Query-Efficient Model Inversion Attacks: An Information Flow View

Yixiao Xu, Binxing Fang, Mohan Li, *Member, IEEE*, Xiaolei Liu, *Member, IEEE*, and Zhihong Tian, *Senior Member, IEEE*

***Abstract*—Model Inversion Attacks (MIAs) pose a certain threat to the data privacy of learning-based systems, as they enable adversaries to reconstruct identifiable features of the training distribution with only query access to the victim model. In the context of deep learning, the primary challenges associated with MIAs are suboptimal attack success rates and the corresponding high computational costs. Prior efforts assumed that the expansive search space caused these limitations, employing generative models to constrain the dimensions of the search space. Despite the initial success of these generative-based solutions, recent experiments have cast doubt on this fundamental assumption, leaving two open questions about the influential factors determining MIA performance and how to manipulate these factors to improve MIAs. To answer these questions, we reframe MIAs from the perspective of information flow. This new formulation allows us to establish a lower bound for the error probability of MIAs, determined by two critical factors: (1) the size of the search space and (2) the mutual information between input and output random variables. Through a detailed analysis of generative-based MIAs within this theoretical framework, we uncover a trade-off between the size of the search space and the generation capability of generative models. Based on the theoretical conclusions, we introduce the Query-Efficient Model Inversion Approach (QE-MIA). By strategically selecting an appropriate search space and introducing additional mutual information, QE-MIA achieves a reduction of 60% ~ 70% in query overhead while concurrently enhancing the attack success rate by 5% ~ 25%.**

***Index Terms*—Model inversion attack, data privacy, deep neural network.**

I. Introduction

DEEP learning algorithms have emerged as a transformative technological breakthrough, finding widespread applications across diverse domains such as medical signal processing [1], [2], smart payments [3], and autonomous driving [4], [5]. However, the rapid expansion of data scale has raised widespread concerns regarding data security and privacy. Beyond the well-established risks of data leakage during collection and transmission, recent research has unveiled a novel category of data leakage risk known as memorization-based data leakage [6], [7], [8]. This risk becomes apparent during the deployment phase of deep learning models, where malicious users exploit the victim model by querying it to reconstruct characteristics of the training data at different levels.

Model Inversion Attacks, among these memorization-based attacks, aim to generate samples revealing privacy features of the training data [6]. For instance, a malicious user might reconstruct images closely resembling the target class in the training data, as illustrated in Fig 1. Fredrikson et al. [6] pioneered Model Inversion Attacks (MIAs) on simple machine learning models and shallow neural networks. They first defined the model inversion process as an optimization problem that maximises the output probability of the target class, and then proposed a gradient-based approach to solve this problem iteratively. However, this approach suffers from low attack success rates (and the corresponding high computational costs) when attacking more complex deep neural networks.

To perform efficient model inversion attacks under deep learning scenarios, subsequent research assumes that it is the expansive search space of MIAs that leads to poor attack performances [9], [10]. Based on this fundamental assumption, these studies concentrated on refining MIAs by narrowing the search space. Zhang et al. [9] first introduced the Generative Adversarial Network (GAN) [11] between the optimization target and the input of the victim model to narrow the search space of MIAs. By optimizing the input variable of the GAN instead of directly manipulating the images, their method significantly reduced the search space by over a factor of 1000. The success of GAN-based solutions inspired the following researchers to leverage more powerful generative models to further improve MIA. Consequently, variations of GAN-based methods become the major research point of recent studies (e.g., α-GAN-based MIA [12], StyleGAN-based MIA [10]).

Despite the initial success of existing generative-based MIAs, there are several problems left unexplored. First, some

Received 14 March 2024; revised 11 September 2024 and 28 October 2024; accepted 9 December 2024. Date of publication 18 December 2024; date of current version 8 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62372126, Grant 62372129, Grant U2436208, Grant 62272119, Grant 62072130, and Grant 62102379; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515030142; in part by the Key Technologies Research and Development Program of Guangdong Province under Grant 2024B0101010002; and in part by the Strategic Research and Consulting Project of the Chinese Academy of Engineering under Grant 2023-JB-13. The associate editor coordinating the review of this article and approving it for publication was Prof. Muhammad Khurram Khan. *(Corresponding author: Mohan Li.)*

Yixiao Xu is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China, also with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China, and also with the Huangpu Research School, Guangzhou University, Guangzhou 510530, China.

Binxing Fang, Mohan Li, and Zhihong Tian are with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China, and also with the Huangpu Research School, Guangzhou University, Guangzhou 510530, China (e-mail: limohan@gzhu.edu.cn).

Xiaolei Liu is with the Institute of Computer Application, Chinese Academy of Engineering Physics, Mianyang 621022, China.

Digital Object Identifier 10.1109/TIFS.2024.3518779

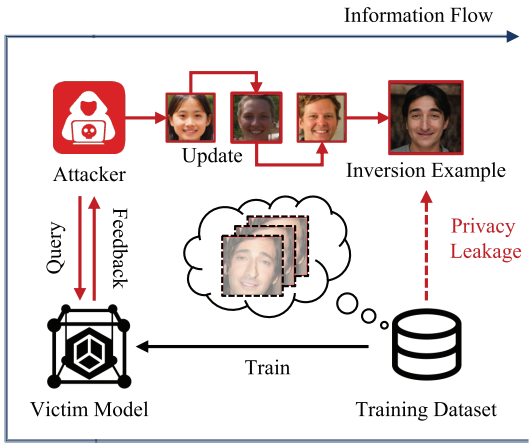

Fig. 1. An illustration of Model Inversion Attacks. The attacker reconstructs identical features of training data by iteratively querying the victim model and updating inversion examples, resulting in privacy leakage.

latest experiments results have cast doubt on the fundamental assumption of existing methods (e.g., the StyleGAN-based MIA [10] has a larger search space than the WGAN-based MIA [9] as well as a better attack performance). Second, even the state-of-the-art solution requires hundreds of attempts to get valid results [10], resulting in unacceptable computational overhead. These problems motivate us to rethink the relationship between the size of search space and the attack performance of MIAs. In other words, there is a lack of formal theoretical analysis of the influencing factors of MIAs and it is meaningful to find more potential optimization directions for MIAs.

In this paper, we tackle the aforementioned challenges by answering two major questions: **(1) What factors determines the attack performance of MIAs? (2) How to manipulate these influencing factors to enhance MIAs?**

Specifically, we conduct a comprehensive theoretical analysis of MIAs from an information flow view. First, we reformulate the inversion process using Fano's inequality [13]. This reformulation helps us establish a lower bound for the error probability in a one-time inversion attack. By utilizing this bound alongside an anticipated success probability, we can calculate the expected number of attempts required to achieve a valid result. Subsequently, we assess previous MIA approaches within our reformulated framework to identify potential optimization directions. We have two key observations: (1) narrowing the search space is not universally advantageous because there exists a trade-off between the dimensions of search space and the generation capability; (2) the output score will no longer provide useful information after attacking and there is a need of additional inductive biases for results selection. In light of these observations, we implement a query-efficient MIA on the basis of existing methods by simply choosing a proper search space and introducing the inductive bias about the inference logic.

In addition, we analyze how different model settings influence the vulnerability of deep learning models against MIAs. From the perspective of information flow, we prove that some specific kinds of deep learning models are more susceptible to MIAs, e.g., face recognition models that use hidden features to make decisions.

Our main contribution can be summarized as follows:

- We reformulate the model inversion process using information theory and demonstrate a lower bound for attack error probability, which provides a deeper understanding of the low success rate of model inversion attacks.
- Following the theoretical analysis, we demonstrate how different model settings influence the success of model inversion attacks and point out that sometimes robustness-enhancing methods (e.g. adversarial training) may introduce additional risk of model inversion attacks.
- On the basis of theoretical analysis, we find potential optimization directions for model inversion attacks and propose a query-efficient model inversion approach, QE-MIA.
- Experiments on three real-world datasets and multiple target models show that QE-MIA significantly reduces the attack overhead and enhances the attack success rate under complicated scenarios.

The rest of this paper is organized as follows: Sec II introduces the risk of privacy leakage in deep learning applications and the use of information theory in deep learning algorithms. We then provide the theoretical foundations and theoretical proofs of our approach in Sec III. Sec IV describes the proposed approach QE-MIA in detail. Subsequently, we evaluate and compare QE-MIA with previous methods in Sec V. Finally, Sec VI summarizes the research and propose potential future work.

II. Related Work

A. Privacy Leakage Risks in Deep Neural Networks (DNNs)

With the widespread application of deep learning models in security-critical domains, the importance of ensuring the security and privacy of training data is increasing. Extensive work has investigated the security of data during collection and transmission [14], e.g., federated learning [15], [16] for data privacy protection during model training. In the other part of the model lifecycle, namely the deployment process of models, there are much fewer considerations for data privacy, since it is difficult for an attacker to get direct access to the training data. However, recent studies have shown that even after extensive training, deep neural networks can store information from the training data individually without obfuscation [7], [8], [17]. Based on this property, malicious users can extract privacy-sensitive information from training data from different perspectives, these memorization-based attacks can be categorized into three classes, Membership Inference Attacks, Model Inversion Attacks, and Training Data Extraction Attacks.

