

# Minimizing Maximum Model Discrepancy for Transferable Black-box Targeted Attacks

Anqi Zhao<sup>1</sup> Tong Chu<sup>1</sup> Yahao Liu<sup>1</sup> Wen Li<sup>2\*</sup> Jingjing Li<sup>1</sup> Lixin Duan<sup>1</sup>

<sup>1</sup>School of Computer Science and Engineering, University of Electronic Science and Technology of China

<sup>2</sup>Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China

{zhaoanqiii, uestcchutong, lyhaolive, liwenbnu, lxduan}@gmail.com, lijjin117@yeah.net

## Abstract

*In this work, we study the black-box targeted attack problem from the model discrepancy perspective. On the theoretical side, we present a generalization error bound for black-box targeted attacks, which gives a rigorous theoretical analysis for guaranteeing the success of the attack. We reveal that the attack error on a target model mainly depends on empirical attack error on the substitute model and the maximum model discrepancy among substitute models. On the algorithmic side, we derive a new algorithm for black-box targeted attacks based on our theoretical analysis, in which we additionally minimize the maximum model discrepancy (M3D) of the substitute models when training the generator to generate adversarial examples. In this way, our model is capable of crafting highly transferable adversarial examples that are robust to the model variation, thus improving the success rate for attacking the black-box model. We conduct extensive experiments on the ImageNet dataset with different classification models, and our proposed approach outperforms existing state-of-the-art methods by a significant margin. The code will be available at <https://github.com/Asteriajojo/M3D>.*

## 1. Introduction

Recently, researchers have shown that Deep Neural Networks (DNNs) are highly vulnerable to adversarial examples [9, 28, 36]. It has been demonstrated that by adding small and human-imperceptible perturbations, images can be easily misclassified by deep-learning models. Even worse, adversarial examples are shown may have *transferability*, *i.e.*, adversarial examples generated by one model can successfully attack another model with a high probability [23, 28, 37]. Consequently, there is an increasing interest in developing new techniques to attack an unseen black-box model by constructing adversarial examples on

a substitute model, which is also known as *black-box* attack [5, 6, 14–16, 20, 39, 40].

While almost all existing black-box attack works implicitly assume the transferability of adversarial examples, the theoretical analysis of the transferability is still absent. To this end, in this work, we aim to answer the question of to what extent the adversarial examples generated on one known model can be used to successfully attack another unseen model. In particular, we are specifically interested in the targeted attack task, *i.e.*, constructing adversarial examples that can mislead the unseen black-box model by outputting a highly dangerous specified class. We first present a generalization error bound for black-box targeted attacks from the model discrepancy perspective, in which we reveal that the attack error on a target model depends on the *attack error on a substitute model* and the *model discrepancy between the substitute model and the black-box model*. Furthermore, the latter term can be bounded by the maximum model discrepancy on the underlying hypothesis set, which is irrelevant to the unseen target model, making it possible to construct adversarial examples by directly minimizing this term and thus the generalization error.

Based on the generalization error bound, we then design a novel method called Minimizing Maximum Model Discrepancy (M3D) attack to produce highly transferable perturbations for black-box targeted attack. Specifically, we exploit two substitute models which are expected to maintain their model discrepancy as large as possible. At the same time, we train a generator that takes an image as input and generates an adversarial example to attack these two substitute models and simultaneously minimize the discrepancy between the two substitute models. In other words, the generator and the two substitute models are trained in an adversarial manner to play a min-max game in terms of the model discrepancy. In this way, the generator is expected to generate adversarial examples that are robust to the variation of the substitute models, thus being capable of attacking the black-box target model successfully with a high chance.

\*The corresponding author

We conduct extensive experiments on the ImageNet dataset using different benchmark models, where our M3D approach outperforms state-of-the-art methods by a significant margin on a wide range of attack settings. Especially, we show impressive improvements in the situations when the black-box model has a large model discrepancy from the substitute model, such as attacking the ResNet [11] model by crafting adversarial examples on a VGG [35] model. The main contributions of this paper are as follows:

- We present a generalization error bound for black-box targeted attacks based on the model discrepancy perspective.
- We design a novel generative approach called Minimizing Maximum Model Discrepancy (M3D) attack to craft adversarial examples with high transferability based on the generalization error bound.
- We demonstrate the effectiveness of our method by strong empirical results, where our approach outperforms the state-of-the-art methods by a significant margin.

## 2. Related Work

**Adversarial Attack:** Many works on the adversarial attack have been proposed, since they reveal that deep neural networks (DNNs) are highly vulnerable to adversarial examples [4–6, 14, 16, 30, 39, 40]. Generally, methods for adversarial attack can be divided into two branches: iterative perturbation methods and generative perturbation methods.

*Iterative Perturbations:* Iterative instance-specific attacks [5, 6, 38–40, 43] perturb a given sample by iteratively using gradient information. Though good performances have been achieved in white-box attack scenarios [9, 18, 25], their black-box transferability is limited. Therefore, many methods have been proposed to improve the transferability against black-box models. Some works are inspired by the model training processes such as introducing momentum items to stabilize optimization [5] or adapting Nesterov accelerated gradient into the iterative attacks [22]. Data augmentation methods have been shown to effectively boost transferability, such as applying random transformations [40], translation [6], and admixing input images with images from other classes [38]. Some works also concentrate on modifying the source model properly such as utilizing skip connections [39] or the back-propagation process [10] to boost transferability. Zhu et al. [43] also boost the transferability by matching the adversarial attacks with the directions which decrease the ground truth density. As for targeted attacks, Li et al. [20] take adversarial examples away from the true label and push towards the target label, while [14, 16] make the source image closer to an image of the target class in feature space, and then Inkawhich

et al. [15] generate targeted perturbations along with classifier information which transfers better than previous iterative methods. Gao et al. [7] align the source and target feature maps by high-order statistics with translation invariance. Zhao et al. [42] use logit loss and generate perturbations by a large number of iterations.

