Black-Box Based Face Model Inversion Attack

Abstract

As artificial intelligence rapidly advances and the increase in private data required for training deep learning models, concerns about private information leakage have arisen. Privacy attacks, including white-box attacks, black-box attacks, and label-only attacks, have become prominent.In previous research, the methods of white-box attacks are inapplicable to machine learning services where model parameters are protected. Recently proposed black-box algorithms search the latent space based on confidence obtained from the target model but exhibit relatively poor performance. Additionally, label-only attacks have achieved success in label-only settings, but they cannot guarantee how many query accesses are needed before finding the first latent vector.To address these issues, we propose a novel approach, namely Reinforcement Learning-based Black-box Model Inversion (RLB-MI). This method integrates reinforcement learning, treating the exploration of latent space in GANs as a Markov Decision Process (MDP). By providing rewards based on confidence scores of generated images, the agent approximates the environment of latent space exploration using replay memory updates, leading to more effective navigation of latent vectors. Ultimately, private data is successfully reconstructed from latent vectors using GAN.We conducted attack experiments on multiple datasets and models, and the results demonstrate that our method successfully recovers meaningful information about private data, outperforming other attack methods in various aspects.

1 Introduction

Deep Neural Networks (DNNs) have been widely applied in various practical scenarios in the field of AI. However, due to the fact that many DNNs utilize private data for training, the security of the data is facing challenges. With the development of research in the privacy attack domain, Model Inversion (MI) attacks have drawn increasing attention concerning privacy. This type of attack can reconstruct training data from public models, and it is the focus of this article. In fact, previous research on privacy attacks has demonstrated the potential for unauthorized information exposure through accessing models [1–4].

The application of Facial Recognition (FR) systems is widespread, encompassing not only classical use cases but also emerging scenarios such as identity-labeled image generation [5–7]. The objective of the FR method, denoted as f, is to obtain an embedding y for a facial image x, such that the embedding of images of the same person is closer than those of different individuals. We refer to this embedding y as the identity or ID vector. In this paper, we propose a technique for exploring and sampling the latent space from $p(x \mid y)$, i.e., generating real facial images from ID vectors. According to the design, the FR method's many-to-one mapping assigns multiple images of the same identity to a single ID vector. The inverse one-to-many problem, generating high-dimensional images from low-dimensional ID vectors, is highly challenging.

Based on past experiences with model inversion attacks, Model Inversion Attacks (MIA) are often modeled as an optimization problem. The goal is to search for a target latent vector in the latent space of a Generative Adversarial Network (GAN) to maximize the likelihood of the target model's output belonging to a specific target category, thereby achieving inversion attack. The training of a generative model for inversion attacks can be divided into the following two key stages:

- Stage One: Training of the Generative Adversarial Network Parameters. The attacker trains a generative model on a public dataset with a structure similar to the private dataset, whether it is of the same or a different distribution. This involves searching for the optimal parameters for the generator.
- Stage Two: Search in the Latent Space of the Generative Adversarial Network. The attacker continually explores the latent space of the pre-trained generative adversarial network until the generated images closely resemble those from the private dataset.

For the white-box attack scenario, where attackers can obtain the parameters of the target model, they can directly use gradient optimization methods to search the latent space of the generative adversarial network and find specific target latent vectors [7–9]. Inspired by semi-supervised GANs [10], KED-MI [11] utilizes a classifier as the discriminator in the GAN. During the training process of the first stage, it leverages the target model to provide soft labels for public data, successfully recovering high-quality private data, including personal information. This achieves state-of-the-art model inversion attack performance.

