Triangle Attack: A Query-efficient Decision-based Adversarial Attack

Xiaosen Wang[1,2], Zeliang Zhang[1], Kangheng Tong[1], Dihong Gong[2], Kun He[1⋆], Zhifeng Li[2⋆], and Wei Liu[2⋆]

[1] School of Computer Science, Huazhong University of Science and Technology
[2] Data Platform, Tencent

Abstract. Decision-based attack poses a severe threat to real-world applications since it regards the target model as a black box and only accesses the hard prediction label. Great efforts have been made recently to decrease the number of queries; however, existing decision-based attacks still require thousands of queries in order to generate good quality adversarial examples. In this work, we find that a benign sample, the current and the next adversarial examples can naturally construct a triangle in a subspace for any iterative attacks. Based on the law of sines, we propose a novel Triangle Attack (TA) to optimize the perturbation by utilizing the geometric information that the longer side is always opposite the larger angle in any triangle. However, directly applying such information on the input image is ineffective because it cannot thoroughly explore the neighborhood of the input sample in the high dimensional space. To address this issue, TA optimizes the perturbation in the low frequency space for effective dimensionality reduction owing to the generality of such geometric property. Extensive evaluations on ImageNet dataset show that TA achieves a much higher attack success rate within 1,000 queries and needs a much less number of queries to achieve the same attack success rate under various perturbation budgets than existing decision-based attacks. With such high efficiency, we further validate the applicability of TA on real-world API, *i.e.*, Tencent Cloud API.

1 Introduction

Despite the unprecedented progress of Deep Neural Networks (DNNs) [27, 24, 25], the vulnerability to adversarial examples [47] poses serious threats to security-sensitive applications, *e.g.*, face recognition [42, 48, 20, 30, 56, 50, 15, 37, 63], autonomous driving [7, 19, 4, 62, 40], *etc*. To securely deploy DNNs in various real-world applications, it is necessary to conduct an in-depth analysis on the intrinsic properties of adversarial examples, which has inspired numerous researches on adversarial attacks [36, 6, 3, 11, 8, 17, 5, 52] and defenses [34, 23, 65, 57, 58, 53]. Existing attacks can be split into two categories: *white-box* attack has full knowledge of the target model (often leveraging the gradient) [21, 6, 34, 17] while *black-box* attack can only access the model output, which is more applicable in real-world

⋆ Corresponding authors. brooklet60@hust.edu.cn, michaelzfli@tencent.com, wl2223@columbia.edu.

scenarios. The black-box attack can be implemented in different ways. *Transfer-based* attack [32, 17, 60, 55] adopts the adversaries generated on the substitute model to fool the target model. *Score-based* attack [9, 26, 2, 31] assumes that the attacker can access the output logits while *decision-based* (*a.k.a.* hard label) attack [5, 11, 10, 29, 35] only has access to the prediction (top-1) label.

Among the black-box attacks, decision-based attack is more challenging and practical due to the minimum information requirement for attack. The number of queries on target model often plays a significant role in decision-based attack, since the access to a victim model is usually restricted in practice. Though recent works manage to reduce the total number of queries from millions to thousands of requests [5, 29, 38], it is still insufficient for most practical applications [35].

Existing decision-based attacks [5, 29, 38, 35] first generate a large adversarial perturbation and then minimize the perturbation while keeping adversarial property by various optimization methods. As shown in Fig. 1, we find that at the t-th iteration, the benign sample x, current adversarial example x_t^{adv}, and next adversarial example x_{t+1}^{adv} can naturally construct a triangle for any iterative attacks. According to the law of sines, the adversarial example x_{t+1}^{adv} at the $(t+1)$-th iteration should satisfy $\beta_t + 2\alpha_t > \pi$ to guarantee that the perturbation decreases, *i.e.*, $\delta_{t+1} = \|x_{t+1}^{adv} - x\|_p < \delta_t = \|x_t^{adv} - x\|_p$ (when $\beta_t + 2 \cdot \alpha_t = \pi$, it would be an isosceles triangle, *i.e.*, $\delta_{t+1} = \delta_t$).

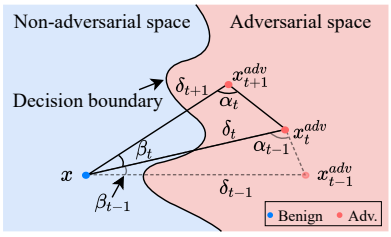


Fig. 1: Illustration of the candidate triangle at an arbitrary iteration of TA. At the t-th iteration, TA constructs a triangle with the learned angle α_t which satisfies $\beta_t + 2\alpha_t > \pi$ in the sampled subspace to find a new adversarial example x_{t+1}^{adv} and update α_t accordingly. Note that different from existing decision-based attacks [5, 38, 35], TA does not restrict x_t^{adv} on the decision boundary but minimizes the perturbation in the low frequency space using the geometric property; making TA itself query-efficient

Based on the above geometric property, we propose a novel and query-efficient decision-based attack, called Triangle Attack (TA). Specifically, at t-th iteration, we randomly select a directional line across the benign sample x to determine a 2-D subspace, in which we iteratively construct the triangle based on the current adversarial example x_t^{adv}, benign sample x, learned angle α_t, and searched angle β_t until the third vertex of the constructed triangle is adversarial. Using the geometric information, we can conduct TA in the low frequency space generated by Discrete Cosine Transform (DCT) [1] for effective dimensionality reduction to improve the efficiency. And we further update α_t to adapt to the perturbation optimization for each constructed triangle. Different from most existing decision-based attacks, there is no need to restrict x_t^{adv} on the decision boundary or estimate the gradient at each iteration, making TA query-efficient.

Our main contributions are summarized as follows:

– To our knowledge, it is the first work that directly optimizes the perturbation in frequency space via geometric information without restricting the adversary on decision boundary, leading to high query efficiency.

