization layer. Subsequently, we trained the modified generator on the FFHQ256 dataset to generate 256x256 pixel images. We then utilized this enhanced GAN as the image prior for RLB-MI and Brep-MI. By adapting the GAN architecture and training on higher-resolution images, we ensured that the comparison between CG-MI and RLB-MI/Brep-MI was carried out on a level playing field.

Table 1 presents the evaluation results of CG-MI and baseline methods in attacking Resnet18 trained on Celeba and Facescrub datasets. The target model Resnet18 achieves test accuracies of 86.38% and 94.22% on Celeba and Facescrub, respectively, while the evaluation model InceptionV3 achieves test accuracies of 93.28% and 96.20% on the same datasets. Based on the data in table 1, CG-MI outperforms other black-box methods in addressing distribution shift issues in black-box attack scenarios. It generates more transferability synthetic images, resulting in higher attack success rates, lower feature distances, and FID values.

The qualitative evaluation results shown in figure 4 demonstrate that compared to previous black-box MIA methods, CG-MI is capable of generating more realistic images in different data distribution scenarios. It overcomes the limitations of distribution shift and produces synthetic images of comparable quality to state-of-the-art white-box methods.

Performance Evaluation on various Models and Datasets. We also evaluated the performance of CG-MI on deeper network architectures and datasets from different categories. Specifically, we trained Resnet152 and Densenet169 models on the CelebA, Facescrub, and Stanford Dogs datasets. Additionally, we trained Resnet18 models on the Facescrub and CelebA datasets, with the recognition accuracy on the respective test datasets indicated in parentheses.

The objective of this experiment was to explore the potential for our method to generalize to other model architectures or deeper network structures. It also aimed to evaluate the performance of

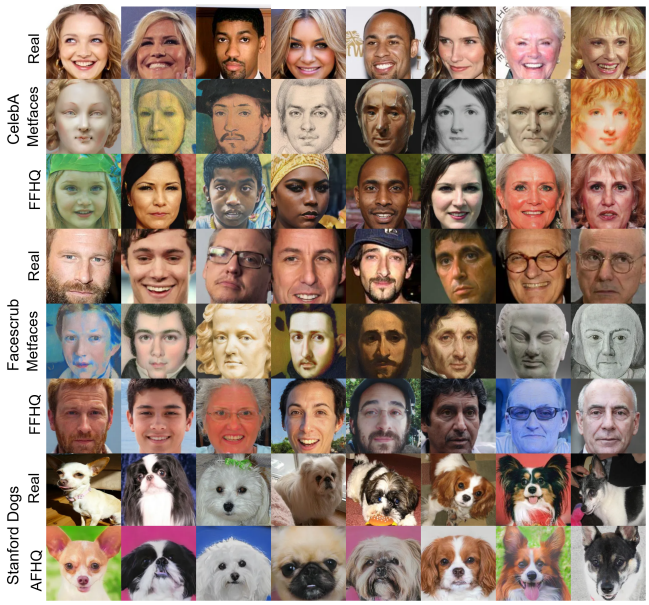

Figure 5: We have visualized the CG-MI attack results on the Densenet169 network architecture for the CelebA, Facescrub, and Stanford Dogs datasets.

$P(X_{prior}) \rightarrow P(X_{target})$	Target Model	(↑acc@1,↑acc@5)	(↓δ_{face}, ↓δ_{eval}, ↓FID)
FFHQ → Celeba	Resnet152 (86.78%)	(**67.42%**,86.16%)	(**0.7773, 319.42, 46.04**)
FFHQ → Celeba	Densenet169 (85.39%)	(65.28%, **87.72%**)	(0.7831, 321.16, 47.91)
Metfaces → Celeba	Resnet18 (86.38%)	(24.42%, 49.26%)	(1.2089, 420.75, 111.40)
FFHQ → Facescrub	Resnet152 (93.74%)	(85.02%, 97.94%)	(0.7998, 122.53, 64.17)
FFHQ → Facescrub	Densenet169 (95.49%)	(**90.78%, 97.95%**)	(**0.7608, 115.02, 63.05**)
Metfaces → Facescrub	Resnet18 (94.22%)	(59.16%, 88.40%)	(1.0286, 133.42, 101.02)
AFHQ.dogs → Stan.dogs	Resnet152 (71.23%)	(**84.80%, 99.04%**)	(-, **60.50, 58.04**)
AFHQ.dogs → Stan.dogs	Densenet169 (74.39%)	(83.06%, 98.02%)	(-, 63.51, 59.03)

Table 2: The attack results of CG-MI on different network architectures and datasets are evaluated. We employ the prior distribution $P(X_{prior})$ = FFHQ to attack the target distribution $P(X_{target})$ = Celeba and Facescrub, $P(X_{prior})$ = AFHQ.dogs to attack $P(X_{target})$ = St.dogs and $P(X_{prior})$ = Metfaces to attack $P(X_{target})$ = Celeba and Facescrub.

