FeatureBA: Hard Label Black Box Attack based on Internal Layer Features of Surrogate Model

Jiaxing Li[a], Yu-an Tan[a], Runke Liu[a], Weizhi Meng[b], Yuanzhang Li[c,*]

[a] *School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, 100081, China*
[b] *School of Computing and Communications, Lancaster University, United Kingdom*
[c] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing, 100081, China*

Abstract

This study revises previous work by emphasizing the integration of surrogate models into query-based black-box adversarial attacks, showcasing their effectiveness in reducing query counts and enhancing robustness. This observation highlights a critical gap in decision-based (hard label) approaches, which have not yet effectively integrated surrogate models. In this paper, we propose a novel decision-based approach to black-box adversarial attacks. By utilizing intermediate layer features of the surrogate network and optimizing the query feedback process, the proposed method achieves competitive results with a significant reduction in query counts (up to 99.73% lower compared to existing methods). Extensive experiments validate its performance across diverse tasks, including image classification, object detection, and face recognition. This work demonstrates the potential for enhancing the practicality of decision-based attacks in real-world scenarios.

Keywords: Deep learning, Adversarial machine learning, Black box attack, Internal layer features

*Corresponding author
Email addresses: jiaxingxx@outlook.com (Jiaxing Li), tan2008@bit.edu.cn (Yu-an Tan), 3220242015@bit.edu.cn (Runke Liu), weizhi.meng@ieee.org (Weizhi Meng), popular@bit.edu.cn (Yuanzhang Li)

1. Introduction

Adversarial machine learning, which includes both white-box and black-box attacks (Reza et al., 2023; Li et al., 2024), has garnered significant attention due to its potential applications in critical areas such as image classification, object detection, and face recognition. In the white-box attack setting (Goodfellow et al., 2015; Carlini and Wagner, 2017; Moosavi-Dezfooli et al., 2016), the attacker has full knowledge of the target neural network and its weights. However, it is usually impractical to exploit information about neural networks in real-world scenarios. Therefore, black-box attacks are a setting that is more in line with real-world attack scenarios. Black-box attacks can be roughly divided into transfer-based, score-based, and decision-based attacks. Score-based and decision-based attacks are collectively referred to as query-based attacks. The black-box setting is the actual setting for adversarial attacks with limited knowledge of neural networks. Transfer-based adversarial attacks (Wang et al., 2021b) use a proxy model to generate adversarial samples, and it does not exploit queries to obtain valid information about the victim model. Score-based attacks (Chen et al., 2017) query the target classifier's predicted probabilities for all classes to estimate the gradient at each step and generate perturbations. However, this attack strategy may not be feasible because in many real-world applications (Ran et al., 2025; Liu et al., 2023; Muthalagu et al., 2025), the classifier only returns the top 1 classification label when responding to a query. Therefore, decision-based adversarial attacks are the most realistic adversarial attacks because they allow the adversary to generate an adversarial example (Liu et al., 2025; Fang et al., 2024; Shen and Li, 2025) by only querying the output label.

Most decision-based attacks are in the field of image classification, such as BoundaryAttack (Brendel et al., 2018), HSJA (Chen et al., 2020), qFool (Liu et al., 2019), GeoDA (Rahmati et al., 2020), QEBA (Li et al., 2020a), and TA (Ma et al., 2021), which are based on finding a normal vector at a point on the decision boundary of the classification task and iteratively reducing the perturbation to search for a new boundary point. Among these attacks, HSJA et al. use the estimated normal vector direction to get a point in the adversarial region, and then apply binary search between the obtained adversarial point and the source to get a new boundary point. qFool Attack (Liu et al., 2019) and GeoDA (Rahmati et al., 2020) approximate the boundary as a hyperplane and find a new boundary point by searching along the direction of the estimated gradient. SurFree (Maho et al., 2021)

also considers the hyperplane boundary and searches along the semicircular path, but it does not use the information of the normal vector, but estimates the attack direction through random trials. These decision-based attacks do not consider the intrinsic characteristics of the image object itself. And in the broader field of visual recognition, such as object detection and face recognition, whether these boundary-based attacks can be directly applied remains a question.

Some recent works (Cheng et al., 2019; Guo et al., 2019; Tashiro et al., 2020; Yang et al., 2020) are based on the combination of transfer-based surrogate model attacks and query-based attacks. Most of them are score-based attacks (Yang et al., 2020) combined with local surrogate models. Query-based attacks and transfer-based attacks are actually complementary. Query-based strategies can benefit from better search directions, while transfer-based strategies can benefit from query feedback, allowing it to dynamically adjust surrogate hypotheses. Score-based attacks combined with local surrogate models are easier to implement because score-based attacks make loose assumptions that they can obtain information such as logits of the victim model. However, decision-based (hard label) attacks have more stringent assumptions and cannot obtain information such as confidence, making it difficult to estimate gradients in a fine-grained manner.

However, existing decision-based attacks often ignore the intrinsic features of images and lack adaptive strategies for optimizing queries. This study addresses these limitations by proposing a decision-based black-box attack method that leverages surrogate models' intermediate layer features to guide query optimization dynamically. As shown in Figure 1, compared with the traditional hard label attack, we used the surrogate model to explore the characteristics of the image object itself. The query hard label is used as feedback information to optimize the process of the surrogate model to generate the characteristics of the image itself. The intrinsic characteristics of the image object generated by the surrogate model can guide the loss update and reduce the number of queries.