- Extensive evaluations on ImageNet dataset show that TA exhibits a much higher attack success rate within 1,000 queries and needs a much less number of queries to achieve the same attack success rate with the same perturbation budget on five models than existing SOTA attacks [11, 12, 8, 29, 38, 35].
- TA generates more adversarial examples with imperceptible perturbations on Tencent Cloud API, showing its industrial-grade applicability.

2 Related Work

Since Szegedy *et al.* [47] identified adversarial examples, massive adversarial attacks have been proposed to fool DNNs. White-box attacks, *e.g.*, single-step gradient-based attack [21], iterative gradient-based attack [36, 28, 34, 14], and optimization-based attack [47, 6, 3], often utilize the gradient and exhibit good attack performance. They have been widely adopted for evaluating the model robustness of defenses [34, 65, 41, 13, 16], but are hard to be applied in real-world with limited information. To make adversarial attacks applicable in practice, various black-box attacks, including transfer-based attack [17, 60, 51, 52, 59, 61], score-based attack [9, 26, 49, 2, 18, 64, 66], and decision-based attack [5, 12, 8, 38, 35], have gained increasing interest. Among them, decision-based attack is most challenging since it can only access the prediction label. In this work, we aim to boost the query efficiency of decision-based attack by utilizing the geometric information and provide a brief overview of existing decision-based attacks.

BoundaryAttack [5] is the first decision-based attack that initializes a large perturbation and performs random walks on the decision boundary while keeping adversarial. Such a paradigm has been widely adopted in the subsequent decision-based attacks. OPT [11] formulates the decision-based attack as a real-valued optimization problem with zero-order optimization. And SignOPT [12] further computes the sign of the directional derivative instead of the magnitude for fast convergence. HopSkipJumpAttack (HSJA) [8] boosts BoundaryAttack by estimating the gradient direction via binary information at the decision boundary. QEBA [29] enhances HSJA for better gradient estimation using the perturbation sampled from various subspaces, including spatial, frequency, and intrinsic components. To further improve the query efficiency, qFool [33] assumes that the curvature of the boundary is small around adversarial examples and adopts several perturbation vectors for efficient gradient estimation. BO [43] uses Bayesian optimization for finding adversarial perturbations in low dimension subspace and maps it back to the original input space to obtain the final perturbation. GeoDA [38] approximates the local decision boundary by a hyperplane and searches the closest point to the benign sample on the hyperplane as the adversary. Surfree [35] iteratively constructs a circle on the decision boundary and adopts binary search to find the intersection of the constructed circle and decision boundary as the adversary without any gradient estimation.

Most existing decision-based attacks restrict the adversarial example at each iteration on the decision boundary and usually adopt different gradient estimation approaches for attack. In this work, we propose Triangle Attack to minimize

the adversarial perturbation in the low frequency space directly by utilizing the law of sines without gradient estimation or restricting the adversarial example on the decision boundary for efficient decision-based attack.

3 Methodology

In this section, we first provide the preliminaries. Then we introduce our motivation and the proposed Triangle Attack (TA).

3.1 Preliminaries

Given a classifier f with parameters θ and a benign sample $x \in \mathcal{X}$ with ground-truth label $y \in \mathcal{Y}$, where $\mathcal{X}$ denotes all the images and $\mathcal{Y}$ is the output space. The adversarial attack finds an adversary $x^{adv} \in \mathcal{X}$ to mislead the target model:

$$f(x^{adv}; \theta) \neq f(x; \theta) = y \quad \text{s.t.} \quad \|x^{adv} - x\|_p < \epsilon,$$

where ϵ is the perturbation budget. Decision-based attacks usually first generate a large adversarial perturbation δ and then minimize the perturbation as follows:

$$\min \|\delta\|_p \quad \text{s.t.} \quad f(x + \delta; \theta) \neq f(x; \theta) = y. \tag{1}$$

Existing decision-based attacks [11, 12, 29] often estimate the gradient to minimize perturbation, which is time-consuming. Recently, some works adopt the geometric property to estimate the gradient or directly optimize the perturbation. Here we introduce two geometry-inspired decision-based attacks in details.

GeoDA [38] argues that the decision boundary at the vicinity of a data point x can be locally approximated by a hyperplane passing through a boundary point x_B close to x with a normal vector w. Thus, Eq. (1) can be locally linearized:

$$\min \|\delta\|_p \quad \text{s.t.} \quad w^\top (x + \delta) - w^\top x_B = 0.$$

Here x_B is a data point on the boundary, which can be found by binary search with several queries, and GeoDA randomly samples several data points for estimating w to optimize the perturbation at each iteration.

Surfree [35] assumes the boundary can be locally approximated by a hyperplane around a boundary point $x + \delta$. At each iteration, it represents the adversary using polar coordinates and searches optimal θ to update the perturbation:

$$\delta_{t+1} = \delta_t \cos\theta(\boldsymbol{u} \cos\theta + \boldsymbol{v} \sin\theta),$$

where $\boldsymbol{u}$ is the unit vector from x to x_t^{adv} and $\boldsymbol{v}$ is the orthogonal vector of $\boldsymbol{u}$.

3.2 Motivation

Different from most decision-based attacks with gradient estimation [11, 12, 29, 38] or random walk on the decision boundary [5, 35], we aim to optimize the

perturbation using the geometric property without any queries for gradient estimation. After generating a large adversarial perturbation, the decision-based attacks move the adversarial example close to the benign sample, *i.e.*, decrease the adversarial perturbation δ_t, while keeping the adversarial property at each iteration. In this work, as shown in Fig. 1, we find that at the t-th iteration, the benign sample x, current adversarial example x_t^{adv} and next adversarial example x_{t+1}^{adv} can naturally construct a triangle in a subspace for any iterative attacks. Thus, searching for the next adversarial example x_{t+1}^{adv} with smaller perturbation is equivalent to searching for a triangle based on x and x_t^{adv}, in which the third data point x' is adversarial and satisfies $\|x' - x\|_p < \|x_t^{adv} - x\|_p$. This inspires us to utilize the relationship between the angle and side length in the triangle to search an appropriate triangle to minimize the perturbation at each iteration. As shown in Sec. 4.4, however, directly applying such a geometric property on the input image leads to poor performance. Thanks to the generality of such a geometric property, we optimize the perturbation in the low frequency space generated by DCT [1] for effective dimensionality reduction, which exhibits great attack efficiency as shown in Sec. 4.4.

