A Closer Look at GAN Priors: Exploiting Intermediate Features for Enhanced Model Inversion Attacks

Yixiang Qiu[1,2†], Hao Fang[2†], Hongyao Yu[1†], Bin Chen[1,3,4#], MeiKang Qiu[5], and Shu-Tao Xia[2,4]

[1] Harbin Institute of Technology, Shenzhen
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[4] Pengcheng Laboratory [5] Augusta University
qiuyixiang@stu.hit.edu.cn, fang-h23@mails.tsinghua.edu.cn
yuhongyao@stu.hit.edu.cn, chenbin2021@hit.edu.cn
qiumeikang@yahoo.com, xiast@sz.tsinghua.edu.cn

Abstract. Model Inversion (MI) attacks aim to reconstruct privacy-sensitive training data from released models by utilizing output information, raising extensive concerns about the security of Deep Neural Networks (DNNs). Recent advances in generative adversarial networks (GANs) have contributed significantly to the improved performance of MI attacks due to their powerful ability to generate realistic images with high fidelity and appropriate semantics. However, previous MI attacks have solely disclosed private information in the latent space of GAN priors, limiting their semantic extraction and transferability across multiple target models and datasets. To address this challenge, we propose a novel method, **I**ntermediate **F**eatures enhanced **G**enerative **M**odel **I**nversion (IF-GMI), which disassembles the GAN structure and exploits features between intermediate blocks. This allows us to extend the optimization space from latent code to intermediate features with enhanced expressive capabilities. To prevent GAN priors from generating unrealistic images, we apply a l_1 ball constraint to the optimization process. Experiments on multiple benchmarks demonstrate that our method significantly outperforms previous approaches and achieves state-of-the-art results under various settings, especially in the out-of-distribution (OOD) scenario. Our code is available at: https://github.com/final-solution/IF-GMI

Keywords: Privacy · Model Inversion · Generative Priors

1 Introduction

In recent years, Deep Neural Networks (DNNs) have experienced unprecedented development and achieved tremendous success in a wide range of applications,

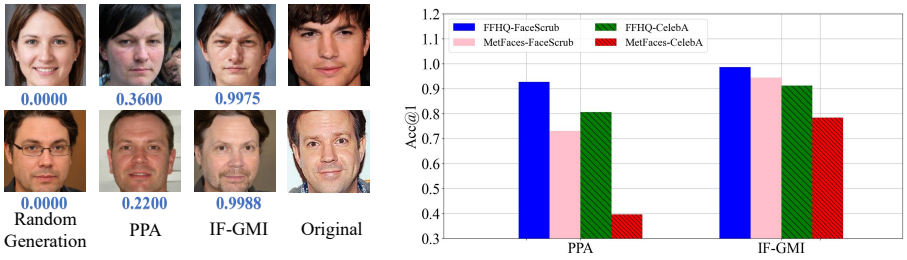
[†]Equal contribution.
[#]Corresponding author.
This work was done while Yixiang Qiu was pre-admitted to Tsinghua University.

including face recognition [17], personalized recommendations [42], and audio recognition [7]. While DNNs bring us many practical benefits, concerns [4,10,11, 46] about privacy and security have also been raised and drawn great attention. Recent studies have demonstrated that there is a certain risk of privacy leakage for DNNs as an adversary could reveal private information from these pre-trained models. Various types of novel privacy attacks [27, 33, 49] have been proposed, such as *membership inference attack* [20, 36] and *gradient inversion attack* [10, 46]. Among the new attack methods, Model Inversion (MI) attack [12] poses a greater threat due to its powerful capability in recovering the privacy-sensitive datasets that are collected and utilized for model training.

[14] proposes the first MI attack to reconstruct sensitive features of genomic data and demonstrate that linear regression models are vulnerable to such privacy attacks. Subsequent studies [13, 37, 43] have extended MI attacks to more Machine Learning (ML) models, but are still limited to models with simple structure and low-dimensional data such as grayscale images. Recent advances in the MI attack field have overcome the challenges in image data recovery by applying Generative Adversarial Networks (GANs) [16], resulting in the extension to DNNs with more complex structure and high-dimensional data such as RGB images. [51] first introduces the GANs to MI attack scenarios, serving as image priors. To better reveal privacy-sensitive information, [51] and subsequent GAN-based methods [5, 41, 47, 48] train GANs with publicly available datasets that have structural similarity with target private datasets. Furthermore, [38] propose to leverage the public pre-trained GAN models (*e.g.*, StyleGAN [24]) as GAN priors, which have a stronger ability to generate high-resolution images and do not require a time-consuming training process.

Although the aforementioned methods have achieved great progress in recovering high-quality and privacy-sensitive images, the effectiveness of GAN-based MI attacks is limited under certain scenarios. One typical challenge is the out-of-distribution (OOD) scenario, where there is a significant distributional shift between the target private dataset and the public dataset used in the training process of GAN priors. Most previous methods [5, 41, 48, 51] merely work well under scenarios with slight distributional shifts. For instance, they split the same dataset into two parts, one used as the public dataset and the other used as the private dataset. In recent years, some studies [3, 8, 35, 40, 45] have demonstrated that there is rich semantic information encoded in the latent code and intermediate features of GANs. Inspired by these works, we empirically observe that the rich semantic information encoded in the intermediate features helps to sufficiently recover high-quality private data under more rigorous settings, as shown in Figure 1. Therefore, it is imperative to explore methods for leveraging the GAN's intrinsic layered knowledge into MI attacks, mitigating the OOD issue.

