Prediction Exposes Your Face: Black-box Model Inversion via Prediction Alignment

Yufan Liu[1,2], Wanqian Zhang[1✉], Dayan Wu[1], Zheng Lin[1,2], Jingzi Gu[1], and Weiping Wang[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{liuyufan,zhangwanqian,wudayan,linzheng,gujingzi,wangweiping}@iie.ac.cn

Abstract. Model inversion (MI) attack reconstructs the private training data of a target model given its output, posing a significant threat to deep learning models and data privacy. On one hand, most of existing MI methods focus on searching for latent codes to represent the target identity, yet this iterative optimization-based scheme consumes a huge number of queries to the target model, making it unrealistic especially in black-box scenario. On the other hand, some training-based methods launch an attack through a single forward inference, whereas failing to directly learn high-level mappings from prediction vectors to images. Addressing these limitations, we propose a novel Prediction-to-Image (P2I) method for black-box MI attack. Specifically, we introduce the Prediction Alignment Encoder to map the target model's output prediction into the latent code of StyleGAN. In this way, prediction vector space can be well aligned with the more disentangled latent space, thus establishing a connection between prediction vectors and the semantic facial features. During the attack phase, we further design the Aligned Ensemble Attack scheme to integrate complementary facial attributes of target identity for better reconstruction. Experimental results show that our method outperforms other SOTAs, e.g., compared with RLB-MI, our method improves attack accuracy by 8.5% and reduces query numbers by 99% on dataset CelebA.

Keywords: Model Inversion · Prediction Alignment · Aligned Ensemble Attack

1 Introduction

Deep neural networks (DNNs) have been widely applied in various scenarios such as finance, healthcare and autonomous driving. Despite the great success on downstream tasks, the collection of DNNs' training data inevitably involve private and sensitive personal information. Malicious people can launch various attacks on DNNs to steal users' private information [30], which may pose serious threats to user privacy [13, 31, 35], especially on sensitive information like face

✉Corresponding author

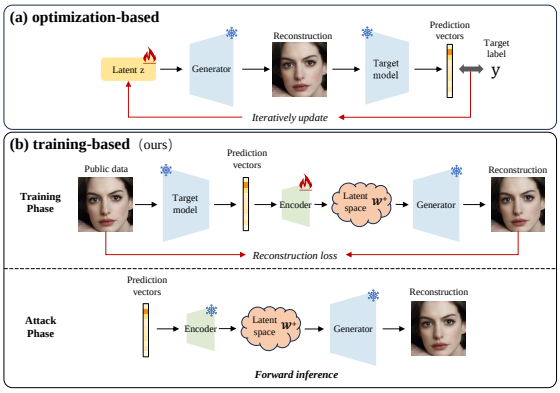

Fig. 1: Previous optimization-based methods iteratively update latent vector z within a fixed prior generator, involving enormous query numbers to target model. Differently, our method works in a training-based manner, optimizing a prediction-to-image inversion model and reconstructing face images through a simple forward inference.

images [4,23]. In this paper, we focus on the model inversion (MI) attack, a representative privacy attack paradigm that reconstructs the training data of target model. Specifically, once obtaining the access to target model and the output predictions, the adversary can attack a face recognition system to reconstruct sensitive face images.

Having access to target model's parameters, white-box MI attacks [8,43,46] recover high-fidelity private images by searching the latent space of generative networks. While in black-box scenario, where only the output prediction scores are available, the adversaries typically adopt genetic algorithm [5] or reinforcement learning [16] to find the optimal latent vector. In the most challenging label-only scenario [18], the attackers only have predicted hard labels, thus resort to estimation of the gradient direction away from decision boundaries. Despite these diversities, common and central to them is an optimization procedure, i.e., searching the input space to find the exact feature value with maximum likelihood under the target model.

However, as in Fig. 1, this optimization-based paradigm updates the input vector in an iterative way, which inevitably involves enormous queries to target model. This is impractical in real-world especially the query-limited scenarios like online machine learning services such as Amazon Rekognition and Google's cloud vision API. For instance, RLB-MI [16] takes 40,000 epochs (nearly 120k queries) to reconstruct only one identity. On the other hand, Yang *et al.* [39] consider the target model as an encoder and train a decoder separately to reconstruct images. This vanilla training scheme introduces a simple-structured model, which fails to directly learn high-level mappings from prediction vectors to images. In other words, it is difficult for a shallow inversion model, usually 3-4 convolutional layers, to provide disentangled image features, thus leading to unsatisfactory attack performance.

In this paper, we propose the Prediction-to-Image (P2I) inversion model, generating target identity's images through a simple forward inference. As in Fig. 1, previous optimization-based methods perform the iterative search for high-likelihood reconstruction from the target model, which could be particularly challenging for high-dimensional search space. Differently, our method reconstructs face images directly in a generative and disentangled manner. Specifically, the P2I model consists of a prediction alignment encoder and a StyleGAN generator, which maps the predictions to the StyleGAN's $\mathcal{W}^+$ space and to the image space. This *prediction-$\mathcal{W}^+$-image* scheme successfully aligns the prediction vector space with the disentangled $\mathcal{W}^+$ space. Subsequently, this alignment provides connections between the prediction vector space and image space, contributing to semantically continuous face embeddings on the target identity. Besides, to provide more prior knowledge for training, we formulate the training set by public images with highest probabilities for each target identity. This public data selection implicitly preserves the facial attribute overlaps between public and private images, preserving rich semantic information for the P2I model. During the attack phase, we further propose the aligned ensemble attack, integrating public images' latent codes of $\mathcal{W}^+$ space and utilizing the contained target's facial attributes for better reconstruction. Empirically, our method shows a significant boost in black-box MI attack accuracy, visual quality and reduced query numbers.

