Towards Efficient Data Free Black-box Adversarial Attack

Jie Zhang[1*] Bo Li[2*‡] Jianghe Xu[2] Shuang Wu [2]
Shouhong Ding[2] Lei Zhang[1] Chao Wu[1‡]

[1]Zhejiang University [2]Youtu Lab, Tencent
{zj_zhangjie, lei.zhang, chao.wu}@zju.edu.cn, {libraboli, jankosxu, calvinwu, ericshding}@tencent.com

Abstract

Classic black-box adversarial attacks can take advantage of transferable adversarial examples generated by a similar substitute model to successfully fool the target model. However, these substitute models need to be trained by target models' training data, which is hard to acquire due to privacy or transmission reasons. Recognizing the limited availability of real data for adversarial queries, recent works proposed to train substitute models in a data-free black-box scenario. However, their generative adversarial networks (GANs) based framework suffers from the convergence failure and the model collapse, resulting in low efficiency. In this paper, by rethinking the collaborative relationship between the generator and the substitute model, we design a novel black-box attack framework. The proposed method can efficiently imitate the target model through a small number of queries and achieve high attack success rate. The comprehensive experiments over six datasets demonstrate the effectiveness of our method against the state-of-the-art attacks. Especially, we conduct both label-only and probability-only attacks on the Microsoft Azure online model, and achieve a 100% attack success rate with only 0.46% query budget of the SOTA method [49].

1. Introduction

Recently, deep neural networks (DNNs) have been employed as a fundamental technique in the advancement of artificial intelligence in both established and emerging fields [24–28, 31–33, 42, 45, 46, 48]. Despite the success of DNNs, recent studies have identified that DNNs are vulnerable to adversarial examples [3, 6, 13, 16, 30, 41]. A virtually imperceptible perturbation to an image can lead a well trained DNN to misclassify. Consequently, the security concerns about DNNs have attracted many researchers' interest in studying the adversarial vulnerability and robustness of networks [29].

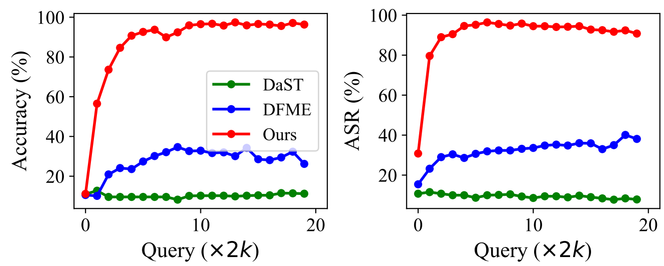

Figure 1. Efficiency comparison with the state-of-the-art methods DaST [49] and DFME [43]. The left subplot shows substitute models accuracy and the right subplot shows untargetd attack successful rate. Attacks are conducted on MNIST in probability-only scenarios with query budget $Q = 40k$ ($1k = 1000$).

Classical works [2, 13, 34] perform attacks in the white-box setting: with full access to the model's parameters and architectures, they can directly use gradient-based optimization to find successful adversarial examples. However, this attack scenario is usually unavailable in real-world deployment due to privacy and security. As a more practical scenario in real-world systems, black-box attacks assume that attackers can only query the target network and obtain its outputs (probability or label) for a given input. By querying the target network with real images, malicious attackers can train substitute models to imitate the target models. Then the substitute models can be used to generate adversarial examples [8, 17, 39] to attack the target model based on the transferability [10, 11, 41] of these adversarial examples. However, substitute models need to be trained by target models' training data, which is hard to acquire due to privacy or transmission reasons.

Recently, some researchers [43, 44, 49] have recognized the limited availability of real data for adversarial queries

[*]Both author contributed equally to this work. Work is completed during Jie Zhang's internship at Tencent Youtu Lab.
[‡]Corresponding author.

and proposed to train substitute models in a data-free black-box scenario. By adopting the principle of generative adversarial networks (GANs), they [43, 49] tried to address this problem with a competition game: A generator is responsible for synthesising some input images, and the substitute model is trained to imitate the target model on these images. In this game, the two adversaries — a substitute model and a generator model, respectively try to minimize and maximize the matching rate between the substitute model and the target model. However, it is very difficult to accurately quantify substitute-target disagreement in a black-box scenario, let alone directly using this object to train the generator. Consequently, this unstable training process makes the models hard to converge. Even after an unlimited number of queries, their approach inevitably leads to model collapse, and can barely reach their ideal Nash equilibrium point in practice (We empirically verify these phenomenons in Section 4). Though the prior art has shed the light on data-free substitute models training, these methods require a large number of queries, which is not practical in real-world settings (*e.g.*, 2M (million) queries to attack the online model on Microsoft Azure [49]). Actually, commercial models are often deployed as pay-per-query prediction APIs for the sake of the protection of data privacy. It remains an open and very challenging problem: *how to effectively learn a substitute model with a limited query budget*?

In this paper, we consider a more stringent yet more practical adversarial scenario, a black-box model with no access to the real data and limited budgets for querying the target model. Rethinking the collaborating relationship between the generator and the substitute model, we design a powerful black-box attack framework. As shown in Figure 1, the proposed method can efficiently imitate the target model through a small number of queries and achieve high attack success rate in both probability and label based black-box settings. Our contributions are as follows:

(1) We revisit the convergence problem of previous data-free attack methods caused by their unstable training process. Instead of training the generator with the inaccurate substitute-target disagreement, we change the game between the generator and the substitute model. The two collaborating players are no longer forced to directly compete in one minimize-maximize game. Instead, we give them different objectives. Especially for the generator, we reset its objective as synthesising surrogate dataset whose distribution is close to the target training data. While, the substitute model aim to efficiently imitate the target model with the generated training examples. In our new game, the generator and the substitute model have relatively independent optimization processes, which allows the substitute model to converge more stably to the target model.