Our contributions are as follows:

- We not only compare with the most advanced algorithms of hard label black box attack on image classification, but also generalize the decision-based Boundary Attack attack in the classification field to object detection and face recognition as a baseline.
- We designed an attack framework based on the surrogate model of the

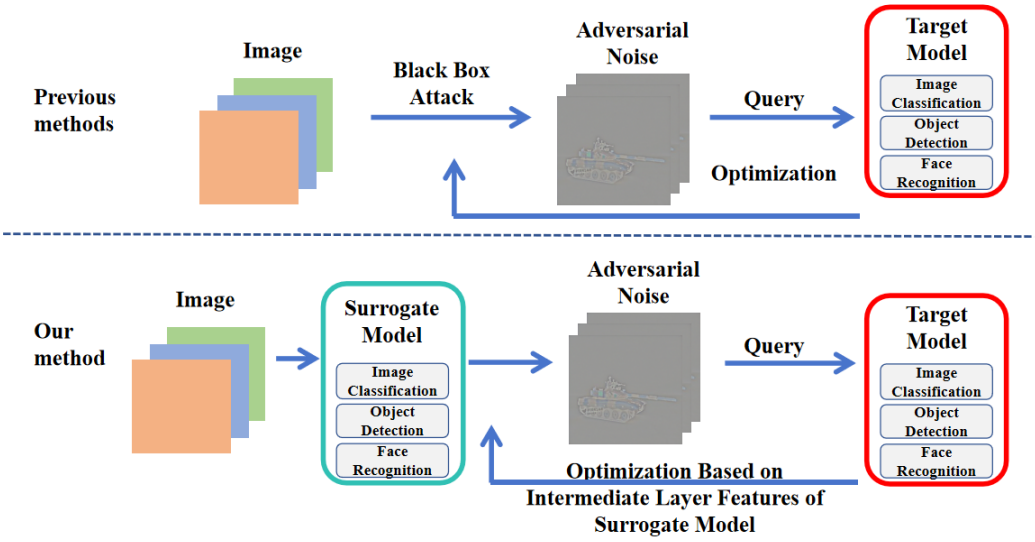

Figure 1: Compared with traditional hard label attacks, we use surrogate models to explore the characteristics of image objects. We use hard label queries as feedback information to optimize the process of generating image characteristics by surrogate model. The intrinsic characteristics of image objects generated by the surrogate model can guide loss updates and reduce the number of queries.

intermediate layer features plus a very small amount of query feedback, and verified it in image classification, object detection and face recognition.

- We designed a new method to obtain the importance of the intermediate layer features, and designed the optimization implementation of L_2 norm and L_∞ norm restrictions based on the algorithm loss

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed method, Section 4 and 5 present experimental results, Section 6 conducts ablation experiments, Section 7 discusses findings and limitations, and Section 8 concludes the study.

2. Related Works

Decision-based black-box attacks are the most end-user-friendly attack setting. The only available information for this type of attack is the output

label of the target neural network. Therefore, decision-based black-box attacks are the most challenging setting to obtain adversarial examples. The Boundary Attack (Brendel et al., 2018) algorithm performs a random walk along the decision boundary to reduce the perturbation of the query. To improve the performance of (Brendel et al., 2018), Biased Boundary Attack (Brunner et al., 2019), OPT (Cheng et al., 2018), and Rays (Chen and Gu, 2020) proposed random search optimal directions such as reducing the search space to reduce perturbations. In (Dong et al., 2019), an evolutionary attack method is proposed, which draws random samples from a normal distribution with a custom covariance in a simplified search space. AHA (Li et al., 2021) generates random samples from a normal distribution using the average of historical queries. Triangle Attack (TA) (Ma et al., 2021) is based on the geometric relationship between benign examples, current and future adversarial examples, forming a triangle in a subspace at each iteration. SurFree (Maho et al., 2021) is a model-free algorithm that claims that queries that bypass normal estimation will improve query efficiency. On the other hand, there are also attacks that rely on estimating normal vectors at boundary points. HSJA (Chen et al., 2020), qFool (Liu et al., 2019), QEBA (Li et al., 2020a), GeoDA (Rahmati et al., 2020), and TA (Ma et al., 2021) use the estimated normal vector direction to obtain a point in the adversarial region, and then apply binary search between the obtained adversarial point and the source to obtain a new boundary point. CGBA (Reza et al., 2023) is a geometry-based normal vector estimation for classification tasks, that is, gradient. However, these algorithms rely on more assumptions than search-based methods and may not be directly applicable to tasks with complex losses such as generalized object detection and face recognition. These decision-based attacks also do not consider the intrinsic characteristics of the image objects themselves. Decision-based black-box attacks have evolved as one of the most challenging paradigms in adversarial machine learning. Early methods, such as Boundary Attack and its variants, rely on random walks along decision boundaries to minimize perturbations. More recent methods, including HSJA and SurFree, aim to enhance query efficiency by refining search strategies. However, these approaches generally fail to consider the intrinsic characteristics of images or the transferability benefits offered by surrogate models. We summarize the existing classic boundary based methods in Table 1.