To this end, we propose a novel MI attack method, **I**ntermediate **F**eatures enhanced **G**enerative **M**odel **I**nversion (IF-GMI), which effectively disassembles the GAN structure and leverages features between intermediate blocks. Specifically, we consider the generator of the GAN as a concatenation of multiple blocks and the vectors produced between the blocks as intermediate features. We first



(a) Visual examples of IF-GMI to PPA (b) Comparison of IF-GMI to PPA

Fig. 1: (a) Comparison of our proposed IF-GMI with baselines. The blue number below the images is the predicted confidence by the evaluation model. The first column shows the randomly generated images and the second column presents the reconstructed results by PPA [38], a typical GAN-based method focusing on directly optimizing the latent code of GAN model. The last two columns exhibit the results of our proposed IF-GMI and the ground truth images in the private dataset, respectively. (b) Top-1 attack accuracy of PPA and IF-GMI (ours) on four OOD scenarios.

optimize the latent code input to the generator and then successively optimize the intermediate features from the start layer to the end layer. To avoid unreal image generation, we utilize a l_1 ball constraint to restrict the deviation when optimizing the intermediate features. In the end, we collect the output images after each intermediate layer optimization process and select the final results with a simple strategy. We conduct comprehensive experiments to evaluate our method in multiple settings, including OOD scenarios, various target models, and different GAN priors. The encouraging experimental results demonstrate that the proposed method outperforms baselines on multiple metrics and achieves high attack accuracy on OOD settings. Finally, we perform extensive experiments and ablation studies to validate the effectiveness of the proposed method. Our main contributions are as follows:

- We propose a novel GAN-based MI attack method, which disassembles the pre-trained generator and successively optimizes the latent code and intermediate features under the l_1 ball constraint.
- We demonstrate that our proposed achieves state-of-the-art performance in a range of scenarios, especially under the challenging OOD settings.
- We conduct extensive experiments to validate the effectiveness and outstanding transferability of our method.

2 Related Work

2.1 GAN as prior knowledge

GANs [15] are a class of deep neural networks that consist of two functional components, a generator and a discriminator, trained concurrently through adversarial processes to generate realistic data. The objective of a GAN is to learn the

distribution of the training dataset and generate more samples from the learned probability distribution [16]. Well-trained GANs are able to generate high-fidelity and diverse images, excellent representative of which are StyleGANs [24,25]. The generator of the StyleGAN consists of a mapping network and a synthesis network. The former maps latent vectors into the intermediate latent space (*i.e.* $\mathcal{W}$ space), and the latter generates images through style vectors. The feature in the $\mathcal{W}$ space is well-disentangled, which means that images sharing similar features correspond to analogous style vectors. Therefore, PPA [38] performs their attacks by searching the style vectors in $\mathcal{W}$ space. The style vectors in the front layers tend to control high-level aspects of the generated images like pose, face shape, and general hair style, while those in the back ones have more influence on details [24], such as smaller scale facial features and eyes open/closed. Moreover, style vectors in $\mathcal{W}$ space do not need to follow the same distribution with the training data, which means that more diverse images can be generated by controlling the vectors [24].

Recent works [10, 32, 52] have shown the richness of intermediate features in GANs, our investigation also tries to explore the potential of leveraging intermediate latent space of different layers to enhance MI attacks. Our findings reveal that this approach significantly improves attack accuracy and obtains high-quality inversion results, particularly under the harder OOD scenario.

2.2 Model Inversion Attacks

Model inversion (MI) attacks aim at reconstructing the private training data from a trained model. Typically, MI attacks can be divided into the white-box scenario [51] and black-box scenario [22]. We only focus on the white-box scenario in this paper, which means that the attacker has full access to the trained model. This kind of attack is initially demonstrated through an attempt to extract genomic markers from a linear regression model, as highlighted in the earliest research by [14]. Building on this foundation, subsequent researches [13, 37, 43] have broadened the scope of MI attacks, applying them to more machine learning models like shallow networks, and simple forms of data, such as low-resolution grayscale images. However, as the scale of both the data and the models increases, the efficacy of MI attack methods diminishes dramatically.

In response to this challenge, a novel approach known as GMI, introduced by [51], employs a GAN-based methodology to enhance the ability of MI attacks with deeper and wider DNNs. This innovative strategy leverages a GAN model trained on publicly available data to encapsulate the distributional characteristics of image data, thereby facilitating the generation of high-quality image reconstructions. The process involves the attackers first generating a set of preliminary images by inputting a batch of randomly sampled latent vectors into the GAN. These generated images are then fed into the target image classifier to obtain initial predictions. To refine the attack, the attackers iteratively optimize the input latent vectors. This optimization process aims to minimize the discrepancy between the classifier's predictions and the intended target class, as measured by the cross-entropy loss, while also reducing the discriminator loss.

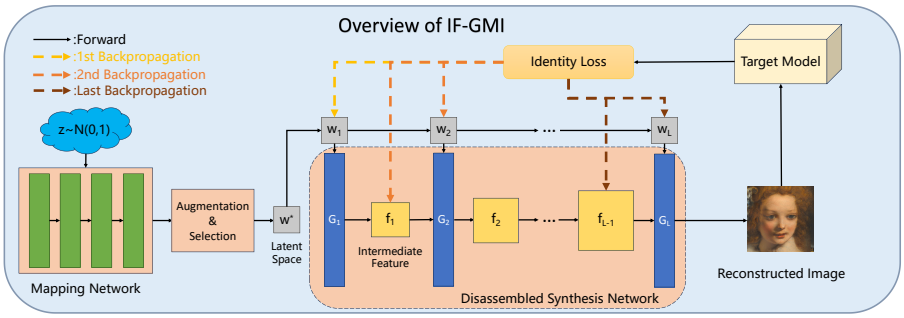


Fig. 2: Overview of our proposed IF-GMI. Firstly, the latent vectors are sampled from standard Gaussian distribution and mapped into disentangled latent codes with semantic meanings by Mapping Network. Then we perform random augmentation on these latent codes to select optimal ones denoted as $\mathbf{w}^*$ for optimization. The Synthesis Network is disassembled into multiple blocks to search the intermediate features, which are successively updated with the identity loss calculated from the target model. Finally, the reconstructed images are generated from the last layer as results.

