Boosting Decision-Based Black-Box Adversarial Attack with Gradient Priors

Han Liu[1], **Xingshuo Huang**[1], **Xiaotong Zhang**[1*], **Qimai Li**[2], **Fenglong Ma**[3], **Wei Wang**[4], **Hongyang Chen**[5], **Hong Yu**[1] and **Xianchao Zhang**[1*]
[1]Dalian University of Technology, Dalian, China
[2]The Hong Kong Polytechnic University, Hong Kong, China
[3]The Pennsylvania State University, Pennsylvania, USA
[4]Shenzhen MSU-BIT University, Shenzhen, China
[5]Zhejiang Lab, Hangzhou, China
liu.han.dut@gmail.com, {xshuang.dut, zxt.dut}@hotmail.com, csqmli@comp.polyu.edu.hk, fenglong@psu.edu, {ehomewang, dr.h.chen}@ieee.org, {hongyu, xczhang}@dlut.edu.cn

Abstract

Decision-based methods have shown to be effective in black-box adversarial attacks, as they can obtain satisfactory performance and only require to access the final model prediction. Gradient estimation is a critical step in black-box adversarial attacks, as it will directly affect the query efficiency. Recent works have attempted to utilize gradient priors to facilitate score-based methods to obtain better results. However, these gradient priors still suffer from the edge gradient discrepancy issue and the successive iteration gradient direction issue, thus are difficult to simply extend to decision-based methods. In this paper, we propose a novel Decision-based Black-box Attack framework with Gradient Priors (DBA-GP), which seamlessly integrates the data-dependent gradient prior and time-dependent prior into the gradient estimation procedure. First, by leveraging the joint bilateral filter to deal with each random perturbation, DBA-GP can guarantee that the generated perturbations in edge locations are hardly smoothed, i.e., alleviating the edge gradient discrepancy, thus remaining the characteristics of the original image as much as possible. Second, by utilizing a new gradient updating strategy to automatically adjust the successive iteration gradient direction, DBA-GP can accelerate the convergence speed, thus improving the query efficiency. Extensive experiments have demonstrated that the proposed method outperforms other strong baselines significantly.

1 Introduction

Deep neural networks have achieved great success on various of tasks, such as image classification [He *et al.*, 2016; Pham *et al.*, 2021], object detection [Wang *et al.*, 2021; Wang *et al.*, 2022] and speech recognition [Chiu *et al.*, 2018; Park *et al.*, 2019]. However, recent researches demonstrate that neural networks are significantly vulnerable to adversarial examples, which are almost indistinguishable from natural data in human perception and yet classified incorrectly by the models [Goodfellow *et al.*, 2015; Cao *et al.*, 2021; Zhong *et al.*, 2022]. This phenomenon probably causes a large risk in many real-world applications, such as spam detection [Wu *et al.*, 2017], automatic drive [Luo *et al.*, 2021; Muhammad *et al.*, 2021] and economic services [Cintas *et al.*, 2020]. Investigating the generation rationale behind adversarial examples seems a promising way to improve the robustness of neural networks, which motivates the research of adversarial attacks. Based on the accessibility level of victim models, adversarial attacks can be categorized into ***white-box attacks*** and ***black-box attacks***. For white-box attacks [Moosavi-Dezfooli *et al.*, 2016; Carlini and Wagner, 2017], the attackers are assumed to have full knowledge about the target model, including training data, model architecture and parameters. Therefore, it is easy to utilize gradient information to lead these methods to generate adversarial examples. However, these attack methods are overly idealistic and even impracticable in real application scenarios, as most model developers are impossible to release all the model and data information in public. For black-box attacks, the attackers only have extremely limited knowledge about the target model, e.g., the predicted labels or confidence scores, so this kind of adversarial attacks seems more promising and practical.

Existing black-box attacks mainly contain transfer-based methods, score-based methods and decision-based methods. Transfer-based methods [Guo *et al.*, 2020; Wu *et al.*, 2020; Qin *et al.*, 2022] aim to train a surrogate model to imitate the behaviors of the target model and then conduct the white-box attacks on it. This kind of attack needs a huge number of training data that are similar to the data used for training the target model, which is difficult to achieve in practice. Score-based attacks [Guo *et al.*, 2019; Li and Chen, 2021; Li *et al.*, 2020b] require that the target models provide the predicted scores, which is also impractical in real-world applications since they may only offer the predicted labels. Compared with transfer-based and score-based approaches, decision-based methods [Chen *et al.*, 2020; Li *et al.*, 2020a] can only use discrete predicted labels to attack the target

*Corresponding author.

model, thus seem more realistic and feasible. However, existing decision-based attack models are not perfect since they usually rely on a large number of queries to generate adversarial examples.

Gradient estimation is the key point in decision-based methods, as it consumes the majority of all the queries. Recently, [Ilyas *et al.*, 2019] attempt to integrate two types of gradient priors, i.e., data-dependent prior and time-dependent prior, into score-based methods to facilitate them to obtain better performance. Nevertheless, these two priors are still difficult to simply extend to decision-based methods, as they suffer from the following drawbacks. (1) The ***data-dependent prior*** follows a strong assumption, that is, if two pixels are spatially close to each other, then their estimated gradients may have similar directions. This prior only takes spatial information into consideration but ignores the importance of the values of pixels. In fact, only the pixels have similar values and are spatially close, their estimated gradients may be similar. The sharp change of pixel values usually appears on the edge of objects. Therefore, to estimate the gradients accurately, it is essential to address the edge gradient discrepancy problem. (2) The ***time-dependent prior*** assumes that the gradients of successive steps are highly correlated and tend to be similar, which is suitable for score-based methods. This is because that in the iterative procedure of score-based methods, the distances between successive adversarial samples keep small, so the gradient direction of successive steps will also be similar. However, in the iterative procedure of decision-based methods, the distances between successive adversarial samples will be relatively large in the beginning, but become small subsequently. This indicates that the gradient direction of successive steps should have a similar tendency. In addition, when the similarity between estimated gradients at current and previous iterations is very large, it means that decision-based methods have fully explored in the estimated gradient direction, so a new gradient direction is needed to accelerate the convergence speed. Based on the above analysis, we need to design a crafty strategy to adjust the successive iteration gradient direction, thus boosting the query efficiency.