(2) Besides the problem of convergence, the previous methods suffer from the model collapse, resulting in low substitute model accuracy and low attack success rate. We attempt to alleviate the mode collapse problem in data-free substitute model training, through the lens of balancing data distribution and promoting data diversity. On one hand, we maximize the information entropy of the synthetic data in each batch. When it maximizes, the categories are evenly distributed. On the other hand, we randomly smooth the pseudo ground-truth labels and steer the generator to synthesis diverse data in each category.

(3) To further improve the training efficiency of the substitute model, we propose to go deeper into the utilization of synthetic data. To achieve higher attack success rate, the substitute model are encouraged to have decision boundaries that are highly consistent with the target model. Accordingly, we argue that there are two types of data that need to be given extra attention. And we design two losses to boost the training of the substitute model.

(4) Our empirical evaluations on six datasets under both untargeted and targeted attacks show that the proposed method can efficiently imitate a target model using a small number of queries and successfully generate adversarial examples using the substitute model. Specifically, we achieve **98.0% untargeted attack success rate** in the label-only scenario on CIFAR10 with **only 3.75% query budget** of the previous SOTA method DFME [43]. Moreover, we conduct both label-only and probability-only attacks on the Microsoft Azure online model, and achieve a **100% attack success rate** with **only 0.46% query budget** of the previous SOTA method DaST [49].

2. Related Work

Black-box Adversarial Attacks In the black-box setting, attackers can only query the target network and obtain its output (probability or label) for a given input. The transferability of adversarial examples was first verified by Szegedy et al. [41], who found that adversarial examples generated by one model are very likely to be misclassified by another. Consequently, in the black-box setting, malicious attackers can train substitute models to imitate the target models. Then the substitute models can be used to generate the adversarial examples [8, 17, 39] to attack the target model based on the transferability [41]. In this paper, we focus on these transfer-based black-box attacks with a more stringent yet more practical adversarial scenario: a black-box model with no access to the real data and limited budgets for querying the target model.

Note that there is another kind of black-box attack, called query-based attack [1, 4, 5, 7] which utilizes inputs query feedback to guide the attack method to generate adversarial examples. Cheng et al. [5] proposed a score-based attack method zeroth order based attack (ZOO) using gradient estimation. Brendel et al. [1] first proposed a decision-based attack. Although these query-based methods also do not require real training data when performing black-box attacksthere are still some notable differences from the data-free transfer-based black-box attacks. The most significant difference is that query-based attack methods generate attacks based on instances (they need to use one original data to access the attacked model numerous times in the evaluation stage to generate each attack). Therefore, the query cost required by their method is linearly related to the number of generated adversarial samples. While, transfer-based black-box attack does not need any query in the evaluation stage but needs queries in the training stage. Such attacks will no longer require an additional query cost to generate attack samples after a substitute model is obtained.

Data-free Knowledge Distillation Data-free knowledge distillation transfers knowledge of a teacher model to a student model without original dataset [35]. A generative model is trained to synthesize data samples for students to query teacher in data-free manner [9, 12, 35]. The success of data-free knowledge distillation hints at the feasibility of data-free adversarial attacks [44, 49]. However, previous works assume that the teacher model is a white-box model, and directly utilized the gradient or feature map information for distillation [12]. The gradient information of teacher model is required to backpropagate to update student model, which is not available in black-box scenarios. [43] utilizes data-free knowledge distillation to extracts model knowledge, which aims to steal the knowledge of target models. Different from previous methods, it approximates the gradient of the target model which is a further step and inspiring to adversarial attack. But the proposed method only takes probability-only output of the target model into consideration and ignores the label-only situation, which is a challenging and practical task in real-world application.

3. Methodology

3.1. Attack Scenarios and Notations

In real-world applications, pretrained models stored on a remote server only provide APIs for inference. Neither the model parameters nor the training data are accessible to users. Assume that attackers can only access the label or probability outputs of the black-box model returned by APIs. We define them as **label-only** and **probability-only** scenarios, respectively. Important notations appeared in this paper are described in Table 1.

Table 1. Important notations and their descriptions

Notation	Description
$\mathcal{T}, \mathcal{S}, \mathcal{G}$	target model, substitute model, generator
X, Z, Y	synthetic data, random noise, label

3.2. Framework Overview

In this section, we illustrate the framework of our proposed data-free adversarial attack method in Figure 2. The procedure of our method consists of two stages: 1) Efficient Data Generation and 2) Substitute Model Distillation. In stage 1, we reset the objective of generator $\mathcal{G}$ as synthesising desired data whose distribution is close to the target training data. $\mathcal{G}$ is not directly involved in substitute model distillation in stage 2. Consequently, the two players are no longer forced to directly compete in one minimize-maximize game. In stage 2, substitute model $\mathcal{S}$ aims to efficiently imitate the target model $\mathcal{T}$ with the generated data. Based on transferability [41] of adversarial examples, these adversarial examples carefully designed by $\mathcal{S}$ can then be transferred to $\mathcal{T}$. The detailed description of our method is shown in Algorithm 1.