Our contributions can be summarized as follows:

- We propose a novel Prediction-to-Image (P2I) method for black-box model inversion attack. By integrating the proposed prediction alignment encoder with StyleGAN's generator, P2I aligns the prediction vector space with the disentangled $\mathcal{W}^+$ space, providing semantically continuous face embeddings for the target identity.
- We propose the aligned ensemble attack scheme to incorporate rich and complementary facial attributes of public images within the $\mathcal{W}^+$ space, further improving the inversion performance.
- Extensive experiments show the effectiveness of our method compared with other baselines across various settings like different target models and target datasets, indicating the superiority of our method.

2 Related Works

2.1 Model Inversion Attack

Model Inversion (MI) attacks leverage the target model's output to reconstruct the training data, putting machine learning models at risk of data privacy leakage. Fredrikson *et al.* [14] first propose the MI attack on pharmacogenetic privacy issues, which however easily sticks into local minima due to direct optimization in the high-dimensional image space [13]. Recent works [5, 8, 16, 18, 26, 27, 33, 36, 42–44, 46] incorporate GANs into the pipeline and optimize GAN's latent space instead of image space. Yuan *et al.* [43] introduce conditional GAN to find a fixed

search space for each category, greatly narrowing down the search space. Han *et al*. [16] focus on black-box MI attack and formalize the latent space search as a Markov Decision Process (MDP) problem. Kahla *et al*. [18] first explore the label-only MI attack and sample multiple points to estimate the gradient of a random vector. Nguyen *et al*. [27] study the issues of optimization objective and overfitting for a generic performance boost of all MI algorithms. However, these methods need enormous iterations and queries to find a latent code when each single attack, which is obviously time-consuming and unrealistic.

In contrast to optimization-based methods, Yang *et al*. [39] first propose a training-based approach. These methods [39, 41, 48] generally requires training an additional inversion model, using the output of the target model as input and the images as output. During the attack phase, attackers only need to input the target model's output representing the target identity, and then reconstruct the image through one forward propagation.

2.2 GAN Inversion

The goal of GAN inversion is to encode a given image into the GAN's latent space, and then invert the latent code to obtain the reconstructed high-fidelity image. GAN inversion is broadly categorized into three types [38]: optimization-based, encoder-based, and the hybrid. The optimization-based methods [1, 2, 6, 11, 15] minimize the error between input image and reconstructed image with gradient descent, while the encoder-based ones [3, 10, 19, 22, 29, 34, 37, 40] train an encoder to map the real image into the latent space as latent code, and the inversion can be achieved by a one-time forward propagation. The hybrid method [47] first trains the encoder to get the latent code and then optimizes it later. Notably, the premise of GAN inversion is that the performer has a real image, which is essential for the task of image reconstruction. In MI attack, however, we aim to utilize the target model's output prediction to reconstruct a representative image of the target identity in the training dataset.

3 Method

3.1 Problem Formulation

Attacker's goal. Consider a target model $T : \mathcal{X} \to \mathcal{P}$ mapping from image space $\mathcal{X}$ to prediction vector space $\mathcal{P}$, which is trained on a private dataset $D_{priv} = \{(\tilde{x}_i, \tilde{p}_i)\}_{i=1}^{N_{priv}}$, where N_{priv} is the total number of private samples, $\tilde{x}_i \in \mathbb{R}^d$ is the input image and $\tilde{p}_i \in \mathbb{R}^C$ is the corresponding prediction vector, C is the total number of classes. In this work, similar as in [16], the target model is specified as the face recognition model, and the attacker aims to recover a representative face image of a given identity. Formally, the objective of our method is to learn an inversion model that can correctly map the output prediction to its corresponding target identity's image.

Attacker's knowledge. Our work focuses on the black-box scenario where the attacker does not know neither the internal structure nor the model parameters,

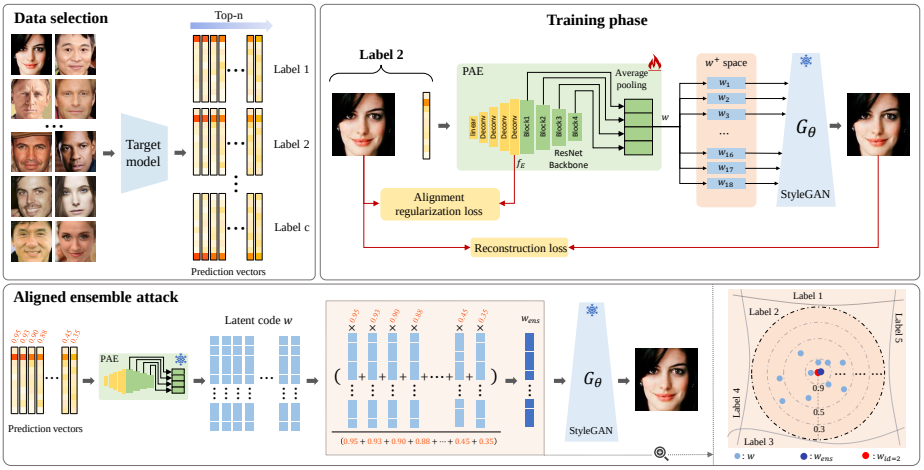

Fig. 2: Overall pipeline of P2I method. We first form training data by selecting top-n public images with highest confidence for each identity. The Prediction Alignment Encoder (PAE) maps prediction vectors into the latent code of disentangled $\mathcal{W}^+$ space, which are then fed into the fixed StyleGAN's generator to reconstruct high-fidelity target image. Furthermore, we introduce aligned ensemble attack to integrate different w, which essentially aims to find the centroid w_{ens} and make it closer to the target identity's w_{id}, contributing to better attack performance.

yet can only obtain the model's output predictions, i.e., the confidence scores for each class. Though the attacker has no access to the private dataset, it is reasonable to assume that he knows what task the model performs, and can easily collect task-related public dataset D_{pub} from the Internet for training [16,43,46]. Note that there is no identity overlaps between the public and private datasets.