Recently, the combination of transferability-based surrogate model attacks and query-based attacks has become a hot topic in black-box attack research. Transferability-based attacks consider improving the transferabil-

Table 1: Comparison of Existing Methods in Boundary-based Adversarial Attacks

Method	**Year**	**Boons (Strengths)**	**Limitations**
Boundary Attack	2018	Simple and efficient, low perturbation.	Slow convergence for complex models.
OPT	2018	Faster convergence with lower perturbations by reducing search space.	Depends on a good search direction, may not suit all models.
HSJA	2020	Efficient boundary exploration with gradient-based normal vector estimation.	Needs a boundary model, not always applicable.
AHA	2021	Adapts using historical queries to minimize perturbations.	Requires many queries for optimization, less effective on complex models.
SurFree	2021	Model-free, avoids surrogate models, and flexible across tasks.	Dependent on query quality, inefficient for complex models.
CGBA	2023	Geometry-based normal vector estimation for efficient boundary exploration.	Limited to tasks with simpler decision boundaries.

ity of adversarial examples generated by local surrogate models on black-box models. Some works (Dong et al., 2018; Lin et al., 2019; Wang and He, 2021; Xie et al., 2019; Li et al., 2020b; Wu et al., 2021; Wang et al., 2021a) enhance the transferability of adversarial examples by perturbing the output layer. Other works (Wang et al., 2021b; Zhang et al., 2022; Li et al., 2024) focus on maximizing internal feature perturbations to improve transferability. The combination of local surrogate models and score-based attacks is the attack setting considered by many black-box attacks (Cheng et al., 2019; Guo et al., 2019; Tashiro et al., 2020; Yang et al., 2020). Query-based attacks and transferability-based attacks are actually complementary. Query-based strategies can benefit from better search directions, while transfer-based strategies can benefit from query feedback, allowing it to dynamically adjust alternative hypotheses. It is easier to combine score-based attacks

with local substitution models because score-based attacks have loose assumptions and can obtain information such as the confidence of the victim model. However, decision-based (hard label) attacks have more stringent assumptions and cannot obtain information such as confidence. Combining hard label attacks with transfer-based methods remains challenging.

Threats to Validity. Potential biases in reviewed studies, such as limited datasets or overly simplistic experimental setups, may impact generalizability. To address these, our work employs comprehensive benchmarks, including ImageNet, COCO, and LFW datasets, under strict black-box conditions.

3. Method

We use the perturbation method to generate the perturbed image $x' = x + \delta$, where δ represents the perturbation vector of the same size as the input image x. To ensure that the perturbation is imperceptible to humans, we usually limit its p-norm to be less than a threshold, i.e., $\|\delta\|_p \leq \epsilon$. For example, the L_2 norm or L_∞ norm is usually adopted. This adversarial attack on the victim model f can be generated by minimizing the so-called adversarial loss function L, so that the output $f(x+\delta)$ is as close as possible to the desired (adversarial) output. Specifically, the attack generation function maps the input image x to the adversarial image x', so that the output $f(x')$ is different from the original output y. The perturbation vector δ is added to the input image x to generate the adversarial image $x' = x + \delta$. The perturbation is constrained using the L_2 or L_∞ norm, so that $\|\delta\|_p \leq \epsilon$.

As shown in Figure 1, in our experiments, we use the L_2 norm or L_∞ norm to limit the maximum level of perturbation. Our goal is to find δ such that the perturbed image $x' = x + \delta$ can destroy the victim visual model f and make incorrect predictions. Assume that the original prediction for the clean image x is $y = f(x)$. The attack target is $f(x') \neq y$, where y is the clean label. For the victim model $f(x)$, "hard label" means that only the final label result is output, and the logit vector of the last layer cannot be used. For classification models, the label $y \in \mathbb{R}$ is a scalar. For object detection, the prediction model can have a more complex output space, generally outputting $y \in \mathbb{R}^{K\times 6}$, where K is the number of detected objects, and each object label and position is encoded in a vector of length 6, which includes the object category, bounding box coordinates, and confidence score.

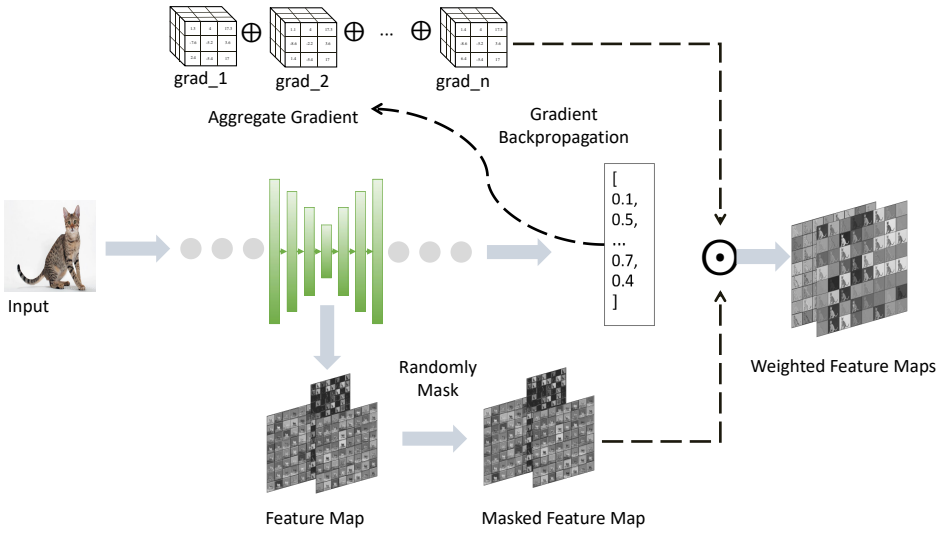

Figure 2: This shows the process of obtaining weighted feature maps by our method. Given an input image, feature maps are extracted from an intermediate layer of alternative models. Then, the feature maps are randomly masked, and the gradients are back-propagated from the calculated output to the feature map, and the gradients are added as the feature importance. After element-wise multiplication of the feature map and the normalized gradient, the perturbed intermediate layer features are point-multiplied with the aggregated gradients.

