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

Anqi Zhao[1] Tong Chu[1] Yahao Liu[1] Wen Li[2*] Jingjing Li[1] Lixin Duan[1]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
{zhaoanqiii, uestcchutong, lyhaolive, liwenbnu, lxduan}@gmail.com, lijin117@yeah.net

In this Supplementary, we first show the complete proof of the generalization error bound in Appendix A. Then, we present the implementation details of the generator and the discriminators in Appendix B. Visual illustrations of adversarial examples crafted by different substitute models are shown in Appendix C. The potential negative impact and the limitations of our work are discussed respectively in Appendix D and E.

A. Proof of the Generalization Error Bound

We give the proof of the generalization error bound for the black-box targeted adversarial attack task. First, we introduce Rademacher Generalization Bound [2] which measures the difference between generalization and empirical errors.

Lemma 1. *(Rademacher Generalization Bound [1, 2]) Suppose that $\mathcal{G}$ is a class of function maps $\mathcal{X} \to [0,1]$. Then for any $\delta > 0$, with probability at least $1-\delta$ and sample size n, the following holds for all $g \in \mathcal{G}$:*

$$|\mathbb{E}_D(g) - \mathbb{E}_{\widehat{D}}(g)| \leq 2\mathfrak{R}_{n,D}(\mathcal{G}) + \sqrt{\frac{log\frac{2}{\delta}}{2n}}, \tag{1}$$

where the $\mathbb{E}_D(g)$ is generalization risk of function g and $\mathbb{E}_{\widehat{D}}(g)$ is empirical risk of function g.

Definition 1. *Given a hypothesis space $\mathcal{H}$ and a Hypothesis Disparity Discrepancy function Γ, the $\mathcal{G}_\Gamma\mathcal{H}$ is defined as [1]:*

$$\mathcal{G}_\Gamma\mathcal{H} = \{\mathbf{x} \to \Gamma(h_1(\mathbf{x}), h_2(\mathbf{x}))|h_1, h_2 \in \mathcal{H}\}. \tag{2}$$

Then, we prove the generalization error bound for the black-box targeted attack task as follows,

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

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

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

Proof.

$$\begin{aligned}\mathcal{E}_\mathcal{Q}(h_b, f_t) &\leq \mathcal{E}_\mathcal{Q}(h_s, f_t) + \mathcal{E}_\mathcal{Q}(h_s, h_b) \\ &\leq \mathcal{E}_\mathcal{Q}(h_s, f_t) + \sup_{h,h' \in \mathcal{H}} \mathcal{E}_\mathcal{Q}(h, h').\end{aligned}$$

According to Lemma 1, for any $h_s \in \mathcal{H}$, we have:

$$\mathcal{E}_\mathcal{Q}(h_s, f_t) \leq \hat{\mathcal{E}}_\mathcal{Z}(h_s, f_t) + 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{H}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Then, considering Lemma 1 and Definition 1, for any $h, h' \in \mathcal{H}$, the following inequality holds [1]:

$$\mathcal{E}_\mathcal{Q}(h, h') \leq \hat{\mathcal{E}}_\mathcal{Z}(h, h') + 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{G}_\Gamma\mathcal{H}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Substituting these inequalities into the generalization bound, we get:

$$\mathcal{E}_\mathcal{Q}(h_b, f_t) \leq \hat{\mathcal{E}}_\mathcal{Z}(h_s, f_t) + \sup_{h,h' \in \mathcal{H}} \hat{\mathcal{E}}_\mathcal{Z}(h, h') + 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{H}) + 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{G}_\Gamma\mathcal{H}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

To simplify the expression, we define $\Omega = 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{H}) + 2\mathfrak{R}_{n,\mathcal{Q}}(\mathcal{G}_\Gamma\mathcal{H}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$. Then, we have:

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

Here we complete the proof. □

*The corresponding author

B. Implementation

B.1. The Implementation Details of Generator

For a fair comparison, we use the same network architecture as in previous generative attacks [3, 4, 6]. The generator is composed of three downsampling blocks, six residual blocks, and two upsampling blocks. The output of the generator is an adversarial sample with the same size as the input image. The details of the network are shown in Fig. 1.