3.2 Prediction Alignment Encoder

Fig. 2 shows the overall framework of our method. We first form the training set by publicly collected top-n images with highest probabilities for each identity. During the training phase, Prediction Alignment Encoder maps prediction vector into disentangled $\mathcal{W}^+$ space, followed by the fixed StyleGAN's generator to reconstruct high-fidelity target image. This *prediction-$\mathcal{W}^+$-image* scheme establishes connections between prediction vector space and the image space, providing semantically continuous face embeddings on the target identity.

Recently, as one of the excellent works on GAN inversions, [40] utilizes StyleGAN to reconstruct images by representing visual attributes with different latent dimensions within the disentangled $\mathcal{W}^+$ space. Here, motivated by this, we raise the following questions for MI attack: Instead of the optimization-based paradigm with high cost and low efficiency, *Can we* directly train an inversion model to reconstruct images of any given identity through a simple forward in-

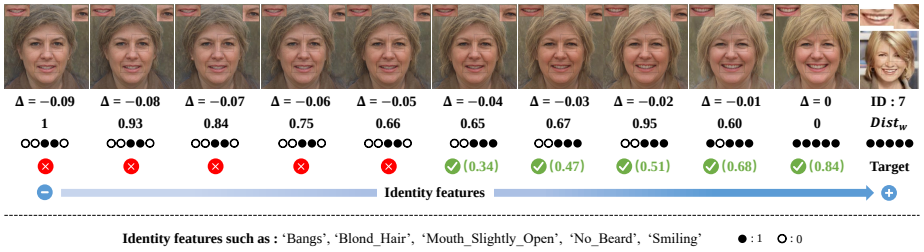

Fig. 3: Visualizations of the interpolation on prediction vectors along the target dimension. As prediction value increases, reconstructed image gradually approaches target visual appearance. This is consistent with the decreasing normalized distance $Dist_w$ between the latent codes of target and reconstructed image. Besides, results on facial attribute classifications and identity recognition (especially the zoom-in parts of mouth and eye) also justify the *prediction-$\mathcal{W}^+$-image* alignment.

ference? *Can we* further align the prediction vector space with the $\mathcal{W}^+$ space, connecting between prediction vectors and disentangled facial attributes?

Thus, we propose the Prediction-to-Image inversion model, consisting of a Prediction Alignment Encoder E and a StyleGAN generator G. Specifically, given the input prediction p of image x, the Prediction Alignment Encoder aims to learn the mapping $E : \mathcal{P} \rightarrow \mathcal{W}^+$, where $p \in \mathcal{P}$ and $w \in \mathcal{W}^+$, such that the output $G(E(p)) \approx x$. Concretely, prediction vector will first pass through one linear layer and several deconvolution layers, generating a feature f_E with the same size of image. Next, it will go through a ResNet backbone and the output features from each block are connected and flattened into a tensor of $1 \times 8,640$ size after the average pooling layer. Then, it is mapped to a $1 \times 18 \times 512$ latent code in $\mathcal{W}^+ \subset \mathbb{R}^{18\times512}$ space through 18 parallel linear layers, which is a more disentangled latent space contributing to style mixing [20, 29] and image inversion [1, 2, 11, 29]. Subsequently, the fixed StyleGAN generator G preserves the capability of generating high-resolution images with various styles and stochastic details. By integrating the proposed E and G, our method aligns the prediction vector space with the disentangled $\mathcal{W}^+$ space, providing semantically continuous face embeddings for the target identity.

To justify this *prediction-$\mathcal{W}^+$-image* alignment achieved by our method, we report empirical visualizations in Fig. 3. Concretely, we select one target private image, interpolate prediction vectors of public images (classified as the target) along the target dimension (maintaining sum of the vector as 1), and show corresponding reconstructed images. Clearly, as the value on target dimension increases, the reconstructed images gradually approach the target image's visual appearance, indicating the alignment between predictions and reconstructed images. This is also consistent with the decreasing distance $Dist_w$ between w of the target image and that of reconstructed image. Intuitively, the image of one certain identity is usually composed of several facial attributes (e.g., the uniqueness combination of face edges, eyes and nose shape, etc). Thus, we further conduct

the classification on five facial attributes and identity recognition tasks, both of which showing the same trending changes to the target identity. This is reasonable since facial attributes should change gradually within a continuous face manifold. By slightly changing the prediction values, we can gradually alter the semantic attributes of the corresponding identity (such as the zoom-in parts of mouth and eye).

3.3 Training Data Selection

Since private data is unavailable during the whole procedure, we resort to the public dataset of same task as training data, which has no identity overlap with private dataset. Following [43], we input all public images into target model and obtain the prediction vectors. For each identity, we select top-n images with the highest prediction scores for training. Meanwhile, we also apply the same selection on the synthesized data [40], forming the same number of top-n images for each identity. Finally, training data D_{pub} can be expressed as $D_{pub} = \{(x_i, p_i)\}_{i=1}^{N}$. Note that the selected training data preserves facial attribute overlaps between public and private images, implicitly formulating a better prior for the training phase.