Moreover, since Brendel *et al.* [5] proposed BoundaryAttack, most decision-based attacks [11, 12, 8, 38, 35] follow the setting in which the adversarial example at each iteration should be on the decision boundary. We argue that such a restriction is not necessary in decision-based attacks but introduces too many queries on the target model to approach the boundary. Thus, we do not adopt this constraint in this work and validate this argument in Sec. 4.4.

3.3 Triangle Attack

In this work, we have the following assumption that the adversarial examples exist for any deep neural classifier f:

Assumption 1. *Given a benign sample x and a perturbation budget ϵ, there exists an adversarial perturbation $\|\delta\|_p \leq \epsilon$ towards the decision boundary which can mislead the target classifier f.*

This is a general assumption that we can find the adversarial example x^{adv} for the input sample x, which has been validated by numerous works [21, 6, 3, 5, 54]. If this assumption does not hold, the target model is ideally robust so that we cannot find any adversarial example within the perturbation budget, which is beyond our discussion. Thus, we follow the framework of existing decision-based attacks by first randomly crafting a large adversarial perturbation and then minimizing the perturbation. To align with previous works, we generate a random perturbation close to the decision boundary with binary search [29, 38, 35] and mainly focus on the perturbation optimization.

In two arbitrary consecutive iterations of the perturbation optimization process for any adversarial attacks, namely the t-th and $(t+1)$-th iterations without loss of generalization, the input sample x, current adversarial example x_t^{adv} and next adversarial example x_{t+1}^{adv} can naturally construct a triangle in a subspace of the input space $\mathcal{X}$. Thus, as shown in Fig. 1, decreasing the perturbation to

generate x_{t+1}^{adv} is equivalent to searching for an appropriate triangle in which the three vertices are x, x_t^{adv} and x_{t+1}^{adv}, respectively.

Theorem 1 (The law of sines). *Suppose a, b and c are the side lengths of a triangle, and α, β and γ are the opposite angles, we have $\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}$.*

From Theorem 1, we can obtain the relationship between the side length and opposite angle for the triangle in Fig. 1:

$$\frac{\delta_t}{\sin\alpha_t} = \frac{\delta_{t+1}}{\sin\left(\pi - (\alpha_t + \beta_t)\right)}. \tag{2}$$

To greedily decrease the perturbation δ_t, the t-th triangle should satisfy that $\frac{\delta_{t+1}}{\delta_t} = \frac{\sin(\pi-(\alpha_t+\beta_t))}{\sin\alpha_t} < 1$, *i.e.*, $\pi - (\alpha_t + \beta_t) < \alpha_t$. Thus, decreasing the perturbation at the t-th iteration can be achieved by finding a triangle constructed by the input sample x, current adversarial example x_t^{adv} and the angles β_t and α_t, which satisfy $\beta_t + 2\alpha_t > \pi$ and the third vertex should be adversarial. We denote such a triangle as candidate triangle and $\mathcal{T}(x, x_t^{adv}, \alpha_t, \beta_t, \mathcal{S}_t)$ as the third vertex, where $\mathcal{S}_t$ is a sampled subspace. Based on this observation, we propose a novel decision-based attack, called Triangle Attack (TA), that searches the candidate triangle at each iteration and adjusts angle α_t accordingly.

Sampling the 2-D subspace $\mathcal{S}$ of frequency space. The input image often lies in a high-dimensional space, such as $224\times224\times3$ for ImageNet [27], which is too large for the attack to explore the neighborhood for minimizing the adversarial perturbation efficiently. Previous works [22, 29, 35] have shown that utilizing the information in various subspaces can improve the efficiency of decision-based attacks. For instance, QEBA [29] samples the random noise for gradient estimation in the spatial transformed space or low frequency space but minimizes the perturbation in the input space with estimated gradient. Surfree [35] optimizes the perturbation in the subspace of the input space determined by a unit vector randomly sampled in the low frequency space. In general, the low frequency space contains the most critical information for images. With the poor performance of TA in the input space as shown in Sec. 4.4 and the generality of the geometric property shown in Fig. 2, we directly optimize the perturbation in the frequency space at each iteration for effective dimensionality reduction. And we randomly sample a d-dimensional line across the benign sample in the low frequency space (top 10%). The sampled line, directional line from benign sample x and current adversary x_t^{adv} can determine a unique 2-D subspace $\mathcal{S}$ of the frequency space, in which we can construct the candidate triangle to minimize the perturbation. The final adversary can be converted into the input space by Inverse DCT (IDCT).

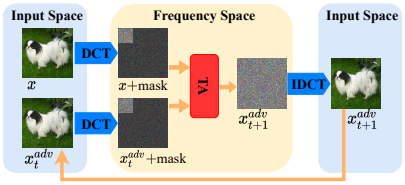


Fig. 2: Illustration of the entire procedure of TA attack at the t-th iteration. We construct the triangle in the frequency space to efficiently craft adversarial examples. Note that here we adopt DCT for illustration but we do not need it for x at each iteration. We still adopt x and x_t^{adv} in the frequency space without ambiguity due to the one-to-one mapping of DCT

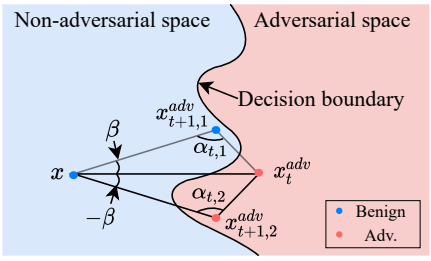


Fig. 3: Illustration of a symmetric candidate triangle (x, x_t^{adv} and $x_{t+1,2}^{adv}$). When the angle β cannot result in adversarial example ($x_{t+1,1}^{adv}$), we would further construct the symmetric triangle based on line $\langle x, x_t^{adv} \rangle$ to check data point $x_{t+1,2}^{adv}$