In this paper, we propose a novel Decision-based Black-box Attack framework with Gradient Priors (DBA-GP). To tackle the edge gradient discrepancy problem, we propose to leverage the data-dependent prior via the joint bilateral filter, which can not only smooth similar gradients for spatially close pixels with similar values, but also diversify gradients for pixels with different values. To deal with the successive iteration gradient direction problem, we simultaneously consider the distance between successive adversarial samples and the gradient direction of successive steps as additional judgment conditions, thus can generate a more appropriate gradient direction to improve the query efficiency. In summary, our contributions are as follows:

- We propose a new decision-based black-box adversarial attack framework with two simple yet effective gradient priors, thus can generate high-quality adversarial examples efficiently.
- We discover two fundamental drawbacks of existing gradient priors, i.e., the edge gradient discrepancy issue and the successive iteration gradient direction issue. To overcome these limitations, we utilize the joint bilateral filter and two specially-designed gradient updating judgement conditions, and integrate them into decision-based attack models seamlessly.
- We conduct extensive experiments against both offline and online models to validate the superiority of the proposed method compared with other strong baselines.

2 Related Work

Decision-based methods only require discrete predicted labels to attack the target model, thus seem more feasible and promising in real-world applications. [Brendel *et al.*, 2018] propose the first decision-based attack method (boundary attack), which starts with a large perturbation and then performs a random walk on the decision boundary to reduce the distance to the target image, but the use of the standard normal distribution affects the efficiency of the attack. Biased boundary attack [Brunner *et al.*, 2019] uses some biases that can significantly reduce the number of queries. SIGN-OPT [Cheng *et al.*, 2020] utilizes the gradient sign estimation to improve the query efficiency. EA [Dong *et al.*, 2019] designs an evolutionary algorithm to carry out the attack. HSJA [Chen *et al.*, 2020] utilizes the binary information on the decision boundary to estimate the gradient direction, which provides a fundamental and powerful framework for decision-based methods. QEBA [Li *et al.*, 2020a] employs three subspace optimization methods that can reduce the number of queries and further improve the performance. PSBA [Zhang *et al.*, 2021] further improves the query efficiency via progressive scaling techniques. However, it requires to train the GAN model with more additional data. SurFree [Maho *et al.*, 2021] attempts to move along diverse directions guided by the geometrical properties of the decision boundary. AHA [Li *et al.*, 2021] utilizes historical query information to improve the random walk optimization. Although decision-based methods have shown to be effective in adversarial attacks, they are complicated and still require a large number of queries.

3 Problem Formulation

Given an input image $\boldsymbol{x} \in [0,1]^{dim}$, considering an m-class image classification model $F : \mathbb{R}^{dim} \rightarrow \mathbb{R}^m$, we can get the prediction result by $y = \text{argmax}_i \left[F\left(\boldsymbol{x}\right)\right]_i$, where $\left[F(\boldsymbol{x})\right]_i$ represents the probability score belonging to the i-th class, and $i \in \{1, 2, ..., m\}$. Given an image $\boldsymbol{x}^*$ with the true label y^*, ***the targeted attack*** aims to find an adversarial image $\boldsymbol{x}_{adv}$ such that the model outputs a pre-specified class y_{adv} under the constraint that $d\left(\boldsymbol{x}^*, \boldsymbol{x}_{adv}\right)$ is minimum, where $d(\cdot)$ is a distance measure function like l_0, l_2 or l_∞ norm. Formally,

$$\min_{\boldsymbol{x}_{adv}} d\left(\boldsymbol{x}^*, \boldsymbol{x}_{adv}\right), \quad s.t., \quad \phi_{\boldsymbol{x}^*}(\boldsymbol{x}_{adv}) = 1, \tag{1}$$

where $\phi_{\boldsymbol{x}^*}(\boldsymbol{x}) : [0,1]^{dim} \longrightarrow \{-1, 1\}$ is a sign function defined as:

$$\phi_{\boldsymbol{x}^*}(\boldsymbol{x}) = \text{sign}\left(S_{\boldsymbol{x}^*}\right) = \begin{cases} 1 & \text{if } S_{\boldsymbol{x}^*}(\boldsymbol{x}) > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

Here $S_{\boldsymbol{x}^*} : \mathbb{R}^{dim} \to \mathbb{R}$ is a real-valued function defined as:

$$S_{\boldsymbol{x}^*}(\boldsymbol{x}) = [F(\boldsymbol{x})]_{y_{adv}} - \max_{y \neq y_{adv}} [F(\boldsymbol{x})]_y. \tag{3}$$

From Eq. (3), it is easy to observe that $\boldsymbol{x}$ is adversarial if and only if $S_{\boldsymbol{x}^*}(\boldsymbol{x}) > 0$. When $S_{\boldsymbol{x}^*}(\boldsymbol{x}) = 0, \boldsymbol{x}$ is exactly on the decision boundary. Note that in the decision-based black-box attack scenario, we can only get the value of function $\phi_{\boldsymbol{x}^*}(\boldsymbol{x})$. For ease of representation, hereinafter we represent the functions $S_{\boldsymbol{x}^*}(\boldsymbol{x})$ and $\phi_{\boldsymbol{x}^*}(\boldsymbol{x})$ as $S(\boldsymbol{x})$ and $\phi(\boldsymbol{x})$ respectively.