During the training phase, we first introduce the pixel-wise $\mathcal{L}_{mse}$ loss which is commonly used in image reconstruction:

$$\mathcal{L}_{mse} = \|G(E(p)) - x\|_2 . \tag{1}$$

In addition, we adopt the perceptual loss [45] to constrain the perceptual similarity between the reconstructed and real images:

$$\mathcal{L}_{lpips} = \|F(G(E(p))) - F(x)\|_2 , \tag{2}$$

where F is the feature extractor [40]. The multi-layer identity loss and face parsing loss are also introduced for identity consistency:

$$\mathcal{L}_{id} = \sum_{j=1}^{5} (1 - \langle \mathbf{R}_j(G(E(p))), \mathbf{R}_j(x) \rangle), \tag{3}$$

$$\mathcal{L}_{parse} = \sum_{j=1}^{5} (1 - \langle \mathbf{P}_j(G(E(p))), \mathbf{P}_j(x) \rangle), \tag{4}$$

where $\mathbf{R}$ is the pre-trained ArcFace network [12], $\mathbf{P}$ the pre-trained face parsing model, j the j-th feature of the pre-trained model, and $\langle\rangle$ the cosine similarity.

To make the intermediate feature f_E better adapt to the ResNet backbone, and to prevent our PAE encoder from meaninglessly overfitting, we further propose the alignment regularization loss formalized as:

$$\mathcal{L}_{align\_reg} = \|f_E - x\|_2 . \tag{5}$$

https: //github.com/zllrunning/face-parsing.PyTorch

In a nutshell, the overall loss function can be summed up as follows:

$$\mathcal{L}_{recon} = \mathcal{L}_{mse} + \lambda_1 \mathcal{L}_{lpips} + \lambda_2 \mathcal{L}_{id} + \lambda_3 \mathcal{L}_{parse}, \tag{6}$$

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_4 \mathcal{L}_{align\_reg}. \tag{7}$$

3.4 Aligned Ensemble Attack

Note that during the attack phase, only the target label (one-hot prediction vector) is available for the adversary. Unfortunately, we empirically find that simply inputting one-hot prediction will lead to a dramatically poor performance (shown in Tab. 6). To tackle this, we argue that if an image is classified as the target identity, it must contain at least partial characteristic facial attributes of this identity. This motivates us to ensemble target's attributes contained in different public images. We thus propose the aligned ensemble attack to integrate the latent codes w encoded by predictions p for better reconstruction. Specifically, given a target identity c, we perform the aligned ensemble as:

$$\mathcal{W}_{ens} = \frac{\sum_{i=1}^{n} \max p_i \cdot w_i}{\sum_{i=1}^{n} \max p_i}, \tag{8}$$

where $\mathcal{W}_{ens}$ denotes the ensembled latent code. Moreover, in light of the prediction-image alignment provided by our method, we can explicitly enhance the prediction vectors of target identity by the following interpolation:

$$S_c(p_i) = S_c(p_i) + m, \tag{9}$$

where m is the enhancement hyper-parameter of the increase on prediction value, $S_c(p_i)$ the prediction score on dimension c of p_i. To ensure the sum of prediction vector as 1, we further adjust each non-target dimension by: $S_k(p_i) = S_k(p_i) - m \cdot S_k(p_i)/(1 - S_c(p_i))$, where $k \neq c$. This is reasonable and consistent with the editability of StyleGAN inversion, i.e., allowing the change of corresponding attributes by manipulating a directional vector in the latent space, thus leading to further improvement on inversion performance.

4 Experiments

4.1 Settings

Datasets. We conduct experiments on three face benchmarks, i.e., CelebA [24], FaceScrub [25] and PubFig83 [28]. CelebA consists of 202,599 face images belonging to 10,177 identities, of which we use 30,027 images of 1,000 identities as the private dataset. FaceScrub consists of 106,863 face images with 530 identities, and 200 of them are randomly selected to form the private dataset. For PubFig83, 50 of the 83 identities are randomly selected to form the private

dataset. All images are center cropped and resized to 64×64. Concretely, we use FFHQ [21] and synthetic data [40] as public dataset for training.

Target models. For fair comparison, we adopt target models with different architectures: VGG16 [32], ResNet-152 [17] and Face.evoLVe [9], which are widely used backbones in previous methods [8, 16, 18, 46].

Evaluation metrics. Similar to [16], we use the attack accuracy (Attack Acc), K-nearest neighbor distance (KNN Dist), feature distance (Feat Dist) and Learned Perceptual Image Patch Similarity (LPIPS) as evaluation metrics, see detailed information in supplementary material.

Implementation details. The StyleGAN generator is pre-trained and fixed, while the PAE encoder is initialized with pre-trained parameters [40]. We train our model for 30 epochs using the Ranger optimizer with an initial learning rate of 10^{-4}, batch size 4, $\beta_1 = 0.95$ and $\beta_2 = 0.999$. We apply the log operator to the input prediction vectors to avoid the dominance effect [7] in the probability distribution. We set $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$ and $\lambda_4 = 1$. We train target model for 50 epochs using the SGD optimizer with an initial learning rate of 0.001, batch size 64, momentum 0.9 and weight decay 10^{-4}.

4.2 Comparison with SOTAs

We compare our method with several state-of-the-art baselines across different settings. Specifically, we include the white-box methods GMI [46], KED-MI [8], LOMMA [27] and PLG-MI [43], considered as the upper bounds. We also implement black-box methods LB-MI [39], MIRROR [5] and RLB-MI [16], as well as lable-only methods BREP-MI [18] and LOKT [26].