1) Membership Inference Attacks: The Membership Inference Attack is one of the most widely-explored memorization-based attacks, it aims to identify the samples appearing in the training set from the test data [7], [18], [19], [20]. The basis of the membership inference attack is that deep learning models respond slightly differently to training and test samples [7]. Therefore, an attacker can train a binary

TABLE I
THE SETUP OF EXISTING MIA RESEARCHES

Method	Target Model	Target Data	Attack Setting		Generative Method
			White-Box	Black-Box	
MI [6]	LR,SNN	Image	✓		
GMI [9]	DNN	Image	✓		✓
KED [27]	DNN	Image	✓		✓
VMI [28]	DNN	Image	✓		✓
Mirror [25]	DNN	Image	✓		✓
PPA [10]	DNN	Image	✓		✓
GraphMI [24]	GNN	Node	✓		✓
UMI [23]	DNN	Image		✓	✓
S-MIA [22]	DNN	Image		✓	✓

classification model to capture this difference for recognizing members and non- members.

2) Training Data Extraction Attacks: Recently, with the development of large language models (LLMs), a new type of memorization-based attack, Data Extraction Attacks [8], [17], [21] are proposed to recover training data from the semantic level. Training Data Extraction Attacks consist of two steps, the attacker first generates extensive test samples using the victim model following the maximum likelihood strategy and then utilizes these samples as a test dataset to perform membership inference attacks against the victim model to determine which samples have appeared in the training dataset [17].

3) Model Inversion Attacks: Fredrikson et al. [6] were among the first to define the concepts of model inversion attacks, as well as to propose basic countermeasures. The attacker reconstructs privacy-sensitive features about the training class by solving an optimization problem that maximizes the output scores of the target class. However, the initial gradient-based approach for the optimization problem can only handle simple machine learning models (e.g. linear regression models and decision trees) and fails to provide meaningful results when dealing with more complex models such as deep neural networks. The following studies attribute the problem to the high dimensions of the search space and attempt to narrow the search space via generative approaches. Zhang et al. [9] introduce the Generative Adversarial Network (GAN) [11] to the inversion process and narrow the search space from the value space of an input image to the value space of an input vector of GAN. Later studies explored the effectiveness of different generative models for inversion attacks (e.g., StyleGAN [10]).

Some recent studies apply model inversion attacks under black-box conditions [22], [23] or on Graph Neural Networks (GNNs) [24]. Nevertheless, even the state-of-the-art inversion methods require thousands of attempts to get valid results, making them less practicable under query-limited scenarios. Meanwhile, some inconsistencies with existing hypotheses appear in the latest researches, for example, enlarging the search space may result in higher attack success rates [10], [25], [26]. These inconsistencies call for further theoretical analysis of the factors influencing model inversion attacks. Tab I lists the setup of several existing MIA research.

In this paper, we focus on model inversion attacks to analyze and improve model inversion attacks from the perspective of information flow.

TABLE II
NOTATIONS USED IN THE PAPER

Notation	Meaning
$\mathbb{X}, \mathbb{Y}$	Sets of real data and corresponding labels
$\overline{\mathbb{X}}, \overline{\mathbb{Y}}$	Sets of training data and corresponding labels
$\hat{\mathbb{X}}$	Set of real reconstructed examples
$\mathcal{F}$	Parameterized mapping function
$\mathcal{L}(.,.)$	Loss function
$\mathbf{X}, \mathbf{Y}$	Samples from $\mathbb{X}, \mathbb{Y}, \overline{\mathbb{X}}, \overline{\mathbb{Y}}$
α	Weight parameter
$I(.;.)$	Mutual information
$\mathbb{E}(.)$	Expectation of a random variable
$\mathcal{D}(.)$	Distance evaluation function
$H(.)$	Information entropy
P_e	Error probability
$\mathcal{G}$	Generative model
$\mathbf{Z}$	Input variable of generative models
$P_a(\mathbf{X})$	Probability that $\exists \mathbf{Z} \in \mathbb{Z}, \mathbf{X} = \mathcal{G}(\mathbf{Z})$

B. Information Theory in DNNs

In the field of deep learning theory, information-theory-based interpretable methods have been one of the important directions in the last decade. Tishby and Zaslavsky [29] first use information bottleneck theory to analyze the training process of DNNs and prove that any DNN can be quantified by the mutual information between the layers and the input and output variables. Schwartz-Ziv and Tishby [30] further investigate the role of hidden layers in DNNs and the interpretation of properties in the training process from an informational perspective.

Following the theoretical analysis, several studies apply information theory to evaluate and enhance the security of DNNs [31], [32]. However, these researches mainly focus on the forward information flow in the inference process (namely the information flow transmitted from the input example to the output probabilities of a DNN), leaving the backward information flow (the information flow transmitted from the output probabilities back to the input example) unexplored.

From the perspective of model inversion defense, Wang et al. [33] establish connection of attack success rate between the mutual information between input examples and output probabilities empirically. They further improve the robustness of DNNs against MIAs by narrowing the mutual information. However, the exact theoretical relationship between model inversion and mutual information remains unclear.

III. PRELIMINARY AND THEORETICAL ANALYSIS

In this section, we provide a basic definition and theoretical analysis of the factors influencing the model inversion process based on information theory. Table II lists the notations used in the paper and their meanings.

A. Model Inversion Attack: Definition

1) Supervised Learning: Consider the generalized supervised learning setting, we can formalize a supervised learning task as a six-tuple $(\mathbb{X}, \mathbb{Y}, \overline{\mathbb{X}}, \overline{\mathbb{Y}}, \mathcal{F}, \mathcal{L})$, where $\mathbb{X}$ is the set of real

data following an unknown distribution, $\mathbb{Y}$ is the set of corresponding labels of real data, $\overline{\mathbb{X}}$ and $\overline{\mathbb{Y}}$ denote the set of samples sampled from $\mathbb{X}$ and the set of corresponding learning targets sampled from $\mathbb{Y}$, $\mathcal{F}$ is a parameterized mapping $\mathcal{F}: \mathbb{X} \to \mathbb{Y}$ which represents the learning model, the loss function $\mathcal{L}$ is a mapping $\mathcal{L}: \mathcal{F}(\mathbb{X}) \circ \mathbb{Y} \to \mathbb{R}$.

A supervised learning process can be represented by the following optimization problem:

$$\arg\min_{\mathcal{F}} \mathop{\mathbb{E}}_{\mathbf{X}\in\overline{\mathbb{X}},\mathbf{Y}\in\overline{\mathbb{Y}}}[\mathcal{L}(\mathbf{X},\mathbf{Y})] \tag{1}$$

From the perspective of information bottleneck [29], $\mathcal{F}$ can also be considered as a series of parameterized mappings $\mathcal{F}: \mathbb{X} \to \mathbb{S}_1 \to \mathbb{S}_2 \ldots \to \mathbb{S}_n \to \mathbb{Y}$, where $\{\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_n\}$ denote the latent features of $\mathbb{X}$. The training process can be represented by the following optimization problem:

$$\arg\min_{\mathcal{F}} \mathop{\mathbb{E}}_{\mathbf{X}\in\overline{\mathbb{X}},\mathbf{Y}\in\overline{\mathbb{Y}}} \sum_{i=1}^{n} [I(\mathbf{S}_i,\mathbf{X}) - \alpha_i I(\mathbf{S}_i,\mathbf{Y})] \tag{2}$$

where $I(.,.)$ denotes the mutual information, α_i is the weight parameter determined by model architecture and settings. Eq 2 indicates that the training process of the model is to find a mapping to optimize the information compression and classification accuracy.

2) Model Inversion: Without loss of generality, we consider supervised learning in image classification scenarios, where $\mathcal{F}$ denotes a deep learning model trained on the image dataset $\overline{\mathbb{X}}$. Model inversion attacks aim to generate synthetic images that reveal information about the training dataset $\overline{\mathbb{X}}$. Let $\hat{\mathbb{X}}$ denote the set of images reconstructed by the attacker, MIA can be formalized as

$$\arg\min_{\hat{\mathbb{X}}} \mathop{\mathbb{E}}_{\mathbf{X}\in\overline{\mathbb{X}},\hat{\mathbf{X}}\in\hat{\mathbb{X}}} [\mathcal{D}(\mathbf{X},\hat{\mathbf{X}})] \tag{3}$$

where $\mathcal{D}$ evaluates the difference between two image sets. However, since the attacker has no access to $\overline{\mathbb{X}}$, it is not practicable to directly perform an attack by solving this optimization problem. Fredrikson et al. [6] adjust the optimization goal by maximizing the output score of the target class, that is,

$$\arg\min_{\hat{\mathbb{X}}} \mathop{\mathbb{E}}_{\hat{\mathbf{X}}\in\hat{\mathbb{X}},\mathbf{Y}\in\overline{\mathbb{Y}}} [\mathcal{L}(\mathcal{F}(\hat{\mathbf{X}}),\mathbf{Y})] \tag{4}$$

Eq. 4 is an intuition based on Eq. 1 that samples with similar output scores should be similar to the training samples. Nevertheless, according to Eq. 2, this intuition is questionable because the mutual information between the output scores and the input samples continues to decrease during training, which means it is difficult to build a solid connection between the similarity of output scores with the similarity of input samples. In practice, methods based on Eq. 4 fail to produce valid results when attacking more complex deep neural networks [6]. Therefore, the first key question in MIAs is:

Q1: What factors determine the low success rate of model inversion attacks?

Subsequent studies use the high dimensionality of the search space to answer this question empirically [6], [9], [28], which does not reveal all deterministic factors of the success rate of model inversion attacks. In this paper, instead, we answer this question from the perspective of information theory.