However, in practical scenarios, information about the target model's architecture and the gradients of its weights is often unavailable, especially in the case of many non-open source model API applications. Therefore, in this paper, we focus on a more universally applicable black-box setting. Unlike white-box attacks, in the inversion process, only the resulting ID vectors are available. Methods are needed to explore the latent space of GANs to leverage them because gradient-based optimization is not feasible.A recently proposed black-box model inversion attack method, Model Inversion for Deep Learning Network (MIRROR) [12], utilizes a genetic algorithm to search the latent space obtained from a black-box target model with confidence score information. In MIRROR attacks, despite the use of confidence scores, its performance is worse than the attack method BREP-MI under only label scenarios [13]. Moreover, current black-box model inversion attacks using GANs lack guarantees of completing the attack process within a predefined number of query accesses.Addressing the aforementioned issues and proposing several improvements, the contributions of this paper can be summarized as follows:

- In this paper, we model the problem of searching the latent space of a GAN as an MDP (Markov Decision Process) and employ a reinforcement learning agent applicable to unknown environments, continuous, high-dimensional image feature spaces to address this problem.
- During the training process, we design several crucial loss terms for model optimization based on the confidence score output of the target model in the black-box scenario. These terms enable the generated images by the GAN to gradually approach real human faces, progressively converge towards the target category, and increase the dissimilarity from other non-target categories.

- In the experimental phase, we conduct comprehensive comparative experiments, including model inversion attacks on different target model networks under the same and different data distributions, to evaluate the effectiveness of the proposed approach.

2 Related works

2.1 White-box Adversarial Attacks

The objective of model inversion attacks is to reverse-engineer the private training data of the target model to extract privacy features [14]. Fredrikson et al. were the first to employ gradient-based optimization methods for model inversion attacks, focusing initially on linear regression models [15]. However, these methods proved ineffective when dealing with complex structured models like deep neural networks. As deep neural networks gained widespread use in the field of AI, researchers began exploring model inversion attacks on DNNs. Zhang et al. introduced Generative Adversarial Networks (GANs), reducing the search space for generated images[16]. Model inversion attacks based on GANs (GMI) pretrain on a public dataset, leveraging meaningful semantic information to produce more effective results [16].Recent studies have predominantly built upon GMI, attempting to enhance it from various perspectives. Struppek et al. increased the robustness of inversion attacks by introducing random augmentation in intermediate results [17]. Knowledge-Enriched Distribution Model Inversion Attacks (KED-MI) [18]improved upon GMI by incorporating a dedicated inversion GAN, where the discriminator performs multi-class inference.

2.2 Black-box Adversarial Attacks

In a black-box scenario, attackers can only obtain confidence scores or categories of the target model's output for a given input image. Szegedy et al. first demonstrated the transferability of adversarial examples [19], showing that an adversarial example generated by one model is likely to be misclassified by another model. Therefore, in a black-box setting, attackers can train substitute models for the target model. Leveraging the transferability of models, they use these substitute models to generate adversarial examples [20–22] to attack the target model [19].Additionally, there are query-based black-box attack methods [23–26], which use feedback from the target model's output to guide the attack method in generating adversarial examples. Cheng et al. [25] proposed a confidence score-based attack method called Zeroth-Order Optimization (ZOO) using gradient estimation. Brendel et al. [23] introduced a decision-based attack. While these query-based methods also do not require real training data during black-box attacks, they still differ significantly from data-free transfer-based black-box attacks.The major distinction lies in the fact that query-based attack methods generate instance-specific attacks and require multiple accesses to the attacked model with the original data during the evaluation phase to generate each attack. Hence, the query cost of their methods is linearly correlated with the number of generated adversarial examples. In contrast, transfer-based black-box attacks do not require any queries during the evaluation phase but involve queries during the training phase. After obtaining substitute models, these attacks no longer require additional query costs to generate adversarial examples.Knowledge-Enriched Distribution Model Inversion Attacks (LB-MI) [27] use a structure similar to an autoencoder to train an inversion model that reverses the target network. The research on Model Inversion for Deep Learning Net-

work (MIRROR) indicates that GANs can be employed for genetic algorithm-based black-box model inversion attacks.