For face recognition, the model determines whether the face image is the same person and outputs the label $y \in \mathbb{R}$.

3.1. Intermediate Layer Feature Extraction

In this section, we introduce how to use the surrogate model to extract intermediate layer features and the corresponding aggregated gradients. The perturbation of the intermediate layer features is inspired by the aggregation gradient methods such as FIA (Wang et al., 2021b) and Smoothgrad (Smilkov et al., 2017). For most DNN-based visual models, it has been experimentally proven that surrogate models (Szegedy et al., 2016), (Wang et al., 2021b), and (He et al., 2016) are inclined to extract semantic features. These features are the basis for the perception of target objects, thereby effectively improving classification accuracy. Disrupting those image attribute perception features that guide all model decisions may be a favorable direction for adversarial generation.

We first obtain the most important area of the image for all model decisions. This part of the area is the area with transfer characteristics, that is, the area where the essential features of the object are located. Our gradient represents the importance of the feature. The direct derivation of the model decision result for the image is similar to the simplest model interpretation method. Because the model interpretation method is for the image. And our method is for the intermediate layer feature maps of different layers.

We use the perturbation re-aggregation method to reduce the influence of the specificity of a single model and highlight the characteristics of the object itself. The aggregated gradient is equivalent to the weight, which is a weight matrix used to reduce important features. Let f represent the surrogate model, the feature map starting from the i-th layer is represented as Z_i, and $J(x)$ represents the original objective function of the model, whose output is logit. Since the importance of a feature is proportional to its contribution to the final decision, an intuitive strategy is to obtain the gradient as follows:

$$\Delta_i = \frac{\partial J(x)}{\partial Z_i} \tag{1}$$

Furthermore, to reduce the influence of single model features and focus on the characteristics of the object itself, we calculate the aggregation gradient within a small neighborhood. The aggregation gradient is calculated as follows:

$$\bar{\Delta}_i = \frac{1}{C} \sum_{n=1}^{N} \Delta_i^{Z_i \odot M_p^n} \tag{2}$$

Here, M_P is a binary matrix of the same size as Z_i, where $\odot$ represents the element-wise product, and C is obtained by the L_2 norm on the corresponding summation terms. N represents the number of random masks applied to the intermediate layer features Z_i. The aggregated gradients $\bar{\Delta}_i$ highlight regions of robust and critical object-aware features that can guide adversarial examples towards more transferable directions. We monitor the training and validation losses during the optimization process to ensure the model converges effectively. The training loss is defined as the average adversarial loss over the training dataset, while the validation loss is computed similarly over the validation set.

Algorithm 1 Generate Adversarial Perturbations of Surrogate Model

Require: Clean image x, loss function L, surrogate model f, original objective function J, maximum perturbation ϵ, number of iterations T

Ensure: Adversarial image x' for surrogate model f

Initialize: $\Delta = 0$, $g_0 = 0$, $\alpha = \epsilon/T$, $x' = x$
for $t = 0$ to $T - 1$ **do**
    $\bar{\Delta}_i^{t+1} = \lambda \cdot \Delta_i^t + \frac{\nabla_{z_i} J(x')}{\|\nabla_{z_i} J(x')\|_1}$
    $g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x'} L(x')}{\|\nabla_{x'} L(x')\|_1}$
    $x'_{t+1} = \mathrm{Clip}_\epsilon \{x'_t - \alpha \cdot \mathrm{sign}(g_{t+1})\}$
    $t = t + 1$
end for
return x'_T

Algorithm 2 FeatureBA Attack Algorithm

Require: Clean image x, loss function L, a victim black-box model v, a surrogate model f, maximum perturbation ϵ, parameter space H of surrogate model

Ensure: Adversarial image x' of the black-box victim model v

Initialize: $\Delta = 0$, $g_0 = 0$, $\alpha = \epsilon/T$, $x' = x$
while True **do**
    Depth First Search on H get $\{i, N, P\}$
    $\Delta_i = \frac{\partial J(x)}{\partial Z_i}$
    $\bar{\Delta}_i = \frac{1}{C} \sum_{n=1}^{N} \Delta_i^{Z_i \odot M_p^n}$
    $L(x') = \bar{\Delta}_i \odot Z_i$
    Update x' with momentum method as Algorithm 1
    Query black-box model v using x'
    if x' is adversarial **then**
        return x'
        break
    else
        Depth First Search H, get new $\{i, N, P\}$
        Repeat line 4 to line 13
    end if
end while

3.2. Query Feedback Optimization Algorithm

In this subsection, we introduce how to combine query information and aggregate gradients and give an attack algorithm. We use the importance of the above features (i.e., aggregate gradients $\bar{\Delta}_i$) to guide the loss function L of the adversarial example x' by explicitly suppressing important features.

$$L(x') = \bar{\Delta}_i \odot Z_i \tag{3}$$

We define the loss of our surrogate model to generate adversarial perturbations as the dot product of the aggregate gradient and the feature map of the intermediate layer. As shown in Figure 2, Important features will produce relatively high values in Δ, which indicates that the model recognizes features to output the correct label. The goal of generating transferable adversarial samples is to reduce important features with high Δ and increase important features corresponding to low Δ. Therefore, by minimizing the loss function equation, adversarial examples can be directly guided to develop in a more transferable direction. We use a momentum-based optimization method to solve the loss. And we give the L_∞ solution Algorithm 1 for the surrogate model under the maximum limit perturbation ϵ. In particular, for L_2 norm bound perturbations, we use $x'_{t+1} = x'_t - \alpha \cdot \frac{g_{t+1}}{\|g_{t+1}\|_2}$ to replace the corresponding update part of the algorithm for x'. We choose the L_2 and L_∞ norms to limit the perturbation size because they offer different trade-offs in terms of smoothness and visibility of the perturbation. The L_∞ norm constrains the maximum change per pixel, ensuring imperceptibility, while the L_2 norm provides a more globally smooth perturbation.