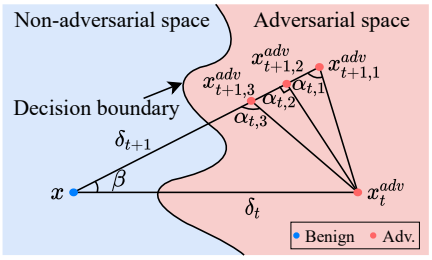


Fig. 4: The effect of magnitude on α for the candidate triangle used in TA. For the same sampled angle β, the larger angle α leads to smaller perturbation but is also more likely to cross over the decision boundary

Searching the candidate triangle. Given a subspace $\mathcal{S}_t$, the candidate triangle only depends on angle β since α is updated during the optimization. As shown in Fig. 3, if we search an angle β without leading to an adversarial example ($x_{t+1,1}^{adv}$), we can further construct a symmetric triangle with the same angle in the opposite direction to check data point $x_{t+1,2}^{adv}$, which has the same magnitude of perturbation as $x_{t+1,1}^{adv}$ but in different direction. We denote the angle as $-\beta$ for the symmetric triangle without ambiguity. Note that with the same angle α, a larger angle β would make the third vertex closer to the input sample x, *i.e.*, smaller perturbation. After determining the subspace $\mathcal{S}_t$, we first check angle $\beta_{t,0} = \max(\pi - 2\alpha, \underline{\beta})$, where $\underline{\beta} = \pi/16$ is a pre-defined small angle. If neither $\mathcal{T}(x, x_t^{adv}, \alpha_t, \beta_{t,0}, \mathcal{S}_t)$ nor $\mathcal{T}(x, x_t^{adv}, \alpha_t, -\beta_{t,0}, \mathcal{S}_t)$ is adversarial, we give up this subspace because it brings no benefit. Otherwise, we adopt binary search to find an optimal angle $\beta^* \in [\max(\pi - 2\alpha, \underline{\beta}), \min(\pi - \alpha, \pi/2)]$ which is as large as possible to minimize the perturbation. Here we restrict the upper bound of β because $\mathcal{T}(x, x_t^{adv}, \alpha_t, \beta, \mathcal{S}_t)$ would be at the opposite direction w.r.t. x for $\beta > \pi/2$ and $\pi - \alpha$ guarantees a valid triangle.

Adjusting angle α. Intuitively, angle α balances the magnitude of perturbation and the difficulty to find an adversarial example.

Proposition 1. *With the same angle β, a smaller angle α makes it easier to find an adversarial example while a larger angle α leads to smaller perturbation.*

Intuitively, as shown in Fig. 4, a smaller angle α results in larger perturbation but is more likely to cross over the decision boundary, making it easier to search an adversarial example, and vice versa. It is hard to consistently find an optimal α for each iteration, letting alone various input images and target models. Thus, we adaptively adjust angle α based on the crafted adversarial example:

$$\alpha_{t,i+1} = \begin{cases} \min(\alpha_{t,i} + \gamma, \pi/2 + \tau) & \text{if } f(x_{t,i+1}^{adv}; \theta) \neq y \\ \max(\alpha_{t,i} - \lambda\gamma, \pi/2 - \tau) & \text{Otherwise} \end{cases} \tag{3}$$

where $x_{t,i+1}^{adv} = \mathcal{T}(x, x_t^{adv}, \alpha_{t,i}, \beta_{t,i}, \mathcal{S}_t)$ is the adversarial example generated by $\alpha_{t,i}$, γ is the change rate, λ is a constant, and τ restricts the upper and lower

Algorithm 1: Triangle Attack

Input: Target classifier $\boldsymbol{f}$ with parameters $\boldsymbol{\theta}$; Benign sample $\boldsymbol{x}$ with ground-truth label $\boldsymbol{y}$; Maximum number of queries Q; Maximum number of iteration N for each sampled subspace; Dimension of the directional line d; Lower bound $\underline{\beta}$ for angle β.

Output: An adversarial example x^{adv}.

Initialize a large adversarial perturbation δ_0;
$x_0^{adv} = x + \delta_0$, $q = 0$, $t = 0$, $\alpha_0 = \pi/2$;
while $q < Q$ **do**
    Sampling 2-D subspace $\mathcal{S}_t$ in the low frequency space;
    $\beta_{t,0} = \max(\pi - 2\alpha, \underline{\beta})$;
    if $f(\mathcal{T}(x, x_t^{adv}, \alpha_{t,0}, \beta_{t,0}, \mathcal{S}_t); \theta) = f(x; \theta)$ **then**
        $q = q + 1$, update $\alpha_{t,0}$ based on Eq. (3);
        if $f(\mathcal{T}(x, x_t^{adv}, \alpha_{t,0}, -\beta_{t,0}, \mathcal{S}_t); \theta) = f(x; \theta)$ **then**
            $q = q + 1$, update $\alpha_{t,0}$ based on Eq. (3);
            Go to line 3; ▷ give up this subspace
    $\overline{\beta}_{t,0} = \min(\pi/2, \pi - \alpha)$;
    for $i = 0 \to N$ **do** ▷ binary search for angle β
        $\beta_{t,i+1} = (\overline{\beta}_{t,i} + \beta_{t,i})/2$;
        if $f(\mathcal{T}(x, x_t^{adv}, \alpha_{t,i}, \beta_{t,i+1}, \mathcal{S}_t); \theta) = f(x; \theta)$ **then**
            $q = q + 1$, update $\alpha_{t,i}$ based on Eq. (3);
            if $f(\mathcal{T}(x, x_t^{adv}, \alpha_{t,i}, -\beta_{t,i+1}, \mathcal{S}_t); \theta) = f(x; \theta)$ **then**
                $\overline{\beta}_{t,i+1} = \beta_{t,i+1}, \beta_{t,i+1} = \beta_{t,i}$;
        $q = q + 1$, update $\alpha_{t,i+1}$ based on Eq. (3);
    $x_{t+1}^{adv} = \mathcal{T}(x, x_t^{adv}, \alpha_{t,i+1}, \beta_{t,i+1}, \mathcal{S}_t)$, $t = t + 1$;
return x_t^{adv}.