2.3 SAC Method

Currently, Reinforcement Learning (RL) has found applications in various domains, ranging from games to robot control [28–30]. However, model-free deep RL methods face challenges in sample complexity and sensitivity to hyperparameters, limiting their applicability in real-world tasks. Among the reasons for the low sample efficiency in deep RL methods is policy learning, as some commonly used algorithms require collecting new samples at each gradient step [30–32]. Researchers have been working on designing efficient and stable model-free deep RL algorithms for continuous state and action spaces.By introducing an entropy maximization term into the standard maximum reward reinforcement learning objective, researchers have improved algorithm performance in exploration and robustness [33, 34]. One such algorithm is the Soft Actor-Critic (SAC) algorithm, a type of off-policy maximum entropy method that combines sample-efficient learning and stability [35]. SAC can easily scale to complex high-dimensional tasks, such as the Humanoid benchmark with 21 action dimensions [36], where non-policy methods like DDPG often struggle to achieve good results [37]. SAC also avoids the complexity and potential instability associated with approximate inference in prior off-policy maximum entropy algorithms based on soft Q-learning [38].

3 Proposed Model

The overall framework of the model proposed in this article is shown in Figure 1:

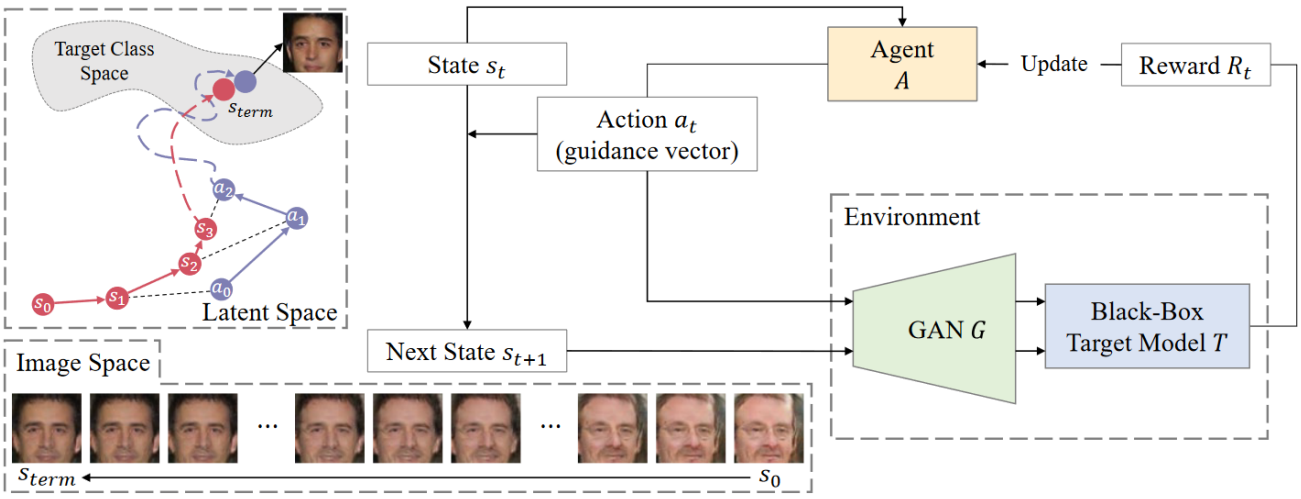

Figure 1. The training process of latent space search and reinforcement learning agents as an MDP problem is outlined. In the latent space, the state s_t moves in the direction of the action a_t according to the distance determined by the diversity factor α at each step. G generates an image from the updated state s_{t+1} and action a_t, and provides rewards to agent A through the target model T. Furthermore, we visualize the state change from the initial state s_0 to the terminal state s_{term} within an episode in image space.

3.1 Black-box Adversarial Attacks Formulation

We consider a target model T and a label y_t corresponding to one of its labels. The purpose of the attack is to characterize the input x_t such that $T(x_t) = y_t$.