How to use the information fed back by black-box queries to optimize the surrogate model and use the intermediate layer features to generate adversarial perturbations is also the main problem we solve. When the surrogate model generates adversarial perturbations, it is affected by layer i, the number of aggregations N, and the probability of mask perturbation P. The perturbations generated by the surrogate model determined by these parameters are different. And these conditions can be guided by the information fed back by black-box queries to optimize. Therefore, we search the space $H_{i\times N\times P}$ spanned by the parameters $\{i, N, P\}$ of the surrogate model. The space spanned by the parameter set of the surrogate model is small, which is why we have fewer queries. Therefore, we can use a simple depth-first search. See Algorithm 2 for details of the proposed FeatureBA.

The computational complexity of FeatureBA is $O(T \cdot N \cdot D)$, where T is the number of iterations, N is the number of random masks, and D is the dimensionality of the intermediate layer features. The memory complexity is $O(D)$, as we store the intermediate layer features and their gradients.

4. Experiments on Image Classification

To evaluate the effectiveness of our approach, we conduct extensive experiments on attacking image classification, object detection, and face recognition models. We first show the attack performance on image classification. All experiments were performed on NVIDIA 3090Ti GPU.

4.1. Datasets and Victim Models

We use the ImageNet dataset (Deng et al., 2009) as the experimental dataset for image classification. The ImageNet dataset consists of 1,000 classes, with over 1,200,000 training images and 50,000 validation images. For each image, we resize it to a fixed size of 224×224, and normalize the pixel values to have a mean of 0 and a standard deviation of 1. We use the ImageNet dataset due to its large size and diversity of object categories, which allows for comprehensive evaluation of our proposed adversarial attack method across a variety of real-world classification tasks. We use 70% of the ImageNet data for training, 15% for validation, and 15% for testing. This standard split ensures sufficient training data while maintaining an independent validation and test set to evaluate the performance of our attack method.

We use 1,000 correctly classified test images. For each victim model (target model), we randomly select 1,000 image-label pairs from the validation set of ILSVRC2012 and are correctly classified by the victim model. Images are resized to $3\times224\times224$ as input to the classifier. The victim models we choose are classifiers pre-trained on the ImageNet dataset. The victim models include VGG16 (Simonyan and Zisserman, 2015), Inception-V3 (Szegedy et al., 2016), and ResNet152 (He et al., 2016). In order to create a rigorous black-box environment, we choose a simple local white-box model ResNet18 (He et al., 2016) that is different from the above victim models as an alternative model. All neural networks used are available through the open-source pre-trained models of PyTorch/torchvision.

We use the perturbation L_2 and L_∞ norms bound and the number of queries as evaluation indicators. We give the same L_2-norm and L_∞ norm perturbation budgets to the attack methods, and the number of queries for

these attack methods determines the effectiveness of the attack. The L_2 norm is computed as follows:

$$L_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

where x_i and y_i are the pixel values of the original and perturbed images, respectively. The L_∞ norm is defined as:

$$L_\infty = \max(|x_i - y_i|) \tag{5}$$

We use these norms to measure the imperceptibility of the perturbations, ensuring that the adversarial examples are visually similar to the original images.
Assume that x is the L_∞ distance (i.e., the maximum change applied to each pixel), the image has a height h, a width w, and 3 color channels (e.g., RGB). The total number of pixels in the image is $h \times w \times 3$. The L_2 distance is computed by summing the square of the L_∞ distance across all pixels and then taking the square root:

$$L_2 = \sqrt{(h \times w \times 3) \times x^2} \tag{6}$$

This equation shows that the L_2 distance is proportional to the square root of the product of the number of pixels and the square of the L_∞ distance. Intuitively, it aggregates the pixel-wise changes represented by the L_∞ distance into a single metric that accounts for the entire image. The number of queries is used as an evaluation criterion to measure the efficiency of the attack method. Fewer queries indicate a more efficient attack, which is crucial for real-world adversarial attacks where query limits may exist.

4.2. Comparison Methods

In image classification, we compare with BoundaryAttack Brendel et al. (2018) and Surfree Maho et al. (2021) algorithms. Boundary Attack Brendel et al. (2018) is the earliest decision-based black-box attack method. It performs random walks along the decision boundary to reduce the perturbation of queries. SurFree Maho et al. (2021) is a model-free algorithm that claims that queries that bypass normal vector estimation will improve query efficiency. We compare with SurFree because SurFree is the most advanced decision-based black-box attack method.

In object detection and face recognition, we compare with BoundaryAttack Brendel et al. (2018) algorithm. Boundary Attack Brendel et al. (2018) is the earliest decision-based black-box attack method. It performs random walks along the decision boundary to reduce the perturbation of queries. We improved BoundaryAttack based on the outputs of object detection and face recognition.

For the comparison method we used, the original prediction for the clean image x is $y = f(x)$. The attack target is $f(x') \neq y$, where y is the clean label. For the victim model $f(x)$, "hard label" means that only the final label result is output, and the logit vector of the last layer cannot be used. For classification models, the label $y \in \mathbb{R}$ is a scalar. For object detection, the prediction model can have a more complex output space, generally outputting $y \in \mathbb{R}^{K\times 6}$, where K is the number of detected objects, and each object label and position is encoded in a vector of length 6, which includes the object category, bounding box coordinates, and confidence score. For face recognition, the model determines whether the face image is the same person and outputs the label $y \in \mathbb{R}$.