With the help of the GAN, GMI seeks to achieve more precise and convincing reconstructions of complex data, thereby representing a significant advancement in the field of MI attacks.

Lots of researches in recent years improve the attack performance on the white-box scenario based on GMI. SecretGen [48] explores the scenario when the attackers know some auxiliary information about the private data. KEDMI [5] improves the discriminator by incorporating target labels and recover the distribution of the input latent vectors for a target class. VMI [41] reformulates the MI attack from the perspective of variational inference and introduce KL-divergence as a regularization to better approximate the target distribution with a variational distribution. PPA [38] employs pre-trained StyleGAN2 to reduce the time cost of attacks and extend the attacks to high-resolution images thanks to the excellent generative ability of StyleGAN2. Moreover, they propose a set of strategies to heighten attack accuracy and robustness, including initial selection, post-selection, and data augmentation. LOMMA [31] introduces model augmentation into MI attacks to reduce overfitting of the target model. They train some surrogate models from the target model via model distillation, co-guiding the optimization process with improved loss function. PLGMI [47] proposes a top-n selection strategy, using target models to generate pseudo labels for publicly available images, thereby directing the training process for the conditional GAN.

3 Methodology

In this section, we begin by explaining the fundamental paradigm of MI attacks and provide a formulation for the MI problem. Subsequently, we present our main components and elaborate the detailed pipeline of the proposed IF-GMI,

which contributes to the improved performance under the OOD scenario. See Figure 2 for an overview of our method.

3.1 Preliminaries

In this paper, we focus on the MI attacks under white-box settings, which means all the parameters and components of target models are available to the attacker. For image classification tasks, the malicious adversary aims to reconstruct privacy-sensitive images by leveraging the output prediction confidence of the target classifier and other auxiliary priors. Early works [44] directly optimize pixels in randomly sampled dummy images $\mathbf{x}$ to approximate target images $\mathbf{x}^*$ given the target model T_θ and target label c, which can be formulated as follows:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathcal{L}(T_\theta(\mathbf{x}), c), \tag{1}$$

where $\hat{\mathbf{x}}$ is the reconstructed image, $\mathcal{L}(\cdot,\cdot)$ denotes the classification loss designed for image optimization and $T_\theta(\mathbf{x})$ represent the output confidence. Due to the full access to the target model in white-box settings, the attacker can calculate loss and directly perform backpropagation to update dummy images.

However, the methods above are no longer functional when $\mathbf{x}$ turns into high-dimensional data which has excessive search space. To tackle such issues, recent studies [5, 38, 47, 51] introduce GANs as image priors due to their superior capability to generate high-fidelity RGB images. They propose to train a specially designed GAN with publicly available datasets that have structural similarities with the private dataset or utilize a public pre-trained GAN before the attack. Furthermore, the optimization objective is replaced with the latent vectors $\mathbf{z}$ of the generator, which has fewer parameters to optimize. With the aforementioned techniques, the MI problem is transformed into the following formulation:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \mathcal{L}_{id}(T_\theta(G(\mathbf{z}), c) + \lambda \mathcal{L}_{aux}(\mathbf{z}), \tag{2}$$

where G represents the trained generator, $\mathcal{L}_{id}(\cdot,\cdot)$ denotes the identity loss calculated from the target model T_θ and $\mathcal{L}_{aux}(\cdot)$ is an optional auxiliary loss (*e.g.*, the discriminator loss) with a hyperparameter λ. By minimizing the Eq.2, the adversary updates the latent vectors $\mathbf{z}$ into the optimal results $\hat{\mathbf{z}}$ and generate final images through $\hat{\mathbf{x}} = G(\hat{\mathbf{z}})$.

Intuitively, directly optimizing the input latent code of GAN priors serves as a natural method to acquire ideal reconstructed images, leading to its widespread application in all the previous works. However, recent studies [3,8,35,40] have indicated that there is fairly rich semantic information in the intermediate features of GANs except for the input latent code. This inspires us to surpass the limitation of merely searching the latent space and propose a novel method focusing on the intermediate feature domains, which are more close to the output.

3.2 Exploiting Intermediate Features for Enhanced MI Attacks

In the following part, we delve into the internal structure of the GAN prior, attempting to explore the hierarchical layers for enhanced utilization of the rich

semantics learned by the generator. Following the pipeline shown in Figure 2, we will elucidate each component in detail.

The GAN prior. Most previous GAN-based attacks [5,31,47,51] require training a specialized GAN with essential auxiliary dataset towards the specific target classifier. However, the prior knowledge of GANs trained under the above setting will be excessively aligned with the target model and the auxiliary dataset, leading to significant reduction in transferability and generalization.