*Generative Perturbations:* Another branch of attack uses generative models to craft adversaries. Compared with iterative methods, generative attacks achieve higher efficiency when attacking large-scale datasets, and they find adversarial patterns on a data distribution independent of a single sample. As long as the distribution of the clean images is fixed, the effectiveness of the adversarial examples depends on how the generator is trained. Therefore, many previous works about generative attacks aim to learn a strong generator, which maps the distribution of clean images to a distribution of adversarial examples that can generalize well on different black-box models. Our method follows this branch. Poursaeed et al. [30] propose novel generative models to produce image-agnostic and image-dependent perturbations. Naseer et al. [27] use a relativistic training objective to boost cross-domain transferability. Zhang et al. [41] enhance the cross-domain transferability of adversarial examples from the data and model perspectives. Naseer et al. [26] achieve state-of-the-art performances in targeted attacks by maximizing the mutual agreement between the given source and the target distribution.

However, to the best of our knowledge, the theoretical analysis of the transferability is still absent, while in this work we make the first attempt to provide a generalization error analysis for black-box targeted attacks based on the model discrepancy. A new approach is also proposed based on our theoretical analysis, which achieves new state-of-the-art results for black-box targeted attacks.

**Hypothesis Discrepancy:** Ben-David et al. [3] define hypothesis discrepancy theory to measure the discrepancy of different models over a given distribution. Based on hypothesis discrepancy theory, the  $\mathcal{H}$ -divergence [2, 3] and  $\mathcal{H}\Delta\mathcal{H}$ -divergence [2, 3] are proposed for measuring the difference between data distributions. These metrics are widely used in Domain Adaptation theory [1–3, 31] to derive upper bounds for generalization error. Many Domain Adaptation methods tried to minimize the generalization error upper bound by minimizing hypothesis discrepancy [19, 21, 24, 33]. These methods usually use multiple models to maximize and minimize the hypothesis discrepancy over the target domain such that the hypothesis discrepancy is small in the entire hypothesis set  $\mathcal{H}$ . Different from these works, in this paper, we exploit the hypothesis discrepancy principle for analyzing black-box targeted attacks.

### 3. Methodology

In this section, we first present the preliminary and generalization error bound for black-box targeted attack in Section 3.1 and 3.2. Then we introduce the Minimizing Maximum Model Discrepancy (M3D) attack in Section 3.3 and its implementation details in Section 3.4.

#### 3.1. Black-Box Targeted Attack

In the black-box targeted attack task, we are given an unseen black-box classification model  $D_b$  and an image dataset  $\mathcal{A} = \{(\mathbf{x}, y)\}$  with  $\mathbf{x}$  being an image and  $y$  being its class label. Normally,  $D_b$  is trained to classify these images with high accuracy, *i.e.*, we expect  $\arg \max_c D_b(\mathbf{x}) = y$ . The goal is, given any image  $\mathbf{x}$ , to produce an adversarial example  $\mathbf{z}$  to mislead the black-box model  $D_b$  towards a specified class  $y^t$  such that  $\arg \max_c D_b(\mathbf{z}) = y^t$ . Without loss of generality, the process of generating an adversarial example can be denoted as  $\mathbf{z} = G(\mathbf{x})$ , where  $G$  can be implemented via an iterative optimization or a generator.

For ease of presentation, we define  $\mathcal{X} = \{\mathbf{x}\}$  as the set of images and the underlying distribution of images as  $\mathcal{P}$ , namely, the image  $\mathbf{x} \in \mathcal{X}$  is deemed being sampled from the distribution  $\mathcal{P}$ , *i.e.*,  $\mathbf{x} \sim \mathcal{P}$ . Correspondingly, the set of adversarial examples is denoted by  $\mathcal{Z} = \{\mathbf{z}\}$ , and  $\mathbf{z} \sim \mathcal{Q}$  where  $\mathcal{Q}$  is the underlying distribution.

Since the black-box model  $D_b$  is inaccessible, existing works usually construct adversarial examples based on a substitute model  $D_s$ . Owing to the transferability, these adversarial examples are often able to attack the black-box  $D_b$  with a decent chance. Formally, constructing adversarial examples on a substitute model for a black-box targeted attack can be written as,

$$\mathbf{z} = \underset{\mathbf{z}}{\operatorname{argmin}} \ell(D_s(\mathbf{z}), y^t), \text{ s.t. } \|\mathbf{z} - \mathbf{x}\|_\infty \leq \epsilon, \quad (1)$$

where the  $l_\infty$ -norm is applied to constrain the imperceptibility of the perturbation, and  $\ell$  is the classification loss (*e.g.*, cross-entropy loss).

#### 3.2. Generalization Error Bound

Let us define  $\mathcal{H}$  as a hypothesis set that contains all classifiers which perform well on the specified classification task. We also reasonably assume that the black-box model  $h_b$  and an arbitrary substitute model  $h_s$  satisfy  $h_b, h_s \in \mathcal{H}$ . Let us define  $f_t$  as the labeling function for the misclassified target category, namely,  $f_t$  classifies all images into the specified target category.

The generalization error  $\mathcal{E}_Q(h_b, f_t)$  for the black-box targeted attack can be defined as:

$$\mathcal{E}_Q(h_b, f_t) = \mathbb{E}_{\mathbf{z} \sim \mathcal{Q}} [\ell(h_b(\mathbf{z}), f_t(\mathbf{z}))], \quad (2)$$

where  $\ell$  is the metric function that measures the prediction difference between  $h_b$  and  $f_t$  for a sample  $\mathbf{z}$ . Since we cannot access the black-box model  $h_b$ , we seek an upper bound of the generalization error  $\mathcal{E}_Q(h_b, f_t)$  as follows.