bounds of α. We adopt $\lambda < 1$ to prevent decreasing the angle too fast considering much more failures than successes during the perturbation optimization. Note that a larger angle α makes it harder to find an adversarial example. However, a too small angle α results in a much lower bound for β, which also makes $\mathcal{T}(x, x_t^{adv}, \alpha_t, \beta_t, \mathcal{S}_t)$ far away from the current adversarial example x_t^{adv}, decreasing the probability to find an adversarial example. Thus, we add bounds for α to restrict it in an appropriate range.

TA iteratively searches the candidate triangle in subspace $\mathcal{S}_t$ sampled from the low frequency space to find the adversarial example and update angle α accordingly. The overall algorithm of TA is summarized in Algorithm 1.

4 Experiments

We conduct extensive evaluations on the standard ImageNet dataset using five models and Tencent Cloud API to evaluate the effectiveness of TA. Code is available at https://github.com/xiaosen-wang/TA.

Table 1: Attack success rate (%) on five models under different RMSE thresholds. The maximum number of queries is set to 1,000. We highlight the highest attack success rate in **bold**

Model	VGG-16			Inception-v3			ResNet-18			ResNet-101			DenseNet-121		
RMSE	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
OPT	76.0	38.5	5.5	34.0	17.0	4.0	67.0	36.0	6.0	51.5	21.0	5.0	51.5	29.0	5.5
SignOPT	94.0	57.5	12.5	50.5	27.0	8.0	84.5	49.5	13.0	69.0	33.0	8.0	69.5	44.0	10.0
HSJA	92.5	58.5	13.0	32.5	14.0	4.0	83.0	51.0	12.5	71.5	37.5	12.0	70.5	43.5	10.5
QEBA	98.5	86.0	29.0	78.5	54.5	17.0	98.0	81.5	34.5	94.0	59.0	20.5	91.0	66.0	24.0
BO	96.0	72.5	17.0	75.5	43.0	10.0	94.5	74.0	16.0	89.5	63.0	16.5	93.0	64.5	16.5
GeoDA	99.0	94.0	35.0	89.0	61.5	23.5	99.5	90.0	30.5	98.0	81.5	22.0	**100.0**	84.5	27.5
Surfree	99.5	92.5	39.5	87.5	67.5	24.5	98.5	87.0	36.0	95.5	76.5	27.0	97.0	78.0	29.0
TA (**Ours**)	**100.0**	**95.0**	**44.5**	**96.5**	**81.5**	**30.0**	**100.0**	**94.0**	**51.5**	**99.0**	**88.5**	**40.0**	99.5	**92.5**	**43.5**

4.1 Experimental Setup

Dataset. To validate the effectiveness of the proposed TA, following the setting of Surfree [39], we randomly sample 200 correctly classified images from the ILSVRC 2012 validation set for evaluation on the corresponding models.

Models. We consider five widely adopted models, *i.e.*, VGG-16 [44], Inception-v3 [45], ResNet-18 [24], ResNet-101 [24] and DenseNet-121 [25]. To validate the applicability in the real world, we evaluate TA on Tencent Cloud API[3].

Baselines. We take various decision-based attacks as our baselines, including four gradient estimation based attacks, *i.e.*, OPT [11], SignOPT [12], HSJA [8], QEBA [29], one optimization based attack, *i.e.*, BO [43], and two geometry-inspired attacks, *i.e.*, GeoDA [38], Surfree [35].

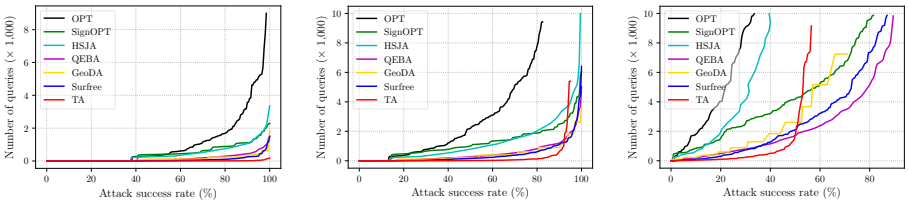
Evaluation metrics. Following the standard setting in QEBA [29], we adopt the root mean squared error ($RMSE$) between benign sample x and adversarial example x^{adv} to measure the magnitude of perturbation:

$$d(x, x^{adv}) = \sqrt{\frac{1}{w \cdot h \cdot c} \sum_{i=1}^{w} \sum_{j=1}^{h} \sum_{k=1}^{c} (x[i,j,k] - x^{adv}[i,j,k])^2}, \tag{4}$$

where w, h, c are the width, height and number of channels of the input image, respectively. We also adopt the attack success rate, the percentage of adversarial examples which reach a certain distance threshold.

Hyper-parameters. For fair comparison, all the attacks adopt the same adversarial perturbation initialization approach as in [35] and the hyper-parameters for baselines are exactly the same as in the original papers. For our TA, we adopt the maximum number of iterations in each subspace $N = 2$, the dimension of directional line $d = 3$ and $\gamma = 0.01$, $\lambda = 0.05$ and $\tau = 0.1$ for updating angle α.

[3] https://cloud.tencent.com/



(a) RMSE=0.1 (b) RMSE=0.05 (c) RMSE=0.01

Fig. 5: Number of queries to achieve the given attack success rate on ResNet-18 for the attack baselines and the proposed TA under various perturbation budgets. The maximum number of queries is 10,000

4.2 Evaluation on Standard Models

To evaluate the effectiveness of TA, we first compare the attack performance on five popular models with different decision-based attacks and report the attack success rate under various $RMSE$ thresholds, namely 0.1, 0.05 and 0.001.