In contrast, ***the untargeted attack*** aims to find an adversarial image with any incorrect category. It is worth noting that by simply treating all classes different from y^* as the class y_{others}, we can convert an untargeted attack to a targeted attack, hence only considering the targeted attack is enough.

In terms of decision-based black-box adversarial attacks, the basic idea is to first select an initial adversarial image $\boldsymbol{x}_{init}$ with predicted category label y_{adv}, then move $\boldsymbol{x}_{init}$ towards $\boldsymbol{x}^*$ as close as possible and keep y_{adv} unchanged simultaneously. In this paper, we focus on this type of method and attempt to improve this procedure with two simple yet effective gradient priors.

4 Decision-Based Black-Box Attack with Gradient Priors

4.1 The Overall Framework

The DBA-GP method utilizes a similar framework with the powerful decision-based boundary attack method HSJA [Chen *et al.*, 2020]. It adopts a sampling-based gradient estimation component to guide the search direction. Specifically, it first selects $\boldsymbol{x}_{init}$ as an initial adversarial image, and then performs an iterative algorithm consisting of the following three parts: (1) Estimating the gradient direction of the current adversarial image; (2) Moving the current adversarial image along the direction of the estimated gradient; (3) Approaching the decision boundary via a binary search strategy.

Estimating the Gradient

Denote by $\boldsymbol{x}_{adv}^{(t)}$ the adversarial image on the decision boundary in the t-th iteration, the direction of the gradient $\nabla S_{\boldsymbol{x}^*}\left(\boldsymbol{x}_{adv}^{(t)}\right)$ can be approximated via the Monte Carlo estimation [Mooney, 1997]:

$$\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right) = \frac{1}{B}\sum_{b=1}^{B} \phi\left(\boldsymbol{x}_{adv}^{(t)} + \delta_t \mathbf{u}_b\right)\mathbf{u}_b, \tag{4}$$

where $\{\mathbf{u}_b\}_{b=1}^{B}$ are d-dimensional random perturbations with the unit length, and δ_t is a small positive parameter. As [Chen *et al.*, 2020] state, this estimate is accurate only if $\boldsymbol{x}_{adv}^{(t)}$ is on the decision boundary.

Moving Along the Gradient Direction

When obtaining the estimated gradient, DBA-GP moves $\boldsymbol{x}_{adv}^{(t)}$ along the gradient direction with the following formula:

$$\widetilde{\boldsymbol{x}}_{adv}^{(t)} = \boldsymbol{x}_{adv}^{(t)} + \xi_t \cdot \frac{\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)}{\left\|\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)\right\|_2}, \tag{5}$$

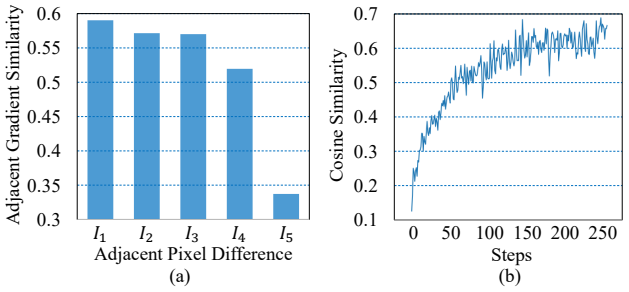

Figure 1: (a) The ratio of similar adjacent pixel gradients in various adjacent pixel difference intervals. $I_1, I_2, \cdots, I_5$ represent the intervals $[0, 0.2]$, $(0.2, 0.4]$, $\cdots$, $(0.8, 1.0]$ respectively. (b) The average cosine similarity of the gradients between the current and previous steps along the optimization trajectory of HSJA.

where ξ_t is the step size at the t-th step, and Eq. (5) should satisfy the constraint that $\phi_{\boldsymbol{x}^*}\left(\widetilde{\boldsymbol{x}}_{adv}^{(t)}\right) = 1$.

Approaching the Boundary

After moving along the gradient direction, $\widetilde{\boldsymbol{x}}_{adv}^{(t)}$ may be far from the decision boundary. To ensure that $\boldsymbol{x}_{adv}^{(t+1)}$ still approaches the decision boundary, DBA-GP pulls $\widetilde{\boldsymbol{x}}_{adv}^{(t)}$ towards the target image $\boldsymbol{x}^*$ by:

$$\boldsymbol{x}_{adv}^{(t+1)} = \alpha_t \cdot \boldsymbol{x}^* + (1 - \alpha_t) \cdot \widetilde{\boldsymbol{x}}_{adv}^{(t)}, \tag{6}$$

where the coefficient $\alpha_t \in [0, 1]$, which can be determined by a binary search method.

4.2 Gradient Estimation with Priors

In the overall framework shown in Section 4.1, estimating the gradient plays the most crucial role as its accuracy will directly affect the query efficiency of the method. Obviously, increasing the value of B is a straightforward way to improve the quality of gradient estimation. However, in adversarial attack scenarios, as the algorithm is iterative and limited by a fixed query budget, an excessively large B is unreasonable and impracticable.

Previous studies [Ilyas *et al.*, 2019; Li *et al.*, 2020a] attempt to use data-dependent and time-dependent gradient priors to obtain more accurate gradient estimation results, which have achieved promising performance. However, they still suffer from ***the edge gradient discrepancy issue*** and ***the successive iteration gradient direction issue***. In the following, we will dissect the reasons behind the above issues and propose the corresponding solutions.

Gradient Estimation with the Data-Dependent Prior

The spatially local similarity (i.e, pixels that are close together tend to be similar) is a well-known prior in the image domain. Inspired by this fact, the data-dependent gradient prior means that the gradients of adjacent pixels tend to be similar. Specifically, if two coordinates (i, j) and (k, l) in $\nabla S_{\boldsymbol{x}^*}(\boldsymbol{x})$ are close, then we have $\nabla S_{\boldsymbol{x}^*}(\boldsymbol{x})_{i,j} \approx \nabla S_{\boldsymbol{x}^*}(\boldsymbol{x})_{k,l}$.