**Theorem 1.** *For any  $\delta \geq 0$ , with probability  $1 - \delta$ , we have the following generalization bound for the black-box classifier  $h_b \in \mathcal{H}$  and any substitute classifier  $h_s \in \mathcal{H}$ ,*

$$\mathcal{E}_Q(h_b, f_t) \leq \hat{\mathcal{E}}_{\mathcal{Z}}(h_s, f_t) + \sup_{h, h' \in \mathcal{H}} \hat{\mathcal{E}}_{\mathcal{Z}}(h, h') + \Omega, \quad (3)$$

where  $h$  and  $h'$  are two classifiers sampled from  $\mathcal{H}$ ,  $\hat{\mathcal{E}}$  is the empirical estimation of the generalization error, and  $\Omega$  is a minor term.

The proof can be derived based on the  $\mathcal{H}$ -divergence theory which was originally proposed for domain adaptation [3]. We sketch the proof here, and the completed proof is provided in the supplementary. By introducing the substitute model  $h_s$ , we have  $\mathcal{E}_Q(h_b, f_t) \leq \mathcal{E}_Q(h_s, f_t) + \mathcal{E}_Q(h_s, h_b)$ . Moreover, from the definition of  $\mathcal{H}$ -divergence, we have  $\mathcal{E}_Q(h_s, h_b) \leq \sup_{h, h' \in \mathcal{H}} \mathcal{E}_Q(h, h')$ . Finally, we transform the generalization error  $\mathcal{E}_Q$  into an empirical estimation error  $\hat{\mathcal{E}}_{\mathcal{Z}}$ , where an additional minor term  $\Omega$  is introduced which usually can be neglected in optimization.

Despite the minor term  $\Omega$ , the above generalization bound mainly consists of two items: 1) the empirical error of attacking the substitute model; 2) the maximum empirical difference between any two models in the hypothesis set  $\mathcal{H}$ . Most existing black-box targeted attack works assume the good transferability of the adversarial samples and mainly focus on designing techniques for attacking the substitute model. This can be seen to minimize the first term only in Eq. (3) while ignoring the second term. However, the adversarial examples found in this way might not be robust enough to the variation of models, especially when there exists a model architecture difference between black-box and substitute models. Next, we will present a new approach for constructing robust adversarial samples by taking the second term into consideration as well.

#### 3.3. Minimizing Maximum Model Discrepancy

As analyzed above, the robust adversarial examples for attacking a black-box model not only successfully attack the substitute model but also should preserve a minimum discrepancy between any two models in the hypothesis set.

To calculate the empirical maximum model discrepancy  $\sup_{h, h' \in \mathcal{H}} \hat{\mathcal{E}}_{\mathcal{Z}}(h, h')$  (*i.e.*, the second term in Eq. (3)), we introduce two discriminator models  $D_1$  and  $D_2$ . Then, the empirical maximum model discrepancy can be calculated by

$$\max_{D_1, D_2} \hat{\mathbb{E}}_{\mathbf{x} \in \mathcal{X}} d[D_1 \circ G(\mathbf{x}), D_2 \circ G(\mathbf{x})], \quad (4)$$

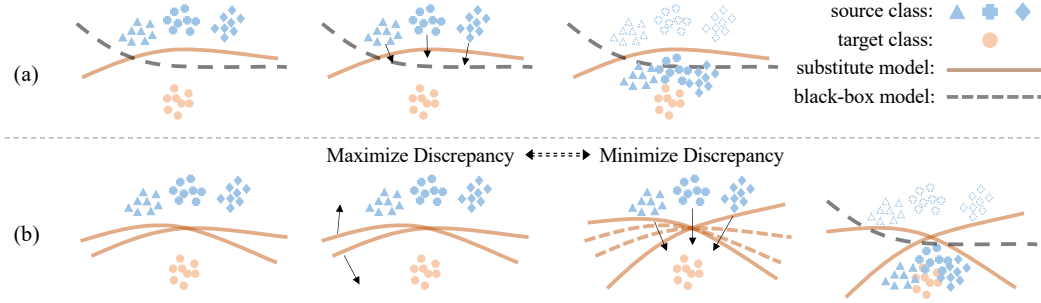


Figure 1. Illustration on our M3D approach: (a) Traditional targeted attack methods craft adversarial examples (moving blue plots towards orange plots) on one substitute model (orange line), which might not be always effective due to the model discrepancy between the substitute model and the black-box model (black dash line); (b) Our M3D approach generates strong adversarial examples by a min-max game to attack two substitute models with the maximum model discrepancy, which are more robust to the model discrepancy (Best viewed in color).

where  $\hat{\mathbb{E}}$  represents the empirical estimator, *i.e.*,  $\hat{\mathbb{E}}_{\mathbf{x} \in \mathcal{X}} d(\cdot, \cdot) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} d(\cdot, \cdot)$ , and  $d(\cdot, \cdot)$  is the discrepancy metric between the  $D_1$  and  $D_2$ . We use the  $\ell_1$  distance as the discrepancy metric. Note that here we rewrite the empirical estimator  $\hat{\mathcal{E}}_{\mathcal{Z}}$  as  $\hat{\mathbb{E}}_{\mathbf{x} \in \mathcal{X}}$ , as each adversarial example  $\mathbf{z} \in \mathcal{Z}$  is generated correspondingly from the clean image  $\mathbf{x} \in \mathcal{X}$ .