Therefore, our method relies on the pre-trained StyleGAN2 [23] instead of training a GAN from scratch. The generator of StyleGAN2 can be simply divided into two components, consisting of a mapping network $G_{map} : \mathcal{Z} \to \mathcal{W}$ which maps the initial latent vectors $\mathbf{z} \in \mathcal{Z}$ into the extended $\mathcal{W}$ space [1], and a synthesis network $G_{syn} : \mathcal{W} \to \mathcal{X}$ which generates images $\mathbf{x}$ with mapped vectors $\mathbf{w} \in \mathcal{W}$. Due to the reduced feature entanglement in $\mathcal{W}$ space that facilitates better style generation, we set $\mathbf{w}$ as the initial optimization objective rather than the commonly used latent code $\mathbf{z}$ in previous works. Specifically, we first randomly sample a batch of latent vectors $\mathbf{z}$ from Gaussian distribution and then map them with G_{map} to acquire $\mathbf{w}$, which will be iteratively updated in the first step of intermediate features optimization. Moreover, the StyleGAN2 is pre-trained without the utilization of the target model T_θ or other auxiliary prior corresponding to the target dataset, ensuring the flexibility and transferability of our method when attacking different target models and datasets.

Initial Selection. Owing to the randomness in sampling latent vectors $\mathbf{z}$, it is potential part of them cannot facilitate the generation of appropriate images, leading to a decrease in attack accuracy. To reduce the risk of generating misleading and low-quality images, previous studies [2, 38, 48] have explored the technique of initial selection and validated its effectiveness in obtaining robust latent vectors. Specifically, we first generate images with the randomly samples $\mathbf{z}$, apply a series of transformations $Aug(\cdot)$ to the images, and feed them into the target classifier T_θ for corresponding prediction confidence. By selecting the latent vectors with higher scores, we can significantly improve the quality of the final images to better approximate the target distribution.

Inspired by these prior studies [2,38,48], we also include the initial selection technique in our method and apply standard image transformations, such as random cropping, resizing and flipping. Different from previous methods, we perform initial selection on the mapped vectors $\mathbf{w}$ instead of latent vectors $\mathbf{z}$. The robust vectors $\mathbf{w}$ are obtained with the following equation:

$$\mathbf{w}_{init} = \arg\max_{\mathbf{w}} \mathrm{Conf}(T_\theta(Aug(G_{syn}(\mathbf{w}))), c), \tag{3}$$

where $\mathrm{Conf}(\cdot, \cdot)$ measures the confidence score for augmented images $Aug(G_{syn}(\mathbf{w}))$ given the specific label c.

Intermediate Features Optimization. According to the research of [24], the front blocks in the generator control the overall characteristics while the back

Algorithm 1 Pseudocode of the core algorithm in our proposed IF-GMI

Input: G_{syn}: a pre-trained generator; L: the number of intermediate features; T_θ: the target classifier; $\mathcal{L}_{id}$: the identity loss; $r[1 \dots L]$: the radius value of l_1 ball for each hierarchical features; N: the number of iterations;

Output: Reconstructed images $\mathbf{x}^*$;

Acquire latent vectors $\mathbf{w}_{init}$ via initial selection process
$\mathbf{w}_{(0)} \leftarrow \arg\min_{\mathbf{w}} \mathcal{L}_{id}(G_{syn}(\mathbf{w}_{init}))$
Decompose the G_{syn} into $G_{L+1} \circ G_L \circ \cdots G_2 \circ G_1$
Obtain the first intermediate feature $\mathbf{f}^0_{(1)} = G_1(\mathbf{w}_{(0)})$
Set $\mathbf{w}^0_{(1)} = \mathbf{w}_{(0)}$
for $i \leftarrow 1$ to L **do**
  Set $G_{remain} = G_{L+1} \circ G_L \cdots \circ G_{i+1}$
  for $j \leftarrow 1$ to N **do**
    $loss = \mathcal{L}_{id}(G_{remain}(\mathbf{f}^{j-1}_{(i)}, \mathbf{w}^{j-1}_{(i)}))$
    $\mathbf{f}^j_{(i)} \leftarrow Adam(\mathbf{f}^{j-1}_{(i)}; loss), ||\mathbf{f}^j_{(i)} - \mathbf{f}^0_{(i)}||_1 \leq r[i]$
    $\mathbf{w}^j_{(i)} \leftarrow Adam(\mathbf{w}^{j-1}_{(i)}; loss), ||\mathbf{w}^j_{(i)} - \mathbf{w}^0_{(i)}||_1 \leq r[i]$
  end for
  $\mathbf{f}^0_{(i+1)} = G_{i+1}(\mathbf{f}^N_{(i)}, \mathbf{w}^N_{(i)})$, $\mathbf{w}^0_{(i+1)} = \mathbf{w}^N_{(i)}$
end for
The final images $\mathbf{x}^* = \mathbf{f}^0_{(L+1)}$
return $\mathbf{x}^*$

ones have more influence on local details. Previous studies [38, 47, 51] neglect the role of the latter, which limits the attack performance. To take advantage of the individual blocks, we propose intermediate features optimization, as shown in the Algorithm 1. We first optimize the selected latent vectors $\mathbf{w}_{init}$ to obtain the optimal ones $\mathbf{w}_{(0)}$ before launching intermediate features optimization. Then we disassemble the pre-trained generator into $L+1$ blocks for hierarchical layer searching, *i.e.*,

$$G_{syn} = G_{L+1} \circ G_L \circ \cdots G_2 \circ G_1. \tag{4}$$

And we can feed $\mathbf{w}_{(0)}$ into block G_1 to attain the first intermediate feature $\mathbf{f}^0_{(1)}$.