Standard setting. We first evaluate our approach using the previous standard setting: the public and private data come from the same dataset, with no identity overlaps. Tab. 1 shows the results of our method and other baselines. Clearly, our method consistently perform best under the same data distribution: 1) Our method achieves optimal attack performance on all three target models. Our attack performance surpasses previous black-box method RLB-MI, with attack accuracy on three target models increasing by 8.5%, 7.6%, and 4.7%, respectively. Besides, our method has significantly narrowed the gap with SOTA white-box methods. 2) Note that although original LOKT performs well, it requires a significant amount of query cost (shown in Tab. 7). To ensure a fair comparison, we limit its number of queries and re-implement it as LOKT*. Obviously, its attack performance has decreased dramatically. 3) Our method is significantly ahead in KNN Dist and Feat Dist, which also indicates that our method reconstructs images that are visually closer to private datasets rather than fitting the evaluation model on Attack Acc.

Distribution shifts. We also consider a more practical setting where the public dataset and private dataset are from different distributions. Generally, as reported in Tab. 2, we can find that: 1) Our method achieves optimal performance on the distribution transfer among three datasets. For instance, when attacking the target model trained on PubFig83, the attack accuracy reaches 82%, which is 32% higher than second best black-box method RLB-MI, and even exceeds

Table 1: Attack performance on different target models trained on CelebA. All public datasets come from CelebA and no identity overlap with private dataset.

Target model	Type	Method	Attack Acc↑	KNN Dist↓	Feat Dist↓
VGG16 (88%)	White-box	GMI	0.185	1693.7	1615.7
		KED-MI	0.703	1363.7	1288.4
		LOMMA	0.903	1147.4	–
		PLG-MI	0.970	1080.5	985.5
	Black-box	LB-MI	0.075	1778.7	1741.6
		MIRROR	0.413	1456.1	1367.5
		RLB-MI	0.659	1310.7	1214.7
		ours	**0.715**	**1039.1**	**920.8**
	Label-only	BREP-MI	0.581	1347.4	1256.5
		LOKT	0.873	1246.7	–
		LOKT*	0.450	1294.7	1230.8
Resnet152 (91%)	White-box	GMI	0.300	1594.1	1503.5
		KED-MI	0.765	1277.3	1184.6
		LOMMA	0.929	1138.6	–
		PLG-MI	1.000	1016.5	910.2
	Black-box	LB-MI	0.041	1800.9	1735.7
		MIRROR	0.570	1360.7	1263.8
		RLB-MI	0.804	1217.9	1108.2
		ours	**0.865**	**1015.8**	**896.3**
	Label-only	BREP-MI	0.729	1277.5	1180.4
		LOKT	0.921	1206.8	–
		LOKT*	0.490	1293.9	1236.0
Face.evoLVe (89%)	White-box	GMI	0.254	1628.6	1541.7
		KED-MI	0.741	1350.8	1261.6
		LOMMA	0.932	1154.3	–
		PLG-MI	0.990	1066.4	972.6
	Black-box	LB-MI	0.111	1776.4	1729.1
		MIRROR	0.530	1379.7	1280.1
		RLB-MI	0.793	1225.6	1112.1
		our	**0.830**	**974.1**	**862.4**
	Label-only	BREP-MI	0.721	1267.3	1164.0
		LOKT	0.939	1181.7	–
		LOKT*	0.600	1262.0	1185.4

the white-box method KED-MI. We believe that the disentangled $\mathcal{W}^+$ space of StyleGAN allows the prediction vector representing the target identity with more general features. 2) The training-based method LB-MI shows both poor attack accuracy and visual quality, which also keeps in line with its shallow structure and less semantic facial features. 3) In addition, the fewer the number of identities in the private dataset, the higher the attack accuracy. We analyze that fewer identities lead to fewer facial features included, thus it is easy to select images with high confidence scores for all identities in public dataset.

Different models. Tab. 3 shows the results of attacking different model architectures with the target and evaluation models provided in [8, 43]. Based on the comparisons, we observe that: 1) Clearly, for all model architectures, the attack accuracy of our method is consistently higher than other baselines. 2) Surprisingly, the attack accuracy is almost 40 times higher than the training-based LB-MI method, and an improvement of 18% compared to RLB-MI. This is reasonable due to the alignment of input predictions and disentangled latent

Table 2: Attack performance on target model Face.evoLVe trained with PubFig83, FaceScrub and CelebA, respectively.

Public→Private	Type	Method	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
FFHQ→PubFig83 (92%)	White-box	GMI	0.20	1536.4	1696.2	0.416
		KED-MI	0.63	1211.0	1353.1	0.368
		PLG-MI	0.91	1211.9	1296.9	0.370
	Black-box	LB-MI	0.42	1392.1	1579.0	0.488
		MIRROR	0.48	1028.4	1195.4	0.325
		RLB-MI	0.62	1193.4	1340.3	0.340
		ours	**0.82**	**840.7**	**992.3**	**0.268**
	Label-only	BREP-MI	0.42	1230.8	1399.5	0.347
FFHQ→FaceScrub (96%)	White-box	GMI	0.23	1585.1	1612.1	0.381
		KED-MI	0.43	1520.9	1546.5	0.381
		PLG-MI	0.70	1344.5	1353.6	0.394
	Black-box	LB-MI	0.14	1530.9	1581.9	0.517
		MIRROR	0.40	1362.4	1367.9	0.266
		RLB-MI	0.49	1451.8	1452.6	0.339
		ours	**0.59**	**1243.8**	**1256.0**	**0.246**
	Label-only	BREP-MI	0.28	1539.0	1565.6	0.332
FFHQ→CelebA (93%)	White-box	GMI	0.17	1648.4	1580.9	0.403
		KED-MI	0.34	1525.9	1460.8	0.379
		PLG-MI	0.85	1332.7	1262.5	0.351
	Black-box	LB-MI	0.07	1660.9	1594.7	0.503
		MIRROR	0.40	1360.9	1267.6	0.282
		RLB-MI	0.33	1519.9	1443.4	0.356
		ours	**0.49**	**1302.9**	**1210.1**	**0.271**
	Label-only	BREP-MI	0.33	1503.2	1428.9	0.350

Table 3: Attack performance on different target models trained on CelebA with FFHQ as the public dataset.