GAN consists of two deep neural networks, namely a generator $G : Z \rightarrow X$ and a proxy model $S : X \rightarrow Y$. The generator $G : Z \rightarrow X$ maps noise $z \in N(0, 1)$ to input and a proxy model $S : X \rightarrow Y$ outputs an estimate $\hat{y}$ of the target model output. GAMIN allows to simultaneously train a surrogate model S and a generator G while performing a model inversion attack on S. Therefore, the purpose of the generator is to learn the distribution y_t of inputs x_t with respect to labels. For each step of GAN training, samples are taken from random noise $N(0, 1)$. The latter is then used by the generator to produce X_G. Then, the target model T is queried using the output X_G of the generator and samples sampled X_S from random noise $N(0, 1)$. Subsequently, the confidence score output by T is used to calculate the proxy loss L_S, which is then used to guide training.

The main task of our method is to search for latent quantities Z_G that generate images with high class confidence from the latent space of G. We use Markov decision process (MDP) to model the search problem of GAN latent space and apply the solution of reinforcement learning. Therefore, this reinforcement learning process has three characteristics of unknown environment, continuous space and high-dimensional space. More specifically, we define the state space of MDP as the latent space of G, then the state S_t at each step t has the same form as the latent vector. The state S_t is guided by an action called the guidance vector and is updated to the next State S_{t+1}, as shown in Figure 1. Finally, the reward is formulated based on the confidence score of the generated image by S_{t+1} and a_t.

3.2 Soft Actor-Critic Method

We address policy learning in a sustained action space. We consider an infinite horizon Markov decision process (MDP), defined by a tuple (S, A, p, r), where the state space S and the action space A are continuous, and the unknown state transition probability $p : S \times S \times A \rightarrow [0, \infty)$ represents the probability density of the next state $S_{t+1} \in S$ given the current state $S_t \in S$ and the action $a_t \in A$. The environment emits a bounded reward $r : S \times A ß[r_{min}, r_{max}]$ on each transition. We will use $\rho_\pi(s_t)$ and $\rho_\pi(s_t, a_t)$ to represent the state and state-action margins of the trajectory distribution caused by the policy $\pi(a_t|s_t)$.

SAC is based on the derivation of soft policy iteration, a general algorithm for learning optimal maximum entropy policies under a maximum entropy framework, which alternates between policy evaluation and policy improvement. For large continuous domains, we are required to derive practical approximations for soft policy iteration. So SAC will use function approximators for both the Q function and the policy, and instead of running evaluation and improving convergence, SAC will alternately optimize both networks using stochastic gradient descent. Consider the parameterized state value function $V_\Psi(s_t)$, the soft Q-function $V_\Psi(s_t, a_t)$ and the tractable policy $V_\Psi(\pi|\varphi)$. The parameters of these networks are Ψ, θ and φ. For example, the value function can be modeled as an expressive neural network, and the policy can be modeled as a Gaussian distribution with mean and covariance given by the neural network.

3.3 Model latent space

In this paper, we model the latent space of GAN as an MDP process. It includes the following components: state, action, state transition and reward.

We define the state during the iteration process as a standard k-dimensional random vector $s_t \sim N_k(0, 1)$, $N_k \in R^k$. where k is the dimension of the latent space, and the state s_t is updated by the action a_t. We hope that these actions will lead a random initial latent vector to a high-payoff final latent vector. We consider the action space as the entire potential space. We define potential vector-shaped actions as guidance vectors. As can be seen in Figure 1, actions are, by definition, selected in the same space as states. This enables extensive exploration of the entire latent space, preventing the agent from getting stuck in local minima and guaranteeing convergence of the agent. The action a_t in state s_t at each step t is determined by agent $A : a_t = A(s_t)$, $a_t \in R^k$.