B. Error Probability of MIA: A Lower Bound

According to Fano's inequality [13], for any estimator $\hat{\mathbf{X}}$ that satisfies $\mathbf{X} \to \mathbf{Y} \to \hat{\mathbf{X}}$, we have:

$$H(P_e) + P_e log|\mathbb{X}| \geq H(\mathbf{X}|\hat{\mathbf{X}}) \geq H(\mathbf{X}|\mathbf{Y}) \tag{5}$$

where H denotes the information entropy, $P_e = Prob\{\mathbf{X} \neq \hat{\mathbf{X}}\}$, $\hat{\mathbf{X}}$ is an estimate of $\mathbf{X}$, $\hat{\mathbf{X}}$ is generated by a certain mapping function $\hat{\mathbf{X}} = \mathcal{G}(\mathbf{Y})$, and $\mathbb{X}$ denotes the value space of $\mathbf{X}$.

Considering the inversion process of the victim classifier $\mathcal{F}$, use random variables $\mathbf{X}, \mathbf{Y}, \hat{\mathbf{X}}$ to represent input examples, output probabilities, and inversion results, respectively. Then $\mathbf{X} \to \mathbf{Y}$ can represent the inference process of the classifier $\mathcal{F}$ and $\mathbf{Y} \to \hat{\mathbf{X}}$ represents the inversion process. Under this definition, $P_e = Prob\{\mathbf{X} \neq \hat{\mathbf{X}}\}$ is the error probability of the inversion result, which provides a lower bound for a one-time inversion attack. Specifically, we have the following theorem:

Theorem 1: Denote the original example as $\mathbf{X} \in \mathbb{X}$, the target model as $\mathcal{F}$, $\mathbf{Y} = \mathcal{F}(\mathbf{X})$ denotes the output score. $\mathbb{X}_\epsilon = \{\mathbf{X}_i | \forall \mathbf{X}_i \in \mathbb{X}, \|\mathbf{X}_i, \mathbf{X}\|_d \leq \epsilon\}$, $n = |\mathbb{X}_\epsilon|$. Then the error probability of a one-time model inversion attack p_e satisfies:

$$P_e \geq \prod_{i=1}^{n} \left(1 - \frac{I(\mathbf{X}_i;\mathbf{Y}) + 1}{log|\mathbb{X}|}\right) \tag{6}$$

where $I(\mathbf{X}_i;\mathbf{Y})$ denotes the mutual information of $\mathbf{X}_i$ and $\mathbf{Y}$.

Proof: According to the definition, $|\mathbb{X}| \geq 2$ and $H(P_e) \leq 1$ ($H(P_e) = 1$ when $P_e = 0.5$). Therefore, Eq. 5 can be weakened to

$$P_e \geq \frac{H(\mathbf{X}|\mathbf{Y}) - 1}{log|\mathbb{X}|} \tag{7}$$

According to the definition of mutual information, Eq. 7 can be transformed to:

$$P_e \geq \frac{H(\mathbf{X}) - I(\mathbf{X};\mathbf{Y}) - 1}{log|\mathbb{X}|} \tag{8}$$

For the model inversion scenario, $\mathbf{X}$ obeys a uniform distribution since each sample point takes place with the same probability in the search space. Under this assumption, $H(\mathbf{X}) = log|\mathbb{X}|$, thus we have:

$$P_e \geq 1 - \frac{I(\mathbf{X};\mathbf{Y}) + 1}{log|\mathbb{X}|} \tag{9}$$

For model inversion tasks, the goal is to reveal semantic features of the training data rather than to reconstruct the training example at the pixel level. For this purpose, any $\mathbf{X}_i$ satisfying $\|\mathbf{X}_i, \mathbf{X}\|_d \leq \epsilon$ will be considered as a successful attack. Therefore, the lower bound of the error probability for a single attack can be calculated by Eq. 6. □

Fig. 2 gives an overview of the overhead of the complete inversion process, where the query effort determines the number of required queries per attempt and the error rate determines the number of required attempts.

C. Quantification: An Example

Given Eq. 6, we can answer Question 1. For better understanding, we use a real-world attack as an example. For a deep recognition model trained on MNIST, we have $\mathbf{X} \in [0, 255]^{28\times28}$, $\mathbf{X}_{step} = 1$ (i.e., 256 different possible values in

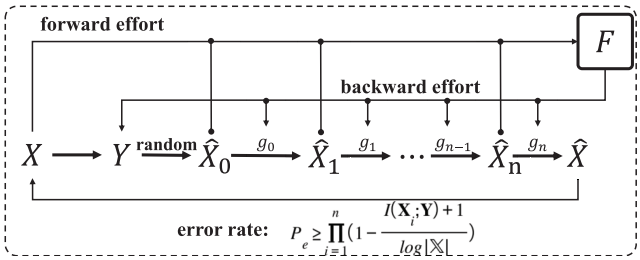

Fig. 2. Overview of the overhead of model inversion attacks, where the query effort determines the number of queries per attempt and consists of forward and backward overhead. The error rate determines the expected number of required attempts.

total for a pixel point). Meanwhile, $I(\mathbf{X};\mathbf{Y})$ can be estimated using information bottleneck theorem [30], for the classifier in the paper, $I(\mathbf{X};\mathbf{Y}) \approx 1$. Therefore, the error probability of a once-inversion attack against the MNIST classifier is approximately equal to 0.99984^n, where n denotes the number of samples in the critical domain of $\mathbf{X}$.

While specific model architectures and training datasets will affect the value of the mutual information [30], similar trends exist in most cases. For datasets that have larger scales, the lower bound of the error rate will be very close to 1, which leads to a low success rate for model inversion attacks. Using information theory, the answer to question 1 is:

A1: The success rate of model inversion attacks is jointly determined by the scale of the search space and the mutual information between input samples and output scores.

D. Generative Solutions

Recent MIA methods introduce the GAN structure to narrow the search space and generate semantic-meaningful images. Denote the GAN as $\mathcal{G}$ and the input variable of the GAN as Z, then the optimization problem in Eq. 4 can be reformulated as:

$$\arg\min_{\mathbb{Z}} \mathop{\mathbb{E}}_{\mathbf{Z}\in\mathbb{Z},\mathbf{Y}\in\mathbb{Y},} [\mathcal{L}(\mathcal{F}(\mathcal{G}(\mathbf{Z})),\mathbf{Y})] \tag{10}$$

Despite the success of GAN-based solutions in practice, there are still several concerns left unexplored:

Q2: Is the effectiveness of generative methods certified or what is the premise?

Following Question 1 and Answer 1, we further answer the above questions using the following corollary:

Corollary 1.1: Let $P_a(\mathbf{X}) = Prob\{\exists \mathbf{Z} \in \mathbb{Z}, \mathbf{X} = \mathcal{G}(\mathbf{Z})\}$*, then after introducing GAN-based solutions,* P_e *satisfies:*

$$P_e \geq \prod_{i=1}^{n}\left(1 - P_a(\mathbf{X}_i) \times \frac{I(\mathbf{Z};\mathbf{Y})+1}{log|\mathbb{Z}|}\right) \tag{11}$$

Proof: GAN-based methods assume that $\mathbf{X}$ can be represented by $\mathcal{G}(\mathbf{Z})$. Then P_e satisfies:

$$P_e \geq \prod_{\exists \mathbf{Z}\in\mathbb{Z},\mathbf{X}=\mathcal{G}(\mathbf{Z})}\left(1 - \frac{I(\mathbf{Z};\mathbf{Y})+1}{log|\mathbb{Z}|}\right) \tag{12}$$

When $\exists \mathbf{Z} \in \mathbb{Z}, \mathbf{X} = \mathcal{G}(\mathbf{Z})$, we have $log|\mathbb{Z}| = log|\mathcal{G}(\mathbb{Z})|$ because the GAN is a one-to-one mapping model. Let $P_a(\mathbf{X}) = Prob\{\exists \mathbf{Z} \in \mathbb{Z}, \mathbf{X} = \mathcal{G}(\mathbf{Z})\}$, then after introducing GAN-based solutions, P_e can be bounded by Eq. 11. □

According to Data Processing Inequality [34], $I(\mathbf{Y};\mathbf{X}) \leq I(\mathbf{Y};\mathcal{G}(\mathbf{Z})) = I(\mathbf{Y};\mathbf{Z})$. So when $\exists \mathbf{Z} \in \mathbb{Z}, \mathbf{X} = \mathcal{G}(\mathbf{Z})$:

$$1 - \frac{I(\mathbf{X};\mathbf{Y})+1}{log|\mathbb{X}|} \geq 1 - \frac{I(\mathbf{Z};\mathbf{Y})+1}{log|\mathbb{Z}|} \tag{13}$$

when $P_a = 1$, according to Eq. 13, the lower error bound of generative methods is lower than baseline methods defined by Eq. 4, which means the expected success rate of generative methods is strictly higher than baseline methods when $\mathbf{X}$ can be reproduced by $\mathcal{G}(\mathbf{Z})$ with 100% probability. However, P_a depends on a variety of factors and is usually less than 1 in practice. Consequently, the answer to Question 2 is as follows:

A2: The effectiveness of generative methods is not certified. There is a trade-off between the generation capacity of the generative model and the scale of search space.

E. Model Vulnerability

Besides the exploration of different model inversion attacks, it is also natural to wonder about the vulnerability of different deep-learning models. On the basis of Eq. 6, we further demonstrate that several specific types of models are more vulnerable to model inversion attacks in this section.

Consider the hidden Markov chain inference process in Eq. 2. According to the Data Processing Inequality [34], we have $I(\mathbf{X},\mathbf{X}_1) \geq \ldots \geq I(\mathbf{X},\mathbf{X}_n) \geq I(\mathbf{X},\mathbf{Y})$. We have demonstrated that the increment in mutual information will decrease the lower error boundary in Eq. 6. Then it is straightforward that using the hidden features to guide MIAs will increase the attack success rate.