Target model	Type	Method	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
VGG16 (88%)	White-box	GMI	0.07	1410.6	1326.5	0.392
		KED-MI	0.30	1363.7	1288.4	0.378
		PLG-MI	0.86	1256.4	1162.0	0.336
	Black-box	LB-MI	0.01	1317.1	1379.0	0.563
		MIRROR	0.19	1281.2	1178.1	0.276
		RLB-MI	0.29	1368.2	1304.0	0.350
		ours	**0.35**	**1238.1**	**1118.0**	**0.258**
	Label-only	BREP-MI	0.26	1367.0	1274.0	0.351
Resnet152 (91%)	White-box	GMI	0.15	1421.6	1341.9	0.395
		KED-MI	0.47	1326.3	1239.2	0.365
		PLG-MI	0.97	1158.4	1047.4	0.352
	Black-box	LB-MI	0.01	1371.4	1324.7	0.468
		MIRROR	0.24	1265.3	1153.4	0.283
		RLB-MI	0.39	1352.6	1259.7	0.359
		ours	**0.46**	**1190.9**	**1073.3**	**0.253**
	Label-only	BREP-MI	0.38	1351.1	1271.9	0.354
Face.evoLVe (89%)	White-box	GMI	0.14	1460.1	1356.5	0.402
		KED-MI	0.44	1316.4	1231.1	0.370
		PLG-MI	0.94	1246.7	1154.2	0.368
	Black-box	LB-MI	0.01	1645.4	1494.1	0.484
		MIRROR	0.19	1242.1	1150.2	0.285
		RLB-MI	0.41	1302.8	1218.4	0.337
		ours	**0.50**	**1181.4**	**1080.7**	**0.254**
	Label-only	BREP-MI	0.41	1346.4	1243.2	0.356

codes. 3) Moreover, our method almost rivals the white-box method KED-MI, and even surpasses all the other methods on all the visual metrics.

4.3 Ablation Study

Effect of loss terms. To evaluate the contribution of proposed loss terms, we train the model by removing each component solely and present the comparison results in Tab. 4.

Note that in the previous works, no checkpoints of Face.evoLVe trained with PubFig83 or FaceScrub are provided. Thus we use three target models trained by ourselves according to previous methods in Tab. 2.

Table 4: Effect of different loss terms.

	Configuration	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
A.	w/o $\mathcal{L}_{mse}$	0.42	1337.5	1246.6	0.264
B.	w/o $\mathcal{L}_{align\_reg}$	0.47	1299.4	1210.9	0.266
C.	w/o $\mathcal{L}_{lpips}$	0.44	1333.6	1250.4	0.278
D.	w/ $\mathcal{L}_{m\_lpips}$	0.39	1295.6	1212.8	0.271
E.	w/o $\mathcal{L}_{id}$	0.03	1580.0	1499.8	0.263
F.	w/o $\mathcal{L}_{parse}$	0.46	1324.2	1246.7	0.268
	our baseline	0.49	1302.9	1210.1	0.271

Table 5: Ablation study on the proposed PAE encoder.

	Configuration	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
G.	w/ $\mathcal{F}$ branch	0.45	1285.4	1202.6	0.264
H.	Random initialization	0.44	1303.9	1221.9	0.266
I.	replace with FC	0.45	1322.6	1246.2	0.266
	our baseline	0.49	1302.9	1210.1	0.271

As can be seen: 1) Configuration (Cfg.) A removes pixel reconstruction loss $\mathcal{L}_{mse}$ during training, which leads to less effective attack. This can be owing to the disentanglement of learned facial features. 2) The decrease of Cfg. B indicates that $\mathcal{L}_{align\_reg}$ promises feature f_E a better adaptation to the ResNet backbone, as well as a better prediction alignment with $\mathcal{W}^+$ space. 3) Cfg. C verifies that LPIPS loss greatly improves the performance, especially the perceived quality of reconstructed images. While Cfg. D shows the performance drop when adding the multi-scale LPIPS loss, which is unsuitable for the inversion process from predictions to images. 4) Cfg. E and F demonstrate that identity loss and parsing loss can better preserve the original identity of reconstructed images, thus further improve their perceived quality.

Analysis on prediction alignment encoder. We comprehensively investigate several mechanisms of the proposed encoder and show the results in Tab. 5. Specifically, Cfg. G adds the feature prediction $\mathcal{F}$ branch to our baseline and results in a decrease in attack accuracy. In [40], the $\mathcal{F}$ branch directly replaces a certain layer of the generator in StyleGAN with the learned feature tensor, while will inevitably disrupt our inversion procedure. Besides, Cfg. H initializes the image encoder randomly, achieving slightly better perceptual quality but worse attack performance. This is acceptable since the pre-trained encoder have some prior knowledge. We also replace the full PAE encoder with fully connected layers (Cfg. I), and the performance decrease indicates the importance of prediction alignment's structure.

Effect of aligned ensemble attack. To evaluate the effectiveness of our aligned ensemble attack, we first replace it with the LB-MI method (using one-hot prediction vector as input). As shown in Tab. 6, the performance of one-hot prediction as input is extremely poor. We believe this is because our PAE encoder maps the prediction vectors into a high-dimensional space. Using one-hot vectors as input will lead to the significant loss of information. We also directly ensemble the prediction vectors instead of the latent codes w, showing an ob-

Table 6: Comparison of attack performance by different input schemes.

Attack Phase	PubFig83	FaceScrub	CelebA
one-hot prediction	0.02	0.005	0
prediction ensemble	0.68	0.45	0.38
aligned ensemble (ours)	0.82	0.59	0.49

Table 7: Comparison with other black-box methods on query number to target model.