Existing works [Ilyas *et al.*, 2019; Li *et al.*, 2020a] attempt to utilize the average-pooling, bilinear interpolation or inverse discrete cosine transform techniques to ameliorate the accuracy of gradient estimation, and have shown to be effective in

black-box adversarial attacks. However, they still suffer from the edge gradient discrepancy problem. Specifically, considering the image edge locations, it is easy to discover a phenomenon that although two coordinates are close, their pixel values are usually different, then their gradients also tend to be different.

To confirm our observation, we randomly select 50 images from ImageNet [Deng *et al.*, 2009]. For each image, we first normalize all pixels into $[0, 1]$, calculate the absolute values of adjacent pixel differences, and assign them into five intervals $[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$ and $(0.8, 1.0]$. Then we calculate the ratio of similar adjacent pixel gradients averaged over all images. Here we treat the gradient values with the same positive and negative signs to be similar. Figure 1 (a) shows the statistic results about the ratio of similar adjacent pixel gradients in various adjacent pixel difference intervals. It can be seen that with the increase of the adjacent pixel difference, the ratio of similar adjacent pixel gradients tends to be small. That is to say, when the adjacent pixel values are close, the corresponding adjacent pixel gradients are inclined to be close. On the contrary, when the adjacent pixel values are large, the corresponding adjacent pixel gradients will be disparate. According to the above analysis, the data-dependent prior should be redefined as: the gradients of adjacent pixels tend to be similar if their pixel values are similar.

The joint bilateral filter [Petschnigg *et al.*, 2004] is a nonlinear, edge-preserving, and noise-reducing smoothing filter, which has the following advantages. If nearby pixels are similar, it can replace the intensity of each pixel with a weighted average of intensity values from nearby pixels. If nearby pixels are diverse, it will not smooth these nearby pixels. To better leverage the data-dependent prior, we propose to use joint bilateral filter to deal with each random perturbation $\mathbf{u}_b$ with $\boldsymbol{x}^*$ as the guide image. Formally,

$$\widetilde{\mathbf{u}}_b = J\left(\mathbf{u}_b, \boldsymbol{x}^*\right), \tag{7}$$

where $\widetilde{\mathbf{u}}_b$ is the filtered perturbation which can be used to replace $\mathbf{u}_b$. $\boldsymbol{x}^*$ is the original target image, which can provide abundant image characteristic information. J is the joint bilateral filter. In the following, we give the concrete computation formula of J.

Given an image A ($n \times n$ size) and a guided image G ($n \times n$ size), after passing through the joint bilateral filter J, the filter output of $A_{i,j}$ can be calculated by:

$$J(A, G)_{i,j} = \frac{1}{w(i,j)} \sum_{(k,l)\in\Omega} g_s(i,j,k,l) g_r(G_{i,j}, G_{k,l}) A_{k,l}, \tag{8}$$

where Ω represents a neighborhood of pixel coordinates (i, j). $w(i, j)$ is a normalization term which can be obtained by:

$$w(i,j) = \sum_{(k,l)\in\Omega} g_s(i,j,k,l) g_r(G_{i,j}, G_{k,l}). \tag{9}$$

The functions $g_s(i, j, k, l)$ and $g_r(G_{i,j}, G_{k,l})$ are computed by:

$$g_s(i,j,k,l) = exp\left(-\frac{(i-k)^2 + (j-l)^2}{2\sigma_s^2}\right), \tag{10}$$

$$g_r(G_{i,j}, G_{k,l}) = exp\left(-\frac{\left(G_{i,j} - G_{k,l}\right)^2}{2\sigma_r^2}\right), \tag{11}$$

where σ_s and σ_r are parameters which can be used to adjust the spatial similarity and the range (intensity/color) similarity respectively.

Gradient Estimation with the Time-Dependent Prior

The time-dependent prior means that the gradients of successive steps are heavily correlated and tend to be highly similar, which is also called the multi-step prior. [Ilyas *et al.*, 2019] attempt to extend the time-dependent prior to score-based attack methods and have achieved impressive performance. However, it is difficult to be directly applied to decision-based attack methods. The reasons are as follows. For score-based attack methods, they start from the original image and then move along the gradient direction until an adversarial image is found. Thus, the distances between successive adversarial samples in the whole iterative procedure will keep small, and the gradient direction of successive steps will also be similar. However, for decision-based attack methods, they first select an adversarial image (outside the decision boundary) which is far away from the original image, and then gradually reduce its distance from the original image. Therefore, the distances between successive adversarial samples in the iterative procedure will be relatively large in the beginning, but become small later on. In the same manner, the gradient direction of successive steps should keep a similar tendency.

To validate our hypothesis, we randomly sample 50 target images from ImageNet and calculate the average cosine similarity of the gradients between successive steps based on the decision-based boundary attack method HSJA [Chen *et al.*, 2020]. Figure 1 (b) shows the average cosine similarity of the gradients between the current and previous steps along the optimization trajectory of HSJA. It can be seen that in the first 50 steps the gradients between successive steps are not very similar, but after that they become closer and closer.

Based on the above analysis, we attempt to use the following formula to estimate the gradient:

$$\begin{aligned}\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right) = \frac{1}{B}\sum_{b=1}^{B} \phi\left(\boldsymbol{x}_{adv}^{(t)} + \delta_t \mathbf{u}_b\right)\mathbf{u}_b \\ + \frac{1}{m}\sum_{x_{adv}^{(j)}\in\mathcal{X}^{(t)}} \widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right),\end{aligned} \tag{12}$$

where $\mathcal{X}^{(t)} = \{\boldsymbol{x}_{adv}^{(j)} : \max(1, t-k) \le j \le t-1, D_{t,j} < \tau\}$ is the set of intermediate-process generated adversarial images that satisfy some conditions. m is the number of images in $\mathcal{X}^{(t)}$. $D_{t,j} = d\left(\boldsymbol{x}_{adv}^{(t)}, \boldsymbol{x}_{adv}^{(j)}\right)$ is the distance between $\boldsymbol{x}_{adv}^{(t)}$ to $\boldsymbol{x}_{adv}^{(j)}$. k represents that there are k iterations before the current iteration step. τ is a threshold parameter.

