A Survey of Model Inversion Attacks on Image Domain

Changjin Kim, Chanwoo Hwang, Sunpill Kim, and Jae Hong Seo*
Department of Mathematics & Research Institute for Natural Sciences, Hanyang University, Seoul, Republic of Korea
{changjinkim, aa5568, ksp0352, jaehongseo}@hanyang.ac.kr

***Abstract*—The widespread deployment of deep learning models in sensitive applications has raised increasing concerns about potential privacy risks. Among them, the Model Inversion Attack (MIA) stands out as a notable threat, aiming to reconstruct data samples representative of the model's private training data. These risks are particularly concerning in the image domain, where successful attacks may lead to the recovery of recognizable faces or sensitive medical images. In response to the growing research interest in this area, this paper presents a comprehensive survey of MIAs applied to image data. We propose a systematic taxonomy that categorizes attacks by threat models, technical attributes, and core methodologies. Additionally, we discuss widely used evaluation metrics and outline the landscape of existing defenses against MIAs.**

***Index Terms*—Model inversion attack, Privacy risk**

I. Introduction

The rapid improvement of deep learning models has unlocked innovation across various fields. However, this progress has also introduced a new landscape of privacy threats. Among them, a Model Inversion Attack (MIA) is gaining attention as a serious privacy breach technique that reversely traces the output of a pretrained model to reconstruct sensitive information from its training data. This attack aims to restore original or similar samples of the training data by leveraging publicly available information, such as the model's prediction results, posing a significant threat, especially to models trained on private data. Since its initial conception by Fredrikson et al. [1, 2], an MIA has steadily evolved and has now reached a level applicable to a wide range of model architectures and domains.

Given that MIAs aim to generate data samples that plausibly belong to the distribution of the train dataset, one might see a parallel with generative models. However, a crucial distinction lies in their core objectives. Generative models aim to *create* new, realistic data by learning the overall distribution of the training data. In contrast, the purpose of a MIA is to exploit this generative capability to *restore* or *infer* specific data that exists within the training dataset. In other words, while generative models are oriented towards producing generalized outputs, MIAs are clearly distinguished by their goal of reproducing specific information that the model has memorized.

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024(RS-2024-00332210).

*Corresponding Author

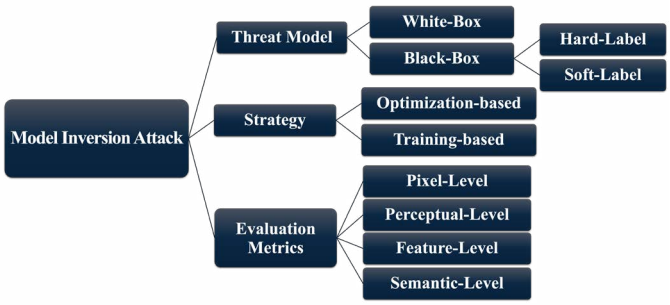

Fig. 1. A taxonomy of the membership inference attacks reviewed in this survey. The classification is based on key criteria such as the threat models, attack strategies, and evaluation metrics

The characteristics of MIAs become clearer when compared with other privacy attack tasks. For instance, a Membership Inference Attack only determines whether a specific data point was included in the training set or not, and a Model Extraction Attack aims to steal the intellectual property of a model itself. On the other hand, MIAs focus on compromising the privacy of the training data by reconstructing the actual content through the model, which poses a more direct threat. Furthermore, the threat of MIAs is becoming a reality in various practical applications that handle sensitive information, such as healthcare, finance, and face recognition. The risk is particularly pronounced in the image domain, which deals with visual data like medical image analysis or face recognition systems. This paper aims to provide a foundation for future research by presenting a survey of MIAs in the image domain.

II. Background

In this section, we provide the formulation of the attacker's objective, and a formal definition of the threat model based on access levels and information availability.

A. Formulating the Objective of Model Inversion Attack

Reconstructing input data solely from a model's output is an inherently ill-posed problem, meaning that multiple different inputs can produce the similar output. This makes accurately recovering the original input quite challenging. Therefore, utilizing prior knowledge is essential to successfully performing the inversion process.