Method	**ours**	RLB-MI	MIRROR	BREP-MI	LOKT
Query number (million)	**0.13**	36	1.85	17.98	12.66
Percentage	–	0.36%	7.02%	0.72%	1.02%

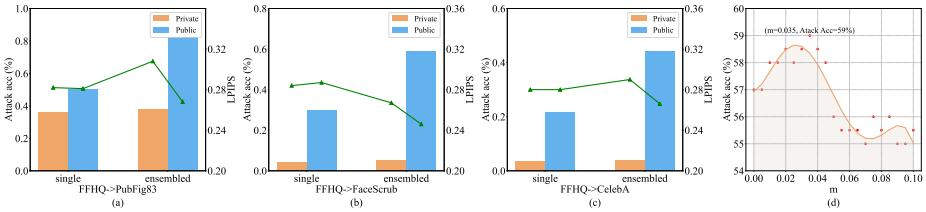

Fig. 4: (a)-(c) show the result comparison of input predictions under single/ensembled and private/public settings. (d) shows the sensitivity of hyper-parameter m.

vious performance drop from our method. We consider that this is due to the difference between the prediction vector space and $\mathcal{W}^+$ space. The ensemble of prediction vectors is merely aiming at finding an numerical average. However, our aligned ensemble aims to enhance the feature information of target identity, leading to more complementary facial attributes and better reconstruction.

In addition, Fig. 4(a)-(c) show the attack performance of four variants on three private datasets, respectively, i.e., single private prediction, single public prediction, ensemble w of private prediction and ensemble w of public prediction. Note that private prediction vector is only presented for thorough analysis, which is unavailable during the attack phase. Obviously, it is rather difficult to achieve high attack accuracy when using single prediction vector as input. We believe that reconstructing high-dimensional images from low-dimensional prediction vectors is inherently challenging. Luckily, the more disentangled nature of $\mathcal{W}^+$ space in StyleGAN makes our ensemble scheme an effective way to compensate for the aforementioned shortcomings. As shown in Fig. 4(d), we compare the attack performance by increasing different values m in the target dimension of public prediction vectors. When $m = 0$, it means no modification is performed on the public prediction vectors. As m increases, the target dimension value of public prediction vectors will also increase, along with the improved attack performance. However, if m continues to increase (from 0.035 to 0.1), the value change on vectors may be so drastic that the altered prediction deviates from the original distribution, resulting in a significant performance decrease.

Query cost. Tab. 7 shows the query costs of our method compared to others. We attack 300 identities in total. It is clear to see that the queries of ours is only 0.13 million, which is approximately only 0.36% of RLB-MI. In real-world scenarios, many MaaS platforms, such as Google and Amazon, limit the number of queries to the model, making black-box MI attack methods that rely on massive queries impractical. Our method only requires a small number of

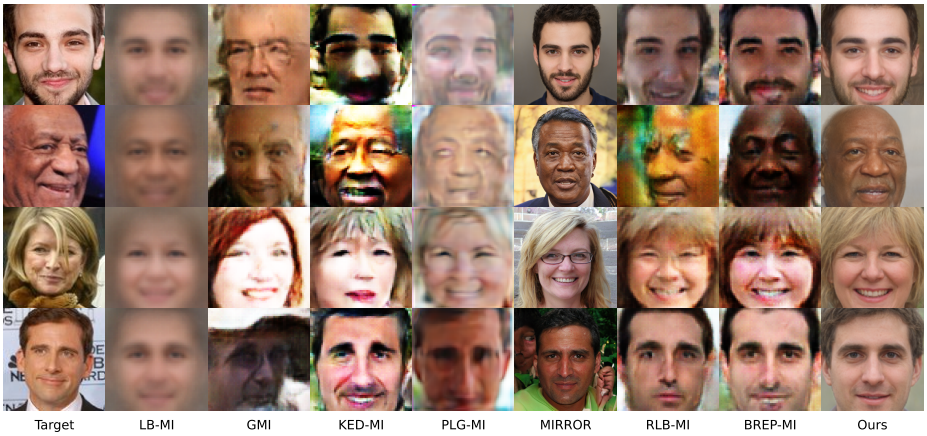

Fig. 5: Visual comparison of different model inversion attacks.

queries for data selection, taking an important step towards practical application of the black-box MI attack.

Visualizations. Fig. 5 shows the qualitative results of different inversion methods. Compared with other baselines, our reconstructed images obviously are more realistic and have higher resolution quality, verifying that the alignment provides more characteristic facial features of the target identity.

5 Conclusion

We propose a novel training-based Prediction-to-Image (P2I) method for black-box model inversion attack. P2I maps the target model's output predictions into the more disentangled latent space of StyleGAN, providing alignment between predictions and reconstructed high-fidelity images. We further design the aligned ensemble attack to enhance the feature information of target identity, leading to more complementary facial attributes and better reconstruction. Extensive experiments show the effectiveness of our method on both attack performance and visual quality. This work highlights that the rich information hidden in the model's prediction can be extracted, leading to data privacy leakage. We hope this will raise the attention of the community on facial privacy protection.

Limitations. One limitation is that we have not fully explored the potential of this StyleGAN-based training paradigm in model inversion task. In the future, we will continue to explore the essence of latent space in model inversion attack to further improve the performance of attacks in black-box or label-only scenario.