For each intermediate block $G_{i+1}, i \in [1, \dots, L]$, the corresponding intermediate features $\mathbf{f}^0_{(i+1)}$ are acquired with following steps. First, we generate images utilizing the remaining blocks (*i.e.*, $\mathbf{x}_i = G_{L+1} \circ G_L \dots G_{i+1}(\mathbf{f}_{(i)}, \mathbf{w}_{(i)})$) and input them into the target classifier T_θ to compute the prediction confidence for loss function. Then, we repeat the aforementioned process to iteratively update both $\mathbf{w}_{(i)}$ and $\mathbf{f}_{(i)}$. During the optimization process, we restrict the $\mathbf{f}_{(i)}$ within the l_1 ball with radius $r[i]$ centered at the initial intermediate feature $\mathbf{f}^0_{(i)}$ to avoid excessive shift that may lead to collapse image generation. Once the iteration process is completed, the optimized $\mathbf{w}^N_{(i)}$ and $\mathbf{f}^N_{(i)}$ are fed into the block G_i to

obtain the next intermediate features $\mathbf{f}^0_{(i+1)}$. Moreover, we denote the optimized $\mathbf{w}^N_{(i)}$ as the initial latent vector $\mathbf{w}^0_{(i+1)}$ before the next layer optimization starts.

Once we finish searching the last intermediate layer, we can generate the final images $\mathbf{x}^*$ from the last intermediate feature $\mathbf{f}^N_{(L)}$, *i.e.*, $\mathbf{x}^* = \mathbf{f}^0_{L+1} = G_{i+1}(\mathbf{f}^N_{(L)})$.

The Overall Loss. While the cross-entropy loss $\mathcal{L}_{CE}$ serves as the identity loss in most early works [5, 48, 51], there is a major drawback of $\mathcal{L}_{CE}$. Specifically, the gradient vanishing problem emerges when the prediction confidence of target label c approaches the ground truth in the one-hot vector. Following the previous study [38], we rely on the Poincaré loss function to overcome this problem. Therefore, the identity loss function utilized in out method is defined as follows:

$$\mathcal{L}_{id} = \operatorname{arccosh}\left(1 + \frac{2||v_1 - v_2||_2^2}{(1 - ||v_1||_2^2)(1 - ||v_2||_2^2)}\right), \tag{5}$$

where $||v||_2$ is the Euclidean norm for the given vector. In our experiments, we denote v_1 as the normalized prediction confidence and v_2 as the one-hot vector for ground truth. Notably, the original number 1 in v_2 is substituted with 0.9999 to avoid division by zero.

4 Experiments

In this section, we first illustrate the details of our experimental settings. Then, we compare our method with state-of-the-art baselines to evaluate the attack performance. Furthermore, we conduct extensive experiments on multiple target datasets and models to further validate the effectiveness of our method in various settings. Finally, the ablation study will be evaluated on the first 100 classes of the whole dataset due to cost concerns.

4.1 Experimental Setup

Datasets. We evaluate our method on two classification tasks, including facial image classification and dog breed classification. For the facial image classification task, we select the FaceScrub [30] and CelebFaces Attributes [28] (CelebA) as private datasets to train the target models. FaceScrub consists of facial images of actors and actresses with 530 identities in total. CelebA contains facial images of 10177 identities with coarse alignment. For FaceScrub, we utilize all the identities in the major experiment. For CelebA, we select the top 1000 identities with the most images for our experiment, consisting of over 30000 images. We use Flickr-Faces-HQ [24] (FFHQ) and MetFaces [23] as public datasets. FFHQ consists of 70000 high-quality human face images. MetFaces is an image dataset of 1336 human faces extracted from the Metropolitan Museum of Art Collection, which has a huge distributional shift with real human faces. For the dog breed classification task, we use Stanford Dogs [9] as a private dataset and Animal Faces-HQ Dogs [6] (AFHQ) as a public dataset. To adapt to the target model, all images in the various datasets are pre-processed to a resolution size of 224×224 pixels in our experiment.

Models. We trained a variety of classification models on the private datasets mentioned above, including various architectures such as ResNet-18 [18], DenseNet-169 [21], ResNet-152 [18], and ResNeSt-101 [50], as target models. Following the settings in the previous work [38], we select Inception-v3 [39] as the evaluation model. For the generative model, we employ publicly released StyleGAN2 pre-trained on the aforementioned public datasets.

Metrics. Following PPA [38], we evaluate the performance of our attack method on various kinds of metrics as follows:

- **Attack Accuracy.** This metric serves as a criterion on how well the generated samples resemble the target class. We use the evaluation model trained on the same dataset with the target model to predict the labels on reconstructed samples and compute the top-1 and top-5 accuracy for target classes, denoted as $Acc@1$ and $Acc@5$ respectively. The higher the reconstructed samples achieve attack accuracy on the evaluation model, the more private information in the dataset can be considered to be exposed [51].
- **Feature Distance.** The feature is defined as the output of the model's penultimate layer. We compute the shortest feature l_2 distance between reconstructed samples and private training data for each class and calculate the average distance. The evaluated feature distances on the evaluation model and a pre-trained FaceNet [34] are denoted as δ_{eval} and δ_{face}, respectively.
- **Fréchet Inception Distance (FID).** FID [19] is commonly used to evaluate the generated images of GANs. It computes the distance between the feature vectors from target private data and reconstructed samples. The feature vectors are extracted by Inception-v3 pre-trained on ImageNet. The lower FID score shows higher realism and overall diversity [41].
- **Sample Diversity.** We compute Precision-Recall [26] and Density-Coverage [29] scores, whose higher values indicate greater intra-class diversity of the reconstructed samples. Our results for these four metrics are stated and analyzed in the Appendix.