On the other hand, the robust adversarial examples are expected to preserve a minimum discrepancy for any two models, which leads to the following min-max objective,

$$\min_G \max_{D_1, D_2} \mathcal{L}_d = \hat{\mathbb{E}}_{\mathbf{x} \in \mathcal{X}} d[D_1 \circ G(\mathbf{x}), D_2 \circ G(\mathbf{x})]. \quad (5)$$

Note the adversarial examples are also expected to successfully attack a given substitute  $h_s \in \mathcal{H}$ . As  $h_s$  can be any classifier, without loss of generality, we use both  $D_1$  and  $D_2$  as the substitute models. Therefore, the generator should be trained to mislead the two discriminators to classify the adversarial example into the specified category  $y^t$ ,

$$\min_G \mathcal{L}_a = \hat{\mathbb{E}}_{\mathbf{x} \in \mathcal{X}} \ell_{ce}(D_1 \circ G(\mathbf{x}), y^t) + \ell_{ce}(D_2 \circ G(\mathbf{x}), y^t), \quad (6)$$

where  $\ell_{ce}(\cdot, \cdot)$  is the cross-entropy loss.

Moreover, as we presume that the hypothesis set  $\mathcal{H}$  consisting of classifiers that perform well on the specified classification task, these two discriminators  $D_1$  and  $D_2$  should maintain a low classification loss on the clean data as well,

$$\min_{D_1, D_2} \mathcal{L}_c = \hat{\mathbb{E}}_{(\mathbf{x}, y) \in \mathcal{A}} [\ell_{ce}(D_1(\mathbf{x}), y) + \ell_{ce}(D_2(\mathbf{x}), y)], \quad (7)$$

where  $\ell_{ce}(\cdot, \cdot)$  is the cross-entropy loss. By taking the losses in Eq. (5), Eq. (6), and Eq. (7) into consideration, the overall objective function of our minimizing maximum model discrepancy (M3D) attack approach can be written as,

$$\min_G \mathcal{L}_a + \max_{D_1, D_2} [\mathcal{L}_d - \mathcal{L}_c]. \quad (8)$$

By optimizing the above objective function, we can learn a strong generator  $G(\mathbf{x})$ , which maps the clean images to the adversarial examples that are robust to the model variation. We illustrate the intuition behind our M3D approach in Fig. 1. The previous targeted attack works mainly concentrate on crafting adversarial examples on a fixed substitute model (orange line), which can be seen as moving the samples from source classes (blue plots) towards the samples from the target class (orange plots) by adding small perturbations on the original images. However, these adversarial examples may fail to attack the black-box model (black line) due to the model discrepancy between the substitute and the black-box model. In contrast, in our M3D approach, we generate strong adversarial examples to attack two substitute models, which are trained dynamically to maintain a maximum discrepancy for predicting these adversarial examples. In this way, the adversarial examples would move closer to the target class thus being more robust to the variation of the black-box model.

### 3.4. Implementation

We depict the training pipeline of our M3D approach in Fig. 2, which consists of a generative model  $G$  and two discriminators  $D_1$  and  $D_2$ . The clean images pass through  $G$  to get their adversarial versions, and then these adversarial examples are respectively fed into  $D_1$  and  $D_2$ . During training, we alternatively train the discriminators and the generator for a number of iterations. For the discriminators, we fix  $G$  and train them to minimize the classification loss on the clean images and simultaneously maximize the discrepancy between the two discriminators. For the generator, we fix  $D_1$  and  $D_2$  and train  $G$  to attack both  $D_1$  and  $D_2$  and at the same time minimize the discrepancy between the two discriminators. The whole training procedure is shown in Algorithm 1. We detail the design of the generator and discriminators as follows.

**Generator:** For the generator, we use the same network

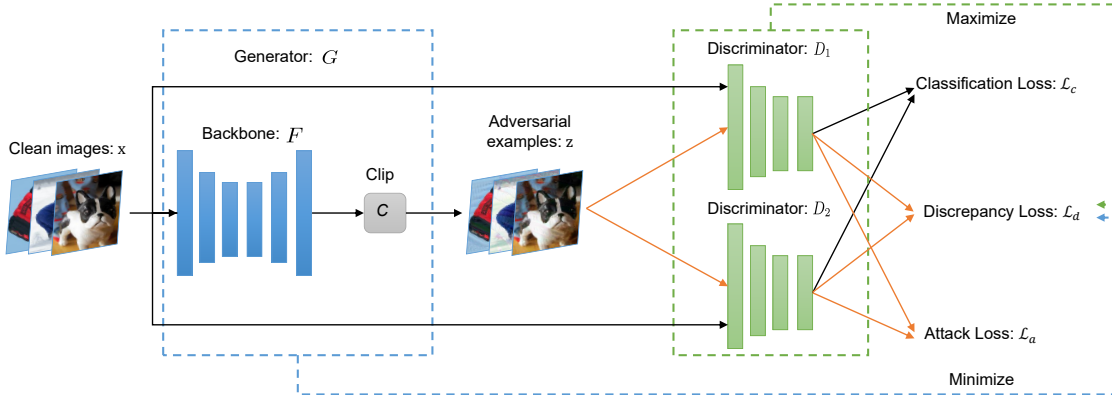


Figure 2. Overview of our M3D approach for black-box targeted attack, which consists of a generator and two discriminators. The discriminators act as substitute models and are trained to keep a large discrepancy on the adversarial examples. The generator is trained to generate robust adversarial examples that are able to attack two discriminators and keep their discrepancy as small as possible. The gray rectangle labeled with  $C$  is the clipping operation. See the description in Section 3.4 for details.