For Eq. (12), we first use $\mathcal{X}^{(t)}$ to filter out the intermediate-process adversarial images with large distances, and then utilize the remaining ones to facilitate the current gradient estimation. Obviously, it will increase the accuracy of the gradient estimation significantly. However, it cannot lead to an

Algorithm 1 Gradient Estimation with Priors

Input: An image $\boldsymbol{x}^*$, the number of random sampling B, the joint bilateral filter J, the time-dependent length k, the decision function ϕ, the perturbation size δ_t, the distance function d, the adversarial image $\boldsymbol{x}_{adv}^{(t)}$ in the t-th iteration, the adversarial image set $\{\boldsymbol{x}_{adv}^{(j)} : \max(1, t-k) \leq j \leq t-1\}$, the gradient estimation result set $\{\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right) : \max(1, t-k) \leq j \leq t-1\}$, the threshold parameters τ and ρ.

Output: The estimated gradient $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$.

Sample B random perturbations $\{\mathbf{u}_b\}_{b=1}^B$
Deal with perturbations using the joint bilateral filter: $\{\widetilde{\mathbf{u}}_b\}_{b=1}^B = \left\{J(\mathbf{u}_b, \boldsymbol{x}_{adv}^{(t)})\right\}_{b=1}^B$
Estimate the gradient with the following formula: $\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right) = \frac{1}{B}\sum_{b=1}^B \phi\left(\boldsymbol{x}_{adv}^{(t)} + \delta_t \widetilde{\mathbf{u}}_b\right)\widetilde{\mathbf{u}}_b$
$\nabla S_p^{(t)} = \mathbf{0}$
for j in $[\max(1, t-k), t-1]$ **do**
  $D_{t,j} = d\left(\boldsymbol{x}_{adv}^{(t)}, \boldsymbol{x}_{adv}^{(j)}\right)$
  $S_{t,j} = \frac{\left\langle \widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right), \widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right)\right\rangle}{\left\|\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)\right\|_2 \left\|\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right)\right\|_2}$
  if $D_{t,j} < \tau$ and $S_{t,j} > \rho$ **then**
    $\nabla S_p^{(t)} = \nabla S_p^{(t)} + \widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right)$
  end if
end for
if $\nabla S_p^{(t)} \neq \mathbf{0}$ **then**
  $\overline{\nabla S}_p^{(t)} = \frac{\nabla S_p^{(t)}}{\left\|\nabla S_p^{(t)}\right\|_2}$
else
  $\overline{\nabla S}_p^{(t)} = \mathbf{0}$
end if
$\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right) = \frac{2\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)}{\left\|\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)\right\|_2} - \overline{\nabla S}_p^{(t)}$
return $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$

improvement in query efficiency. This is because in each iteration, decision-based attack methods will move a step along the estimated gradient direction as large as possible, which means that decision-based attack methods have fully explored in that gradient direction. Therefore, if the similarity between estimated gradients at current and previous iterations is very large, we need to generate a new gradient direction to speed up the algorithm convergence, thus improving the query efficiency.

To achieve the above goal, we define the set $\mathcal{A}^{(t)} = \{\boldsymbol{x}_{adv}^{(j)} : \max(1, t-k) \leq j \leq t-1, D_{t,j} < \tau, S_{t,j} > \rho\}$, which contains intermediate-process generated adversarial images that satisfy some conditions. $D_{t,j} = d\left(\boldsymbol{x}_{adv}^{(t)}, \boldsymbol{x}_{adv}^{(j)}\right)$ is the distance between $\boldsymbol{x}_{adv}^{(t)}$ to $\boldsymbol{x}_{adv}^{(j)}$. $S_{t,j}$ is the cosine similarity between $\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$ and $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right)$, where $\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$ is the estimated gradient in the t-th iteration with Eq. (4), and $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right)$ is the final estimated gradient in the j-th iteration. τ, ρ and k are hyperparameters. Based on the above definitions, in the t-th iteration we can first estimate the gradient $\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$ with Eq. (4), and then obtain the final gradient $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$ with:

$$\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right) = \frac{2\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)}{\left\|\widetilde{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)\right\|_2} - \overline{\nabla S}_p^{(t)}, \tag{13}$$

where $\overline{\nabla S}_p^{(t)}$ is the average estimated gradient of adversarial images in $\mathcal{A}^{(t)}$, and it can be calculated by:

(1) If $\mathcal{A}^{(t)}$ is not an empty set, we first compute $\nabla S_p^{(t)}$ by:

$$\nabla S_p^{(t)} = \sum_{x_{adv}^{(j)} \in \mathcal{A}^{(t)}} \widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(j)}, \delta_j\right), \tag{14}$$

and then compute $\overline{\nabla S}_p^{(t)}$ by:

$$\overline{\nabla S}_p^{(t)} = \frac{\nabla S_p^{(t)}}{\left\|\nabla S_p^{(t)}\right\|_2}. \tag{15}$$

(2) If $\mathcal{A}^{(t)}$ is an empty set, we simply set $\overline{\nabla S}_p^{(t)} = \mathbf{0}$. By using $\widehat{\nabla S}\left(\boldsymbol{x}_{adv}^{(t)}, \delta_t\right)$ as the estimate of the gradient direction, we can better leverage the time-dependent prior to improve the query efficiency of decision-based attack methods.