B.2. The Implementation Details of Discriminators

The discriminators can be any classification models sampled from the whole hypothesis set. For a fair comparison, we follow the previous targeted attack works [3] to attack black-box models trained on ImageNet. The two discriminator models D_1 and D_2 in our method are derived from the same naturally-trained ImageNet model. Since the training data input to D_1 and D_2 are the same, the two models need to be initialized differently to ensure the model discrepancy loss works. We simply use a pre-trained model [5] for one discriminator, and a model fine-tuned for a few steps on ImageNet for another discriminator.

B.3. Data Augmentation

To further enhance the transferability of our adversarial examples, we apply the same data augmentation methods that TTP [3] uses during training, such as random rotation, random flipping, color jittering, and so on.

C. Visualization

In Fig. 2, 3, 4, 5, 6, we illustrate different perturbations produced by our M3D methods trained against ResNet50, DenseNet121 and VGG19$_{BN}$. The first column indicates the clean images and their original category label. The other columns represent different adversaries before and after clip operation which are crafted against different models and the target label.

D. Societal Impact

The threats of adversarial examples have raised great concerns in the deep learning community as they can be used maliciously in numerous security-sensitive applications, such as face recognition and autonomous driving. While the targeted black-box attacks, aiming at misleading the black-box models by outputting a highly dangerous specified class, can cause more harmful results. Understanding the strength and mechanism of adversarial attacks can reveal the vulnerability of real-world systems and motivate the community to design stronger defenses in the future. Though it is possible that our method can be misused maliciously, we believe that the help of our paper to researchers can outweigh the help of malicious attackers.

E. Limitation

Although our method outperforms existing state-of-art methods by a large margin, currently, our method needs to learn a generative model for each target class. Thus, efficiency is limited when facing an increasing number of targets, such as hundreds or thousands of classes. In the future, we plan to overcome this limitation by designing a conditional generative method by considering the target class as input.

References

[1] 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. 1

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 1

[3] 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

[4] 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] 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. 2

[6] 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

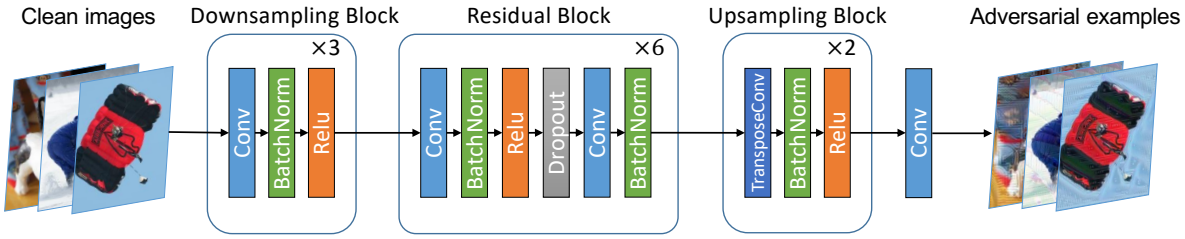

Figure 1. The network architecture of generator.

Figure 2. Targeted adversaries crafted by M3D. Generators are trained against ResNet50, Densenet121 and Vgg19$_{BN}$ to target 'Goose' distribution. The first row shows unrestricted outputs of an adversarial generator while the second row shows adversaries after clip operation. Perturbation budget is set to $l_\infty \leq 16$.