---

**Algorithm 1** Minimizing Maximum Model Discrepancy for Black-box Targeted Attack

---

**Require:** Training data  $\mathcal{A}$ , pretrained substitute models  $D_1, D_2$ , perturbation budget  $\epsilon$ .

- 1: Randomly initialize the generator  $G$
  - 2: **repeat**
  - 3: Randomly sample a mini-batch of training data  $\tilde{\mathcal{A}} = \{(\mathbf{x}, y)\} \subset \mathcal{A}$ , and denote mini-batch of images as  $\tilde{\mathcal{X}} = \{x\}$ .
  - 4: Forward-pass all images  $\mathbf{x} \in \tilde{\mathcal{X}}$  through the generator and generate bounded adversaries  $\{\mathbf{z}\}$  using Eq. (9).
  - 5: Forward pass all  $\mathbf{z}$  through  $D_1, D_2$ .
  - 6: **Update  $G$ :**
  - 7: Calculate the loss  $\mathcal{L}_a + \mathcal{L}_d$  and perform backward pass to update  $G$  while keeping  $D_1, D_2$  frozen.
  - 8: **Update  $D_1$  and  $D_2$ :**
  - 9: Forward pass all clean images  $\mathbf{x}$  through  $D_1, D_2$ .
  - 10: Calculate the loss  $\mathcal{L}_c - \mathcal{L}_d$ , and perform backward pass to update  $D_1, D_2$ , while keeping  $G$  fixed.
  - 11: **until** The stop criterion is reached.
- 

architecture as in previous generative attack works [26, 27, 30]. The generator consists of a backbone  $F$  and a clip operation  $C$ . The generator backbone consists of downsampling, residual and upsampling blocks whose output is an adversarial sample with the same size as of input. The detail of the generator network is provided in the supplementary. As the adversarial noise  $\delta = \|G(x) - x\|$  is strictly constrained under a norm distance for the imperceptibility of perturbation *i.e.*,  $\|\delta\|_\infty \leq \epsilon$ , we follow [26] to scale the output of  $F$  by using a differentiable clipping operation.

$$\mathbf{z} = \text{clip}(\min(\mathbf{x} + \epsilon, \max(\mathcal{W} * F(\mathbf{x}), \mathbf{x} - \epsilon))), \quad (9)$$

where  $\mathcal{W}$  is a smoothing operator with fixed weights that reduces high frequencies without violating the  $l_\infty$  distance constraint.

**Discriminators:** The discriminators can be any classification model. We follow the previous targeted attack works [26] to attack a black-box model trained on ImageNet, so we initialize the discriminators with the ImageNet pre-trained classification models using different architectures such as ResNet [11], DenseNet [13], VGGNet [35] for an extensive evaluation (see Section 4.2 for details).

## 4. Experiments

### 4.1. Experiments Setup

For a fair comparison, we follow the recent state-of-the-art method TTP [26] to evaluate our method. We adopt the same architecture for the generator  $G$  and vary the substitute model with different architecture as similar as in TTP [26]. Note that in each round of experiments (*e.g.*, using ResNet50 as the substitute model), the substitute model in all methods shares the same network architecture to ensure a fair comparison. Moreover, the two discriminator models  $D_1$  and  $D_2$  in our M3D approach also share the same network architecture. For the training data, we use 49.95K images sampled from ImageNet train set [32]. For the test data, we adopt the original ImageNet validation set (50k samples). We use Adam optimizer [17] with a learning rate of 0.0002 and the values of the exponential decay rate for the first and second moments are set to 0.5 and 0.999, respectively. We conduct all the experiments on GeForce RTX 3090 GPUs with a PyTorch [29] implementation.

We evaluate our approach on both subset-source setting [14] and all-source setting [26]. In the subset-source setting, 10 classes in the test data are used for evaluation. Each round, we select one of these 10 classes as the target class (50 samples), and the remaining 9 classes (450 samples) are the samples that need to be attacked. In the all-source setting, 1000 classes in the test data are used

Table 1. Target Transferability Against Unknown Models: Top-1(%) target accuracy in all-source and sub-source setting . Perturbation budget:  $l_\infty \leq 16$  . In comparison to previous state-of-the-art methods, our method outperforms them by a large margin. ‘\*’ indicates white-box attack. In subset-source setting, only black-box attack results are shown.

Attack	Naturally Trained									
	All-source					Subset-source				
	Method	VGG19 <sub>BN</sub>	Dense121	ResNet50	ResNet152	WRN-50-2	Method	VGG19 <sub>BN</sub>	Dense121	ResNet50
VGG19 <sub>BN</sub>	MIM [5]	<b>99.91*</b>	0.92	0.68	0.36	0.47	AA [16]	–	0.8	0.6
	DIM [40]	99.38*	3.10	2.08	1.02	1.29	FDA-fd [14]	–	3.0	2.1
	GAP [30]	98.23*	16.19	15.83	5.89	7.78	FDA <sup>N</sup> [15]	–	6.0	5.4
	CDA [27]	98.30*	16.26	16.22	5.73	8.35	CDA [27]	–	17.82	17.09
	TTP [26]	98.54 *	45.77	45.87	27.18	32.63	TTP [26]	–	48.29	47.07
	Ours	99.22*	<b>79.46</b>	<b>81.91</b>	<b>68.41</b>	<b>68.43</b>	Ours	–	<b>80.47</b>	<b>81.44</b>
Dense121	MIM [5]	1.85	<b>99.90*</b>	2.71	1.68	1.88	AA [16]	0.0	–	0.0
	DIM [40]	7.31	98.81*	9.06	5.78	6.29	FDA-fd [14]	34.0	–	34.0
	GAP [30]	39.01	97.30*	47.85	39.25	34.79	FDA <sup>N</sup> [15]	42.0	–	48.3
	CDA [27]	42.77	97.22*	54.28	44.11	46.01	CDA [27]	44.84	–	53.73
	TTP [26]	58.90	97.61*	68.72	57.11	56.80	TTP [26]	61.75	–	69.60
	Ours	<b>92.73</b>	98.60*	<b>94.23</b>	<b>90.06</b>	<b>90.75</b>	Ours	<b>93.02</b>	–	<b>95.22</b>
ResNet50	MIM [5]	1.58	3.37	98.76*	3.39	3.17	AA [16]	1.1	2.0	–
	DIM [40]	9.14	15.47	<b>99.01*</b>	12.45	12.61	FDA-fd [14]	16.0	21.0	–
	GAP [30]	58.47	71.72	96.81*	64.89	61.82	FDA <sup>N</sup> [15]	32.1	48.3	–
	CDA [27]	64.58	73.57	96.30*	70.30	69.27	CDA [27]	68.55	75.68	–
	TTP [26]	78.15	81.64	97.02*	80.56	78.25	TTP [26]	79.04	84.42	–
	Ours	<b>92.41</b>	<b>94.39</b>	98.33*	<b>93.85</b>	<b>93.87</b>	Ours	<b>92.24</b>	<b>95.36</b>	–