4.2 The Number of Optimized Layers

To obtain the highest attack performance, the number of intermediate features L should be explored before conducting the major experiments. When L takes a small value, there is a risk of underfitting as we merely optimize the intermediate features of the previous few layers to reconstruct the target images, especially in the OOD scenario. In contrast, when L is too large, the latter layers have a greater influence on the local details [24], which may lead to overfitting to the target model in some details and produce unrealistic images. Therefore, we must balance underfitting and overfitting when choosing L. We conduct a simple attack on only 10 classes for each combination of public and private datasets to select L according to the results. For instance, Figure 3(a) shows the $Acc@1$ result for GAN prior pre-trained on FFHQ against the target DenseNet-169 trained on CelebA. The $Acc@1$ reaches the highest when $L = 3$. Hence, we keep this configuration in conducting the following experiments.

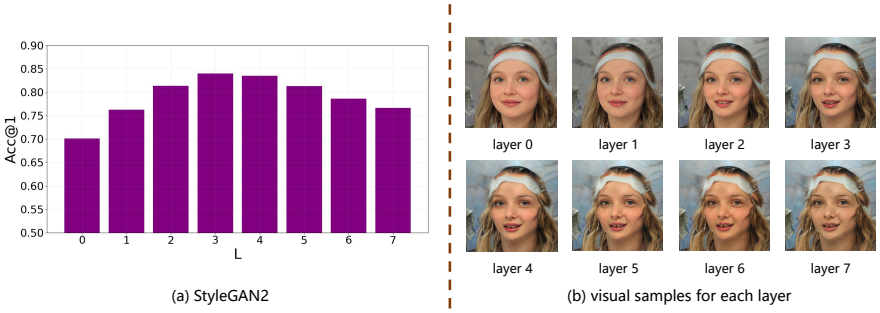


Fig. 3: (a) Comparison of $Acc@1$ metric under various settings of L (*i.e.*, the number of intermediate features). (b) Visual results generated from different end layers. We define $L = 0$ as a special case that our method degenerates into merely optimizing the latent vectors $\mathbf{w}$.

Table 1: Comparison of our method with state-of-the-art methods against ResNet-18 trained on FaceScrub.

Public Dataset	Method	↑ $Acc@1$	↑ $Acc@5$	↓ δ_{face}	↓ δ_{eval}	↓FID
FFHQ	GMI [51]	0.131	0.339	1.260	149.530	77.800
	KEDMI [5]	0.127	0.317	1.155	186.409	144.195
	PPA [38]	0.962	**0.996**	0.707	117.834	41.688
	LOMMA+GMI [31]	0.828	0.945	0.784	126.178	55.840
	LOMMA+KEDMI [31]	0.549	0.814	0.916	217.991	114.045
	PLGMI [47]	0.758	0.928	0.676	214.978	154.497
	IF-GMI(ours)	**0.979**	**0.996**	**0.667**	**112.915**	**40.581**
MetFaces	GMI [51]	0.038	0.136	1.361	161.036	114.648
	KEDMI [5]	0.003	0.017	1.651	212.952	347.468
	PPA [38]	0.628	0.854	1.035	146.749	**62.518**
	LOMMA+GMI [31]	0.160	0.361	1.220	156.297	101.600
	LOMMA+KEDMI [31]	0.002	0.020	1.623	214.883	333.572
	PLGMI [47]	0.438	0.731	**0.796**	205.222	245.208
	IF-GMI(ours)	**0.949**	**0.992**	0.838	**120.354**	68.107

4.3 Comparison with Previous State-of-the-art Attacks

We compare our method with state-of-the-art MI attack methods, including GMI [51], KEDMI [5], PPA [38], LOMMA [31] and PLGMI [47]. Note that LOMMA [31] is a plug-and-play technique designed to augment existing attack methods. We use their original setup where LOMMA is integrated with GMI and KEDMI as our baselines.

The GAN structures employed by GMI, KEDMI, and PLGMI are inherently limited to generating images at a resolution of 64×64 pixels. To ensure a fair comparison, we adopt the same operation used in PPA [38], which modifies the architecture of the generators and discriminators to enable the generation of images at an enhanced resolution of 256×256 pixels, *i.e.*, adding two ex-

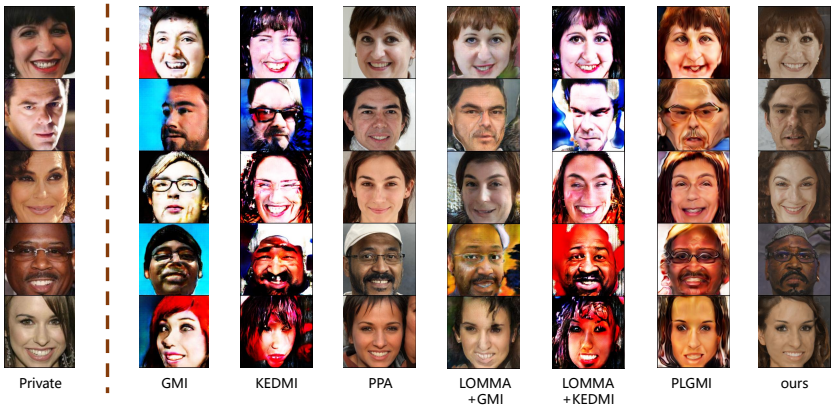


Fig. 4: Visual comparison of reconstructed images from different methods against the ResNet-18 trained on FaceScrub. The first column shows ground truth images of the target class in the private dataset.

tra upsampling layers for the generator and two downsampling layers for the discriminator respectively.