3.3. Efficient Data Generation

Firstly, given a batch of random noise $Z = \{z_1, z_2, \cdots, z_n\}$ and pseudo label $Y = \{y_1, y_2, \cdots, y_n\}$, the generator $\mathcal{G}$ is utilized to maps Z to the desired data $X = \mathcal{G}(Z)$. The distribution of synthetic data X is expected to be similar to the real data. If the images generated by $\mathcal{G}$ have the same distribution as the training dataset, their predictions should also be similar. Thus, Then we optimize $\mathcal{G}$ as follows:

$$L_G = CE\left(\mathcal{T}\left(X\right), Y\right), \tag{1}$$

where CE denotes cross-entropy loss function. However, the back propagation of this loss requires the gradient information of $\mathcal{T}$, which violates the principles of black-box attacks. Therefore, we use $\mathcal{S}$ to approximate $\mathcal{T}$ in Equation 1 as (We empirically verify the feasibility of this replacement in experiment):

$$L_G = CE\left(\mathcal{S}\left(X\right), Y\right). \tag{2}$$

Note that the pseudo label y can be randomly generated or provided by $\mathcal{T}$. However, continuously querying $\mathcal{T}$ during data generation will greatly consume the limited query budget. As a result, we randomly sample Y as the pseudo ground-truth labels.

As discussed in the Introduction, the previous method suffers from the model collapse, resulting in low substitute model accuracy and low attack success rate. We attempt to alleviate the mode collapse problem, through the lens of balancing the generated data distribution and promoting the data diversity. In order to make the generated samples covering all categories in our method, we introduce information entropy to measure the degree of chaos for labels. Assuming that there are k categories in total, and $\mathcal{H}_{infor} = -\frac{1}{k}\sum_{i=1}^{k} p_i \log p_i$ is the information entropy loss for a given probability vector $P = \{p_1, p_2, \ldots, p_k\}$.

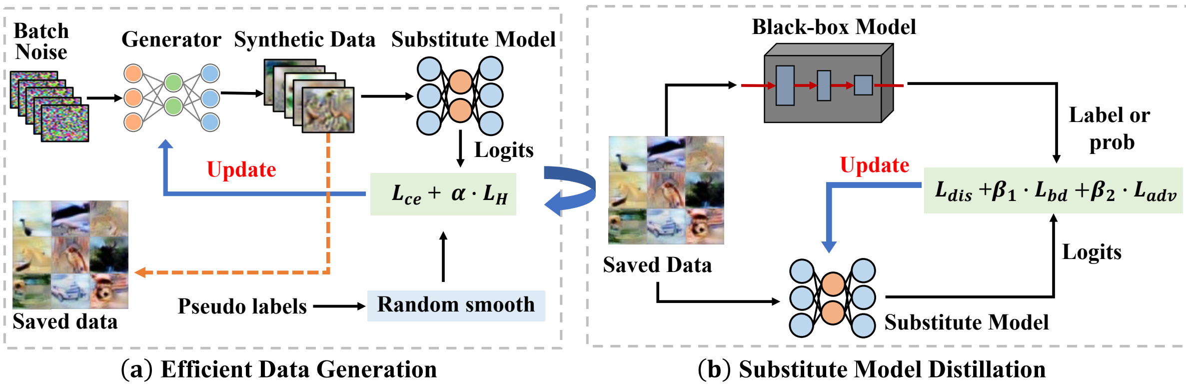

Figure 2. The illustration of our proposed data-free adversarial attack method.

Then the information entropy loss $\mathcal{L}_H$ for synthetic data is formulated as a regularization term:

$$\mathcal{L}_H = -\mathcal{H}_{infor}(\frac{1}{n}\sum_{i=1}^{k} S(X^i)) \tag{3}$$

When $\mathcal{L}_H$ reaches the maximum value, the categories are evenly distributed. To further promote the data diversity, we randomly smooth [40] the pseudo ground-truth labels and steer the generator to synthesis diverse data in each category.

In summary, we minimize the following loss function to update $\mathcal{G}$:

$$\mathcal{L}_G = CE\left(\mathcal{S}\left(X\right), \hat{Y}\right) + \alpha\mathcal{L}_H, \tag{4}$$

where α denotes the hyperparameter to adjust the value of regularization and $\hat{Y}$ is the smoothed label. For each epoch, we run t iterations to synthesize X. As opposed to previous research, our approach does not rely on the adversarially trained $\mathcal{G}$. Actually, we randomly initialize $\mathcal{G}$ at each epoch. In this case, $\mathcal{G}$ is only responsible for synthetic data X generated in this epoch, and $\mathcal{G}$ does not directly participate in the model distillation stage.

3.4. Substitute Model Distillation

Once we obtain the synthetic data X, the outputs of $\mathcal{T}(X)$ and $\mathcal{S}(X)$ are expected to be as consistent as possible. Inspired by knowledge distillation [15], $\mathcal{S}$ can imitate the outputs of $\mathcal{T}$ as follows:

$$L_{dis} = d(\mathcal{T}(X), \mathcal{S}(X)), \tag{5}$$

where d is a metric to measure the distance. In detail, for label-only scenario, this measurement can be the cross-entropy loss, and for probability-only scenario, d can be L_2 Norm.

To achieve higher attack success rate, the substitute model are encouraged to have decision boundaries that are highly consistent with the target model. However, as shown in Figure 3, the left sub-figure illustrate the poor data generated by previous methods with low diversity. They are far from the classification boundary. These data are very easy to learn for $\mathcal{S}$ and can easily lead to overfitting. To further improve the training efficiency of substitute mod-

Algorithm 1 The proposed data-free black-box attack.