In practice, face recognition models are usually pre-trained on a large-scale open-source dataset while deployed on a rather small set of registered identities (e.g., the employees of an institute). Therefore, deployed face recognition models often use the cosine similarity between the hidden features of input images and registered ones. This will make deployed face recognition models more vulnerable to model inversion attacks according to our analysis.

More importantly, adversarial training, which is often used to enhance model robustness, will increase the mutual information between the training samples and the output probabilities [32], [35] and thus may introduce additional risk about model inversion attacks.

IV. Query-Efficient Model Inversion Attack

Following theoretical analysis, we improve model inversion attacks from two directions and propose the Query-Efficient Model Inversion Attack (QE-MIA) in this section.

A. Search Space Selection

Eq. 11 and Answer 2 show that there is a trade-off between the generation capacity of the generative model and the scale of search space. Then finding the optimal search space will strictly reduce the lower error bound. Existing generative model inversion attacks [9], [10] concentrate on reducing the scale of the search space while neglecting the generation capacity of generative models on the search space. This motivates us to perform search space selection before performing model inversion attacks.

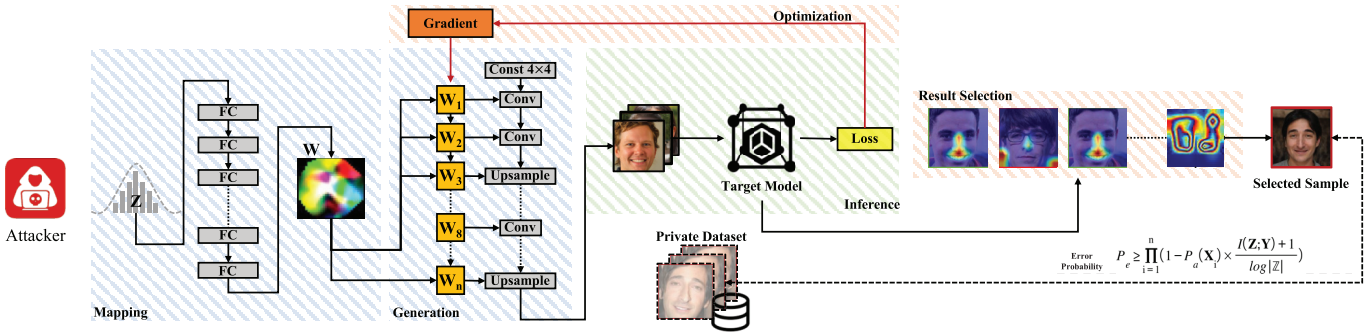

Fig. 3. High-level overview of QE-MIA. In the first phase, the attacker generates a series of inversion examples by optimizing the latent variables of the generative model. In the second phase, the attacker performs a voting-based ranking of the results, selecting the more robust example as the final result.

To achieve this goal, a basic problem is how to quantify the generative capacity of a generative model given a specific search space. Some previous studies about embedding images to the hidden layers of the StyleGAN [36] provide a practical solution. For a certain test example $\mathbf{X}$, we can find a corresponding $\mathbf{Z}$ which satisfies $\mathcal{D}(\mathbf{X}, \mathcal{G}(\mathbf{Z})) \leq \epsilon$ using gradient-based updating, where ϵ is a small constant value. The iterations cost by the process can represent the overhead of generating $\mathbf{X}$ using $\mathcal{G}$. Therefore, given a set of examples $\mathbb{X}$ and a certain interation number N, we can estimate the generation capacity of P_a using the following equation:

$$P_a = \mathop{\mathbb{E}}_{\mathbf{X}\in\mathbb{X},\mathbf{Z}\in\mathbb{Z}} Prob[Iter(\mathcal{G}, \mathbf{Z}, \mathbf{X}) < N] \tag{14}$$

where $\mathbb{Z}$ is the given search space. Then we can estimate the lower error bound of different generative models on different search spaces.

We perform a toy experiment on the Stanford Dogs [37] dataset to compare the estimated P_a, $I^*(\mathbf{Z}, \mathbf{Y})$, and P_e of different generative models on different search spaces. The WGAN [11] and StyleGAN2 [38] are trained on the AFHQ dataset [39], following [36], we randomly select 1000 examples from the Stanford Dogs dataset and embed them into different latent layers of the two generative models. We set the max number of iterations as 50 to keep up with the setting for model inversion attacks. Then we use a pre-trained classification model on the Stanford Dogs dataset to evaluate the embedding success rate and use the success rate as an estimation of P_a. Then we randomly generate 1000 samples for each optimization target, and estimate the mutual information $I^*(\mathbf{Z}, \mathbf{Y})$. Given the estimated P_a^* and $I^*(\mathbf{Z}, \mathbf{Y})$, we can calculate the estimated P_e^*. Tab III lists the results, where $1 - \sqrt[n]{P_e^*}$ is positively correlated with the attack success rate. Denote the $1 - \sqrt[n]{P_e^*}$ of WGAN by o, we find that performing an attack on $W+$ achieves the best attack success rate.

We also provide an intuitive explanation of this observation from the perspective of searching. Fig 4 illustrates how different optimization variables influence the attack results: (1) Optimizing $\mathbf{Z}$ will fall into local minima with a high probability because the distribution bias between $\mathbf{Z}$ and real samples, a small perturbation on $\mathbf{Z}$ may cause a significant difference in generated images. (2) It is easier to find a global optimum for $\mathbf{W}$ than $\mathbf{Z}$ because $\mathbf{W}$ has a more similar distribution with real samples [38]. However, since the number of dimensions of $\mathbf{W}$ is still much smaller than that of real samples, several non-targets will share the same projection on $\mathbb{W}$ as target samples, which leads to a decrement on P_a. (3) Compared with $\mathbf{W}$, $\mathbf{W}^+$ has more dimensions and thus contains more detailed information about the generated images. By optimizing $\mathbf{W}^+$, we balance the trade-off between the dimension of search space and the representation capabilities of the optimization variable.

TABLE III

THE ESTIMATED P_a, $I^*(\mathbf{Z}, \mathbf{Y})$, AND P_e OF DIFFERENT GENERATIVE MODELS ON DIFFERENT SEARCH SPACES. k DENOTES THE EFFECTIVE PRECISION OF THE SEARCH SPACE

Model	Search Space	$log\|\mathbb{Z}\|$	P_a^*	$I^*(\mathbf{Z}, \mathbf{Y})$	$1 - \sqrt[n]{P_e^*}$
WGAN	$\mathbb{Z}$	klog128	0.053	$\leq 1e-4$	o
StyleGAN2	$\mathbb{Z}$	klog512	0.145	$\leq 1e-4$	2.13o
	$\mathbb{W}$	klog512	0.429	$\leq 1e-4$	5.98o
	$\mathbb{W}^+$	klog15*512	0.714	$\leq 1e-4$	6.93o

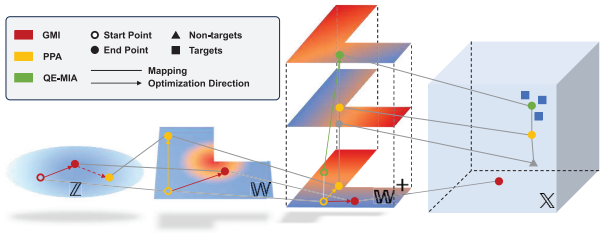

Fig. 4. An overview of how different search spaces influence the attack effectiveness.

B. Result Sorting

To get valid inversion results given the high error probability of a single model inversion attack, attackers tend to make multiple attempts. After several inversion attempts, the attacker will get a set of reconstructed images $\hat{\mathbb{X}} = \{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \ldots, \hat{\mathbf{X}}_n\}$. Then how to further evaluate these results and filter unsuccessful ones is another important problem in MIAs. Previous work [10], [28] performs random image transformations on generated images and selects images with more robust scores as the final results. However, these results sorting algorithms

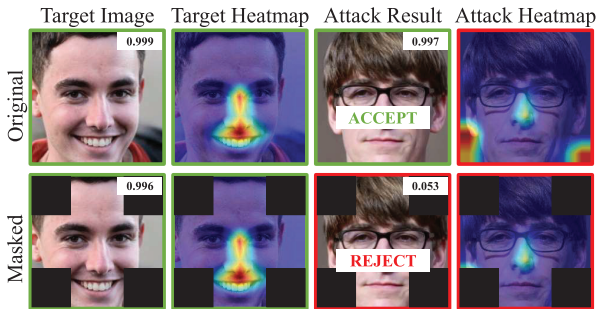

Fig. 5. An example of the mismatch between the image salient region and the decision salient region.

are query-consuming. Consequently, we reanalyze the result sorting process and propose a query-efficient solution.

Assume these reconstructed images are sampled from a certain set $\hat{\mathbb{X}}$ and denote the corresponding set of $\hat{\mathbf{Y}}$ as $\hat{\mathbb{Y}}$, we can update Eq. 9 as follows:

$$P_e \geq 1 - \frac{I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) + 1}{log|\hat{\mathbb{X}}|} \tag{15}$$

where $I(\hat{\mathbf{X}};\hat{\mathbf{Y}}) \approx 0$ because the $\hat{\mathbf{Y}}$ corresponding to each $\hat{\mathbf{X}}$ is the same (i.e., $\hat{\mathbf{Y}}$ converges to a constant value with probability of about 100%). Then the lower bound of error probability is only determined by $log|\hat{\mathbb{X}}|$, which means the random variable $\mathbf{Y}$ will not provide any additional information for results filtering after attacking. This observation motivates us to introduce additional information for results sorting.