for evaluation. Each round, we select one of these 1000 classes as the target (50 samples), and the remaining 999 classes (49,950 samples) are the samples that need to be attacked. For both settings, we evaluate our model at each round and report the average Top-1(%) target accuracy over all rounds, where the Top-1(%) target accuracy means the proportion of adversarial examples that the black-box model predicts to the target class.

## 4.2. Comparison with State-of-the-art Methods

### 4.2.1 Attacking Unknown Models

We first evaluate our proposed approach for attacking unseen classification models pre-trained on ImageNet with both subset-source and all-source settings. Following [26], we respectively conduct experiments using the VGG19<sub>BN</sub>, DenseNet121, ResNet50 to initialize our substitute models, and attack VGG19<sub>BN</sub>, DenseNet121, and ResNet50 models. The ResNet152 and WRN-50-2 are additionally used as black-box models in the all-source setting. We follow [26] to compare with a bunch of baselines including MIM [5], AA [16], FDA [14], DIM [40], GAP [30], CDA [27] and TTP [26].

The visualization of our adversarial examples is shown in Fig. 3. The results of all methods on different settings are shown in Table 1, where the results of baselines are from [26]. We observe that our proposed approach outperforms all existing methods for black-box targeted attacks with a clear margin. Specifically, in the all-source setting, our method outperforms the recent state-of-the-art TTP [26] for black-box targeted attack by 12.75% to 41.23%. When the substitute models are VGG19<sub>BN</sub>, DenseNet121, ResNet50,

our method averagely outperforms TTP [26] by 36.69%, 31.56%, 13.98% respectively. It is worth noting that our method shows impressive improvements when the black-box model has a large model discrepancy from the substitute model.

For example, when the substitute model is VGG19<sub>BN</sub> and the black-box model is ResNet152, our method can reach 68.41% target accuracy while the TTP method is 27.18%. A similar tendency can also be observed in the subset-source setting. These results clearly demonstrate the effectiveness of our methods for black-box targeted attacks.

### 4.2.2 Attacking Unknown Robust Models

To further evaluate the robustness of the adversarial examples, we follow TTP [26] to attack unknown robust models. Three kinds of robust models are tested under the all-source setting: 1) models trained with data augmentation methods to be robust to natural corruptions, such as Augmix [12]; 2) models trained on stylized ImageNet to have a higher shape bias so that they are inherently more robust to image style distortions, such as SIN and SIN-IN [8]; 3) adversarially trained models which are shown to be able to defend against adversarial examples to certain extent [25, 34]. We craft adversarial examples by using the ResNet50 architecture for the substitute models and then attack the ResNet50 models trained by different robust training mechanisms.

The results of all methods are shown in Table 2, where the numbers for baseline methods are from [26]. We observe that our results outperform previous generative attacks in all settings. Particularly, we can see a significant improvement of 17.0%, 35.19%, 21.73% when attacking

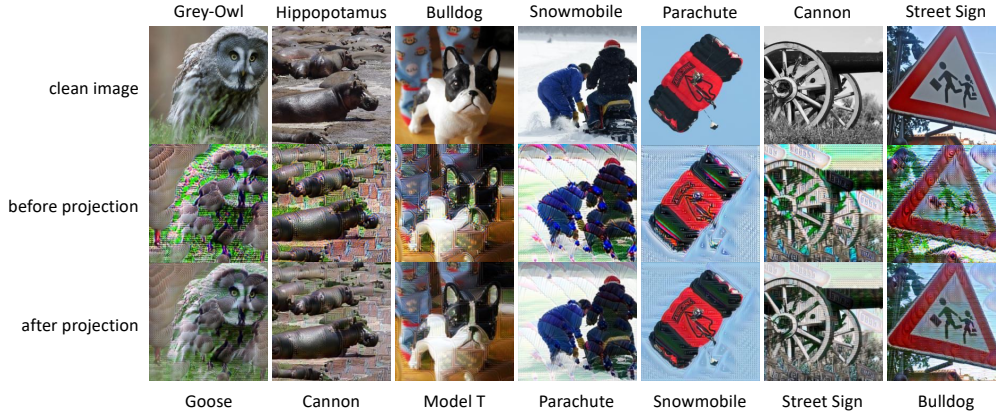


Figure 3. Targeted adversaries produced by a generator trained against ResNet50. The 1st row shows original images while the 2nd row shows unrestricted outputs of an adversarial generator, and the 3rd row shows adversaries after valid projection. The caption at the top represents the original category. The caption below represents the target category. Perturbation budget is set to  $l_\infty \leq 16$ .