Require: random noise Z, generator $\mathcal{G}$, target model $\mathcal{T}$, substitute model $\mathcal{S}$, synthetic data X, epochs E, iterations per epoch t, parameters θ_G, θ_S and learning rate γ_1, γ_2.

for each $e \in E$ **do**
  // *Efficient data generation:*
  for each $i \in t$ **do**
    Generate a batch of data $X \leftarrow \mathcal{G}(Z)$
    Compute $\mathcal{L}_G = CE\left(\mathcal{S}\left(X\right), y\right) + \alpha\mathcal{L}_H$
    Update $\theta_G \leftarrow \theta_G - \gamma_1 \bigtriangledown_{\theta_G} \mathcal{L}_G(\theta_G)$
    Save X **to** $D = \{X^1, \cdots, X^t\}$
  // *Substitute model distillation:*
  for x in D **do**
    Compute $\mathcal{L}_S = \mathcal{L}_{dis} + \beta_1 \cdot \mathcal{L}_{bd} + \beta_2 \cdot \mathcal{L}_{adv}$
    Update $\theta_S \leftarrow \theta_S - \gamma_2 \bigtriangledown_{\theta_S} \mathcal{L}_S(\theta_S)$
return θ_S

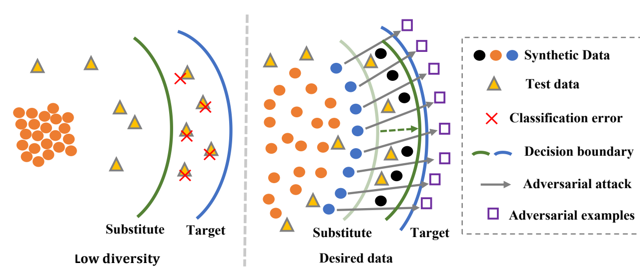

Figure 3. **Left:** Low diversity data and easy to learn. **Right:** Desired data for distillation.

els, we propose to go deeper into the utilization of synthetic data. Accordingly, we argue that there are two types of data that need to be given extra attention. The first type refers to data where there are decision disagreements between $\mathcal{S}$ and $\mathcal{T}$ (black circles). This type of data mainly exists between the decision boundaries of the target and substitute models. Giving more weight to these data helps to bridge the gap between the two decision boundaries. We pay more attention to those samples and introduce a boundary support loss:

$$\mathcal{L}_{bd} = d(\mathcal{T}(X), \mathcal{S}(X))\cdot 1 \left\{\arg\max \mathcal{T}(X) \neq \arg\max \mathcal{S}(X)\right\}. \tag{6}$$

The function 1 is an indicator when $\mathcal{T}$ and $\mathcal{S}$ produce inconsistent predictions on the given data.

Data whose adversarial samples can easily transfer from $\mathcal{S}$ to $\mathcal{T}$ are considered as another important type. The presence of this type of data means that near it, the decision boundary of $\mathcal{S}$ and $\mathcal{T}$ are relatively close. Giving more attention to this type of data ensures $\mathcal{S}$ continues to move in the right direction close to the boundary of $\mathcal{T}$. Then we introduce an adversarial samples support loss:

$$\mathcal{L}_{adv} = d(\mathcal{T}(X), \mathcal{S}(X))\cdot 1\left\{\arg\max \mathcal{T}(\hat{X}) = \arg\max \mathcal{S}(\hat{X})\right\}. \tag{7}$$

$\hat{X}$ is the adversarial examples generated by PGD [34] attack. Note that this loss will cost additional query cost. Note that this loss requires us to query the target model again.

In summary, we update the loss of $\mathcal{S}$ as:

$$\mathcal{L}_S = \mathcal{L}_{dis} + \beta_1 \cdot \mathcal{L}_{bd} + \beta_2 \cdot \mathcal{L}_{adv}, \tag{8}$$

where β_1 and β_2 control the value of different loss functions and are set to 1 by default.

4. Experiments

4.1. Experiment Setup

Datasets and models We evaluate our method on popular datasets: MNIST [23], FMNIST [47], SVHN [36], CIFAR10 [19], CIFAR100 [19] and Tiny-ImageNet [22]. Following the setting in [49], for MNIST and FMNIST, we employ a lightweight CNN model as the target model. A small CNN is used as the substitute model. Besides, we utilize ResNet-34 [14] for SVHN and CIFAR-10 as the target model, and use ResNet-18 [14] as the substitute model. Following the architecture in [12], we use the same generator in StyleGAN [18].

Training details The substitute models are trained with a batch size of 256 with SGD, with an initial learning rate of 0.01, a momentum of 0.9 and no weight-decay. The generator is also trained with a same batch size of 256, but using an Adam optimizer with a fixed learning rate of 0.001. As there are more categories in CIFAR100 and Tiny-ImageNet (100 classes in CIFAR100 and 200 classes in Tiny-ImageNet), we set the size to 1024 to maintain the diversity of data generated by $\mathcal{G}$. The training epoch is 400, and we train the generator 10 rounds at each epoch. The default query budget $Q = 20k$ for MNIST, FMNIST and SVHN, and $Q = 250k$ for CIFAR-10, CIFAR-100 and Tiny-ImageNet in our experiments.

Baselines To ensure fair comparisons, we compare our method with three types of state-of-the-art approaches: 1) black-box attacks that require for training data, e.g. JPBA [38] and Knockoff [37]; 2) data-free black-box attacks, e.g. DaST [49] and Del [44]; 3) data-free model extration attacks based on probability returned by the target model, e.g. DFME [43]. Note that this method is not designed for label-only scenarios. To facilitate comparison, we extend this method to label-only scenarios based on the framework of DaST. We conduct all experiments under a same query budget Q.