Looking back into the model inference and inversion process $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{\mathbf{X}}$, attackers introduce an important inductive bias here that the output score of the target class should be 1 for any training example. This inductive bias provides all information contained in $\mathbf{Y}$. Therefore, if we can introduce more inductive biases about training examples, we can use them to sort and filter the attack results.

In this paper, we introduce another inductive bias about the inference logic of the target classifier: the inference logic of examples in the same category should be similar. This prior information is ignored during the inversion process. Fig 5 provides a visualized example to illustrate that inversion results that have different inference logic from others are unsuccessful results with high probability.

Specifically, for N unselected inversion results $\{\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \ldots, \hat{\mathbf{X}}_n\}$ of the same target class, we calculate corresponding decision heat-maps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_n\}$. Then the confidence score c_i of the $i-th$ result can be calculated using the following equation:

$$c_i = \|\mathbf{H}_i, \frac{1}{N}\sum_{k=1}^{n}\mathbf{H}_k\|_2 \tag{16}$$

Intuitively, samples that are far from the group's decision logic are more likely to be local optimal or adversarial examples. Compared with transformation-based result selecting methods [10], [28], this decision-logic-based method only requires one forward-backward query for an example, which is query-efficient.

Combining the search space selection algorithm with the result sorting method, we propose QE-MIA, the Query-Efficient Model Inversion Attack. Fig 3 illustrates the overview of QE-MIA.

V. Experiments

In this section, we assess the Query-Efficient Model Inversion Attack (QE-MIA) across various image classification scenarios and gauge its performance in comparison to earlier MIA techniques. We first outline the specific settings used in our experiments. Then we compare the overall performance of QE-MIA with previous methods and show that QE-MIA achieve the best attack success rate and attack efficiency. We also conduct ablation studies to affirm the accuracy of our theoretical analysis. Additionally, we delve into the vulnerability introduced by varying model architectures and configurations and find that some widely-used implementation strategy and algorithms make models more susceptible to model inversion attacks.

A. Experimental Setup

1) Dataset: Following recent MIA research studies [10], [22], we utilize three distinct datasets to assess and compare our approach with previous MIAs. Specifically, we employ two datasets containing human facial images: FaceScrub [40] and CelebA [41], as well as a dataset featuring dog images known as Stanford Dogs [37]. FaceScrub comprises a total of 106,863 face images featuring 530 different celebrities, while CelebA offers a more extensive dataset, with 10,177 unique identities and 202,599 facial images. In contrast, the Stanford Dogs dataset encompasses 120 distinct dog breeds and comprises a grand total of 20,580 images. We utilize these datasets to train the target model following the implementation details specified in PPA [10].

In addition to the target datasets, we leverage three prior knowledge datasets: FFHQ [42], MetFaces [43], and AFHQ [39]. These prior knowledge datasets consist of high-resolution images, providing a visual contrast to the target datasets, which predominantly consist of low-resolution images. This approach serves to validate the transferability of MIA methods across datasets with varying feature distributions. Furthermore, it's worth noting that the target dataset images are primarily sourced from the web and often lack strict alignment and preprocessing, reflecting real-world scenarios. In contrast, the images in the prior knowledge dataset are better organized, enhancing the generative model's capacity for producing high - quality synthetic data.

2) Model Selection: We employ three groups of pre-trained models for the target datasets as our target classification models. Each group includes a ResNet [44] model, a ResNeSt [45] model, and a DenseNet [46] model. To ensure a fair comparison, we utilize the pre-trained classification models made available in [10] instead of training new models from scratch. These models are widely used for face recognition and object classification in real-world scenarios. Threats against these models present specific challenges to data privacy in deep learning contexts. For performance evaluation, we also

utilize the pre-trained Inception-v3 [47] models released in PPA [10] for a fair comparison.

Regarding generative models, we opt for publicly accessible StyleGAN2 [38] models that have been trained on FFHQ [42], MetFaces [43], and AFHQ [39]. While more recent and potent generative models like StyleGAN3 [48] and diffusion models [49] are available, our primary focus in this paper is on enhancing generative MIAs within the context of a specific generative model.

3) Baseline Methods: We introduce four white-box model inversion attack methods as baseline methods for performance comparison. GMI [9] is the first generative model inversion attack using WGAN [11]. KED [27] trains a separate GAN for each target model, while VMI [28] fits a separate variational model for each target category and model. PPA [10] improves the transferability of generative model inversion attacks by introducing the powerful generative model StyleGAN [42] and performs transformations during the attack process.

For analyzing the model vulnerability, we introduce S-BMI [22] to perform model inversion attacks under black-box constraints and compare its performance under different model settings.

4) Evaluation Metrics: The evaluation of model inversion attacks remains an open question, primarily centered on the challenge of effectively gauging image similarity. While prior research has proposed various evaluation metrics, the reliability of these metrics remains largely empirical. These metrics can be broadly categorized into decision-based and statistical criteria. Decision-based criteria entail employing a third-party model to determine whether a reconstructed sample belongs to the target class. While this approach simulates human judgment, it can be influenced by the capabilities of the third-party model, including factors like local minima and adversarial examples. Statistical criteria, on the other hand, utilize statistical features to calculate the similarity between the reconstructed samples and real data distributions. However, the connection between these statistical features and the semantic content of images remains an under-explored area. To enhance the accuracy of model inversion attack assessments, we combine both decision-based and statistical metrics, following previous research. Furthermore, we conduct a manual survey experiment to evaluate the similarity between reconstructed samples and target samples.

For the initial comparison, we employ five metrics to evaluate our method. Firstly, for decision-based evaluation, we utilize Inception-v3 models [47] as independent evaluation models alongside the target models, calculating the Top-1 and Top-5 accuracy of the inversion results. In terms of statistical evaluation, we compute the feature distance δ_{eval} between the inversion results and the target training examples. To assess image quality, we employ FID [50] to evaluate the quality of the inversion results. Finally, we compare the average query numbers (ANQ) required for a successful attack by different methods. For the extended comparison between PPA and QE-MIA, we provide both the original sample and several reconstructed samples generated by various methods without labels. A group of observers then manually assigns scores based on the similarity between these samples. Table IV presents the evaluation metrics and perspectives used in the experiments conducted in this paper.

TABLE IV
EVALUATION METRICS USED IN THIS PAPER

Metric	Quality			Efficiency
	Decision-based	Statistical	Manual	
$\uparrow Acc@1$	✓			
$\uparrow Acc@5$	✓			
$\downarrow \delta_{eval}$		✓		
$\downarrow FID$		✓		
$\downarrow ANQ$				✓
$\uparrow Score_m$			✓	

TABLE V
EXPERIMENTAL RESULTS ON RESNET18 TRAINED ON FACESCRUB. ALL GENERATIVE MODELS ARE TRAINED ON FFHQ. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{eval}$	$\downarrow FID$	$\downarrow ANQ$
GMI [9]	12.46%	28.84%	155.25	85.30	12038.52
KED [27]	05.02%	09.94%	163.71	229.65	29880.48
VMI [28]	57.98%	68.66%	151.57	62.10	2587.10
PPA [10]	<u>87.60%</u>	<u>95.90%</u>	<u>125.01</u>	<u>43.77</u>	<u>684.93</u>
QE-MIA(Ours)	**93.10%**	**98.74%**	**119.52**	**42.68**	**219.12**

5) Attacker's Capability: In this paper, we mainly consider model inversion attacks under white-box constraints, where attackers have access to calculate gradients through the target model, which is the same as the baseline methods [9], [10], [27], [28].

6) Implementation Details: We build our algorithm mainly on the basic of PPA [10] using the PyTorch platform [51]. All experiments are carried out on 16×4352 CUDA cores.

B. Initial Comparison

We first compare our method with four baseline methods on a ResNet-18 model pre-trained on the FaceScurb dataset. We perform different MIAs on the model following the same implementation details in [10]. Table V displays the performances of these five methods. The experimental results demonstrate that QE-MIA outperforms the four baseline methods across all five metrics. Additionally, we make several observations based on the data presented in Table V:

(1) In terms of decision-based metrics, both PPA and QE-MIA exhibit significantly higher Top-1 and Top-5 accuracy. This improvement cannot be attributed to a reduction in the search space, as suggested by earlier studies [9], [10], since the search spaces of PPA and QE-MIA encompass more dimensions than the other baseline methods. According to our theoretical analysis, the inclusion of the more potent StyleGAN enhances the generative capability of the attack, leading to heightened attack accuracy. Consequently, by expanding the search space to a balanced extent, QE-MIA further improves the attack performances compared to PPA.

(2) When assessing statistical metrics, we observe similar trends as those seen in the decision-based metrics, but with significant variations in the extent of change. For example,

when compared with GMI, VMI significantly improves Top-1 accuracy by 45.52%, while only enhancing δ_{eval} by 3.68 out of 155.25. This observation underscores the disparities between existing evaluation metrics and underscores the need for more comprehensive metrics.

(3) Regarding efficiency metrics, GMI and KED demand over 10,000 queries on average to the target model for a successful attack, rendering them computationally intensive and less practical under query-limited conditions. QE-MIA, on the other hand, only requires an average of 219.12 queries to obtain a valid result, with a query overhead of 31.99% compared to PPA and 8.47% compared to VMI, respectively. This improvement is attributed to two factors: QE-MIA enhances the attack success rate by selecting an appropriate search space, leading to a reduction in the number of attempts required for a valid outcome. Additionally, QE-MIA reduces the number of queries needed for a single attempt by introducing additional inductive biases.