We first evaluate the attack within 1,000 queries, which is widely adopted in recent works [8, 38, 35]. The attack success rate is summarized in Table 1, which means the attack would fail to generate adversarial example for the input image if it takes 1,000 queries without reaching the given threshold. We can observe that TA consistently achieves much higher attack success rate than existing decision-based attacks under various perturbation budgets on five models with different architectures. For instance, TA outperforms the runner-up attack with a clear margin of 1.0%, 7.5% and 13.0% under the $RMSE$ threshold of 0.1, 0.05, 0.01 on ResNet-101, which is widely adopted for evaluating the decision-based attacks. In particular, the proposed TA significantly outperforms the two geometry-inspired attacks, *i.e.*, GeoDA [38] and Surfree [35], which exhibit the best attack performance among the baselines. This convincingly validates the high effectiveness of the proposed TA. Besides, among the five models, Inception-v3 [46], which is rarely investigated in decision-based attacks, exhibits better robustness than other models under various perturbation budgets against both baselines and TA. Thus, it is necessary to thoroughly evaluate the decision-based attacks on various architectures instead of only ResNet models.

To further verify the high efficiency of TA, we investigate the number of queries to achieve various attack success rates under the $RMSE$ threshold of 0.1, 0.05 and 0.01, respectively. The maximum number of queries is set to 10,000 and the results on ResNet-18 are summarized in Fig. 5. As shown in Fig. 5a and 5b, TA needs much less number of queries to achieve various attack success rates with $RMSE$ threshold of 0.1 and 0.05, showing the high query efficiency of our method. For the smaller threshold of 0.01, as shown in Fig. 5c, our TA still needs less number of queries when achieving the attack success rate smaller than 50% but fails to achieve the attack success rate higher than 60%. Note that as shown in Fig. 6 and Table 1, $RMSE$ threshold of 0.01 is very rigorous so that the perturbation is imperceptible but is also hard to generate the adversarial examples for decision-based attacks. Since we mainly focus on the query

Table 2: The number of adversarial examples successfully generated by various attack baselines and the proposed TA on Tencent Cloud API within 200/500/1,000 queries. The results are evaluated on 20 randomly sampled images from the correctly classified images in ImageNet due to the high cost of online APIs

RMSE	OPT	SignOPT	HSJA	QEBA	GeoDA	Surfree	TA (**Ours**)
0.1	4/6/6	8/8/9	7/8/8	12/12/12	15/15/15	13/13/13	**17/17/17**
0.05	1/3/3	4/4/7	6/6/8	11/11/12	13/14/14	12/12/13	**15/17/17**
0.01	1/1/2	1/1/3	2/5/6	3/8/9	3/7/12	5/8/10	**8/12/13**

efficiency of attack only based on geometric information, the attack performance under the $RMSE$ threshold of 0.01 is acceptable because it is impractical for such high number of queries when attacking real-world applications.

Besides, since TA aims to improve the query efficiency by utilizing the triangle geometry, the global optima might be worse than existing gradient estimation based attacks when more queries are allowed. Other geometry-inspired methods also perform poorer than QEBA [29] in this case without gradient estimation. However, it is not the goal of TA and can be easily solved using gradient estimation. With the high efficiency of TA, we can achieve higher attack performance with lower number of queries by taking the TA as warm-up for the precise gradient estimation attacks, such as QEBA [29], if the high number of queries is acceptable. We integrate the gradient estimation used in QEBA [29] into TA after 2,000 queries, dubbed **TAG**. For the perturbation budget of 0.01, TAG achieves the attack success rate of 95% using 7,000 queries, which is better than the best baseline with the attack success rate of 92% using 9,000 queries.

4.3 Evaluation on Real-world Applications

With the superior performance and unprecedented progress of DNNs, numerous companies have deployed DNNs for a variety of tasks and also provide commercial APIs (Application Programming Interfaces) for different tasks. Developers can pay for these services to integrate the APIs into their applications. However, the vulnerability of DNNs to adversarial examples, especially the prosperity of decision-based attack which does not need any information of target models, poses severe threats to these real-world applications. With the high efficiency of TA, we also validate its practical attack applicability using Tencent Cloud API. Due to the high cost of commercial APIs, we randomly sample 20 images from ImageNet validation set and the maximum number of queries is 1,000.

The numbers of successfully attacked images are summarized in Table 2. We can observe that TA successfully generates more adversarial examples than the attack baselines within 200, 500 and 1,000 queries under various $RMSE$ thresholds. In particular, TA can generate even more adversarial examples within 500 queries than the best attack baselines within 1,000 queries, showing the superiority of TA. We also visualize some adversarial examples generated by TA in Fig. 6. As we can see, TA can successfully generate high quality adversarial examples for various classes with few queries (≤ 200), validating the high ap-

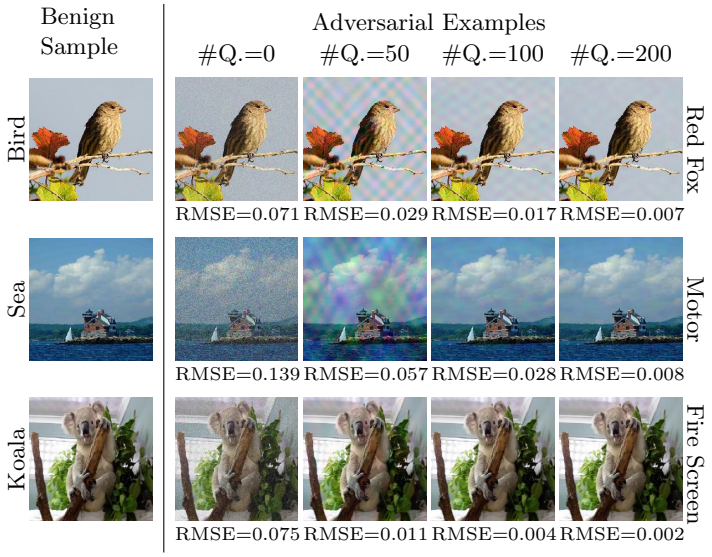


Fig. 6: The adversarial examples crafted by TA against Tencent Cloud API. #Q. denotes the number of queries for attack and RMSE denotes the RMSE distance between the benign sample and adversarial example. We report the correct label and the predicted label on the leftmost and rightmost columns, respectively (Zoom in for details.)

plicability of TA in real-world. Especially when the number of queries is 200, the adversarial examples generated by TA are almost visually imperceptible for humans, highlighting the vulnerability of current commercial applications.