We update the state by using the diversity factor α to move the state towards the action at each step: $s_{t+1} = \alpha \cdot s_t + (1 - \alpha) \cdot \alpha_t$. After updating the state through actions, the agent receives rewards from the environment. G uses the updated latent vector to generate the image, and we can obtain the confidence score of the target category y of the image through inference using the target network T. Therefore, we compose the reward with a state score and an action score, which are calculated as the logarithm of the confidence score of the image created by each vector. Scores are calculated as follows:

$$\begin{aligned} r_1 &= \log\left[T_y(G(s_{t+1}))\right] \\ r_2 &= \log\left[T_y(G(a_t))\right] \end{aligned} \tag{1}$$

In addition, we want the reconstructed image to have the characteristics of the target class, so as to distinguish it from other kinds of images. Therefore, we propose an additional term r_3 to punish the high confidence scores of other types of images:

$$r_3 = \log\left[max\left\{\epsilon, T_y(G(s_{t+1})) - \max_{i \neq y} T_i(G(s_{t+1}))\right\}\right] \tag{2}$$

Therefore, the reward score at each step in the iterative process $R_t = w_1 \cdot r_1 + w_2 \cdot r_2 + w_3 \cdot r_3$.

The MDP process modeled above consists of G and T, so we have the following requirements for the reinforcement learning agent: it needs to be robust in complex environments; it must be able to handle continuous action spaces; it must be able to handle high-dimensional spaces.Therefore, this paper uses the soft actor evaluation method (SAC) that meets all the above conditions to solve the MDP.

4 Experimental Aanalysis

4.1 Setting

Model. We use several popular network structures as target models to compare inversion attacks. Similar to previous studies, we used two network structures: VGG16 [39] and Face.evoLVe [40] for experiments.

Datasets. Select representative face data sets CelebFaces Attributes Dataset (CelebA) [41], FaceScrub Dataset [42]. We split each dataset into a private dataset for training the target classifier and a public dataset for training the generative model. There is no class intersection between the public and private datasets, so the

generative model cannot learn class-specific information for the target classifier. In addition, we use Flickr-Faces-HQ Dataset (FFHQ) [43] as a public data set to conduct additional inversion attack experiments under different data distributions.

Implementation details. Use SGD to train the target classifier for 50 epochs, with a learning rate of 0.01, a batch size of 64, a momentum of 0.9, and a weight attenuation of 1×10^{-4}. The target model is selected based on accuracy. Use Adam to train GAN for 300 rounds, learning rate 0.004, batch size 64, β_1=0.5, β_2=0.999. The SAC agent is trained using Adam with discount factor γ=0.99, soft update factor τ=0.01, learning rate 5×10^{-4}, replay memory size 1×10^{6}, batch size 256, maximum step size and diversity of 1 per episode Factor α=0 for 40,000 episodes. The reward weights w_1, w_2 and w_3 are set to 2, 2 and 8 respectively. The value of ε in these experiments was 1×10^{-7}.

Evaluation indicators. We choose Attack Accuracy, K-nearest neighbor distance (KNN Dist) as the evaluation indicators for the experiment. Among them, KNN Dist is a measure of the average L_2 distance between the features of the reconstructed image and the features of the image closest to the sample in the target label image.

4.2 Experimental Results

In the experiment, we conducted experiments with the same data distribution and different data distribution.

Experiments under the same data distribution. The VGG16 and Face.evoLVe models were used as target models respectively and trained on the private set of CelebA. The pre-training of GAN is carried out in the CelebA public collection. As can be seen from the results in Table 1, the RLB-MI method we proposed achieved the highest attack accuracy when attacking two different target models. In addition, the image reconstructed by the RLB-MI method has the lowest KNN Dist between it and the real image, indicating that our method better inverts the feature information of the target face image.