It's worth noting that the computational overhead encompasses both the preparation process and the attack process. In the methods mentioned, GMI necessitates the training of a WGAN for a specific target dataset, while KED and VMI require training the corresponding generative model for each model and class, respectively. Only PPA and QE-MIA utilize publicly available generative models for attacks without the need for additional training, further widening the gap in computational overhead between the various approaches.

C. Extended Comparison

We further evaluate QE-MIA on more datasets and models. For extended evaluation, we only compared QE-MIA with the Plug&Play Attack [10], as it had already outperformed other existing approaches by a large margin as shown in Table V. First, we compared the performance of PPA and QE-MIA using FFHQ as prior knowledge for human facial images and AFHQ for animal images. Table VI presents the experimental results across three different target models on three distinct datasets. In general, a similar trend to Table V can be observed in Table VI. QE-MIA demonstrates a significant improvement in decision-based metrics and a slight improvement in statistical metrics compared to PPA. Moreover, QE-MIA substantially enhances attack efficiency compared to PPA. We can made the following observations based on Table VI:

(1) Across different datasets, the decision-based metrics (attack success rate) exhibit significant variations. The effectiveness of the attack is observed to be negatively correlated with the size of the training dataset. Large-scale datasets contain more information (bits) compared to smaller ones. When employing the same deep learning model to create classifiers for different datasets, the information loss rate is higher in large-scale datasets, resulting in less mutual information between input samples and output scores. Our theoretical analysis suggests that a reduction in mutual information leads to a decrease in the success rate of the attack, which aligns with the experimental results.

(2) For different target models trained on the same dataset, the attack success rates also have a large variance even when the capabilities of these models are very close. For example, the test accuracy of the target ResNeSt-101, ResNet-152, and DenseNet-169 on CelebA are 87.35%, 86.78%, and 85.39%, respectively. While the Top-1 accuracy rate of QE-MIA against these three different models are 97.57%, 91.81%, and 81.49%, respectively. This observation motivates us to further analyze the vulnerability of different models in the following sections.

TABLE VI

EXPERIMENTAL RESULTS ON RSENET-152, DENSENET-169, AND RESNEST-101. UTILIZING FFHQ AS HUMAN FACIAL PRIOR KNOWLEDGE AND AFHQ AS ANIMAL FACE PRIOR KNOWLEDGE. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

		Model	↑ $Acc@1$	↑ $Acc@5$	↓δ_{eval}	↓FID	↓ANQ
FaceScrub	PPA	ResNeSt-101	93.83%	99.68%	122.14	46.35	639.45
		RseNet-152	92.90%	99.54	120.73	46.58	645.86
		DenseNet-169	94.77%	99.71%	116.42	46.42	633.11
	Ours	ResNeSt-101	**98.51%**	**99.93%**	118.79	45.84	**203.03**
		ResNet-152	96.42%	99.88%	**115.64**	**45.17**	207.43
		DenseNet-169	97.59%	99.91%	117.69	46.52	204.94
CelebA	PPA	ResNeSt-101	81.85%	94.46%	302.70	44.19	733.05
		ResNet-152	80.14%	94.80%	308.65	40.69	748.69
		DenseNet-169	73.76%	89.47%	310.83	41.25	813.45
	Ours	ResNeSt-101	**93.57%**	99.57%	**297.50**	40.37	**204.98**
		ResNet-152	91.81%	**99.81%**	305.27	**40.16**	217.84
		DenseNet-169	81.49%	95.91%	311.14	42.63	245.43
St. Dogs	PPA	ResNeSt-101	91.29%	98.83%	60.87	33.51	657.25
		ResNet-152	95.33%	99.55%	59.90	32.15	629.39
		DenseNet-169	93.86%	99.47%	60.31	32.07	639.25
	Ours	ResNeSt-101	94.67%	99.28%	60.35	32.18	211.26
		ResNet-152	**97.43%**	99.35%	**58.44**	**31.99**	**205.28**
		DenseNet-169	95.82%	**99.64%**	60.17	32.37	208.72

(3) In contrast to decision-based metrics, QE-MIA only marginally enhances statistical metrics compared to PPA. Moreover, we observe that sometimes inconsistent trends arise between decision-based and statistical metrics, such as a higher success rate alongside a higher FID, highlighting the limitations of existing evaluation metrics for model inversion attacks.

To further validate the quality of reconstructed images, we conducted additional manual evaluations on the attack results. Figure 6 provides visualized examples of attack results from Table VI. Based on these visual results, we performed manual assessment experiments. Specifically, we randomly selected 50 target classes from CelebA and FaceScrub, generated and selected reconstruction samples using both PPA and QE-MIA. We then invited five observers to independently score the target samples and their corresponding reconstructed samples. The attack methods corresponding to the reconstructed samples were not disclosed during the experiment. Each group of samples was awarded one point for the reconstructed sample that visually resembled the original sample more closely, with no points awarded to the other sample. Table VII presents the statistics for individual observers and the average scores.

Our observations indicate that the visual quality of reconstructions produced by PPA and QE-MIA is very similar on small-scale datasets, which aligns with the higher attack success rates of both methods on FaceScrub. However, on CelebA, QE-MIA improves the visual similarity between reconstruction samples and target samples while simultaneously achieving higher attack success rates.

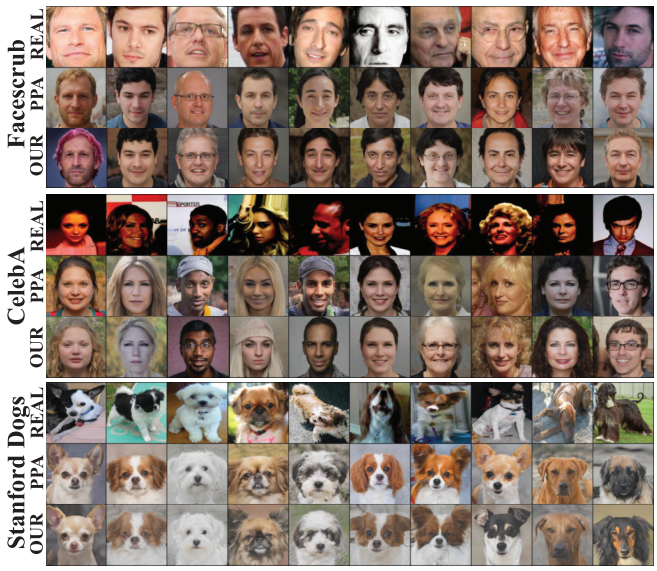

Fig. 6. A visualization example of Tab VI.

TABLE VII
MANUAL SCORES OF ATTACK RESULTS OF PPA AND QE-MIA

Dataset	Method	W_1	W_2	W_3	W_4	W_5	$Score_m$
CelebA	PPA	30	26	27	28	29	0.56
	QE-MIA	20	24	23	22	21	0.44
Facescrub	PPA	26	27	22	26	26	0.52
	QE-MIA	24	23	28	24	24	0.48

Fig. 7. A visualization example of Tab VIII.

TABLE VIII
EXPERIMENTAL RESULTS ON RSENET-152, DENSENET-169, AND RESNEST-101. UTILIZING METFACES AS HUMAN FACIAL PRIOR KNOWLEDGE. THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED RESPECTIVELY

		Model	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow\delta_{eval}$	$\downarrow FID$	$\downarrow ANQ$
FaceScrub	PPA	ResNeSt-101	75.58%	93.26%	135.44	89.06	793.86
		RseNet-152	72.89%	91.75%	141.24	**68.12**	823.16
		DenseNet-169	80.14%	95.33%	126.08	78.02	748.69
	Ours	ResNeSt-101	85.64%	98.72%	130.45	82.52	233.54
		RseNet-152	78.23%	95.96%	133.10	68.86	255.66
		DenseNet-169	**88.20%**	**99.11%**	**121.35**	75.13	**226.76**
CelebA	PPA	ResNeSt-101	36.45%	60.84%	385.73	74.69	1646.09
		RseNet-152	40.02%	65.80%	385.09	74.28	1499.25
		DenseNet-169	30.72%	55.81%	394.05	82.11	1953.13
	Ours	ResNeSt-101	61.73%	88.28%	**366.43**	**70.30**	323.99
		RseNet-152	**67.21%**	**92.59%**	368.57	71.28	**297.57**
		DenseNet-169	55.48%	76.05%	386.80	74.92	360.49

D. Transferability

Another important capability of model inversion attack methods is the ability to migrate between different distributions. Since the attacker lacks access to the target dataset, a substantial gap often exists between the distribution of the prior knowledge dataset and the distribution of the victim dataset. In line with previous methods [10], we employed StyleGAN2 trained on the Metfaces dataset to evaluate the transferability of QE-MIA.

As presented in Table VIII, QE-MIA markedly enhances attack accuracy compared to PPA. Utilizing the same generative model as PPA, QE-MIA better strikes a balance between generative capability and search complexity through the selection of an appropriate search space. Consequently, it achieves improved transferability between distinct distributions. When compared to PPA, QE-MIA enhances the attack success rate by an average of 7.82% on the FaceScrub dataset. For the larger - scale CelebA dataset, QE-MIA improves the attack success rate by an average of 25.74%. From an attack efficiency standpoint, QE-MIA requires only 30.26% and 19.26% of the query times needed by PPA on FaceScrub and CelebA, respectively. Figure 7 provides visual examples of the results from Table VIII. As shown in the figure, the images generated by QE-MIA exhibit highly recognizable and similar features, despite undergoing a substantial stylistic shift from the original distribution.