4.4 Ablation Study

In this section, we conduct a series of ablation studies on ResNet-18, namely the subspace chosen by TA, the ratio for low frequency subspace and the change rate γ and λ for updating angle α. The parameter studies on the dimension of sampled line d and the bound τ for α are summarized in Appendix B.

On the subspace chosen by TA. Different from existing decision-based attacks, the generality of geometric property used by TA makes it possible to directly optimize the perturbation in the frequency space. To investigate the effectiveness of frequency space, we implement TA in various spaces, namely input space (TA_I), sampling the directional line in the frequency space but optimizing the perturbation in the input space (TA_FI) used by Surfree [35] and full frequency space (TA_F). As shown in Table 3, due to the high-dimensional input space, TA_I cannot effectively explore the neighborhood of the input sample

Table 3: Ablation study on ResNet-18 for different spaces, *i.e.* input space (TA_I), frequency space for line sampling but input space for perturbation optimization (TA_FI), and full frequency space without mask (TA_F)

RMSE	TA_I	TA_FI	TA_F	TA
0.1	39.5	97.5	98.5	**100.0**
0.05	17.5	73.0	85.0	**94.0**
0.01	3.0	22.5	25.5	**51.5**

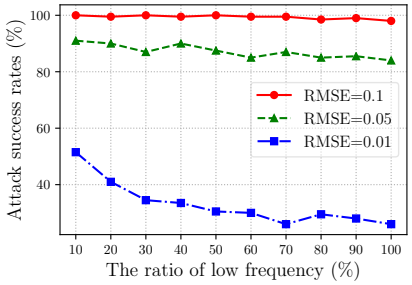


Fig. 7: Attack success rate (%) of TA on ResNet-18 within 1,000 queries with various ratios for the low frequency subspace under three $RMSE$ thresholds

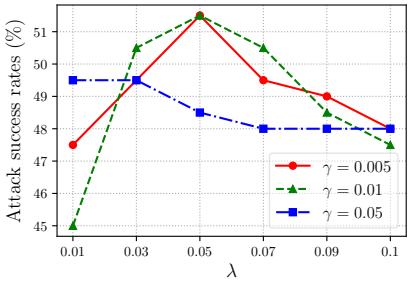


Fig. 8: Attack success rate (%) of TA on ResNet-18 within 1,000 queries with various γ and λ used for updating α under $RMSE = 0.01$

to find good perturbation and shows very poor performance. With the information from frequency space to sample the subspace, TA_{FI} exhibits much better results than TA_{I}. When optimizing the perturbation in the full frequency space, TA_{F} can achieve higher attack success rate than TA_{FI}, showing the benefit of frequency space. When sampling the subspace using the low frequency information, TA achieves much better performance than all the other attacks, supporting the necessity and rationality of the subspace chosen by TA.

On the ratio for low frequency subspace. The low frequency domain plays key role in improving the efficiency of TA. However, there is no criterion to identify the low frequency since it corresponds to high frequency, which is usually determined by the lower part of the frequency domain with a given ratio. Here we investigate the effect of this ratio on the attack performance of TA. As shown in Fig. 7, the ratio has more significant influence on the attack success rate under the smaller $RMSE$ threshold. In general, increasing the ratio roughly decreases the attack performance because it makes TA focus more on the high frequency domain, containing less critical information of the image. Thus, we adopt the lower 10% parts as the low frequency subspace for high efficiency, which also helps TA effectively reduce the dimension, making it easier for attack.

On the change rate γ and λ for updating angle α. As stated in Sec. 3.3, the angle α plays a key role in choosing a better candidate triangle but it is hard to find a uniformly optimal α for different iterations and input images. We assume that the larger angle α makes it harder to find a candidate triangle but leads to smaller perturbation. As in Eq. (3), if we successfully find a triangle, we would increase α with γ. Otherwise, we would decrease α with $\lambda\gamma$. We investigate the impact of various γ and λ in Fig. 8. Here we only report the results for $RMSE = 0.01$ for clarity and the results for $RMSE = 0.1/0.05$ exhibit the same trend. In general, $\gamma = 0.01$ leads to better attack performance than $\gamma = 0.05/0.005$. When we increase λ with $\gamma = 0.01$, the attack success rate increases until $\lambda = 0.05$ and then decreases. We also investigate the impact of τ which controls the bound for α in Eq. (3), which shows stable performance within 1,000 queries but takes effect for 10,000 queries and we simply adopt $\tau = 0.1$. In our experiments, we adopt $\gamma = 0.01$, $\lambda = 0.05$ and $\tau = 0.1$.