In the following, the model inversion attack is defined under the setting where the attacker has access to the classifier, regardless of whether the attacker operates in a white-box

or black-box setting. The target classifier M_T is defined as follows:

$$M_T(\mathbf{x}) : \mathcal{X} \longrightarrow \Delta^{C-1} \tag{1}$$

where $\mathcal{X}$ denotes the data domain (e.g., images or faces), and Δ^{C-1} represents the $(C-1)$-dimensional probability simplex. We assume the target classifier M_T is trained on private dataset $\mathcal{X}_{\mathsf{train}}$. The goal of the attacker is to reconstruct inputs that resemble the training data, using access to the target model M_T and some prior knowledge $\mathcal{K}$. Depending on the approach, this can be formulated either as learning an approximate inverse function (e.g., training-based methods), or as directly estimating reconstructed samples through optimization (e.g., optimization-based methods). Formally, we express the reconstruction objective as:

$$\mathcal{X}_{\mathsf{recon}} = M_T^{-1}(M_T,\ \mathcal{K}) \approx \mathcal{X}_{\mathsf{train}}, \tag{2}$$

where $\mathcal{X}_{\mathsf{recon}}$ is the reconstructed dataset recovered from M_T and prior knowledge $\mathcal{K}$—which may include statistical properties of the data distribution, auxiliary datasets, or partial access to the training data. If the attacker has access to the parameters and architecture of M_T, it is referred to as a white-box setting; otherwise, it is referred to as a black-box setting. In Section III, we explore various approaches to model inversion, depending on the attacker's level of access to the target model.

B. Threat Model

In real-world applications, accurately assessing potential threats requires a well-defined threat model. Since the goal of an MIA is to infer training data from the pre-trained model's output, we assume the attacker does not have access to the training dataset itself. Depending on the attacker's access to the model, we categorize MIAs as follows:

a) White-box vs. Black-box Settings: In the *white-box setting*, the attacker has full access to the model's architecture and parameters except training data. This enables the attacker to utilize internal representations such as gradients and intermediate features for reconstructing input training data. For instance, [2, 3] apply gradient-based optimization to reconstruct images, starting from random noise. This technique is conceptually similar to conventional adversarial example generation algorithms [4, 5], but differs in its objective and initial condition. Meanwhile, [6] trains a generative model using intermediate feature outputs to reconstruct high-fidelity inputs. In contrast, the *black-box setting* assumes that the attacker only has query access to the target model's input-output pair. This includes typical scenarios like public API-based attacks, where attackers iteratively query the model with crafted inputs and observe outputs to infer information about the training data. Black-box MIAs are significantly constrained and typically require thousands to tens of thousands of queries to succeed [7].

b) Soft-labels vs. Hard-labels: Another important aspect of the threat model is how much output information the attacker can obtain. The attack strategy of an attacker can vary significantly depending on the level of access to the target model[1]. In the *soft-label* setting, the attacker has access to confidence scores or full probability distributions over the classes. In contrast, the *hard-label* setting provides only discrete class predictions, such as the top-1 or top-k labels. In such cases, the attacker may directly use a one-hot vector for a known target class, or select a target label using prior knowledge about the model's domain (e.g., assuming the model classifies faces, the attacker may choose a likely identity to target).

III. Model Inversion Attack Types

MIAs have been widely explored in various data domains, most notably in image [2, 8, 9] and text [10, 11]. In this paper, we focus on MIAs in the image data domain. As previously described, MIAs attempt to find the inverse mapping that recovers input data from the model's output. In general, existing approaches to finding this inverse can be categorized into two main types: optimization-based and training-based methods.

A. Optimization-based Inversion

In optimization-based approaches, the attacker attempt a point estimation by applying optimization methods in the input space $\mathcal{X}$ (e.g., gradient descent [2, 6, 8], genetic algorithm [8], zero-order optimization algorithm [12] etc.). That is, for some distance function , the attacker's goal is to find $\mathbf{x}^*$ by minimizing the following objective function:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \quad (M_T(\mathbf{x}),\ \mathbf{y}_T) + \lambda\mathcal{R}(\mathbf{x}) \tag{3}$$

where $\mathcal{R}$ represents regularization terms such as the p-norm, total variation etc., and $\mathbf{y}_T$ denotes the target label, which can be either a hard-label or soft-label. The regularization term serves to encourage the $\mathbf{x}^*$ to resemble a more plausible or natural image.