Combining the results from Table VI and Table VIII, we observe that QE-MIA achieves significant performance improvements, particularly on larger datasets and when the prior distribution is significantly different from the target distribution. This makes QE-MIA applicable to a wide range of attack scenarios.

E. Ablation Study

To provide a more comprehensive analysis of QE-MIA, we conducted a series of ablation studies to validate the correctness of our theoretical analysis and the effectiveness of the proposed algorithms. Specifically, we delved into the effectiveness of applying different search spaces to QE-MIA under varying prior knowledge. Furthermore, we examined how different decision heat-map generation algorithms and different group sizes influence the effectiveness of the proposed result selection method.

In our evaluation of the search space selection process, we conducted QE-MIA on the CelebA dataset using generative models trained on FFHQ and Metfaces. We considered three baseline search spaces: Z, W, and $W+$. For $W+$, which comprises multiple components with the same structure in StyleGAN2, we explored different subsets, as each style controller vector contains 18 components, represented by $\{W_1, W_2, \ldots, W_{18}\}$. These components control the generation process at various levels. We gradually incorporated more components into the search space, moving from lower to

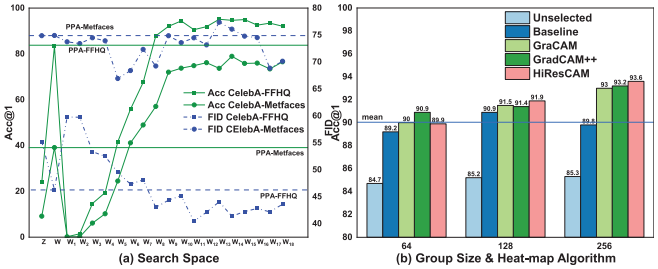

Fig. 8. Experimental results of ablation studies. (a) Search space; (b) group sizes and heat-map algorithms.

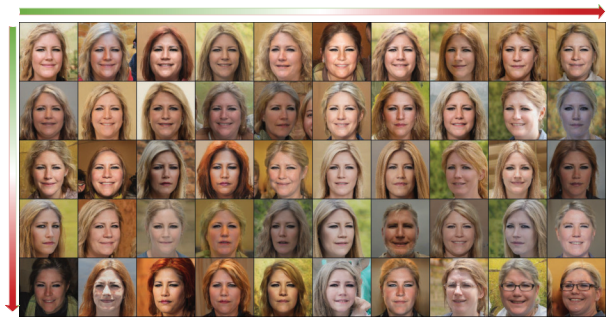

Fig. 9. A visualized example for results selection.

higher levels, leading to a range of attack success rates and FID scores.

Figure 8 (a) presents the results of the ablation experiments on the search space selection process. The findings are as follows:

(1) Variations in the search space have a notable impact on both decision-based metrics (attack success rates) and statistical metrics (FID). For the experimental conditions, QE-MIA achieves the best attack performance when the search space includes 13 to 15 components of $W+$. As we gradually reduce the number of components in the search space, the attack performance of QE-MIA gradually decreases, as the generative capability of generative models is not fully utilized.

(2) For different prior knowledge datasets, the same generative model achieves optimal performance in a close search space (e.g., generative models trained on FFHQ and Metfaces both achieve the best performance near $W14$), which suggests that we can select and migrate the search space on an existing dataset to an unknown dataset, improving the utility of QE-MIA.

In our exploration of the result selection algorithm's impact on QE-MIA's performance, we employed decision heat-map generation algorithms (e.g., Grad-CAM [52]) to generate decision heat-maps corresponding to each result for each target class. We then computed an average heat-map and compared the similarity between individual heat-maps and the average heat-map. This process resembles a form of collective voting. In the ablation study, we assessed the influence of two key factors on the result selection process: the decision heat-map generation algorithm and the group size. We conducted experiments on the CelebA dataset and recorded the results in Figure 8 (b). Here are our observations regarding the result selection process:

(1) As the group size increases, the success rates of unselected results and results chosen by the baseline method (transformation-based selection) remain largely constant, as the selection process is unrelated to group size. Conversely, the success rates of results chosen by different heat-map-based methods steadily rise, indicating that an increase in group size enhances the accuracy of the group members' voting process.

(2) With smaller group sizes, the accuracy of decision-heat-map-based methods closely aligns with that of the transformation-based method, but with significantly reduced computational overhead. As the group size grows, the accuracy and efficiency of decision-heat-map-based methods far outstrip the transformation-based method.

(3) The choice of heat-map generation methods has a limited impact on result selection because the multi-member voting process blurs the subtle differences between heat maps generated by different methods.

Figure 9 provides a visual example of QE-MIA's result selection process. For an attack with a group size of 50, QE-MIA ranks the group members using the decision-heat-map-based algorithm. As shown in Figure 9, the top-ranked samples exhibit robust similarity features, while the lower-quality generated samples are ranked lower in the hierarchy.

TABLE IX

EXPERIMENTAL RESULTS OF MODEL INVERSION ATTACKS AGAINST DIFFERENT MODEL SETTINGS

	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{eval}$	$\downarrow FID$	$\downarrow ANQ$
ResNeSt-101 (white-box)	94.28%	99.35%	299.03	41.97	212.13
ResNeSt-101 (white-box+deployed)	97.50%	99.81%	297.85	40.24	205.13
ResNeSt-101 (black-box)	71.64%	90.88%	312.43	45.57	5583.47
ResNeSt-101 (black-box+deployed)	85.25%	94.16%	308.27	43.35	4692.08

F. Model Vulnerability

As per our theoretical analysis, the efficacy of model inversion attacks is linked to the mutual information of input and output variables, a value influenced by specific model settings. We sought to validate this conclusion by conducting experiments. As displayed in Table IX, under both white-box and black-box conditions, deployed face recognition models (models utilizing hidden features for prediction) are more susceptible to model inversion attacks. Specifically, for black-box conditions, the deployment process of face recognition models increased the attack success rate by an average of 13.61% and reduced the query overhead by an average of 15.96%. This finding introduces a new threat, suggesting that specific settings during real-world applications may lead to model privacy leakage.

We also examined the impact of adversarial training on model inversion attacks. Using an adversarial training method with acceptable computational overhead [53], we trained the DenseNet-169 model on the Stanford Dogs dataset. We then compared the attack performance of QE-MIA on the original and adversarially pre-trained models. As indicated in Table X,

TABLE X
EXPERIMENTAL RESULTS OF MODEL INVERSION ATTACKS AGAINST ADVERSARIAL ROBUST MODELS

Model (Test Acc)	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{eval}$	$\downarrow FID$	$\downarrow ANQ$
Densenet-169 (74.39%)	95.82%	99.67%	60.17	32.37	208.72
ADV-Densenet-169 (66.23%)	98.15%	99.86%	60.06	32.54	203.77

although adversarial training slightly decreased the classification accuracy for clean test examples and the attack success rate for adversarial attacks, the attack success rate of model inversion attacks increased. Intuitively, the adversarial training process reduces the probability $P(\mathcal{F}(\hat{\mathbf{X}}) = \mathcal{F}(\mathbf{X})$ *and* $\hat{\mathbf{X}} \neq \mathbf{X})$, leading to an increase in $I(\mathbf{X}, \mathbf{Y})$. According to our theoretical analysis, the adversarial training process diminishes the error probability of model inversion attacks.

From a model vulnerability standpoint, we demonstrate that specific model settings, including those considered to enhance robustness, may elevate the risk of model inversion attacks in practical scenarios. This finding offers a new perspective for assessing model vulnerability.

G. Model Inversion Defense

To alleviate the privacy leakage risk introduced by model inversion attacks, recent studies proposed several model inversion defense methods [6], [33], [54], [55]. These methods can be divided into two main categories: (1) reducing the information available to model inversion attackers [6], [33]; (2) misleading the model inversion process [54], [55]. Therefore, we further evaluate the attack performances of different model inversion attacks under these defense methods to compare the robustness of different attack methods.

Specifically, we compare the attack success rates (ASR) and the top-5 attack success rates (ASR-5) of four baseline attacks and QE-MIA under No Defense, Differential Privacy (DP) [6], and GAN-ID [55]. Of the two defense methods, DP and GAN-ID represent the first type and the second type of defense methods, respectively. DP utilizes differential privacy techniques to reduce recognizable features of training examples contained in classification information. GAN-ID uses generative models to generate and inject several fake targets in the training dataset and misleads the attackers to reconstruct fake targets rather than protected targets.

Tab. XI provides the experimental results of attacks under model inversion defenses. Generally, QE-MIA maintains the highest attack success rate under different situations, despite the varying degrees of competence of DP and GAN-ID in defending against different attack methods. By investigating the results of the experiment, we can draw several further conclusions:

(1) QE-MIA can bypass the first type of defense methods: Although DP slightly reduced the attack success rate of QE-MIA, there exists a trade-off between defense ability and model usability (e.g., to effectively reduce the attack success rate, DP will cause a 20% ∼ 30% decrease in the clean accuracy of protected models). In order to ensure the availability of the target model, the first type of defense methods has a maximum percentage of limitations on the available information, and QE-MIA can still utilize the remaining information to achieve model inversion attacks in practice.