Negative social impacts. Our work might be adopted by malicious users to expose the privacy leaks that exist in online models. However, in light of this work, we hope to explore effective defense methods to counter different types of MI attacks, mitigating the underlying negative impact.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2022YFB3103500, the National Natural Science Foundation of China under Grants 62202459 and 62106258, and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001) and BAST (NO.BYESS2023304), and Beijing Natural Science Foundation QY23179.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4432–4441 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8296–8305 (2020)
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6711–6720 (2021)
4. An, J., Zhang, W., Wu, D., Lin, Z., Gu, J., Wang, W.: Sd4privacy: Exploiting stable diffusion for protecting facial privacy. In: 2024 IEEE International Conference on Multimedia and Expo (ICME) (2024)
5. An, S., Tao, G., Xu, Q., Liu, Y., Shen, G., Yao, Y., Xu, J., Zhang, X.: Mirror: Model inversion for deep learning network with high fidelity. In: Proceedings of the 29th Network and Distributed System Security Symposium (2022)
6. Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J.Y., Torralba, A.: Semantic photo manipulation with a generative image prior. arXiv preprint arXiv:2005.07727 (2020)
7. Chen, P., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.: ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: AISec@CCS. pp. 15–26. ACM (2017)
8. Chen, S., Kahla, M., Jia, R., Qi, G.J.: Knowledge-enriched distributional model inversion attacks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16178–16187 (2021)
9. Cheng, Y., Zhao, J., Wang, Z., Xu, Y., Karlekar, J., Shen, S., Feng, J.: Know you at one glance: A compact vector representation for low-shot learning. In: ICCV Workshops. pp. 1924–1932. IEEE Computer Society (2017)
10. Collins, E., Bala, R., Price, B., Susstrunk, S.: Editing in style: Uncovering the local semantics of gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5771–5780 (2020)
11. Creswell, A., Bharath, A.A.: Inverting the generator of a generative adversarial network. IEEE transactions on neural networks and learning systems **30**(7), 1967–1974 (2018)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
13. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322–1333 (2015)

14. Fredrikson, M., Lantz, E., Jha, S., Lin, S.M., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: USENIX Security Symposium. pp. 17–32. USENIX Association (2014)
15. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3012–3021 (2020)
16. Han, G., Choi, J., Lee, H., Kim, J.: Reinforcement learning-based black-box model inversion attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20504–20513 (2023)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778. IEEE Computer Society (2016)
18. Kahla, M., Chen, S., Just, H.A., Jia, R.: Label-only model inversion attacks via boundary repulsion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15045–15053 (2022)
19. Kang, K., Kim, S., Cho, S.: Gan inversion for out-of-range images with geometric transformations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13941–13949 (2021)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410. Computer Vision Foundation / IEEE (2019)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
22. Liu, H., Song, Y., Chen, Q.: Delving stylegan inversion for image editing: A foundation latent space viewpoint. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10072–10082 (2023)
23. Liu, Y., An, J., Zhang, W., Wu, D., Gu, J., Lin, Z., Wang, W.: Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. arXiv preprint arXiv:2405.20584 (2024)
24. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
25. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). pp. 343–347. IEEE (2014)
26. Nguyen, B.N., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.M.: Label-only model inversion attacks via knowledge transfer. Advances in Neural Information Processing Systems **36** (2024)
27. Nguyen, N.B., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.: Rethinking model inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16384–16393 (2023)
28. Pinto, N., Stone, Z., Zickler, T., Cox, D.: Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In: CVPR 2011 workshops. pp. 35–42. IEEE (2011)
29. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2287–2296 (2021)
30. Rigaki, M., Garcia, S.: A survey of privacy attacks in machine learning. arXiv preprint arXiv:2007.07646 (2020)

31. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
33. Struppek, L., Hintersdorf, D., Correia, A.D.A., Adler, A., Kersting, K.: Plug & play attacks: Towards robust and flexible model inversion attacks. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 20522–20545. PMLR (2022)
34. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
35. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16). pp. 601–618 (2016)
36. Wang, K.C., Fu, Y., Li, K., Khisti, A., Zemel, R., Makhzani, A.: Variational model inversion attacks. Advances in Neural Information Processing Systems **34**, 9706–9719 (2021)
37. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11379–11388 (2022)
38. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3121–3138 (2022)
39. Yang, Z., Zhang, J., Chang, E., Liang, Z.: Neural network inversion in adversarial setting via background knowledge alignment. In: CCS. pp. 225–240. ACM (2019)
40. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A style-based gan encoder for high fidelity reconstruction of images and videos. In: European conference on computer vision. pp. 581–597. Springer (2022)
41. Ye, D., Chen, H., Zhou, S., Zhu, T., Zhou, W., Ji, S.: Model inversion attack against transfer learning: Inverting a model without accessing it. arXiv preprint arXiv:2203.06570 (2022)
42. Ye, Z., Luo, W., Naseem, M.L., Yang, X., Shi, Y., Jia, Y.: C2fmi: Corse-to-fine black-box model inversion attack. IEEE Transactions on Dependable and Secure Computing (2023)
43. Yuan, X., Chen, K., Zhang, J., Zhang, W., Yu, N., Zhang, Y.: Pseudo label-guided model inversion attack via conditional generative adversarial network. arXiv preprint arXiv:2302.09814 (2023)
44. Yuan, Z., Wu, F., Long, Y., Xiao, C., Li, B.: Secretgen: Privacy recovery on pre-trained models via distribution discrimination. In: European Conference on Computer Vision. pp. 139–155. Springer (2022)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595. Computer Vision Foundation / IEEE Computer Society (2018)
46. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: Generative model-inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 253–261 (2020)
47. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: European conference on computer vision. pp. 592–608. Springer (2020)

48. Zhu, T., Ye, D., Zhou, S., Liu, B., Zhou, W.: Label-only model inversion attacks: Attack with the least information. IEEE Transactions on Information Forensics and Security **18**, 991–1005 (2022)