4.3. Results

We selected Boundary Attack and Surfree algorithms and conducted comparative experiments with our attack methods. We attacked the current 7 mainstream image classification models ResNet34, ResNet152, VGG16, DenseNet201, EfficientNet-b0, Inception-v3, and AlexNet, and measured the number of queries to the black-box model when 100 adversarial samples were successfully generated under the conditions of L_∞ norm of 10/255, 20/255, 30/255 and L_2 norm of 15.21, 30.42, and 45.64. The experimental results are shown in Table 2. The visualization of adversarial examples is shown in Figure 3. Table 2 shows the average number of queries required to generate 100 adversarial samples for different models under various perturbation constraints. Our method (FeatureBA) consistently outperforms Boundary Attack and Surface, demonstrating its efficiency in generating adversarial examples.

Whether under the L_2 norm or the L_∞ norm, for the selected classification model, the order of magnitude of the query of our attack method is significantly smaller than that of Boundary Attack and Surfree. For example, under the condition of L_∞ norm of 20/255, the average number of queries of the black-box by Boundary Attack against the DenseNet201 model is as high as 10460.76 times, and the number of queries of Surfree also reaches

Table 2: The average number of queries of our algorithm and the comparison algorithm.

L_∞ 10/255 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	7443.12	10378.03	6091.38	9504.29	8940.2	5269.26	2826.96
Surfree	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00
FeatureBA(Ours)	**37.14**	**117.26**	**54.43**	**152.09**	**147.76**	**112.85**	**19.46**
L_∞ 20/255 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	8331.92	8650.54	7022.26	10460.76	10013.96	6297.57	3020.03
Surfree	5000.00	5000.00	3042.87	5000.00	5000.00	2908.36	5000.00
FeatureBA(Ours)	**13.18**	**45.05**	**7.21**	**70.94**	**38.24**	**57.75**	**12.68**
L_∞ 30/255 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	6061.55	7278.09	5377.12	9258.11	8505.13	4787.53	3865.32
Surfree	2398.12	5000.00	1272.49	5000.00	3166.35	1341.19	1064.45
FeatureBA(Ours)	**8.82**	**25.73**	**6.16**	**27.42**	**19.53**	**29.61**	**6.73**
L_2 15.21 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	5302.26	6232.55	2920.70	7263.15	6780.65	4876.74	2984.56
Surfree	727.18	1854.37	585.47	1878.71	1304.46	630.19	798.54
FeatureBA(Ours)	**13.08**	**48.21**	**12.02**	**84.76**	**58.12**	**64.83**	**8.11**
L_2 30.42 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	2196.02	2807.74	1203.18	3362.74	3463.76	2155.22	1043.34
Surfree	346.30	658.08	113.91	965.46	477.78	273.33	32.57
FeatureBA(Ours)	**6.32**	**19.32**	**5.39**	**20.96**	**13.65**	**17.42**	**2.52**
L_2 45.64 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
Boundary	1211.90	1343.32	570.68	1921.52	1856.78	1218.64	410.36
Surfree	45.46	275.54	57.78	291.25	181.14	65.15	31.11
FeatureBA(Ours)	**6.11**	**12.19**	**4.02**	**16.74**	**7.96**	**10.17**	**4.05**

5000 times, which is the maximum number of attacks we set. In contrast, our attack method only needs 70.94 queries; under the condition of L_∞ norm 20/255 perturbation constraint, the attack on ResNet34 is much smaller than that of Boundary Attack. Our attack method reduces the number of queries to the black box by 99.84%, which is 99.73% lower than that of Surfree. Com-

pared with the query times of ResNet152 and ResNet34, the query times of ResNet152 are higher than those of ResNet34 regardless of the attack method. The more complex the model architecture is, the more queries are required for the black-box model. Our method significantly reduces the number of queries compared to Boundary Attack and Surface, as shown in Table 2. This is because FeatureBA leverages intermediate layer features of surrogate models, which guide the attack process more effectively. The reduction in queries has practical implications for real-world adversarial attacks, where query limits are often imposed. The query counts for ResNet152 are higher than those for ResNet34, indicating that more complex model architectures require more queries to generate adversarial examples. This suggests that our method is particularly effective for simpler models, but may require further optimization for more complex architectures.

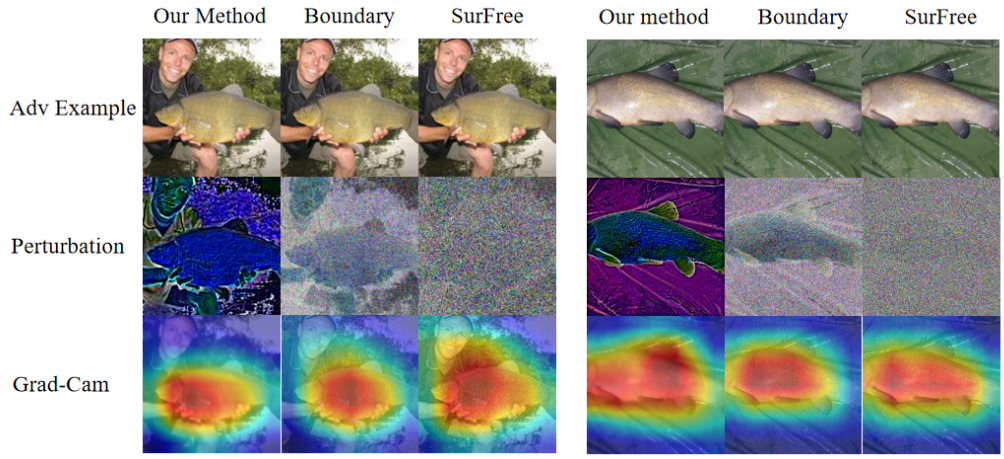

Figure 3: Adversarial examples, perturbation noise, and attention generated by our algorithm and comparison algorithms on image classification.