When $\mathbf{x}^*$ is high-dimensional data, the search space becomes huge, causing the optimization to fail. To address this issue, leveraging Generative Adversarial Networks (GANs) is a promising approach. GAN can produce high-fidelity images from a relatively low-dimensional latent space compared to the input space dimension [6]. Given the GAN's generator $G : \mathcal{Z} \longrightarrow \mathcal{X}$ and discriminator $D : \mathcal{X} \longrightarrow \{0, 1\}$, the objective function corresponding to (3) is formulated as follows:

$$\mathbf{z}^* = \underset{\mathbf{z}\in\mathcal{Z}}{\operatorname{argmin}} \quad (M_T(G(\mathbf{z})), \mathbf{y}_T) + \lambda\mathcal{R}(G(\mathbf{z})) \tag{4}$$

where typically $\mathcal{R}(\mathbf{z}) = -D(G(\mathbf{z}))$, which serves as a penalty for unrealistic images. In general, GAN model is trained on a public training dataset $\mathcal{X}_{\mathsf{public}}$, after which a latent vector $\mathbf{z}^*$

[1]In this paper, we assume that the target model is a *classification model (i.e., classifier)* by default, and we use the term "model" to refer to such a classifier throughout the paper.

is obtained through an optimization process. The final output can then be generated as $\mathbf{x}^* = G(\mathbf{z}^*)$.

In GAN-based model inversion, the generator G aims to produce images that the discriminator D classifies as real, which often results in limited diversity in the reconstructed samples. Recently, diffusion models have gained attention for their ability to generate high-quality images while mitigating such diversity issues. Leveraging these benefits, Li et al. [13] propose an MIA framework based on Conditional Diffusion Models (CDMs), which address these limitations of GAN-based approaches. Their method also follows a two-step process similar to GAN-based methods for point estimation. Precisely, these methods leverage a CDM trained on the (image, label) pairs of the public training dataset $\mathcal{X}_{\mathsf{public}}$, where the label used for conditioning is the pseudo label– the predicted class label assigned by the target model M_T to the corresponding public image. This conditioning allows the CDM to synthesize class-consistent samples for inversion. Afterwards, to obtain a target-specific CDM for the target label $\mathbf{y}_T$ of target class, fine-tuning is performed by minimizing the following loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_T \sim \mathcal{N}(0,I)} \left[\mathcal{L}_{\mathsf{cls}}(M_T(S(\mathbf{x}_T, \mathbf{y}_T)), \mathbf{y}_T) \right]. \tag{5}$$

where $\mathcal{L}_{\mathsf{cls}}$ denotes the classification loss, and S represents the sampling process of the CDM (denoising process). That is, fine-tuning is performed on a pre-trained CDM to adapt it more closely to the distribution of the specified target class. For the next step, the image reconstruction is carried out initial image $\mathbf{x}_0 \in \mathcal{N}(0, I)$. Then, given a fine-tuned conditional denoiser ϵ_θ and a target model M_T, we optimize the reconstructed image $\mathbf{x}_0$ with respect to the target label $\mathbf{y}_T$ by minimizing the following diffusion prior loss:

$$\mathcal{L}_{\mathsf{prior}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I), t_i} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_{t_i}, \mathbf{y}_T, t_i)\|_2^2 \right] \tag{6}$$

where $\mathbf{x}_{t_i} = \sqrt{\alpha_{t_i}}\mathbf{x}_0 + \sqrt{1-\alpha_{t_i}}\epsilon$, α_{t_i} follows the noise schedule from the original formulation [14]. The denoising steps $\{t_i\}_{i=1}^N$ are annealed from high to low values over N iterations. In the process of minimizing the $\mathcal{L}_{\mathsf{prior}}$, $\mathbf{x}_i$ is iteratively updated and the final $\mathbf{x}_N$ is the reconstructed image.