We provide quantitative results against ResNet-18 [18] trained on the FaceScrub dataset in Table 1. We can observe that our method achieves significant improvements over previous methods. Especially when the generator is trained on MetFaces, IF-GMI remarkably improves the $Acc@1$ by 15.1% and the $Acc@5$ is nearly to 100%. Moreover, our method generally achieves a lower feature distance than baselines between reconstructed samples and private data. For instance, we reduce the distance by more than 10% compared to the PPA on the MetFaces dataset. Notably, the MetFaces dataset is composed of artworks and thus has a larger distributional shift with real human faces compared with the FFHQ dataset. We note that this severely reduces the reconstruction performance of previous attack methods, while our proposed method still exhibits outstanding performance, highlighting the excellent generalization ability of our approach. Visualization results of the recovered images using generators trained on FFHQ are shown in Figure 4. Compared with previous methods, our reconstructed images have higher fidelity and realism, demonstrating the superiority of exploiting GAN's intermediate features.

4.4 Comparison under different target datasets and models

To validate the effectiveness of the proposed method, we conducted extensive experiments on various datasets using different target models with different architectures. We chose the PPA method as our baseline for comparison due to its comprehensive performance in both accuracy and fidelity. Additional experimental results are in the Appendix.

Table 2: Comparison results against ResNet-152 trained on CelebA.

Public Dataset	Method	↑ Acc@1	↑ Acc@5	↓ δ_{face}	↓ δ_{eval}	↓FID
FFHQ	PPA	0.806	0.946	0.736	**312.580**	40.430
	IF-GMI(ours)	**0.912**	**0.982**	**0.678**	314.392	**30.685**
MetFaces	PPA	0.396	0.643	1.063	387.810	**74.030**
	IF-GMI(ours)	**0.784**	**0.929**	**0.835**	**340.894**	74.504

Table 3: Comparison results against different target models trained on FaceScrub with the public dataset being MetFaces.

Target Model	Method	↑ Acc@1	↑ Acc@5	↓ δ_{face}	↓ δ_{eval}	↓FID
ResNet-152	PPA	0.731	0.920	0.966	139.380	**68.540**
	IF-GMI(ours)	**0.904**	**0.984**	**0.882**	**138.752**	69.937
ResNeSt-101	PPA	0.750	0.927	0.979	137.170	88.660
	IF-GMI(ours)	**0.922**	**0.983**	**0.884**	**132.609**	**76.195**
DenseNet-169	PPA	0.798	0.948	0.938	129.440	**77.520**
	IF-GMI(ours)	**0.933**	**0.987**	**0.851**	**125.050**	82.123

As shown in Table 2, our proposed IF-GMI maintains superiority in most metrics against the ResNet-152 trained on the CelebA. Our method achieves a remarkable increase of 10.6% in Acc@1 and significantly reduces the FID value using the StyleGAN2 trained on FFHQ. When utilizing the MetFaces StyleGAN2, our method still achieves much better results than the baseline despite a larger distributional shift, including a 38.8% increase in Acc@1 and competitive feature distance. In addition to ResNet-18, we evaluate the performance of the proposed method on more target models trained on FaceScrub, including ResNet-152, ResNeSt-101, and DenseNet-169. Benefiting from the fully utilized generative prior, our method achieves $13\% \sim 17\%$ improvement in Acc@1 metrics than the baselines and also achieves better results in most of the other metrics, as illustrated in Table 3.

The results presented above demonstrate that our method maintains outstanding attack performance in a variety of settings, exhibiting excellent generalizability and transferability. We also provide additional experimental results on more datasets and architectures in the Appendix.

4.5 Ablation Studies

To estimate the contributions from each component in our method, we conduct ablation studies on the ResNet-152 trained on the CelebA dataset using the StyleGAN2 trained on FFHQ. The results are presented in Table 4. More ablation studies are listed in the Appendix.

Intermediate Features Optimization. We merely remove the intermediate features optimization from our pipeline while keeping the remaining param-

Table 4: Ablation study performed on ResNet-152 trained on CelebA dataset with FFHQ as the public dataset. IF-GMI-i removes the intermediate feature optimization and only searches the latent space. IF-GMI-l removes the l_1 ball constraint compared to IF-GMI.

Method	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{face}$	$\downarrow \delta_{eval}$	$\downarrow$FID
IF-GMI-i	0.803	0.928	0.732	**314.275**	43.576
IF-GMI-l	0.945	0.992	0.678	315.278	37.528
IF-GMI	**0.947**	**0.993**	**0.677**	315.032	**37.461**

eters unchanged. As shown in the first row of Table 4, it leads to degradation up to 14% in $Acc@1$ and much worse FID without this technique, demonstrating the superiority of utilizing the hierarchical features of intermediate layers.

l_1 **Ball Constraint.**To avoid unreal image generation, we introduce the l_1 ball constraint into the intermediate features optimization. By observing the results shown in the second row of Table 4, the l_1 ball is beneficial in improving the performance in all metrics. Thus, we demonstrate the necessity of restricting the intermediate features within the l_1 ball constraint.

5 Conclusion

We proposed IF-GMI, a novel model inversion attack that performs effective attack in the OOD scenario. Surpassing the limitation of treating the generator as a black-box, we studied the structure and decomposed the generator into hierarchical layers, extending the optimization space from latent code to intermediate features to generate stable and high-quality images. Moreover, to avoid generating low-fidelity images, we applied a l_1 ball constraint to the optimization process. Through our extensive experiments, we demonstrated that the proposed IF-GMI achieves the state-of-the-art attack accuracy while generating samples with high fidelity and diversity.