Algorithm 1 summarizes the details about how to integrate data-dependent and time-dependent priors into the gradient estimation procedure.

5 Experiments

5.1 Datasets

For offline experiments, we first conduct preliminary experiments on a simple dataset MNIST [LeCun, 1998]. Then we make a comprehensive evaluation on ImageNet [Deng *et al.*, 2009] and Celeba [Liu *et al.*, 2015] datasets. For different datasets, we exactly follow [Li *et al.*, 2020a] to randomly select 50 pairs of correctly classified images from the validation set of each dataset as the target images and the initial adversarial images. For online experiments, we attack the commercial face recognition API Face++[1].

5.2 Victim Models

For MNIST, we train a neural network consisting of two convolutional layers and two fully connected layers as the victim model. For ImageNet, we choose two well-known pre-trained models ResNet50 [He *et al.*, 2016] and VGG16 [Simonyan and Zisserman, 2015] as the victim models. For Celeba, we utilize samples from 100 people to fine-tune the pre-trained models ResNet50 and VGG16 on ImageNet, and take the fine-tuned models as the victim models.

[1]https://www.faceplusplus.com/face-comparing/.

5.3 Baselines

We compare the proposed method with the state-of-the-art targeted decision-based black-box attacks: EA [Dong *et al.*, 2019], SIGN-OPT [Cheng *et al.*, 2020], HSJA [Chen *et al.*, 2020], QEBA [Li *et al.*, 2020a], SURFREE [Maho *et al.*, 2021], and AHA [Li *et al.*, 2021].

5.4 Implementation Details

Evaluation metrics. The first evaluation metric is the average mean squared error (**MSE**) between the generated adversarial image and the target image as the number of queries increases. A smaller MSE means that the adversarial image is closer to the target image and also indicates that the attack quality is better. Under the same query budget, the lower the achieved MSE, the higher the query efficiency of the attack. The second evaluation metric is the attack success rate (**ASR**) of reaching a specified MSE threshold under a limited budget of queries. For the same query budget, a higher ASR indicates better attack quality.

Parameter settings. We develop our framework based on the FoolBox library [Rauber *et al.*, 2017; Rauber *et al.*, 2020]. The image size of MNIST is 28×28, we set the spacial sensitivity $\sigma_s = 2$ and the range sensitivity $\sigma_r = 8/255$. For other datasets, we resize their image size to $3 \times 224 \times 224$, and set $\sigma_s = 8$ and $\sigma_r = 32/255$. We set the time-dependent length $k = 5$, the MSE threshold $\tau = 0.2$ and the cosine similarity threshold $\rho = 0.1$ respectively. We set $B = 100$, which is the number of perturbations selected in each gradient estimation. We use the l_2 norm as the distance measure function $d(\cdot)$. In addition, we set the step size $\xi_t = \left\| \boldsymbol{x}_{adv}^{(t)} - \boldsymbol{x}^* \right\|_2 / \sqrt{t}$ and the perturbation size $\delta_t = \left\| \boldsymbol{x}_{adv}^{(t)} - \boldsymbol{x}^* \right\|_2 / dim$, where t is the iteration number and dim is the input dimension.

5.5 Experimental Results

Contribution analysis of each gradient prior. To make a comprehensive analysis of different gradient priors, we conduct the ablation study on a simple dataset MNIST, which has a smaller image resolution of 28×28, thus can converge faster and demonstrate the contribution of each gradient prior clearly. Specifically, we compare the following five cases.

- **DBA-GP** means the DBA model with both time-dependent prior and data-dependent prior.
- **Without data-dependent prior (w/o DP)** means the DBA model with only time-dependent prior.
- **Without time-dependent prior (w/o TP)** means the DBA model with only data-dependent prior.
- **Without DP and TP (w/o DP & TP)** means the DBA model without both gradient priors.
- **With only naive time-dependent prior (Naive TP)** means the DBA model with only the naive time-dependent prior described in Eq. (12).

Figure 2 shows the curves of MSE versus the number of queries when using different gradient priors on MNIST. X-axis represents the number of queries, and Y-axis denotes the average MSE value. From the results, we can get that w/o DP and w/o TP perform better than w/o DP & TP, which indicates that using arbitrary gradient prior properly could improve the performance to some extent. We can also observe that the naive time-dependent prior described in Eq. (12) will bring some side effects on performance. The reason is that Eq. (12) ignores the gradient direction change at the current and previous iteration. In addition, it also can be seen that DBA-GP using both data-dependent and time-dependent gradient priors can achieve the best performance. The reason is that DBA-GP utilizes the joint bilateral filter and two specially-designed judgement conditions to better leverage data-dependent and time-dependent gradient priors respectively, thus avoiding the edge gradient discrepancy issue and the successive iteration gradient direction issue.

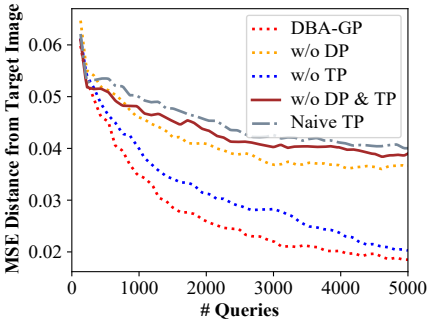

Figure 2: The curves of MSE versus the number of queries when using different gradient priors (lower is better).

Comparison with baselines. Table 1 shows the performance against different baselines on ImageNet and Celeba with 1K, 3K, and 5K queries respectively. The main results in the table are MSE values, and the results in parentheses represent ASR values when the MSE threshold is 0.001. The best results are highlighted in bold. It can be observed that DBA-GP performs much better than other strong baselines. Specifically, in terms of MSE under 3K queries, no matter attacking ResNet50 or VGG16, DBA-GP can have less than one-half of MSE values than other models on ImageNet, and even less than one-third of MSE values on Celeba. In terms of ASR with 3K queries, when attacking ResNet50 on ImageNet and Celeba, DBA-GP can improve the ASR values by 16% and 40% respectively compared with the best baseline. The reason is that DBA-GP utilizes more gradient prior information, which enables it can obtain better adversarial images within less number of queries. In addition, as the query number increases, each method tends to achieve a lower MSE and a higher ASR. But the convergence speed of DBA-GP is the fastest, which makes it more promising in real applications.