Optimization-based methods typically search in the image domain or latent space, gradually updating the input in a step-by-step manner to find an image that yields a high confidence score from the target classifier. In addition to such methods, some other works take a different approach. For example, a generator takes the target label as input generates multiple candidate images, and the the final reconstruction image is selected as the one that yields the highest confidence score from the target classifier [15]. These examples demonstrate that model inversion attacks are not always required to optimization-based strategies.

The next subsection briefly summarizes training-based inversion methods, which represents a more general form of inversion beyond point estimation approaches.

B. Training-based Inversion

In the case of a classifier M_T trained on CIFAR-10, the input dimension is $3 \times 32 \times 32 = 3,072$ and the output dimension is 10 classes. For ImageNet, the input dimension is usually $3 \times 224 \times 224 = 150,528$ and the output dimension is 1000 classes. This means that the input is compressed into a low-dimensional representation as it passes through M_T. Moreover, the architecture of M_T generally includes non-invertible layers such as ReLU, Pooling, and Dropout making the inverse mapping difficult in practice. Thus, instead of finding an exact inverse, one can approximate the inverse mapping of M_T using a deep neural network g_θ. From this perspective, attempting MIA by training an inverse mapping g_θ is referred to as a training-based MIA. In addition, the information required to train g_θ consists of the (input-output) pairs of the target classifier M_T, e.g., (image-label), which can be obtained solely through inference queries to the target classifier. Therefore, the threat model for training-based MIAs is commonly assumed to be in the black-box setting [9, 16, 17].

Formally, the deep neural network g_θ is trained on public dataset $\mathcal{X}_{\mathsf{public}}$, which has no intersection of the target model's private training dataset $\mathcal{X}_{\mathsf{train}}$ (assume that $\mathcal{X}_{\mathsf{public}}$ have a distributionally similar to $\mathcal{X}_{\mathsf{train}}$).

For all pairs $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$ where $\mathbf{y}_i = M_T(\mathbf{x}_i) \in [0,\ 1]^N$ denotes the model's output as either hard-label or soft-label encoding over N classes. The attacker's goal is to find the optimal parameters θ^* by minimizing the following objective:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{\mathbf{x} \in \mathcal{X}_{\mathsf{public}}} \frac{1}{N} \cdot \ (g_\theta(\mathbf{y}),\, \mathbf{x}). \tag{7}$$

This objective is interpreted as the reconstruction loss, and additional regularization terms are often used to improve the quality, naturalness, and fidelity of the reconstructed images. A classification loss, such as minimizing $(M_T\left(g_\theta(\mathbf{y})\right), \mathbf{y})$ can be used instead of the reconstruction loss.

If the architecture of g_θ consist only of layers that simply match input and output dimensions, it cannot effectively learn the inverse mapping. Nevertheless, Yang et al. [18] designed g_θ using transposed convolution layers for upsampling purposes and attempted inversion on face images. He et al. [16] constructed their model architecture with convolutional layers and ReLU activations to perform inversion on relatively simple datasets such as MNIST and CIFAR-10. However, these approaches exhibited limitations in terms of image quality, and inversion remains a challenging problem for more complex datasets like ImageNet.

IV. Evaluation of MIA

In this section, we define the metrics to evaluate the performance of various MIAs. Based on these metrics, we then analyze and summarize the performance of previous works.

A. Evaluation Metrics

Now, we describe commonly used evaluation metrics to assess how well the reconstructed input $\mathcal{X}_{\mathsf{recon}}$ aligns with the original training dataset $\mathcal{X}_{\mathsf{train}}$ in terms of data distribution. In