4.4. More Experimental Results

In image classification, we selected Boundary Attack and Surfree algorithms and conducted comparative experiments with our attack method. We attacked the current 7 mainstream image classification models ResNet34, ResNet152, VGG16, DenseNet201, Efficientnet-b0, Inception-v3, and Alexnet, and measured the number of queries on the black box model. In order to more intuitively show the experimental results, we show in Figure 4 - 10, the

number of queries for each of the 20 adversarial examples that were successfully attacked under the L_2 norm condition of 30.42.

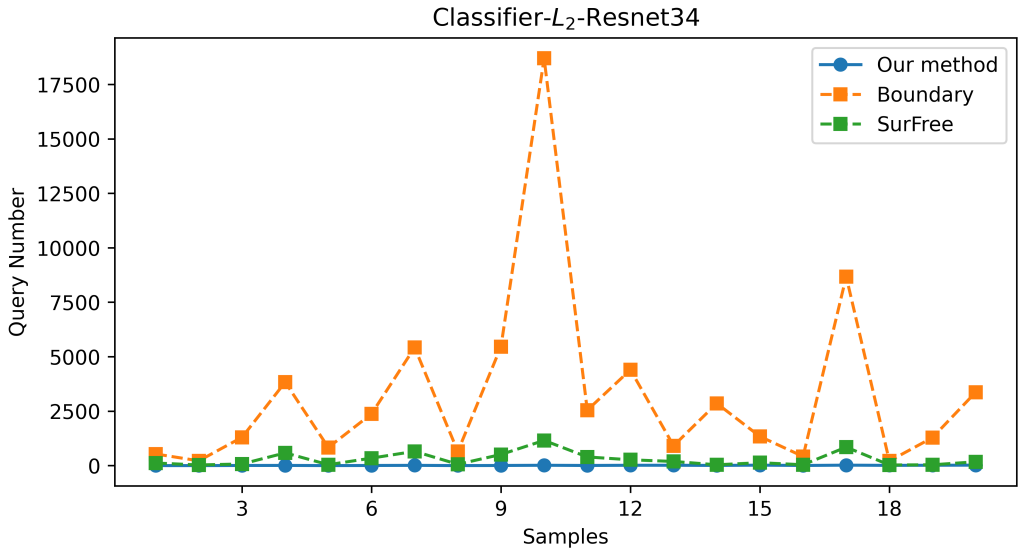

Figure 4: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

Figure 11 shows the adversarial examples, noise, and attention maps of our algorithm and the comparison algorithm when attacking the black-box model in the image classification task. All attack methods successfully attack the image. Given adversarial examples from different attack methods, the attention map is calculated according to different target models. Adversarial examples are successfully attacked images randomly selected from the test set.

5. Experiments on Object Detection and Face Recognition

In this section, we evaluate the effectiveness of our method on object detection and face recognition, and compare it with other algorithms.

5.1. Object Detection Dataset and Victim Model

We choose the widely used COCO 2017 dataset (Lin et al., 2014) and various model architectures and weights pre-trained on the COCO 2017 dataset to evaluate the effectiveness of our attack method. COCO (Common Objects