Attack the real-world API Face++. We also attack the real-world face recognition API Face++ from MEGVII. The API Face++ could give the prediction confidence score of whether two images contain the same person. In the experiment, when the returned confidence score is greater than 60%, we think the corresponding images are labeled as the same person. Note that since the pixel values of the images uploaded to the API are 8-bit floating point numbers that are not continuous in $[0, 1]$, we follow QEBA [Li *et al.*, 2020a] to discretize the images. Figure 3 shows the results of at-

Dataset		ImageNet			Celeba		
Method	Model	Q=1K	Q=3K	Q=5K	Q=1K	Q=3K	Q=5K
EA	ResNet50	0.0426 (0%)	0.0179 (0%)	0.0104 (0%)	0.0693 (0%)	0.0353 (0%)	0.0243 (2%)
	VGG16	0.0423 (0%)	0.0195 (0%)	0.0123 (0%)	0.0574 (0%)	0.0433 (0%)	0.0398 (2%)
SIGN-OPT	ResNet50	0.0335 (0%)	0.0179 (0%)	0.0117 (2%)	0.0409 (0%)	0.0241 (2%)	0.0171 (6%)
	VGG16	0.0334 (0%)	0.0183 (0%)	0.0119 (2%)	0.0448 (0%)	0.0314 (2%)	0.0254 (6%)
HSJA	ResNet50	0.0299 (0%)	0.0143 (0%)	0.0081 (6%)	0.0377 (0%)	0.0193 (6%)	0.0121 (12%)
	VGG16	0.0308 (0%)	0.0140 (2%)	0.0080 (6%)	0.0452 (4%)	0.0332 (10%)	0.0286 (12%)
QEBA	ResNet50	0.0260 (0%)	0.0112 (4%)	0.0064 (36%)	0.0124 (10%)	0.0024 (48%)	0.0015 (76%)
	VGG16	0.0266 (0%)	0.0135 (12%)	0.0096 (36%)	0.0149 (18%)	0.0039 (48%)	0.0015 (68%)
SURFREE	ResNet50	0.0433 (0%)	0.0176 (2%)	0.0116 (6%)	0.0268 (0%)	0.0146 (4%)	0.0089 (6%)
	VGG16	0.0421 (0%)	0.0233 (0%)	0.0134 (6%)	0.0277 (0%)	0.0152 (4%)	0.0081 (8%)
AHA	ResNet50	0.0243 (0%)	0.0119 (4%)	0.0067 (32%)	0.0119 (12%)	0.0038 (42%)	0.0015 (74%)
	VGG16	0.0248 (0%)	0.0133 (4%)	0.0099 (30%)	0.0138 (18%)	0.0043 (46%)	0.0021 (60%)
DBA-GP	ResNet50	**0.0190 (0%)**	**0.0056 (20%)**	**0.0025 (48%)**	**0.0047 (30%)**	**0.0007 (88%)**	**0.0004 (92%)**
	VGG16	**0.0214 (2%)**	**0.0075 (14%)**	**0.0036 (36%)**	**0.0049 (30%)**	**0.0009 (66%)**	**0.0005 (86%)**

Table 1: Attack performance against different target models on different datasets. The main results in the table are the MSE values between the adversarial and target images, and the ASR values are shown in parentheses.

Method	Q=1K	Q=3K	Q=5K
EA	0.0926 (0%)	0.0910 (0%)	0.0893 (0%)
SIGN-OPT	0.0833 (0%)	0.0801 (0%)	0.0764 (0%)
HSJA	0.0684 (0%)	0.0651 (0%)	0.0634 (0%)
QEBA	0.0583 (4%)	0.0493 (4%)	0.0427 (8%)
SURFREE	0.0799 (0%)	0.0741 (0%)	0.0689 (0%)
AHA	0.0621 (0%)	0.0542 (2%)	0.0468 (6%)
DBA-GP	**0.0415 (6%)**	**0.0272 (12%)**	**0.0205 (22%)**

Table 2: Attack performance of defending with adversarial training on ImageNet.

tacking Face++ API. The first column is the target image and the initial adversarial image, and the last four columns are the adversarial images produced by different attack methods under different query numbers. We can observe that the adversarial image attempts to get close to the target image gradually and keep the label unchanged. For HSJA, the MSE values do not decrease with increasing the number of queries, which indicates that it is difficult to find a better adversarial image. In addition, the right side of the generated adversarial image always contains the facial feature of the initial adversarial image, which also indicates HSJA works not well. QEBA performs much better, but it requires 5K queries to get a clean-looking adversarial image. DBA-GP could generate high-quality adversarial images at only 1K queries, and the MSE value decreases continually as the number of queries increases. All these phenomena validate the superiority of our proposed DBA-GP.

Attack results of defending with adversarial training. Adversarial training [Madry *et al.*, 2018] has shown to be effective to defend against adversarial attacks. Therefore, we further compare the performance of different attacking models when attacking the ResNet-152 model with adversarial training on ImageNet. Table 2 gives the results, the main results are MSE values, and the results in parentheses represent

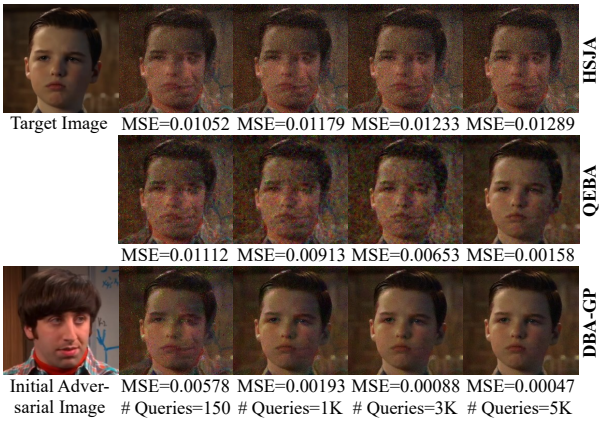

Figure 3: Comparison of different attack models when attacking against real-world API Face++.