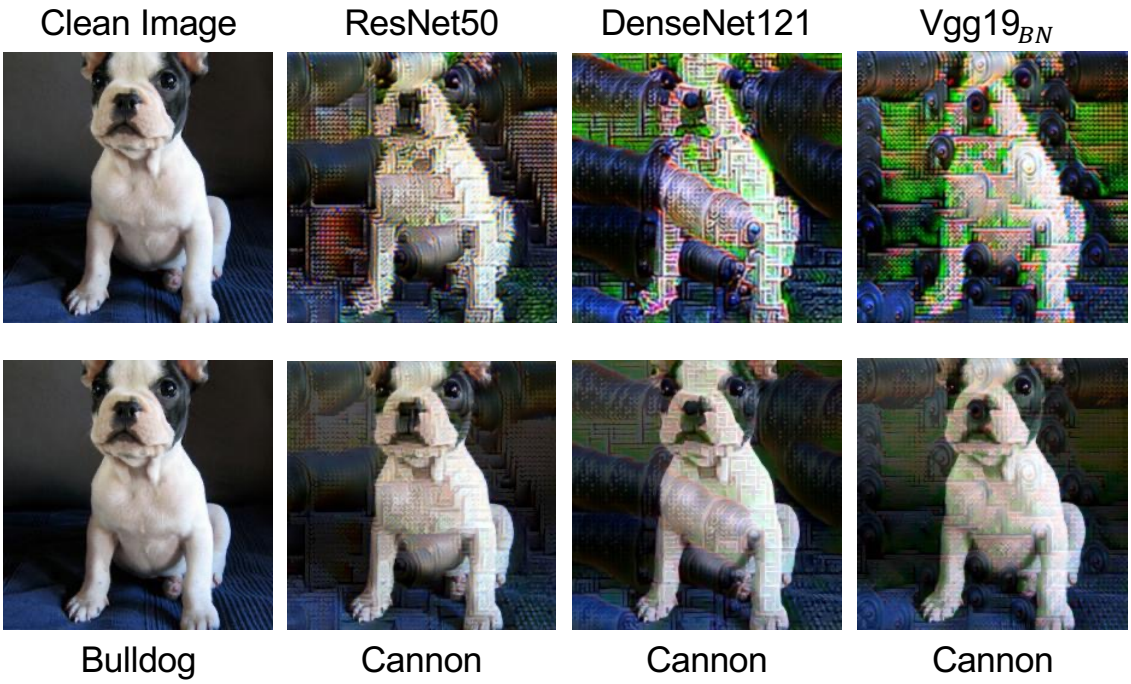

Figure 3. Targeted adversaries crafted by M3D. Generators are trained against ResNet50, Densenet121 and Vgg19$_{BN}$ to target 'Cannon' distribution. The first row shows unrestricted outputs of an adversarial generator while the second row shows adversaries after clip operation. Perturbation budget is set to $l_\infty \leq 16$.