TABLE I
COMPARISON OF MODEL INVERSION ATTACK PERFORMANCE ON FACE RECOGNITION CLASSIFIER

Method	Threat Model	M_T	Strategy	D_{priv}	D_{pub}	Acc1	Acc5	FID	KNN.	Feat.	δ_{Face}	δ_{Eval}
PPA[19]	White-Box	RN-101	Selection / StyleGAN2[20]	CelebA	FFHQ	82.96	95.44	44.04	-	-	0.7506	299.73
		RN-101		FaceScrub	FFHQ	93.95	99.21	46.3	-	-	0.7199	119.9
		RN-152		CelebA	FFHQ	80.61	94.58	40.43	-	-	0.7362	312.58
		RN-152		FaceScrub	FFHQ	92.73	98.91	46.69	-	-	0.7163	123.25
PL-GMI[21]		evoLVe	Optim. / GAN	CelebA	FaceScrub	55.2	77.12	27.99	1474.22	-	-	-
		evoLVe		CelebA	FFHQ	95.04	99.01	25.57	1241.41	-	-	-
IF-GMI[22]		RN-18	Optim. / GAN	CelebA	FaceScrub	97.9	99.6	40.581	-	-	0.667	112.915
		RN-152		CelebA	FFHQ	94.7	99.3	37.461	-	-	0.677	315.032
Diff-MI[13]		evoLVe	Optim. / Diffusion	CelebA	FFHQ	92.6	98.6	37.73	1204.6	-	-	-
		RN-152		CelebA	FFHQ	94.73	99.67	37.82	1140.09	-	-	-
RLB-MI[23]	Black-Box	evoLVe	Optim. / GAN	FaceScrub	FFHQ	38.5	-	-	2204.1	2278.5	-	-
		evoLVe		CelebA	FFHQ	43.3	-	-	1481.9	1361.8	-	-
FHtoT[7]		evoLVe	Optim.&Surr. / GAN	VGGFace2	CelebA	44.22	61.9	-	339.16	306.98	-	-
		evoLVe		VGGFace2	FFHQ	68.71	93.2	-	262.56	237.75	-	-
SDM[24]		evoLVe	Selection / Diffusion	CelebA	FFHQ	71.23	90.17	35.04	1368.14	-	-	-

addition to these distribution-based metrics, we also include *attack accuracy* as a semantic-level evaluation, which reflects how likely the reconstructed inputs are to be recognized as the intended target class by the target model. The evaluation metrics can be categorized as follows:

- **Pixel-Level**: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR). These metrics measure low-level, pixel-wise similarity between the original and reconstructed images. Lower MSE and higher PSNR values indicate greater similarity.
- **Perceptual-Level**: Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS). These metrics assess perceptual similarity based on human visual perception. Higher SSIM and lower LPIPS indicate stronger perceptual similarity.
- **Feature-Level**: Fréchet Inception Distance (FID), Feature Distance, K-Nearest Neighbor Distance (KNN .). These metrics evaluate the similarity between reconstructed and original images in the feature space of a pretrained model. FID compares the overall feature distributions; Feature Distance (Feat.) measures the Euclidean distance (L_2 distance) between the reconstructed feature and the class centroid; and KNN Distance quantifies the proximity of reconstructed features to training data embeddings. Note that these metrics also capture semantic information through high-level features but primarily focus on distributional similarity and smaller values suggest that the reconstructed inputs lie close to real training examples, which may indicate potential privacy leakage.
- **Semantic-Level**: Attack Accuracy. This metric measures whether the reconstructed input is classified as the target class by the model. Evaluation is typically based on Top-1 or Top-5 accuracy, where higher values indicate better semantic alignment with the target class and a stronger model inversion attack. (Note that the specific evaluation networks may differ across methods.)

B. Attack Performance Summary of MIA

The values in TABLE I are aggregated from the result reported in each individual paper. Due to differences in experimental setups and the evaluation networks used to measure accuracy, direct comparisons may not entirely reasonable. Nevertheless, we present results from recent works with comparable settings for easier comparison. Older works with significantly lower performance were excluded for a fairer comparison. For training the M_T, the train dataset D_{priv} as follows:

- CelebA [25] : 202,599 Face Images of 10,177 People
- FaceScrub [26] : 100,000 Face Images of 530 People
- VGGFace2 [27] : 3.31M Face Images of 9,131 People
- FFHQ [28] : 70,000 Face Images

δ_{Face} and δ_{Eval} are face-specific evaluation metrics. δ_{Face} measures feature distance using pre-trained FaceNet [29] and δ_{Eval} measures the average of the shortest L_2 feature distances from each generated image to any training sample in the target class.