Our exploration of enhanced utilization of intermediate features in the GAN prior contributes to advances in MI attack field, paving the way to more practical employment for MI attacks. We hope this paper can raise concerns about privacy leakage risk of released pre-trained models and facilitate more response to the threat of MI attacks.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant 62171248, 62301189, Guangdong Basic and Applied Basic Research Foundation under grant 2021A1515110066, the PCNL KEY project (PCL2021A07), and Shenzhen Science and Technology Program under Grant JCYJ20220818101012025, RCBS20221008093124061, GXWD20220811172936001.

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. 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: NDSS (2022)
3. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:1811.10597 (2018)
4. Chen, B., Feng, Y., Dai, T., Bai, J., Jiang, Y., Xia, S.T., Wang, X.: Adversarial examples generation for deep product quantization networks on image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1388–1404 (2022)
5. Chen, S., Kahla, M., Jia, R., Qi, G.J.: Knowledge-enriched distributional model inversion attacks. In: ICCV (2021)
6. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
7. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979 (2020)
8. Daras, G., Dean, J., Jalal, A., Dimakis, A.G.: Intermediate layer optimization for inverse problems using deep generative models. arXiv preprint arXiv:2102.07364 (2021)
9. Dataset, E.: Novel datasets for fine-grained image categorization. In: First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer. Citeseer (2011)
10. Fang, H., Chen, B., Wang, X., Wang, Z., Xia, S.T.: Gifd: A generative gradient inversion method with feature domain optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4967–4976 (2023)
11. Fang, H., Kong, J., Yu, W., Chen, B., Li, J., Xia, S., Xu, K.: One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. arXiv preprint arXiv:2406.05491 (2024)
12. Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., Chen, B., Wang, X., Xia, S.T.: Privacy leakage on dnns: A survey of model inversion attacks and defenses. arXiv preprint arXiv:2402.04013 (2024)
13. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: CCS. pp. 1322–1333 (2015)
14. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In: USENIX Security. pp. 17–32 (2014)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
20. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) **54**(11s), 1–37 (2022)
21. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
22. Kahla, M., Chen, S., Just, H.A., Jia, R.: Label-only model inversion attacks via boundary repulsion. In: CVPR (2022)
23. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
24. 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)
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
26. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
27. Li, C., Qiu, M.: Reinforcement learning for cyber-physical systems: with cybersecurity case studies. Chapman and Hall/CRC (2019)
28. 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)
29. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
30. 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)
31. Nguyen, N.B., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.: Rethinking model inversion attacks against deep neural networks. In: CVPR. pp. 16384–16393 (2023)
32. Park, J.Y., Smedemark-Margulies, N., Daniels, M., Yu, R., van de Meent, J.W., HAnd, P.: Generator surgery for compressed sensing. In: NeurIPS 2020 Workshop on Deep Learning and Inverse Problems (2020), `https://openreview.net/forum?id=s2EucjZ6d2s`
33. Qiu, H., Dong, T., Zhang, T., Lu, J., Memmi, G., Qiu, M.: Adversarial attacks against network intrusion detection in iot systems. IEEE Internet of Things Journal **8**(13), 10327–10335 (2020)
34. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

35. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020)
36. 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)
37. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: CCS. pp. 587–601 (2017)
38. Struppek, L., Hintersdorf, D., Correira, A.D.A., Adler, A., Kersting, K.: Plug & play attacks: Towards robust and flexible model inversion attacks. In: ICML (2022)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
40. Tewari, A., Elgharib, M., Bernard, F., Seidel, H.P., Pérez, P., Zollhöfer, M., Theobalt, C.: Pie: Portrait image embedding for semantic control. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020)
41. Wang, K.C., Fu, Y., Li, K., Khisti, A., Zemel, R., Makhzani, A.: Variational model inversion attacks. In: NeurIPS (2021)
42. Wu, C., Yan, M.: Session-aware information embedding for e-commerce product recommendation. In: Proceedings of the 2017 ACM on conference on information and knowledge management. pp. 2379–2382 (2017)
43. Yang, Z., Zhang, J., Chang, E.C., Liang, Z.: Neural network inversion in adversarial setting via background knowledge alignment. In: CCS (2019)
44. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8715–8724 (2020)
45. Yu, W., Chen, B., Zhang, Q., Xia, S.T.: Editable-deepsc: Cross-modal editable semantic communication systems. arXiv preprint arXiv:2310.10347 (2023)
46. Yu, W., Fang, H., Chen, B., Sui, X., Chen, C., Wu, H., Xia, S.T., Xu, K.: Gi-nas: Boosting gradient inversion attacks through adaptive neural architecture search. arXiv preprint arXiv:2405.20725 (2024)
47. Yuan, X., Chen, K., Zhang, J., Zhang, W., Yu, N., Zhang, Y.: Pseudo label-guided model inversion attack via conditional generative adversarial network. In: AAAI (2023)
48. Yuan, Z., Wu, F., Long, Y., Xiao, C., Li, B.: Secretgen: Privacy recovery on pretrained models via distribution discrimination. In: ECCV (2022)
49. Zeng, Y., Pan, M., Just, H.A., Lyu, L., Qiu, M., Jia, R.: Narcissus: A practical clean-label backdoor attack with limited information. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 771–785 (2023)
50. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2736–2746 (2022)
51. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: Generative model-inversion attacks against deep neural networks. In: CVPR (2020)
52. Zhong, X., Fang, H., Chen, B., Gu, X., Dai, T., Qiu, M., Xia, S.T.: Hierarchical features matter: A deep exploration of gan priors for improved dataset distillation. arXiv preprint arXiv:2406.05704 (2024)