Table 2. Target Transferability Against Unknown Robust Models: Top-1 (%) target accuracy in all-source setting. Generators are trained against naturally trained ResNet50. Perturbations are then transferred to ResNet50 trained using different methods including Augmix [12], Stylized [8] or adversarial [34].

$\epsilon$	Attack	Augmix [12]	Stylized [8]		Adversarial [34]			
			SIN-IN	SIN	$l_\infty$		$l_2$	
					$\epsilon=.5$	$\epsilon=1$	$\epsilon=.1$	$\epsilon=.5$
16	GAP [30]	51.57	76.92	12.96	1.88	0.34	23.41	0.92
	CDA [27]	59.79	75.93	9.21	2.10	0.39	23.89	1.18
	TTP [26]	73.09	87.40	30.17	4.63	0.56	45.40	1.99
	Ours	<b>90.09</b>	<b>95.57</b>	<b>65.36</b>	<b>7.82</b>	<b>0.85</b>	<b>67.13</b>	<b>4.10</b>
32	GAP [30]	54.86	81.15	28.07	26.32	6.36	59.04	16.53
	CDA [27]	63.18	76.81	19.65	27.60	6.74	57.54	16.07
	TTP [26]	78.66	91.27	41.52	46.82	16.35	75.97	30.94
	Ours	<b>94.26</b>	<b>97.53</b>	<b>81.71</b>	<b>65.20</b>	<b>22.84</b>	<b>92.19</b>	<b>44.48</b>

Augmix [12] trained model, stylized ImageNet [8] models, adversarially [25, 34] trained models, respectively. These results prove that our adversarial examples are robust not only to the variation of model architecture but also to popular defense mechanisms.

### 4.2.3 Attacking Real-world System

We also evaluate our methods in a more challenging task, *i.e.*, to fool a real-world computer vision system Google Cloud Vision. As for each image, the Google Cloud Vision API will return a list of the ten most likely classes. For a targeted attack, as long as the target class is in the returned list, we consider it a success. As the semantic labels predicted by the API don't correspond to the 1k ImageNet labels, we consider semantically consistent classes as the same class (The criteria of "*semantically consistent*" is shown in Table 3 with our results). As for evaluation, we randomly sample 100 adversarial examples crafted by TTP [26] and our M3D method, respectively. Specifically, the adversarial examples are generated on ResNet50. From

Table 3, we observe that even the real-world vision system Google Cloud Vision can be attacked successfully by our approach with a certain chance. This reveals the vulnerability of real-world systems. Our M3D approach outperforms TTP [26] on all classes by a significant margin.

Table 3. Target Transferability on Google Cloud Vision: Top-1(%) target accuracy averaged across 100 images. Perturbation budget:  $l_\infty \leq 16$ .

Target Class	Similar Class	TTP	Ours
Grey-Owl	Owl, Screech owl, great grey owl	0	24
Goose	Goose, geese	4	37
Bulldog	Dog, Companion dog	25	61
Parachute	Parachute	4	23
Street sign	Signage, Street sign	26	47

### 4.3. Additional Analysis

**Perturbation Robustness:** We experimentally find that the generated adversarial semantic pattern itself generalizes well among the different models. We illustrate and corroborate our claim by directly feeding scaled adversarial perturbations into different classifiers in the subset-source setting. The results in Fig. 4 and Table 4 reveal that the perturbations generated by our method can be considered as the target class and have a higher attack success rate compared with those by TTP [26].

Table 4. Target Transferability of Perturbations: Top-1 (%) target accuracy in the subset-source setting.

Src.	Attack	VGG19 <sub>BN</sub>	Dense121	ResNet50
VGG19 <sub>BN</sub>	TTP [26]	75.11	9.36	10.58
	Ours	<b>78.78</b>	<b>28.18</b>	<b>17.62</b>
Dense121	TTP [26]	19.44	58.58	16.20
	Ours	<b>42.53</b>	<b>69.98</b>	<b>31.56</b>
ResNet50	TTP [26]	25.02	33.96	40.69
	Ours	<b>45.62</b>	<b>53.73</b>	<b>54.56</b>

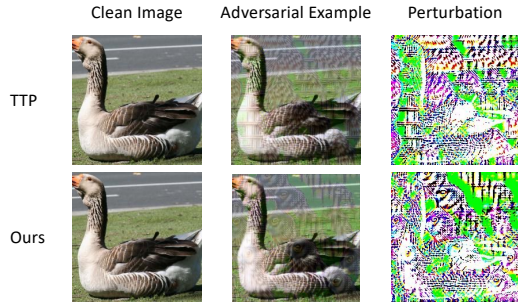


Figure 4. The perturbation comparison with TTP [26]. The 1st column displays the clean images, the 2nd column shows the clipped adversarial examples and the 3rd column displays the perturbations.

**Perturbation Effectiveness:** In order to understand the improvements of our method, we put a further analysis by taking the classification error rate into consideration. The classification error rate is defined as the proportion of adversarial examples that are not classified as their original classes by the black-box model, which can be used to measure the ability to move samples away from their original classes. However, this does not necessarily mean the targeted attack is successful since we expect the black-box model could classify these samples to the specified class. It can be seen from Fig. 5 that, though both TTP [26] and our method have a similar ability to mislead the target black-box model to predict wrong categories, our method outperforms TTP [26] by a significant margin at attack success. This confirms our analysis in the last paragraph of Section 3.3 that our M3D approach helps to draw the adversarial examples closer to the target class.