Evaluations We utilize three common attack methods to generate adversarial examples, which include FGSM [20], BIM [21], PGD [2] *. For FMNIST and FMNIST, we set perturbation bound $\epsilon = 32/255$, and step size $\alpha = 0.031$. And for SVHN, CIFAR10 and CIFAR100, we set the perturbation bound $\epsilon = 8/255$, step size $\alpha = 2/255$. In the untargeted attack scenario, we only generate adversarial examples on the images classified correctly by the attacked model. In targeted attacks, we only generate adversarial examples on the images which are not classified to the specific wrong labels. The attack success rate (ASR) are calculated by n/m, where n and m are the number of adversarial examples which can fool the attacked model and the total number of adversarial examples, respectively. To evaluate the performance of the proposed method in real-world tasks, we further apply our method to attack the online model of Microsoft Azure.

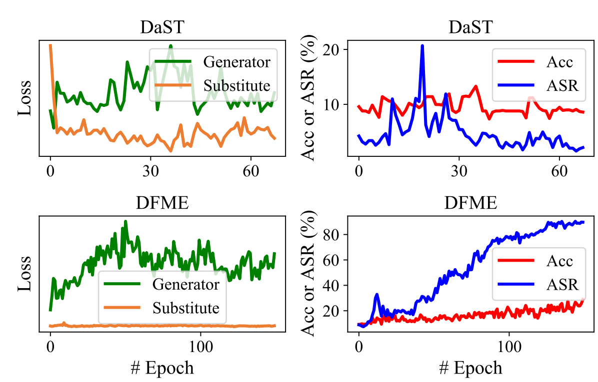

Figure 4. Training flaws of previous SOTA methods.

*We use AdverTorch for implementation

Table 2. ASR(%) comparisons between our proposed method and baselines over MNIST and FMNIST under a same query budget $Q = 20k$.

Dataset	Type / Method	Targeted, label-only FGSM	BIM	PGD	Untargeted, label-only FGSM	BIM	PGD	Targeted, probability-only FGSM	BIM	PGD	Untargeted, probability-only FGSM	BIM	PGD
MNIST	JPBA	3.89	6.89	5.31	18.14	23.56	20.18	4.29	7.02	5.49	18.98	25.14	21.98
	Knockoff	4.18	6.03	4.66	19.55	27.32	22.18	4.67	6.86	5.26	21.35	28.56	23.34
	DaST	4.33	6.49	5.17	20.15	27.45	27.13	4.57	6.41	5.34	25.36	29.56	29.14
	Del	6.45	9.14	6.13	22.13	25.69	23.18	6.97	9.67	6.24	24.56	25.35	25.28
	DFME	10.45	14.28	6.38	50.14	68.89	63.38	11.67	16.32	7.93	54.16	70.18	66.32
	Ours	**14.45**	**28.71**	**9.86**	**66.21**	**95.90**	**87.89**	**16.99**	**36.82**	**14.55**	**60.45**	**97.46**	**80.76**
FMNIST	JPBA	6.45	8.46	7.57	24.22	30.56	30.11	6.89	8.56	7.56	26.23	31.35	31.11
	Knockoff	6.34	8.35	7.32	28.19	36.88	35.92	6.65	8.98	8.23	30.21	36.94	36.22
	DaST	5.38	7.18	6.53	30.45	36.17	34.23	5.33	7.46	7.84	32.14	37.34	34.91
	Del	3.89	8.19	7.47	28.14	34.14	32.45	3.23	8.59	8.11	31.43	36.26	33.87
	DFME	7.18	22.45	24.58	60.45	74.29	72.19	9.44	26.89	25.74	62.15	78.56	77.89
	Ours	**30.08**	**76.46**	**32.42**	**91.41**	**100.00**	**98.83**	**31.15**	**79.3**	**35.45**	**91.99**	**99.90**	**98.93**

4.2. Empirical Studies of Previous Methods

To better illustrate the training flaws of the previous SOTA methods (DaST [49] and DFME [43]) that we mentioned in the introduction, in this section we provide an empirical analysis of their proposed min-max competition game. As shown in Figure 4, in the top subplot, we can see the loss for the generator (green) and the loss for the substitute (orange) both oscillating sharply over time. Meanwhile, the accuracy (red) of the substitute model fluctuates around a low level (10%) and the transferable attack successful rate (blue) gradually decreases in violent oscillations. This unstable training process caused by the inaccurate substitute-target disagreement makes the models hard to converge. In the bottom subplot, we can see a relatively stable substitute (orange) loss due to a more accurate estimation of the substitute-target disagreement. However, with the increase in the number of training epochs, the substitute model loss stays close to zero, despite the increasing generator loss (green). This suggests that the generator is poor at generating examples in some consistent way that makes the substitute model hard to learn any more knowledge from the target model. The substitute accuracy (red) and the attack success rate (blue), which remain low and no longer increase, also indicate the emergence of model collapse.

