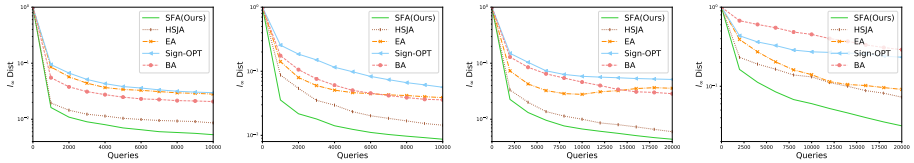
# Supplementary Material: Boosting Decision-based Black-box Adversarial Attacks with Random Sign Flip

Weilun Chen[1,2], Zhaoxiang Zhang[1,2,3⋆], Xiaolin Hu[4], and Baoyuan Wu[5,6]

[1] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS
[4] Tsinghua University
[5] The Chinese University of Hong Kong, Shenzhen
[6] Tencent AI Lab
{chenweilun2018, zhaoxiang.zhang}@ia.ac.cn, xlhu@mail.tsinghua.edu.cn, wubaoyuan1987@gmail.com

## A Median $l_\infty$ distances versus queries

We provide the curves of the median $l_\infty$ distances of various methods for both untargeted and targeted settings. We use our trained ResNet-18 as the targeted model for CIFAR-10 and ResNet-50 for ImageNet. From Fig.10, we can see that our method has smaller median $l_\infty$ distances than baseline methods under different query budgets.




(a) (b) (c) (d)

**Fig. 10.** The median $l_\infty$ distances over the number of queries. (a) CIFAR-10 untargeted. (b) CIFAR-10 targeted. (c) ImageNet untargeted. (d) ImageNet targeted. The target models for CIFAR-10 and ImageNet are ResNet-18 and ResNet-50, respectively.

## B A detailed comparison with HSJA and EA

We quantify the performance in terms of three dimensions: attack success rate, average queries and median queries. Average queries and median queries are

⋆ Corresponding Author

calculated over successful trails, which means they can be influenced by the attack success rate. For targeted attacks on ImageNet, we provide the results achieved by our method on the images that HSJA and EA attack successfully in Table 6. It can be clearly seen that our method surpasses HSJA and EA on both attack success rate and query efficiency.

**Table 6.** Additonal Results of **targeted** attacks on **ImageNet**. EA+ (or HSJA+) denotes results achieved by our method on images that EA (or HSJA) attacks successfully.

| Method | ResNet-50 | | | VGG-16 | | | DenseNet-121 | | | Inception-v3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | AQ | MQ | ASR | AQ | MQ | ASR | AQ | MQ | ASR | AQ | MQ |
| EA | 9.6% | 19,682 | 20,435 | 8.5% | 12,126 | 8,534 | 12.2% | 17,820 | 7,195 | 4.3% | 18,164 | 15,362 |
| EA+ | 100% | 7,612 | 5,284 | 100% | 5,037 | 4,906 | 100% | 7,042 | 4,328 | 100% | 6,657 | 5,604 |
| HSJA | 84.5% | 44,188 | 41,205 | 77.8% | 39,400 | 36,172 | 79.7% | 41,319 | 36,964 | 0.0% | - | - |
| HSJA+ | 99.2% | 22,196 | 19,382 | 100% | 14,468 | 13,650 | 100% | 17,361 | 15,776 | - | - | - |
| **SFA(Ours)** | **99.3%** | 22,538 | 19,380 | **99.2%** | 16,627 | 15,008 | **98.6%** | 20,331 | 17,762 | **95.8%** | 36,681 | 32,210 |

# C Ablation study

## C.1 Analysis of hyperparameter adjustment

For experiments in the main paper, we dynamically adjust the hyperparameters as described in Section 3.3. To verify the good performance of our method is mostly derived from the random sign flip strategy itself, we fix $\boldsymbol{p}$ which controls the sign flip probability of each coordinate (see Eq.6) as 0.001, and conduct experiments on CIFAR-10 following the setting in Section 4.1. The attack success rate, median and average queries of untargeted attacks are 91.5%, 646 and 540, respectively, while those of targeted attacks are 99.6%, 2,089 and 1,632. These results are still better than the compared methods.

## C.2 Analysis of dimensionality reduction

We consider different $d'$ of $d$, $d/4$, $d/16$, $d/64$, corresponding to $224 \times 224 \times 3$, $112 \times 112 \times 3$, $56 \times 56 \times 3$, $28 \times 28 \times 3$. We use a pre-trained ResNet-50 as the target model. The results are reported in Table 7. With a suitable $d'$, it is able to achieve a higher attack success rate with fewer queries. However, a too small $d'$ harms the performance.