For each threat model, the methods are listed in chronological order, with more recent works appearing further down the list. The target model M_T is selected from widely used architectures, including RN-101 (ResNet-101), RN-152 (ResNet-152), and evoLVe (Face.evoLVe). In the distributional shift setting, we adapt the standard configuration where the private dataset D_{priv} is CelebA and the public dataset D_{pub} is FFHQ. When a directly comparable setting is not available, we report configurations where one of the distributions overlaps.'Optim.' refers to an optimization-based, while 'Selection' denotes the generator outputs multiple candidate images and selects the one with the highest confidence score. 'Surr.' refers to utilize a surrogate model to reduce the number of queries required. Although it is difficult to directly compare the performance of each model, it can be observed that recent attacks in the white-box setting achieve considerably high success rates. In contrast, due to the limited amount of information available in the black-box setting, attack success rates tend to be significantly lower compared to white-box setting.

V. DISCUSSION

A. Robustness and Defense Mechanisms

An adversary may attempt a Model Inversion Attack (MIA) by accessing information obtained from feeding data into the model, such as intermediate features, gradients, or confidence scores. If such attacks are unlikely to succeed, the model

is robust. Differential Privacy (DP) is a representative defense mechanism that mitigates information leakage by adding noise to the values accessible to the adversary. In addition, certain obfuscation [30, 31] approaches intentionally impair prior knowledge, leading to perceptual degradation of the reconstructed images, and these methods are typically heuristic and lack formal theoretical guarantees. Such strategies, while enhancing robustness against inversion attacks, naturally give rise to an inherent trade-off between robustness and overall model performance. These observations highlight that defending against model inversion attacks inherently requires balancing privacy and task performance. Model designers of sensitive applications, such as medical imaging or face recognition, must carefully manage this trade-off, and the limitations of current approaches indicate the need for theory-driven, scalable defenses that preserve both privacy and task performance.

B. Inverse model on Face Recognition

Similar to MIA, several studies [32, 33, 34] have investigated model inversion that reconstructs images from face recognition model outputs, using face features instead of confidence score or class labels. These studies indicate that face features contain a substantial amount of information about the original face image. From a privacy leakage perspective, reconstructing the original face image from model outputs shares the same fundamental concerns as MIAs. Moreover, the development of inversion techniques from using deconvolution methods to GANs and more recently to diffusion models closely parallels the improvement process observed in MIA research.

C. Other Approaches with Privacy Concerns

While this paper focuses on MIAs, other approaches such as membership inference attacks and model extraction attacks also pose significant privacy threats. These attacks differ in their objectives and methods. Membership inference attacks, introduced by Shokri et al. [35], aim to determine if a specific data sample was in the training set, thus threatening data privacy. For a comprehensive comparison of MIA techniques, we refer readers to the MIBench framework [36].

Besides, model extraction attacks target the model's intellectual property. The adversary's goal is to reconstruct a functionally equivalent copy of a proprietary model by querying it multiple times. While MIAs focus on training data leakage, model extraction compromises the model itself and may also enable further downstream privacy attacks. Recent work has demonstrated the feasibility of model extraction even for image classification models. For instance, Jagielski et al. [37] show that high-fidelity copies of commercial image classification APIs can be extracted through adaptive querying, threatening both intellectual property and privacy.

VI. Conclusion

Deep learning is used in many different areas, but because it strongly depends on data. This has led to frequent application in fields that handle sensitive information. As a result, privacy issues have become critical, and attacks such as MIA pose serious threats. To better understand and address these risks, this survey provides a systematic overview of existing approaches. We hope it serves as a clear and accessible foundation for new researchers interested in MIA.

References

[1] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing," in *23rd USENIX security symposium (USENIX Security 14)*, pp. 17–32, 2014.

[2] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

[3] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "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.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, Ieee, 2017.