Optimizer	↑acc@1	↑acc@5	↓δ_{face}	↓δ_{eval}	↓FID
BO	54.65%	85.26%	0.9511	135.06	72.67
PSO	64.18%	89.32%	0.8976	130.10	68.32
CMA-ES	**90.92%**	**99.34%**	**0.7570**	**111.75**	**62.24**

Table 3: Attack performance with various gradient-free optimizer.

CG-MI in scenarios involving significant data distribution shifts. From the data presented in table 2, it is evident that as the target model becomes structurally deeper, CG-MI encounters increased

difficulty when attacking CelebA and Facescrub. Consequently, this results in a slight reduction in attack effectiveness. In settings with significant data distribution shifts, such as using Metfaces to attack Celeba and Facescrub, CG-MI still maintains a certain level of attack effectiveness. In figure 5, the visual experimental results illustrate that CG-MI continues to produce meaningful attack results, even when dealing with deeper network architectures, different classification tasks, and scenarios involving significant data distribution shifts.

Experiments with Various Gradient-free Optimizer. We compared different gradient-free optimization algorithms, namely PSO[26] and BO[8], under the setting of $P(X_{prior}) = FFHQ$ and $P(X_{target}) = Facescrub$, with a target model architecture of Resnet18. Bayesian Optimization (BO) is a global optimization algorithm that utilizes probabilistic models to efficiently search for the optimal solution of an expensive black-box function. PSO is an optimization technique inspired by social behaviors of swarms, used for finding optimal solutions efficiently. The CMA-ES algorithm exhibited the superior attack performance in comparison with other optimization algorithms.

5.3 Ablation Study

In the ablation study, we considered two probability distributions: $P(X_{target})$, which represents the distribution of images from the CelebA dataset, and $P(X_{prior})$, which represents the distribution of images from the FFHQ dataset. We first compared three different loss functions for the model inversion attack: Poincaré loss[37], max-margin loss[43], and cross-entropy loss[4]. Poincaré loss achieved the best performance in terms of generating effective synthesis images. We also evaluated the performance of the PSO algorithm on the CelebA dataset.

The experiments involved also replacing the StyleGAN2 architecture with the DCGAN architecture combine our proposed attack CG-MI. Ablation study results indicated that CG-MI is compatible with other GAN architectures. Our findings demonstrate that, even when replacing StyleGAN2 with the DCGAN, CG-MI still yields favorable evaluation results. We also conducted ablation experiments by employing the objective function proposed in [37] in combination with the gradient-free optimization algorithm CMA-ES. From the experimental results(**No Mapping**), we observed that in the black-box scenario, where gradient information is unavailable, directly using the objective function from PPA did not yield favorable attack results. By using our newly proposed objective function, we significantly improved our attack capabilities in the black-box scenario.

We also investigated the influence of transformation-based selection techniques[37] on the attack outcomes(**No Trans. Selection**). By employing selection transformations, we can enhance the stability of attack outcomes and, to some extent, increase the success rate of attacks. In the process of adjusting the hyperparameters of the CMA-ES, setting the maximum iterations to 200 resulted in suboptimal outcomes compared to 300 iterations. Furthermore, a population size of 50 showed similar optimization results to 25 but with significantly increased computational time. Reducing the population size to 20 diminished the attack effectiveness. A population size of 25 optimally trades off between performance and computational efficiency.

	↑acc@1	↑acc@5	$\downarrow\delta_{\text{face}}$	$\downarrow\delta_{\text{eval}}$	↓FID
Poincare Loss	**77.86%**	**94.16%**	**0.7465**	**292.14**	**46.66**
Cross-Entropy Loss	62.77%	85.67%	0.8550	329.71	51.81
Max-Margin Loss	71.50%	91.25%	0.7827	331.48	49.02
PSO	42.04%	72.40%	0.8954	339.78	51.37
DCGAN	32.80%	66.50%	0.9526	322.88	91.32
No Mapping	**00.03%**	**00.09%**	1.4894	435.82	201.02
No Trans. Selection	72.79%	89.31%	0.7688	300.11	47.28
Max. Iterations 200	72.04%	90.22%	0.7678	299.09	47.10
Population Size 50	77.32%	**94.24%**	0.7495	293.71	48.50
Population Size 20	75.90%	93.04%	0.7576	296.29	**45.07**

Table 4: Ablation study performed on a Resnet18 trained on CelebA using the FFHQ StyleGAN2 as image prior.

6 DISCUSSION, LIMITATIONS AND CONCLUSION

Our paper proposes a novel black-box attack method called CG-MI. Unlike existing black-box methods, we focus on considering more realistic scenarios without making assumptions about the data distribution of the target model. Our approach only requires knowledge about the model's classification task and leverages pre-trained publicly available images' prior knowledge to attack various target models and generate high-resolution synthetic images. Furthermore, we introduce the concept of synthetic images transferability and investigate its impact on in MIAs. By designing a novel objective function and combining gradient-free optimization methods, we achieve MIAs in black-box scenarios and enhance the transferability of the synthesized images. Experimental results demonstrate that CG-MI outperforms existing black-box MIAs in more realistic scenarios, achieving state-of-the-art attack performance.