## C.3 Analysis of the project step

To investigate the effects of the project step, we fix the maximum distortion magnitude and solely apply the random sign flip step. Here, we present the attack

**Table 7.** Effects of dimentionality reduction.

| | Untargeted | | | Targeted | | |
|---|---|---|---|---|---|---|
| | ASR | AQ | MQ | ASR | AQ | MQ |
| $d$ | 96.9% | 3,193 | 1,570 | 98.6% | 29,440 | 26,208 |
| $d/4$ | **98.2%** | 2,712 | 1,288 | **99.3%** | 22,538 | 19,382 |
| $d/16$ | 97.5% | 2,539 | 1,150 | 97.3% | 23,572 | 19,558 |
| $d/64$ | 94.1% | 2,457 | 960 | 89.9% | 27,552 | 20,724 |

results of this setting on CIFAR-10. The attack success rates of the untargeted and targeted attacks are 27.7% and 7.45%, respectively. The performance drop is expected, as the method with only the random sign flip step (without the project step) can be viewed as a random exploration of the vertices of the $l_\infty$ ball. Thus, both the project and the random sign flip step make key contributions to the attack performance of our method. The project step ensures the continuous decreasing of the $l_\infty$ distortion along a trajectory of successful attacks, while the sign flip step could accelerate the searching for the trajectory.

## D Details about Defenses

To investigate the effectiveness of decision-based black-box attacks against defenses, we conduct experiments on 7 defense mechanisms: Adversarial Training, Thermometer Encoding, Bit Depth Reduction, FeatDenoise, FeatScatter, KWTA, and TRADES. We use the official codes and models for Adversarial Training[5], FeatDenoise[6], FeatScatter[7], KWTA[8], and TRADES[9]. For Thermometer Encoding, the number of levels of discretization is 16. For Bit Depth Reduction, we use a pre-trained ResNet-50 as the target model and reduce the bit depth to 3. All the experiments on CIFAR-10 are conducted on the first 1,000 images from the validation set. As for ImageNet, we randomly select 1,000 images. The corresponding cumulative attack success rates of the number of queries are presented in Fig. 11.

## E Attacks on Real-world Applications

We conduct experiments on the face verification API and food API in Tencent AI Open Platform.

For face verification, there are two types of attacks. One is the dodging attack where the adversary tries to make the original image with an imperceptible perturbation not recognized. The other is the impersonation attack corresponding to finding an adversarial image classified as a specific identity. We have provided

[5] https://github.com/MadryLab/cifar10_challenge
[6] https://github.com/facebookresearch/ImageNet-Adversarial-Training
[7] https://github.com/Haichao-Zhang/FeatureScatter
[8] https://github.com/a554b554/kWTA-Activation
[9] https://github.com/yaodongyu/TRADES

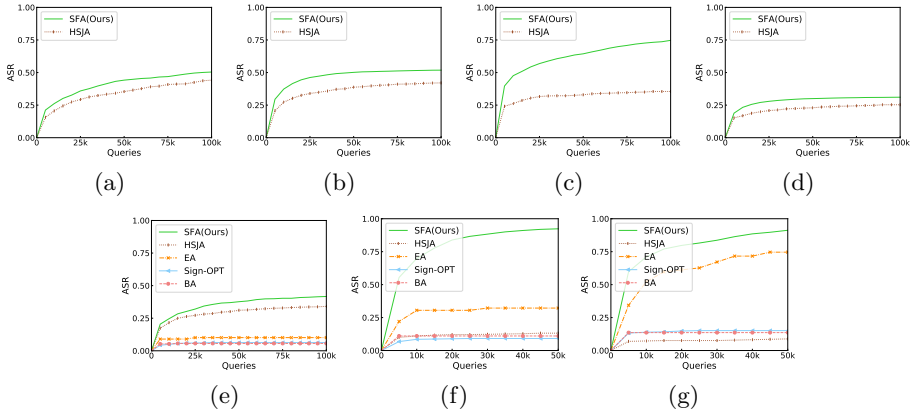



**Fig. 11.** Attack success rates versus numbers of queries on defensive models. (a) Feat-Denoise (b) FeatScatter (c) KWTA (d) TRADES (e) Adversarial Training (f) Thermometer Encoding (g) Bit Depth Reduction.

examples of these two types of attacks in Fig. 1. In our experiments, we choose 10 pairs of images where each pair of them are from different identities and perform impersonation attacks. As there are only 2 classes for the food API, untargeted and targeted attacks are the same. Our method can craft an adversarial image that is hard to distinguish from the original image in very few queries. Examples are shown in Fig. 12.

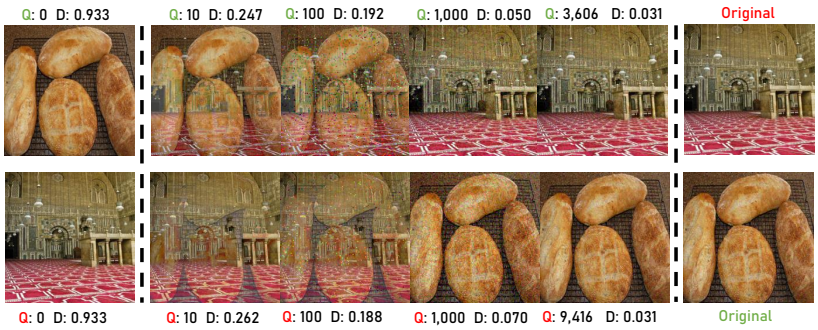



**Fig. 12.** Attacking the food API in Tencent AI Open Platform. **Q** denotes the query number. **D** denotes the $l_\infty$ distance towards the original image. Green means the API classifies the image as food, red is otherwise. Best viewed in color with zoom in.