ASR values when the MSE threshold is 0.01. The results in Table 2 show that our method DBA-GP can also obtain the best adversarial performance.

6 Conclusion

In this paper, we propose DBA-GP, a novel decision-based black-box attack framework with gradient priors. To better leverage the data-dependent gradient prior, DBA-GP exploits the joint bilateral filter to process each perturbation, which can mitigate the edge gradient discrepancy to some extent. To seamlessly integrate with the time-dependent gradient prior, DBA-GP introduces a new successive iteration gradient direction update strategy, which can speed up the convergence significantly. Extensive experiments confirm the superiority of our proposed method over other strong baselines, especially in query efficiency. In future work, we plan to investigate the theoretical underpinnings of the proposed method and extend it to other types of black-box adversarial attack methods.

Acknowledgments

The authors are grateful to the reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China (No. 62106035, 62206038, 61972065) and Fundamental Research Funds for the Central Universities (No. DUT20RC(3)040, DUT20RC(3)066), and supported in part by Key Research Project of Zhejiang Lab (No. 2022PI0AC01), National Key Research and Development Program of China (2022YFB4500300) and CAAI-Huawei Mindspore Open Fund. We also would like to thank Dalian Ascend AI Computing Center and Dalian Ascend AI Ecosystem Innovation Center for providing inclusive computing power and technical support.

References

[Brendel *et al.*, 2018] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[Brunner *et al.*, 2019] Thomas Brunner, Frederik Diehl, Michael Truong-Le, and Alois C. Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *ICCV*, pages 4957–4965, 2019.

[Cao *et al.*, 2021] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 176–194, 2021.

[Carlini and Wagner, 2017] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017.

[Chen *et al.*, 2020] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1277–1294, 2020.

[Cheng *et al.*, 2020] 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 *ICLR*, 2020.

[Chiu *et al.*, 2018] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *ICASSP*, pages 4774–4778, 2018.

[Cintas *et al.*, 2020] Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, and Edward McFowland. Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error. In *IJCAI*, pages 876–882, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Dong *et al.*, 2019] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, pages 7714–7722, 2019.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[Guo *et al.*, 2019] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *ICML*, pages 2484–2493, 2019.

[Guo *et al.*, 2020] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. In *NeurIPS*, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Ilyas *et al.*, 2019] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019.

[LeCun, 1998] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[Li and Chen, 2021] Nannan Li and Zhenzhong Chen. Toward visual distortion in black-box attacks. *IEEE Transactions on Image Processing*, pages 6156–6167, 2021.

[Li *et al.*, 2020a] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: query-efficient boundary-based blackbox attack. In *CVPR*, pages 1218–1227, 2020.

[Li *et al.*, 2020b] Jie Li, Rongrong Ji, Hong Liu, Jianzhuang Liu, Bineng Zhong, Cheng Deng, and Qi Tian. Projection & probability-driven black-box attack. In *CVPR*, pages 359–368, 2020.

[Li *et al.*, 2021] Jie Li, Rongrong Ji, Peixian Chen, Baochang Zhang, Xiaopeng Hong, Ruixin Zhang, Shaoxin Li, Jilin Li, Feiyue Huang, and Yongjian Wu. Aha! adaptive history-driven attack for decision-based black-box models. In *ICCV*, pages 16148–16157, 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Luo *et al.*, 2021] Chenxu Luo, Xiaodong Yang, and Alan L. Yuille. Self-supervised pillar motion learning for autonomous driving. In *CVPR*, pages 3183–3192, 2021.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[Maho *et al.*, 2021] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: A fast surrogate-free black-box attack. In *CVPR*, pages 10430–10439, 2021.

[Mooney, 1997] Christopher Z Mooney. *Monte Carlo Simulation*. SAGE, 1997.

[Moosavi-Dezfooli *et al.*, 2016] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[Muhammad *et al.*, 2021] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C. de Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2021.

[Park *et al.*, 2019] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, pages 2613–2617, 2019.

[Petschnigg *et al.*, 2004] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael F. Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics*, pages 664–672, 2004.

[Pham *et al.*, 2021] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *CVPR*, pages 11557–11568, 2021.

[Qin *et al.*, 2022] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. In *NeurIPS*, 2022.

[Rauber *et al.*, 2017] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *ICML workshop*, 2017.

[Rauber *et al.*, 2020] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and JAX. *Journal of Open Source Software*, page 2607, 2020.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Wang *et al.*, 2021] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. In *CVPR*, pages 15849–15858, 2021.

[Wang *et al.*, 2022] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *CoRR*, abs/2207.02696, 2022.

[Wu *et al.*, 2017] Tingmin Wu, Shigang Liu, Jun Zhang, and Yang Xiang. Twitter spam detection based on deep learning. In *ACSW*, pages 3:1–3:8, 2017.

[Wu *et al.*, 2020] 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 *ICLR*, 2020.

[Zhang *et al.*, 2021] Jiawei Zhang, Linyi Li, Huichen Li, Xiaolu Zhang, Shuang Yang, and Bo Li. Progressive-scale boundary blackbox attack via projective gradient estimation. In *ICML*, pages 12479–12490, 2021.

[Zhong *et al.*, 2022] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *CVPR*, pages 15324–15333, 2022.