However, the current black-box MIAs, including our work, still have some limitations. While black-box methods do not require full access to the target model, frequent queries to the target model may hinder the progress of the attack in real-world settings. Therefore, exploring how to control the query count and ensure attack success rate is a worthwhile research direction.

It is worth noting that our research may have negative implications. However, the purpose of revealing vulnerabilities in existing systems is to promote the development of better defense mechanisms. Our work aims to call for attention from the academic and technical community to research on machine learning privacy protection. We believe that the positive impact of these efforts will outweigh the potential negative risks.

ACKNOWLEDGEMENTS

REFERENCES

[1] Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. 2022. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*.

[2] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther. 2019. GAMIN: An Adversarial Approach to Black-Box Model Inversion. *Cornell University - arXiv,Cornell University - arXiv* (Sep 2019).

[3] AndrewS. Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *International Conference on Learning Representations* (Sep 2018).

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

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00821

[6] Swagatam Das and Ponnuthurai Nagaratnam Suganthan. 2011. Differential Evolution: A Survey of the State-of-the-Art. *IEEE Transactions on Evolutionary Computation* (Feb 2011), 4–31. https://doi.org/10.1109/tevc.2010.2059031

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2009.5206848

[8] David Eriksson, Michael Pearce, JacobR. Gardner, Ryan Turner, and Matthias Poloczek. 2019. Scalable Global Optimization via Local Bayesian Optimization. *Neural Information Processing Systems(NeurIPS)* (Jan 2019).

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

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

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2017. Generative Adversarial Nets. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics* (Oct 2017), 177–177. https://doi.org/10.3156/jsoft.29.5_177_2

[12] Gyojin Han, Jaehyun Choi, Haeil Lee, and Junmo Kim. 2023. Reinforcement Learning-Based Black-Box Model Inversion Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20504–20513.

[13] Nikolaus Hansen. 2016. The CMA evolution strategy: A tutorial. *Towards a new evolutionary computation* (2016), 75–102.

[14] Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. 2019. CMA-ES/pycma on Github. Zenodo, DOI:10.5281/zenodo.2559634. https://doi.org/10.5281/zenodo.2559634

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Neural Information Processing Systems (NeurIPS)* (Jan 2017).

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.243

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

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. *Cornell University - arXiv* (Jun 2020).

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. *Neural Information Processing Systems,Neural Information Processing Systems* (Jan 2020).

[21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.00453

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00813

[23] Aditya Khosla, N Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel Dataset for Fine-Grained Image Categorization. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*.

[24] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V Vasilakos. 2020. Privacy and security issues in deep learning: A survey. *IEEE Access* 9 (2020), 4566–4593.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.425

[26] Federico Marini and Beata Walczak. 2015. Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems* 149 (2015), 153–165.

[27] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*. https://doi.org/10.1109/icip.2014.7025068

[28] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. [n. d.]. Label-Only Model Inversion Attacks via Knowledge Transfer. ([n. d.]).

[29] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. 2023. Re-thinking Model Inversion Attacks Against Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16384–16393.

[30] Gege Qi, YueFeng Chen, Xiaofeng Mao, Binyuan Hui, Xiaodan Li, Rong Zhang, and Hui Xue. 2023. Model Inversion Attack via Dynamic Memory Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5614–5622.

[31] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)* (Jan 2016).

[32] Anton Razzhigaev, Klim Kireev, Edgar Kaziakhmedov, Nurislam Tursynbek, and Aleksandr Petiushko. 2020. Black-box face recovery from identity features. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 462–475.

[33] Maria Rigaki and Sebastián García. 2020. A Survey of Privacy Attacks in Machine Learning. *Cornell University - arXiv,Cornell University - arXiv* (Jul 2020).

[34] Maria Rigaki and Sebastián García. 2020. A Survey of Privacy Attacks in Machine Learning. *Cornell University - arXiv* (Jul 2020).

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7298682

[36] Junzhe Song and Dmitry Namiot. 2022. A Survey of the Implementations of Model Inversion Attacks. In *International Conference on Distributed Computer and Communication Networks*. Springer, 3–16.

[37] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. 2022. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks. In *Proceedings of the 39th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research)*. PMLR, 20522–20545.

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.308

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

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 8798–8807.

[41] Marius Wernke, Pavel Skvortsov, Frank Dürr, and Kurt Rothermel. 2014. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing* 18 (2014), 163–175.

[42] Ziqi Yang, Ee-Chien Chang, and Zhenkai Liang. 2019. Neural network inversion in adversarial setting via background knowledge alignment. *In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19* (2019), 225–240.

[43] Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. 2023. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. *AAAI Conference on Artificial Intelligence(AAAI)* (2023).

[44] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 253–261.