[6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

[7] Z. Li, H. Zhang, J. Wang, M. Chen, H. Hu, W. Yi, X. Xu, M. Yang, and C. Ma, "From head to tail: Efficient black-box model inversion attack via long-tailed learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29288–29298, 2025.

[8] S. An, G. Tao, Q. Xu, Y. Liu, G. Shen, Y. Yao, J. Xu, and X. Zhang, "Mirror: Model inversion for deep learning network with high fidelity," in *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022.

[9] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 225–240, 2019.

[10] J. X. Morris, W. Zhao, J. T. Chiu, V. Shmatikov, and A. M. Rush, "Language model inversion," *arXiv preprint arXiv:2311.13647*, 2023.

[11] J. X. Morris, V. Kuleshov, V. Shmatikov, and A. M. Rush, "Text embeddings reveal (almost) as much as text," *arXiv preprint arXiv:2310.06816*, 2023.

[12] M. Kahla, S. Chen, H. A. Just, and R. Jia, "Label-only model inversion attacks via boundary repulsion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15045–15053, 2022.

[13] O. Li, Y. Hao, Z. Wang, B. Zhu, S. Wang, Z. Zhang, and F. Feng, "Model inversion attacks through target-specific conditional diffusion models," *arXiv preprint arXiv:2407.11424*, 2024.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] R. Liu, D. Wang, Y. Ren, Z. Wang, K. Guo, Q. Qin, and X. Liu, "Unstoppable attack: Label-only model inversion via conditional diffusion model," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3958–3973, 2024.

[16] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th annual computer security applications conference*, pp. 148–162, 2019.

[17] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 682–692, 2021.

[18] Z. Yang, E.-C. Chang, and Z. Liang, "Adversarial neural network inversion via auxiliary knowledge alignment," *arXiv preprint arXiv:1902.08552*, 2019.

[19] L. Struppek, D. Hintersdorf, A. D. A. Correia, A. Adler, and K. Kersting, "Plug & play attacks: Towards robust and flexible model inversion attacks," *arXiv preprint arXiv:2201.12179*, 2022.

[20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "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.

[21] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang, "Pseudo label-guided model inversion attack via conditional generative adversarial network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 3349–3357, 2023.

[22] Y. Qiu, H. Fang, H. Yu, B. Chen, M. Qiu, and S.-T. Xia, "A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks," in *European Conference on Computer Vision*, pp. 109–126, Springer, 2024.

[23] G. Han, J. Choi, H. Lee, and J. Kim, "Reinforcement learning-based black-box model inversion attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20504–20513, 2023.

[24] X. Peng, B. Han, F. Yu, F. Liu, T. Liu, and M. Zhou, "Single-step diffusion model-based generative model inversion attacks,"

[25] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[26] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE international conference on image processing (ICIP)*, pp. 343–347, IEEE, 2014.

[27] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[28] T. Karras, S. Laine, and T. Aila, "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.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "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.

[30] S. V. Dibbo, A. Breuer, J. Moore, and M. Teti, "Improving robustness to model inversion attacks via sparse coding architectures," in *European Conference on Computer Vision*, pp. 117–136, Springer, 2024.

[31] S. Jin, H. Wang, Z. Wang, F. Xiao, J. Hu, Y. He, W. Zhang, Z. Ba, W. Fang, S. Yuan, *et al.*, "{FaceObfuscator}: Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition," in *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 6849–6866, 2024.

[32] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 5, pp. 1188–1202, 2018.

[33] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy, "Vec2face: Unveil human faces from their blackbox features in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6132–6141, 2020.

[34] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2face: A foundation model for id-consistent human faces," in *European Conference on Computer Vision*, pp. 241–261, Springer, 2024.

[35] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18, IEEE, 2017.

[36] J. Niu, X. Zhu, M. Zeng, G. Zhang, Q. Zhao, C. Huang, Y. Zhang, S. An, Y. Wang, X. Yue, *et al.*, "Comparing different membership inference attacks with a comprehensive benchmark," *IEEE Transactions on Information Forensics and Security*, 2025.

[37] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in *29th USENIX security symposium (USENIX Security 20)*, pp. 1345–1362, 2020.