TABLE XI
EXPERIMENTAL RESULTS OF DIFFERENT MODEL INVERSION ATTACKS AGAINST TWO DEFENSE METHODS

Attack	No Defense		DP		GAN-ID	
	ASR	ASR-5	ASR	ASR-5	ASR	ASR-5
GMI	0.12	0.29	0.05	0.13	0.00	0.08
KED	0.05	0.10	0.02	0.08	0.00	0.05
VMI	0.58	0.69	0.39	0.57	0.00	0.10
PPA	0.88	0.96	0.51	0.74	0.13	0.32
QE-MIA	**0.93**	**0.99**	**0.68**	**0.87**	**0.27**	**0.44**

TABLE XII
EXPERIMENTAL RESULTS ON HUMAN FACIAL DATASETS USING AFHQ AS PRIOR KNOWLEDGE

Dataset	Method	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{eval}$	$\downarrow FID$	$\downarrow ANQ$
CelebA	PPA	0.71%	5.89%	406.54	97.75	84,507.04
	QE-MIA	4.82%	18.39%	398.08	93.86	4149.38
Facescrub	PPA	1.07%	3.57%	144.93	88.45	56074.77
	QE-MIA	10.06%	15.32%	139.55	85.60	1988.07

(2) QE-MIA is more robust against the second type of defense methods: As can be obsered in Tab. XI, GAN-ID can successfully defend GMI, KED, and VMI with 100% probabilities. This is because these attack methods are based on the same or a similar prior data distribution with the target data distribution, thus GAN-ID can take advantage of the same prior knowledge to generate fake targets. The transferability of QE-MIA supports it to utilize a prior data distribution that has a large shift from the target data distribution, which makes QE-MIA more robust against the second type of defense methods.

H. Limitations and Future Work

While the proposed method QE-MIA achieves superior performance in terms of attack success rate and efficiency, it shares some limitations with existing methods. Generative-based model inversion attacks, including QE-MIA, make a crucial prior assumption: the attacker already knows the data type of the target model. When this assumption is not met, existing methods are highly likely to fail. Intuitively, if the attacker uses a GAN trained on animal images to conduct model inversion attacks on face recognition models, the probability P_a in Eq. 11 will approach zero, resulting in a high error probability. Table XII provides experimental results for PPA and QE-MIA when attacking ResNeSt-101 trained on different human facial datasets using the AFHQ dataset as prior knowledge. As shown in Table XII, while QE-MIA still has a slightly higher attack success rate than PPA, both methods have significantly lower success rates than when using human facial datasets as prior knowledge. Furthermore, as depicted in Figure 10, even the successful attack results exhibit a substantial offset from the identifiable features of real samples, potentially leading to misclassification of the target data type. These limitations of generative-based model inversion attacks warrant further research to address scenarios

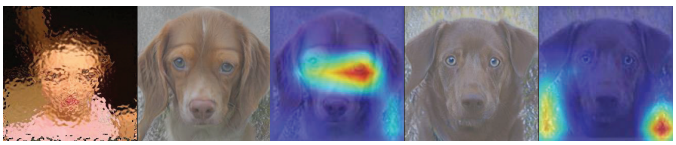

Fig. 10. A visualized example of Tab XII: (a) original image (we blurred it to avoid ethical issues); (b) attack results share similar features but with large style offset; (c) decision heat-map of b; (d) attack results behave as adversarial examples; (e) decision heat-map of d.

with limited or even incorrect prior knowledge, which also represents a potential direction for future work.

VI. Conclusion

In this paper, we theoretically analyze the model inversion process from the perspective of information flow and demonstrate the lower boundary of the error probability of a single attack. Based on the theoretical analysis, we propose QE-MIA, a query-efficient model inversion attack. By choosing the proper optimization variable and introducing additional inductive biases, QE-MIA reduced the attack overhead by a large margin (60% $\sim$ 70%) while achieving a significantly better attack success rate (5% $\sim$ 25%). Additionally, we analyze the vulnerability caused by different model settings from the perspective of information flow and demonstrate that specific model settings will increase the risk of model inversion attacks in practice. We hope our method can provide a new perspective for analyzing model inversion attacks and motivate future work. The most relevant future work might be performing valid model inversion attacks under limited prior knowledge, where the attacker has no knowledge about the target data type.

References

[1] F. Jiang et al., "Medical image semantic segmentation based on deep learning," *Neural Comput. Appl.*, vol. 29, no. 5, pp. 1257–1265, 2018.

[2] S. Wang et al., "Annotation-efficient deep learning for automatic medical image segmentation," *Nature Commun.*, vol. 12, no. 1, p. 5915, Oct. 2021.

[3] A. Shukla, P. Bhattacharya, S. Tanwar, N. Kumar, and M. Guizani, "DwaRa: A deep learning-based dynamic toll pricing scheme for intelligent transportation systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12510–12520, Nov. 2020.

[4] Y. Almalioglu, M. Turan, N. Trigoni, and A. Markham, "Deep learning-based robust positioning for all-weather autonomous driving," *Nature Mach. Intell.*, vol. 4, no. 9, pp. 749–760, Sep. 2022.

[5] X. Liu, Y. Zhou, and C. Gou, "Learning from interaction-enhanced scene graph for pedestrian collision risk assessment," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 9, pp. 4237–4248, Sep. 2023.

[6] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Denver, CO, USA, I. Ray, N. Li, and C. Kruegel, Eds., Oct. 2015, pp. 1322–1333.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, May 2017, pp. 3–18.

[8] N. Carlini et al., "Extracting training data from large language models," in *Proc. 30th USENIX Secur. Symp.*, M. Bailey and R. Greenstadt, Eds., Aug. 2021, pp. 2633–2650.

[9] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 250–258.

[10] L. Struppek, D. Hintersdorf, A. D. A. Correia, A. Adler, and K. Kersting, "Plug & play attacks: Towards robust and flexible model inversion attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 162, Baltimore, MD, USA, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., 2022, pp. 20522–20545.

[11] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 214–223.

[12] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 357–372, 2022.

[13] R. Fano, "Class notes for transmission of information," in *Course 6.574*. Cambridge, MA, USA: MIT, 1952.

[14] M. Li, Z. Tian, X. Du, X. Yuan, C. Shan, and M. Guizani, "Power normalized cepstral robust features of deep neural networks in a cloud computing data privacy protection scheme," *Neurocomputing*, vol. 518, pp. 165–173, Jan. 2023.

[15] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, Jun. 2021.

[16] Z. Li, L. Wang, Z. Gu, Y. Lv, and Z. Tian, "Labels are culprits: Defending gradient attack on privacy," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6007–6019, Feb. 2024.

[17] W. Yu et al., "Bag of tricks for training data extraction from language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 202, Honolulu, HI, USA, Jan. 2023, pp. 40306–40320.

[18] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1964–1974.

[19] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *Proc. IEEE Symp. Security Privacy (SP)*, Jun. 2022, pp. 1897–1914.

[20] C. Wu et al., "Rethinking membership inference attacks against transfer learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6441–6454, 2024.

[21] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. Wallace, "Does BERT pretrained on clinical notes reveal sensitive data?," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 946–959.

[22] Y. Xu, X. Liu, T. Hu, B. Xin, and R. Yang, "Sparse black-box inversion attack with limited information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2023, pp. 1–5.

[23] R. Liu et al., "Unstoppable attack: Label-only model inversion via conditional diffusion model," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3958–3973, 2024.

[24] Z. Zhang, Q. Liu, Z. Huang, H. Wang, C.-K. Lee, and E. Chen, "Model inversion attacks against graph neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8729–8741, Sep. 2022.

[25] S. An et al., "MIRROR: Model inversion for deep learning network with high fidelity," in *Proc. 29th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Apr. 2022, pp. 1–18.

[26] Z. Zhang, X. Wang, J. Huang, and S. Zhang, "Analysis and utilization of hidden information in model inversion attacks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4449–4462, 2023.

[27] S. Chen, M. Kahla, R. Jia, and G.-J. Qi, "Knowledge-enriched distributional model inversion attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16158–16167.

[28] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. S. Zemel, and A. Makhzani, "Variational model inversion attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., Dec. 2022, pp. 9706–9719.

[29] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[30] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[31] R. Ning, J. Li, C. Xin, H. Wu, and C. Wang, "Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 9, pp. 10309–10318.

[32] D. Zhou et al., "Improving adversarial robustness via mutual information estimation," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 27338–27352.

[33] T. Wang, Y. Zhang, and R. Jia, "Improving robustness to model inversion attacks via mutual information regularization," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11666–11673.
[34] N. J. Beaudry and R. Renner, "An intuitive proof of the data processing inequality," 2011, *arXiv:1107.0740*.
[35] M. Atsague, O. Fakorede, and J. Tian, "A mutual information regularization for adversarial training," in *Proc. Asian Conf. Mach. Learn.*, 2021, pp. 188–203.
[36] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4431–4440.
[37] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *Proc. 1st Workshop Fine-Grained Vis. Categorization, IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, p. 2.
[38] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8107–8116.
[39] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8185–8194.
[40] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 343–347.
[41] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3730–3738.
[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.
[43] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Jan. 2020, pp. 1–11.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[45] H. Zhang et al., "ResNest: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jul. 2022, pp. 2735–2745.
[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[48] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 852–863.
[49] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.
[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 6626–6637.
[51] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 8024–8025.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[53] A. Shafahi et al., "Adversarial training for free," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Vancouver, BC, Canada, Jan. 2019, pp. 3353–3364.
[54] Z. Yang, B. Shao, B. Xuan, E.-C. Chang, and F. Zhang, "Defending model inversion and membership inference attacks via prediction purification," 2020, *arXiv:2005.03915*.
[55] X. Gong, Z. Wang, S. Li, Y. Chen, and Q. Wang, "A GAN-based defense framework against model inversion attacks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4475–4487, 2023.