Model	Method	Attack Acc	KNN Dist
VGG16	KED-MI	0.689	1321.8
	MIRROR	0.401	1441.1
	BREP-MI	0.577	1335.4
	Ours	0.645	1312.1
Face.evoLVe	KED-MI	0.738	1348.3
	MIRROR	0.531	1383.8
	BREP-MI	0.715	1269.1
	Ours	0.788	1229.5

Table 1. The attack performance of the model inversion attack on target models with different structures trained on CelebA, with the test accuracy in parentheses.

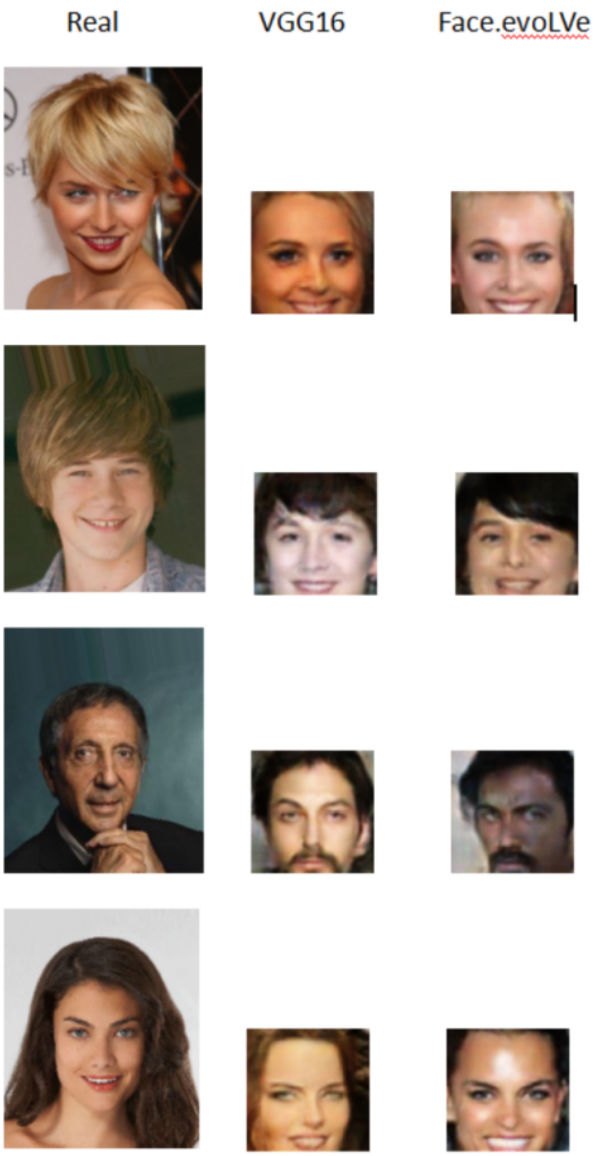

Figure 2. Comparison of the inversion attack results of the RLB-MI method on the same target face in VGG16 and Face.evoLVe.

Experiments under different data distributions. Taking Face.evoLVe as the target model, it is trained on the private sets of CelebA and FaceScrub respectively. The pre-training of GAN is performed in the FFHQ data set. As can be seen from the results in Table 2, when Face.evoLVe was trained on two different data sets, the RLB-MI method we proposed achieved the highest attack accuracy under different data distributions, and the reconstructed image has The lowest KNN Dist with the real image.

Private	Method	Attack Acc	KNN Dist
CelebA	KED-MI	0.416	1523.9
	MIRROR	0.261	1611.3
	Ours	0.429	1488.0
FaceScrub	KED-MI	0.279	2390.1
	MIRROR	0.248	2282.4
	Ours	0.381	2211.5

Table 2. The attack performance of the model inversion attack when the public data distribution is different from the private data distribution.