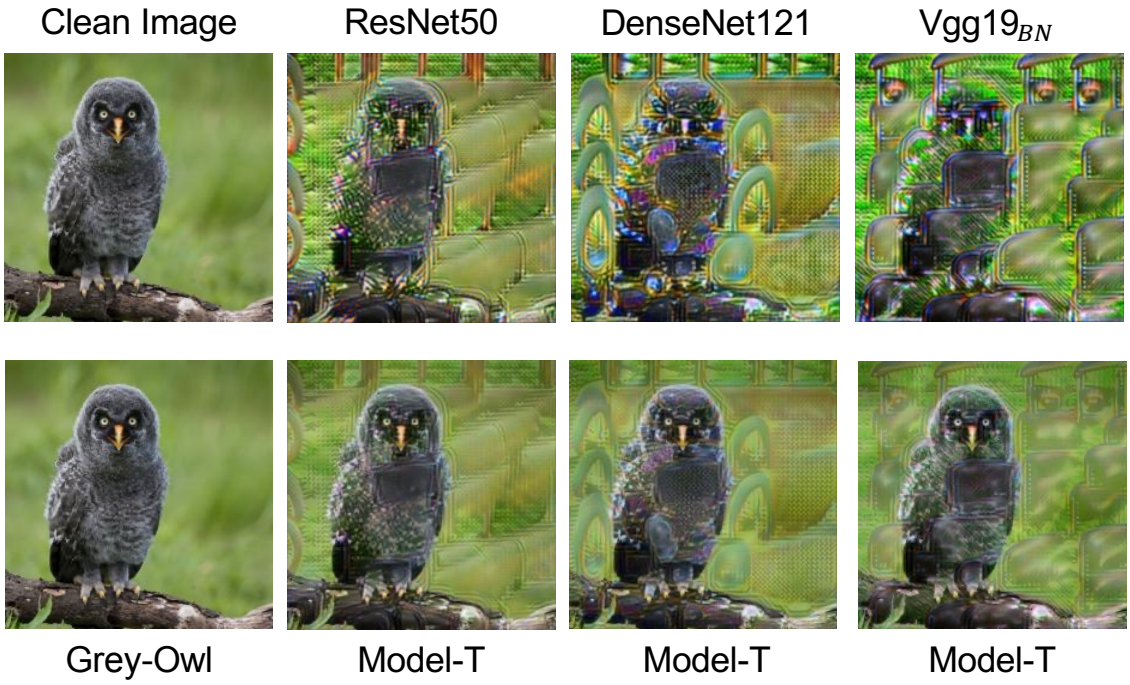

Figure 4. Targeted adversaries crafted by M3D. Generators are trained against ResNet50, Densenet121 and Vgg19$_{BN}$ to target 'Model-T' distribution. The first row shows unrestricted outputs of an adversarial generator while the second row shows adversaries after clip operation. Perturbation budget is set to $l_\infty \leq 16$.

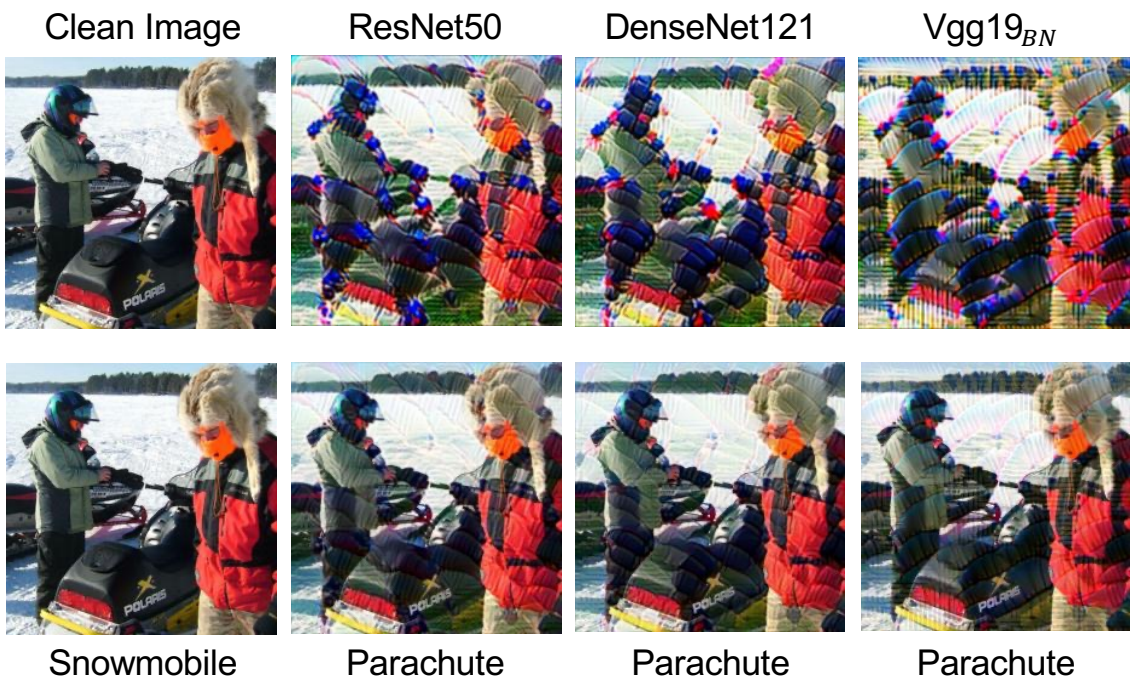

Figure 5. Targeted adversaries crafted by M3D. Generators are trained against ResNet50, Densenet121 and Vgg19$_{BN}$ to target 'Parachute' distribution. The first row shows unrestricted outputs of an adversarial generator while the second row shows adversaries after clip operation. Perturbation budget is set to $l_\infty \leq 16$.

Figure 6. Targeted adversaries crafted by M3D. Generators are trained against ResNet50, Densenet121 and Vgg19$_{BN}$ to target 'Street-Sign' distribution. The first row shows unrestricted outputs of an adversarial generator while the second row shows adversaries after clip operation. Perturbation budget is set to $l_\infty \leq 16$.