4.3. Black-box Attack Results

Experiments on MNIST and FMNIST We report the attack success rate under targeted and untargetd attack for label-only and probability-only scenarios. As shown in Table 2, the attack success rate of our method is much higher than other state-of-the-art baselines on all datasets. We remark that our proposed method can achieve a very high attack success rate with a small number of queries, while other methods perform poorly. Compared to targeted attacks, all of these methods show a higher ASR for untargeted attacks. The reason is that untargeted attacks attempt to misdirect the model to predict any incorrect class, whereas targeted attacks attempt to misguide the model to a particular class. Obviously, our method can even obtain higher ASR improvements than other baselines in targeted attacks. Moreover, we found that other methods can not achieve a satisfactory attack success rate with a small number of queries $Q = 10k$. This unstable training process caused by the inaccurate substitute-target disagreement makes the models hard to converge. Because their generators are trained with the inaccurate substitute-target disagreement, which is difficult to converge at an early stage. Consequently, these methods require a large number of queries, which is not practical in real-world applications.

Table 3. Attack success rate on MNIST. A large query budget $Q =$ 10M for all baselines, and a very small query budget $Q = 10k$ for our proposed method.

Type	Dataset	Attack	DaST (10M)	Del (10M)	DFME (10M)	Ours (10k)
Label-only	MNIST	FGSM	35.75	34.30	37.20	**66.21**
		BIM	38.58	38.65	70.85	**95.91**
		PGD	36.12	36.95	56.46	**87.89**
	FMNIST	FGSM	39.47	37.02	63.30	**91.99**
		BIM	42.65	42.66	74.08	**99.91**
		PGD	39.24	40.42	59.31	**98.93**
probability-only	MNIST	FGSM	55.64	53.34	58.44	**60.45**
		BIM	58.55	58.27	90.36	**97.46**
		PGD	55.89	56.92	75.88	**80.76**
	FMNIST	FGSM	59.13	56.97	82.43	**91.41**
		BIM	62.37	61.76	93.76	**100.00**
		PGD	58.90	60.20	79.26	**98.83**

To futher demonstrate the advantages of our method, we report the best results of other data-free adversarial attack methods with a large number of queries $Q = 10$M. As shown in Table 3, our method still outperforms other baselines by a large margin with only a small number of

Table 4. ASR(%) comparisons between our proposed method and baselines over several datasets. The default query budget $Q = 250k$.

Dataset	Type	Targeted, label-only			Untargeted, label-only			Targeted, probability-only			Untargeted, probability-only		
	Method	FGSM	BIM	PGD	FGSM	BIM	PGD	FGSM	BIM	PGD	FGSM	BIM	PGD
SVHN	JPBA	4.13	5.18	5.03	22.15	27.43	26.23	4.67	5.85	5.52	23.11	27.72	26.82
	Knockoff	3.89	4.98	4.82	23.78	26.05	24.75	4.43	5.50	5.15	24.51	26.94	24.99
	DaST	4.28	5.19	5.12	22.16	28.94	21.36	5.19	5.82	5.96	22.29	29.29	21.95
	Del	4.67	5.01	4.45	20.14	25.44	24.78	5.53	5.81	4.81	20.88	25.79	25.74
	DFME	9.78	15.38	14.11	34.18	36.82	35.11	10.12	15.88	14.45	34.23	37.54	35.54
	Ours	**21.58**	**31.25**	**21.88**	**55.76**	**76.37**	**74.51**	**19.34**	**32.81**	**24.02**	**58.01**	**76.37**	**75.59**
CIFAR10	JPBA	6.32	7.70	7.92	27.82	33.23	31.70	7.28	8.56	7.64	28.77	33.38	31.96
	Knockoff	6.26	7.02	7.04	29.61	31.86	30.68	6.46	8.27	7.35	30.02	31.98	30.35
	DaST	6.54	7.81	7.41	27.61	34.43	26.99	8.15	8.40	8.26	27.58	34.75	27.47
	Del	7.14	7.44	6.95	25.33	30.45	30.34	7.86	8.29	7.17	26.38	31.53	31.47
	DFME	12.62	18.32	16.76	39.66	42.07	40.51	12.58	18.70	16.80	39.43	43.33	40.69
	Ours	**34.57**	**76.95**	**72.27**	**86.13**	**99.22**	**99.41**	**31.54**	**73.93**	**69.14**	**83.89**	**99.32**	**99.02**
CIFAR100	JPBA	4.35	6.20	6.17	33.58	38.54	37.08	5.73	7.50	6.41	34.21	39.12	37.31
	Knockoff	4.40	5.86	5.25	34.84	36.92	36.34	4.88	7.05	6.18	36.01	37.61	35.47
	DaST	4.97	6.19	5.92	33.57	39.86	32.71	6.38	7.04	7.01	32.80	40.34	32.78
	Del	5.38	5.72	5.69	30.80	35.63	36.15	6.30	6.53	5.23	31.64	36.63	37.44
	DFME	11.23	17.02	15.58	45.66	47.26	46.22	10.62	17.62	15.17	44.76	48.73	46.51
	Ours	**26.64**	**46.88**	**42.77**	**78.61**	**91.31**	**91.21**	**7.91**	**56.15**	**52.54**	**83.69**	**94.53**	**94.14**

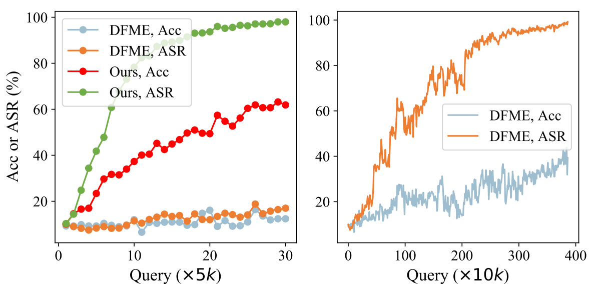

Figure 5. **Left:** ASR and accuracy of BIM attacks generated by our method and DFME with a limited query budget $Q = 150k$ on CIFAR10. **Right:** With a large number of queries $Q = 4000k$, DFME obtains a comparable performance.

queries $Q = 10k$. According to Table 2 and Table 3, these benchmark methods can achieve better performance when the number of queries is large. This is due to the gradual stabilization training of GAN in the later stages. On the contrary, in our method the generator and the substitute model are no longer forced to directly compete in one minimize-maximize game. As a result, our method can converge rapidly at the early stage. This is another proof that it is feasible and effective for us to replace $\mathcal{T}$ with $\mathcal{S}$ in generator training.