5 Conclusion

This paper proposes a black-box model inversion attack (RLB-MI) method based on reinforcement learning, which treats the exploration of the latent space as a Markov decision process (MDP). By allowing the agent to obtain rewards from the confidence scores of the generated images, this method achieves effective guidance of latent vectors in unknown environments and high-dimensional image feature spaces, and ultimately successfully reconstructs private data through a generative adversarial network (GAN). Experimental results show that this method successfully recovers meaningful private information on multiple data sets and models, and outperforms other attack methods in all aspects.

References

[1] Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In IEEE Symposium on Security and Srivacy, 318.

[2] Gopinath, D.; Converse, H.; Pasareanu, C.; and Taly, A. 2019. Property inference for deep neural networks. In 34th IEEE/ACM International Conference on Automated Software Engineering, 797–809.

[3] Tram‘ er, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing Machine Learning Models via Prediction APIs. In 25th USENIX Security Symposium, 601618.

[4] Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322–1333.

[5] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smoothswap: a simple enhancement for face-swapping with smoothness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1077910788, 2022. 1

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2003–2011, 2020. 1

[7] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. arXiv preprint arXiv:2005.07728, 2020. 1, 2, 3

[8] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3703–3712, 2017. 1, 2, 3

[9] Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. arXiv preprint arXiv:1606.04189, 2016. 1, 2, 3, 4

[10] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. Advances in Neural Information Processing Systems, 29.

[11] Chen, S.; Kahla, M.; Jia, R.; and Qi, G.-J. 2021. KnowledgeEnriched Distributional Model Inversion Attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 16178–16187.

[12] Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In Proceedings of the 29th Network and Distributed System Security Symposium, 2022. 1, 2, 3, 6

[13] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1504515053, June 2022. 1, 2, 3, 5, 6

[14] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon M.Lin, David Page, and Thomas Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in USENIX Security Symposium. 2014, pp. 17–32, USENIX Association.

[15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in CCS. 2015, pp.1322–1333, ACM.

[16] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in CVPR. 2020, pp. 250–258, Computer Vision Foundation / IEEE.

[17] Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting, "Plug & play attacks: Towards robust and flexible model inversion attacks," in ICML. 2022, vol.162 of Proceedings of Machine Learning Research, pp.20522–20545, PMLR.

[18] Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 16178–16187, October 2021. 1, 2, 3, 5, 6

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. 1, 2, 3

[20] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alch' e-Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 10932–10942, 2019. 1, 2

[21] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. 1, 2

[22] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 6519–6527. Computer Vision Foundation / IEEE, 2019. 1, 2

[23] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. 2

[24] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020, pages 12771294. IEEE, 2020. 2

[25] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based blackbox attacks to deep neural networks without training substitute models. In Bhavani M. Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha, editors, Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017, pages 15–26. ACM, 2017. 2

[26] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. 2

[27] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19, pages 225–240, New York, NY, USA, 2019. ACM. 2, 6

[28] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[29] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game

of go with deep neural networks and tree search. Nature, 529(7587):484–489, Jan 2016. ISSN 0028-0836. Article.

[30] Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. Trust region policy optimization. In International Conference on Machine Learning (ICML), pp. 1889–1897, 2015.

[31] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017b.

[32] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning (ICML), 2016.

[33] Ziebart, B. D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Carnegie Mellon University, 2010.

[34] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In International Conference on Machine Learning (ICML), pp. 1352–1361, 2017.

[35] Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." International conference on machine learning. PMLR, 2018.

[36] Duan, Y., Chen, X. Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In International Conference on Machine Learning (ICML), 2016.

[37] Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., and Levine, S. Q-prop: Sample-efficient policy gradient with an off-policy critic. arXiv preprint arXiv:1611.02247, 2016.

[38] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In International Conference on Machine Learning (ICML), pp. 1352–1361, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In In ternational Conference on Learning Representations (ICLR), 2015. 5

[40] Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Oct 2017. 5

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015. 5

[42] Hongwei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. 2014 IEEE International Conference on Image Processing (ICIP), pages 343–347, 2014. 5

[43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2, 5, 6