#### 4.4. Ablation Studies

To gain a better understanding of the effect of our M3D approach, we present an ablative study in Table 5. We consider three cases: A) using one fixed substitute model; B) using two trainable models without considering their model discrepancy; C) the full model of our M3D method. For all cases, the substitute models are initialized using the ResNet50 pre-trained on ImageNet, and we attack four normal models and one robust model. As shown in Table 5, compared to using one fixed substitute model, using two trainable models gains improvements due to the model ensemble. However, it is still inferior to the full model with a clear margin. This clearly demonstrates the effectiveness of minimizing the maximum model discrepancy (i.e., the second term in Eq. (3)) in our M3D approach.

## 5. Conclusion

In this paper, we study the black-box targeted attack problem from the model discrepancy perspective. We first

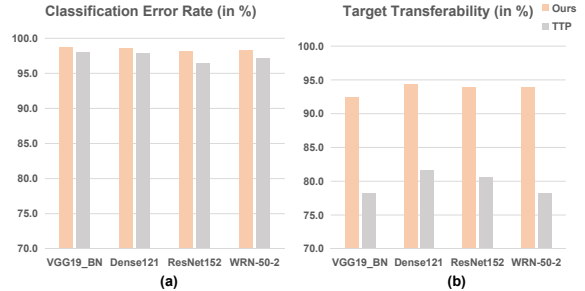


Figure 5. Classification Error Rate and Target Transferability in the all-source setting under the perturbation budget  $l_\infty \leq 16$ . Though having a similar classification error rate, our method outperforms TTP in target transferability.

Table 5. Ablation study in the all-source setting. 'A' indicates training the generator with one fixed substitute model. 'B' indicates training the generator with two trainable substitute models without considering model discrepancy. 'C' indicates training the generator with our method M3D. The substitute model is chosen as ResNet50.

	A	B	C
Number	1	2	2
Trainable	-	✓	✓
Discrepancy	-	-	✓
VGG19 <sub>BN</sub>	71.26	76.86	<b>92.41</b>
Dense121	81.73	85.07	<b>94.39</b>
ResNet152	78.36	79.51	<b>93.85</b>
WRN-50-2	78.88	80.19	<b>93.87</b>
SIN	24.18	31.48	<b>65.36</b>

present a generalization error bound based on the model discrepancy for the black-box targeted attack. Based on this bound, we design a novel approach called Minimizing Maximum Model Discrepancy (M3D) attack, which is capable of crafting highly transferable adversarial examples by minimizing the maximum model discrepancy between two substitute models. We show that our M3D approach can be easily implemented with deep neural networks, which are trained in the adversarial training manner between the generator and the substitute models. Extensive experiments demonstrate that our proposed M3D approach outperforms existing state-of-the-art methods with a significant margin for the black-box targeted attack using various models on the ImageNet dataset.

**Acknowledgement:** This work is supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (No. 62176047), Shenzhen Fundamental Research Program (No. JCYJ20220530164812027), and the Fundamental Research Funds for the Central Universities Grant (No. ZYGX2021YGLH208).



## References

- [1] David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75, 2021. [2](#)
- [2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. [2](#)
- [3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. [2](#), [3](#)
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017. [2](#)
- [5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [1](#), [2](#), [6](#)
- [6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. [1](#), [2](#)
- [7] Lianli Gao, Yaya Cheng, Qilong Zhang, Xing Xu, and Jingkuan Song. Feature space targeted attacks by statistic alignment. In *International Joint Conferences on Artificial Intelligence*, 2021. [2](#)
- [8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations*, 2019. [6](#), [7](#)
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ArXiv preprint arXiv:1412.6572*, 2014. [1](#), [2](#)
- [10] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *Advances in Neural Information Processing Systems*, 33:85–95, 2020. [2](#)
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2](#), [5](#)
- [12] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations*, 2020. [6](#), [7](#)
- [13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. [5](#)
- [14] Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *Proceedings of the International Conference on Learning Representations*, 2020. [1](#), [2](#), [5](#), [6](#)
- [15] Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *ArXiv preprint arXiv:2004.14861*, 2020. [1](#), [2](#), [6](#)
- [16] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7066–7074, 2019. [1](#), [2](#), [6](#)
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ArXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [18] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. [2](#)
- [19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. [2](#)
- [20] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 641–649, 2020. [1](#), [2](#)
- [21] Shuang Li, Fangrui Lv, Binhui Xie, Chi Harold Liu, Jian Liang, and Chen Qin. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. [2](#)
- [22] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *ArXiv preprint arXiv:1908.06281*, 2019. [2](#)
- [23] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ArXiv preprint arXiv:1611.02770*, 2016. [1](#)
- [24] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. [2](#)
- [25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018. [2](#), [6](#), [7](#)
- [26] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2021. [2](#), [5](#), [6](#), [7](#), [8](#)

- [27] Muzammal Naseer, Salman H Khan, Harris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 2019. 2, 5, 6, 7
- [28] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *ArXiv preprint arXiv:1605.07277*, 2016. 1
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5
- [30] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 2, 5, 6, 7
- [31] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. *Advances in domain adaptation theory*. Elsevier, 2019. 2
- [32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5
- [33] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2
- [34] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020. 6, 7
- [35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representation*, 2015. 2, 5
- [36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ArXiv preprint arXiv:1312.6199*, 2013. 1
- [37] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *ArXiv preprint arXiv:1704.03453*, 2017. 1
- [38] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, October 2021. 2
- [39] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *International Conference on Learning Representation*, 2020. 1, 2
- [40] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6
- [41] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In *Proceedings of the International Conference on Learning Representations*, 2022. 2
- [42] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34, 2021. 2
- [43] Yao Zhu, Jiacheng Sun, and Zhenguo Li. Rethinking adversarial transferability from a data distribution perspective. In *Proceedings of the International Conference on Learning Representations*, 2021. 2