Experiments on SVHN and CIFAR-10, CIFAR-100 Since grayscale image datasets with a simple style (i.e. MNIST and FMNIST) are easily to learn for neural networks, underlying representations can be easily learnt when queried over synthetic data. Therefore, we futher investigate the performance of our method on more complex datasets. We remark that we have discussed in Figure 1 that the performance of DaST is very poor on MNIST with a small query budget, and it is difficult to scale to large datasets. Consequently, we first compare our method with the best baseline DFME on CIFAR10 dataset. As shown in Figure 5, a small query budget leads to extremely unstable performance for DFME. Our method is still able to obtain a much higher success rate and accuracy than DFME. Actually, the accuracy and ASR of our proposed method is 61.9% and 98.0% when $Q = 60k$, respectively. In sufficient queries ($Q = 400$M), DFME can obtain a comparable ASR (97.8%) to ours, but the test accuracy is much lower than our method (43.9%).

As shown in Table 4, we conduct extensive comparisons with multiple methods for each dataset under both probability-only and label-only scenarios. For both untarget and target attacking settings, our method achieves the best ASR over probability-only and label-only scenarios under all datasets. In addition, compared to the strong baseline DFME, our method significantly outperforms it with a large margin. Note that the number of categories directly affects the training of substitute model. Experiments on larger datasets (CIFAR-100 and Tiny-ImageNet) are all conducted with a large batch size (1024). Obviously, our method still achieves a very high ASR on CIFAR100 dataset, which has 100 categories of images. Experiments on Tiny-ImageNet can be found in the supplementary.

Attacks on the Microsoft Azure Online Model To investigate the effectiveness of our method in a real-world setting, we conduct experiments for attacking the online model on Microsoft Azure in two scenarios. Following the setting in [49], we employ the example MNIST model of the machine learning tutorial on Azure as the target model and make it available as a web service. The black-box scenario

Table 5. Attack results of the Microsoft Azure online model.

Type	Attack	DaST	Del	DFME	Ours
label-only	FGSM	66.46	65.22	80.24	**98.12**
	BIM	74.16	73.95	84.26	**100.00**
	PGD	72.55	71.28	83.16	**98.35**
probability-only	FGSM	71.32	70.05	84.72	**99.32**
	BIM	78.91	78.54	88.66	**100.00**
	PGD	77.49	76.00	87.34	**99.56**

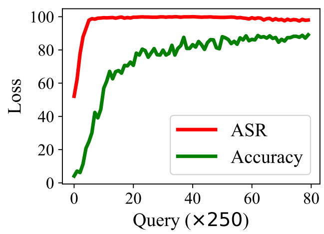

Figure 6. ASR of BIM attacks generated by our method for attacking the online model.

does not provide any information regarding this model, including its structure and parameters. We can only obtain information from the outputs of this model. The target model achieves 91.80% accuracy on MNIST test set. We report the untargetd attack results for both probability-only and label-only scenarios in Table 5, and all experiments are conducted with the query budget Q = 10k. Our method achieves a 100% attack success rate for both label-only and probability-only attacks.

We remark that as reported in [49], DaST is trained by 20M queries for the attacked model in the training stage. However, the attacked Azure model is too simple to attack for our method. We show the curve of attack success rate of BIM attacks generated by our method in the training stage of Azure experiments, which is shown in Figure 6. Obviously, our method can achieve a high ASR of 100% with a very small queries $Q = 9.2$k (much lower than DaST), and the accuracy of substitute model is 89.11%.

4.4. Comprehensive Understanding of our method

Contribution of different loss To begin with, we investigate the contribution of different loss functions introduced in our method, including the boundary support loss $\mathcal{L}_{bd}$ and the adversarial loss $\mathcal{L}_{adv}$ described in Section 3.4, as well as the information entropy loss $\mathcal{L}_H$. As shown in Table 6, cutting off both $\mathcal{L}_{bd}$ and $\mathcal{L}_{adv}$ will lead to poor performance, but it seems cutting off the $\mathcal{L}_H$ may lead to more severe degradation. According to our previous discussion in Section 3.4, if we do not control the distribution of the labels on the generated data, the generator can produce skewed data with an extreme distribution, i.e. label imbalance. Besides, the boundary support loss $\mathcal{L}_{bd}$ and the adversarial loss $\mathcal{L}_{adv}$ are also important for substitute model training.

Table 6. Ablation Study by cutting of different modules.

Method	SVHN	CIFAR10	CIFAR100
Ours	**77.13**	**99.26**	**94.36**
w/o $\mathcal{L}_H$	72.45	93.78	90.16
w/o $\mathcal{L}_{bd}$	73.69	96.02	91.58
w/o $\mathcal{L}_{adv}$	74.06	97.34	90.12

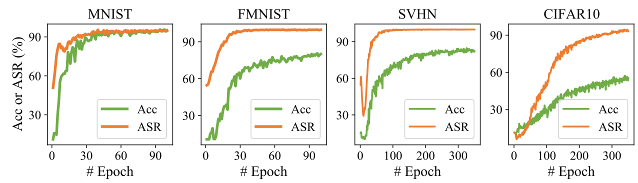

Figure 7. Accuracy and ASR of substitute models on various datasets for untargetd attack and probability-only scenarios.