4.5 Further Discussion

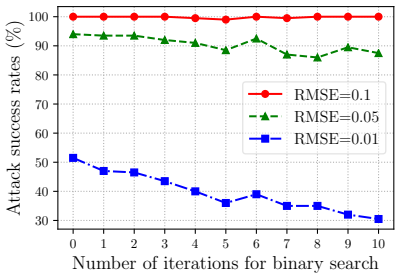


Fig. 9: Attack success rate (%) of TA using various number of iterations for binary search (N_{bs}) to restrict the adversary on the decision boundary at each iteration

BoundaryAttack [5] adopts random walk on the decision boundary to minimize the perturbation for decision-based attack and the subsequent works often follow this setting to restrict the adversarial example on the decision boundary. We argue that such a restriction is not necessary and do not adopt it in our TA. To validate this argument, we also conduct binary search to move the adversarial example towards the decision boundary at each iteration after we find the candidate triangle to investigate the benefit of this restriction. As illustrated in Fig. 9, when the number of iterations for binary search (N_{bs}) is 0, it is vanilla TA that exhibits the best attack success rate. When we increase N_{bs}, the binary search takes more queries in each iteration which degrades the total number of iterations under the given total number of queries. In general, the attack success rate stably decreases when increasing N_{bs} especially for $RMSE = 0.01$, which means the cost (*i.e.*, queries) for binary search to restrict the adversarial example on the decision boundary is not worthy. Such restriction might not be reliable and rational either for most decision-based attacks, especially for geometry-inspired attacks. We hope this would inspire more attention to discuss the necessity of restricting the adversarial examples on the decision boundary and shed new light on the design of more powerful decision-based attacks.

5 Conclusion

In this work, we found that the benign sample, the current and next adversarial examples can naturally construct a triangle in a subspace at each iteration for any iterative attacks. Based on this observation, we proposed a novel decision-based attack, called Triangle Attack (TA), which utilizes the geometric information that the longer side is opposite the larger angle in any triangle. Specifically, at each iteration, TA randomly samples a directional line across the benign sample to determine a subspace, in which TA iteratively searches a candidate triangle to minimize the adversarial perturbation. With the generality of geometric property, TA directly optimizes the adversarial perturbation in the low frequency space generated by DCT with much lower dimensions than the input space, and significantly improves the query efficiency. Extensive experiments demonstrate that TA achieves a much higher attack success rate within 1,000 queries and needs much less queries to achieve the same attack success rate. The practical applicability on Tencent Cloud API also validates the superiority of TA.

Acknowledgement This work is supported by National Natural Science Foundation (62076105), Intertational Coorperation Foundation of Hubei Province, China(2021EHB011) and Tencent Rhino Bird Elite Talent Training Program.

References

1. Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 1974.
2. Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.
3. Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.
4. Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, and Beat Flepp et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
5. Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.
6. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.
7. Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *International Conference on Computer Vision*, pages 2722–2730, 2015.
8. Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy*, 2020.
9. 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 *ACM Workshop on Artificial Intelligence and Security*, 2017.
10. Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *European Conference on Computer Vision*, 2020.
11. 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 *International Conference on Learning Representations*, 2019.
12. Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020.
13. Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.
14. Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.
15. Zhongying Deng, Xiaojiang Peng, Zhifeng Li, and Yu Qiao. Mutual component convolutional neural networks for heterogeneous face recognition. *IEEE Transactions on Image Processing*, 28(6):3102–3114, 2019.
16. Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Conference on Computer Vision and Pattern Recognition*, 2020.
17. Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Conference on Computer Vision and Pattern Recognition*, 2018.

18. Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020.
19. Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Conference on Computer Vision and Pattern Recognition*, 2018.
20. Dihong Gong, Zhifeng Li, Jianzhuang Liu, and Yu Qiao. Multi-feature canonical correlation analysis for face photo-sketch image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 617–620, 2013.
21. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
22. Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *Uncertainty in Artificial Intelligence*, 2019.
23. Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.
24. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.
25. Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.
26. Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, 2018.
27. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
28. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations, Workshop*, 2017.
29. Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: query-efficient boundary-based blackbox attack. In *Conference on Computer Vision and Pattern Recognition*, 2020.
30. Zhifeng Li, Dihong Gong, Yu Qiao, and Dacheng Tao. Common feature discriminant analysis for matching infrared face images to optical face images. *IEEE Transactions on Image Processing*, 23(6):2436–2445, 2014.
31. Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. *arXiv preprint arXiv:2201.08970*, 2022.
32. Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.
33. Yujia Liu, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. A geometry-inspired decision-based attack. In *International Conference on Computer Vision*, pages 4890–4898, 2019.
34. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

35. Thibault Maho, Teddy Furon, and Erwan Le Merrer. SurFree: A fast surrogate-free black-box attack. In *Conference on Computer Vision and Pattern Recognition*, 2021.
36. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2016.
37. Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
38. Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: a geometric framework for black-box adversarial attacks. In *Computer Vision and Pattern Recognition*, pages 8446–8455, 2020.
39. 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. In *International booktitle of computer vision*, 2015.
40. Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.
41. Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, 2019.
42. Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM SIGSAC conference on Computer and Communications Security*, 2016.
43. Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1461–1469, 2021.
44. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
45. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition*, 2016.
46. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition*, 2016.
47. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.
48. Xiaoou Tang and Zhifeng Li. Video based face recognition using multiple classifiers. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 345–349. IEEE, 2004.
49. Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI Conference on Artificial Intelligence*, 2019.
50. Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition*, 2018.

51. Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Conference on Computer Vision and Pattern Recognition*, 2021.
52. Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *International Conference on Computer Vision*, 2021.
53. Xiaosen Wang, Hao Jin, Yichen Yang, and Kun He. Natural language adversarial defense through synonym encoding. *Conference on Uncertainty in Artificial Intelligence*, 2021.
54. Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. In *British Machine Vision Conference*, 2021.
55. Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.
56. Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.
57. Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.
58. Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021.
59. Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Conference on Computer Vision and Pattern Recognition*, 2021.
60. Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Conference on Computer Vision and Pattern Recognition*, 2019.
61. Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
62. Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Conference on Computer Vision and Pattern Recognition*, pages 2174–2182, 2017.
63. Xiaolong Yang, Xiaohong Jia, Dihong Gong, Dong-Ming Yan, Zhifeng Li, and Wei Liu. Larnet: Lie algebra residual network for face recognition. In *International Conference on Machine Learning*, pages 11738–11750. PMLR, 2021.
64. Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, and Michael W. Mahoney. Trust region based adversarial attack on neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2019.
65. Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.
66. Pu Zhao, Pin-Yu Chen, Siyue Wang, and Xue Lin. Towards query-efficient black-box adversary with zeroth-order natural gradient descent. In *AAAI Conference on Artificial Intelligence*, 2020.