Convergence Process In this section, we show detailed accuracy and ASR curves in Figure 7. The training accuracy increases smoothly during the whole training phrase, and converge to local optima at around 40 epochs for MNIST and 50 epochs for FMNIST. In each epoch, the black-box model is queried for 256 times. With such small queries, it strongly demonstrate the effectiveness of our method in stealing the target model in a data-free manner. Additionally, we can see from Figure 7 that the accuracy and ASR are highly correlated, and the training curve for accuracy fluctuates slightly. Due to the transferability of adversarial examples, ASR tends to be higher than the accuracy.

5. Conclusion

In this paper, we consider a more stringent yet more practical adversarial scenario, a black-box model with no access to the real data and limited budgets for querying the target model. Though the prior art has shed the light on data-free black-box attack, their GANs based framework suffers from the convergence failure and the model collapse, resulting in low efficiency. Rethinking the collaborating relationship between the generator and the substitute model, we design a powerful new black-box attack framework. The comprehensive experiments over the six datasets and one online machine learning platform demonstrate the proposed method can efficiently imitate the target model with a small query budget and achieve high attack success rate.

6. Acknowledgement

This work was supported by the National Key Research and Development Project of China (2021ZD0110400 No. 2018AAA0101900), National Natural Science Foundation of China (U19B2042), Zhejiang Lab (2021KE0AC02), Academy Of Social Governance Zhejiang University, Fundamental Research Funds for the Central Universities.

References

[1] 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

[2] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017. 1, 5

[3] Chen Chen, Jie Zhang, and Lingjuan Lyu. Gear: A margin-based federated adversarial training approach. 1

[4] 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 1277–1294. IEEE, 2020. 2

[5] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box 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

[6] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. *arXiv preprint arXiv:2203.08519*, 2022. 1

[7] 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

[8] 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'Alché-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

[9] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 3

[10] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[11] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, June 2020. 1

[12] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019. 3, 5

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5

[15] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 4

[16] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. *CoRR*, abs/2105.10872, 2021. 1

[17] 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

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 5

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017. 5

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5

[22] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[24] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8569–8576. AAAI Press, 2019. 1

[25] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8518–8527. IEEE, 2019. 1

[26] Bo Li, Zhengxing Sun, Lv Tang, and Anqi Hu. Two-b-real net: Two-branch network for real-time salient object detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 1662–1666. IEEE, 2019. 1

[27] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 818–825. ijcai.org, 2019. 1

[28] Bo Li, Zhengxing Sun, Quan Wang, and Qian Li. Co-saliency detection based on hierarchical consistency. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1392–1400. ACM, 2019. 1

[29] Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Jilin Li, and Feiyue Huang. Detecting adversarial patch attacks through global-local consistency. In Dawn Song, Dacheng Tao, Alan L. Yuille, Anima Anandkumar, Aishan Liu, Xinyun Chen, Yingwei Li, Chaowei Xiao, Xun Yang, and Xianglong Liu, editors, *ADVM '21: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia, Virtual Event, China, 20 October 2021*, pages 35–41. ACM, 2021. 1

[30] Siao Liu, Zhaoyu Chen, Wei Li, Jiwei Zhu, Jiafeng Wang, Wenqiang Zhang, and Zhongxue Gan. Efficient universal shuffle attack for visual object tracking. *arXiv preprint arXiv:2203.06898*, 2022. 1

[31] Lingjuan Lyu and Chen Chen. A novel attribute reconstruction attack in federated learning. *arXiv preprint arXiv:2108.06910*, 2021. 1

[32] Lingjuan Lyu, Yitong Li, Karthik Nandakumar, Jiangshan Yu, and Xingjun Ma. How to democratise and protect ai: Fair and differentially private decentralised deep learning. *IEEE Transactions on Dependable and Secure Computing*, 2020. 1

[33] Lingjuan Lyu, Han Yu, Xingjun Ma, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S Yu. Privacy and robustness in federated learning: Attacks and defenses. *arXiv preprint arXiv:2012.06337*, 2020. 1

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 5

[35] Paul Micaelli and Amos J. Storkey. Zero-shot knowledge transfer via adversarial belief matching. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-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 9547–9557, 2019. 3

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[37] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. *CoRR*, abs/1812.02766, 2018. 5

[38] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*, abs/1602.02697, 2016. 5

[39] 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

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 4

[41] 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

[42] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3560–3570. IEEE, 2021. 1

[43] Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-free model extraction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4771–4780. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 6

[44] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4761–4770. Computer Vision Foundation / IEEE, 2021. 1, 3, 5

[45] Chuhan Wu, Fangzhao Wu, Ruixuan Liu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Fedkd: Communication efficient federated learning via knowledge distillation. *arXiv preprint arXiv:2108.13323*, 2021. 1

[46] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Fedctr: Federated native ad ctr prediction with cross platform user behavior data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2022. 1

[47] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 5

[48] Yijie Zhong, Bo Li, Lv Tang, Hao Tang, and Shouhong Ding. Highly efficient natural image matting. *CoRR*, abs/2110.12748, 2021. 1

[49] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 231–240. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 5, 6, 7, 8