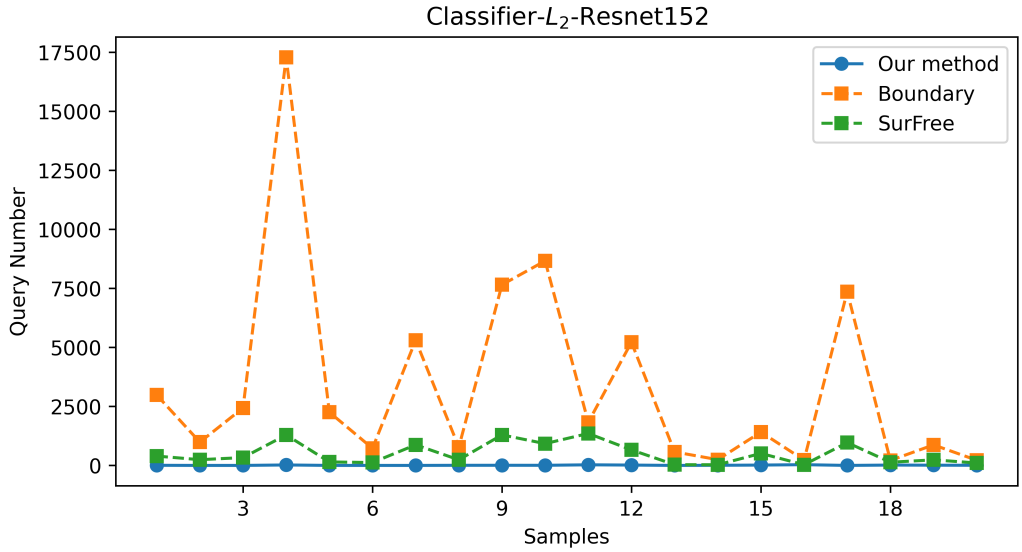

Figure 5: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

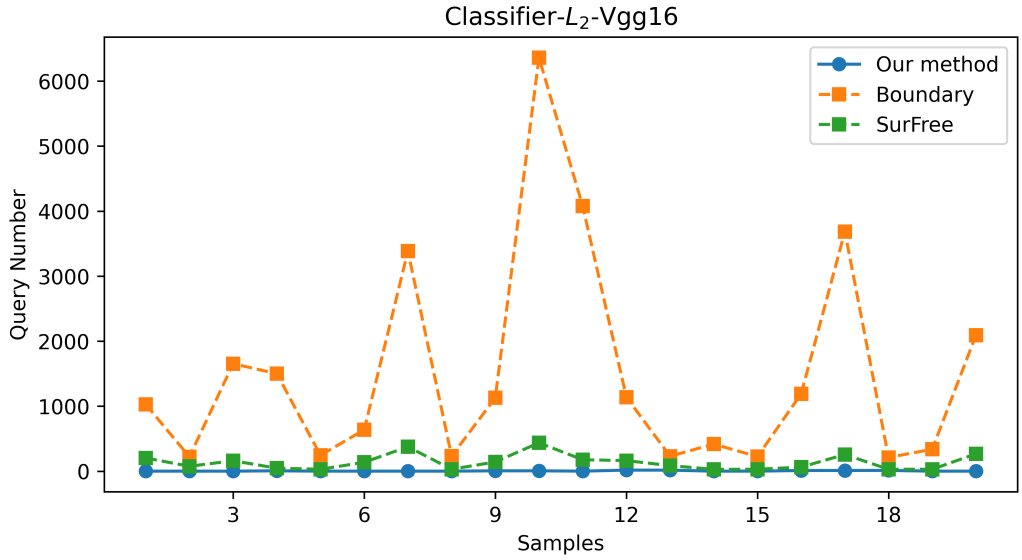

Figure 6: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

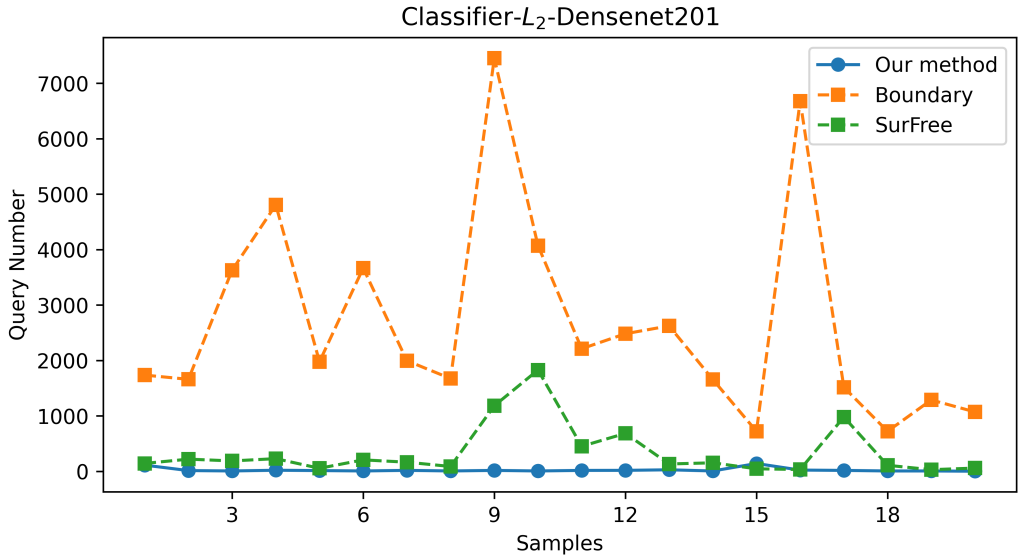

Figure 7: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

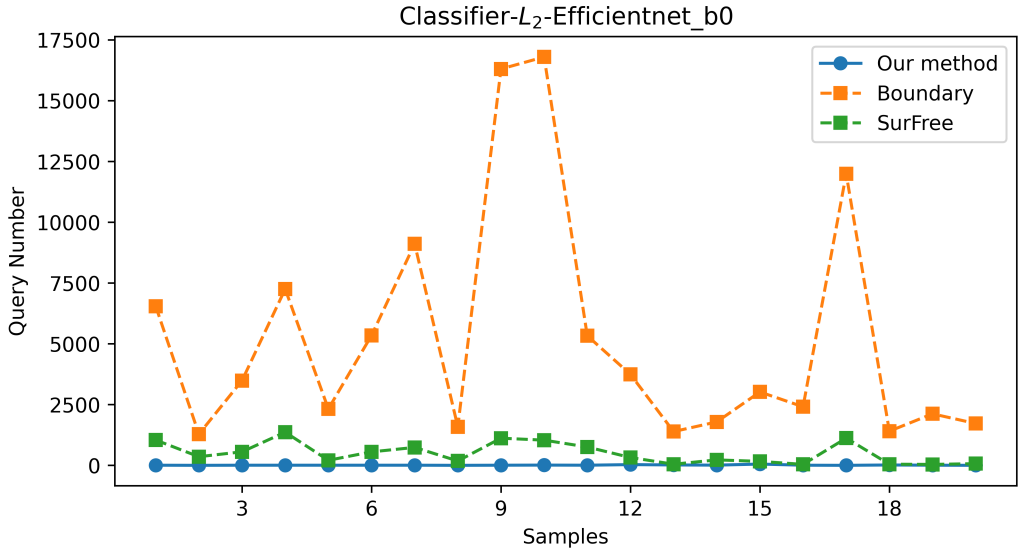

Figure 8: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

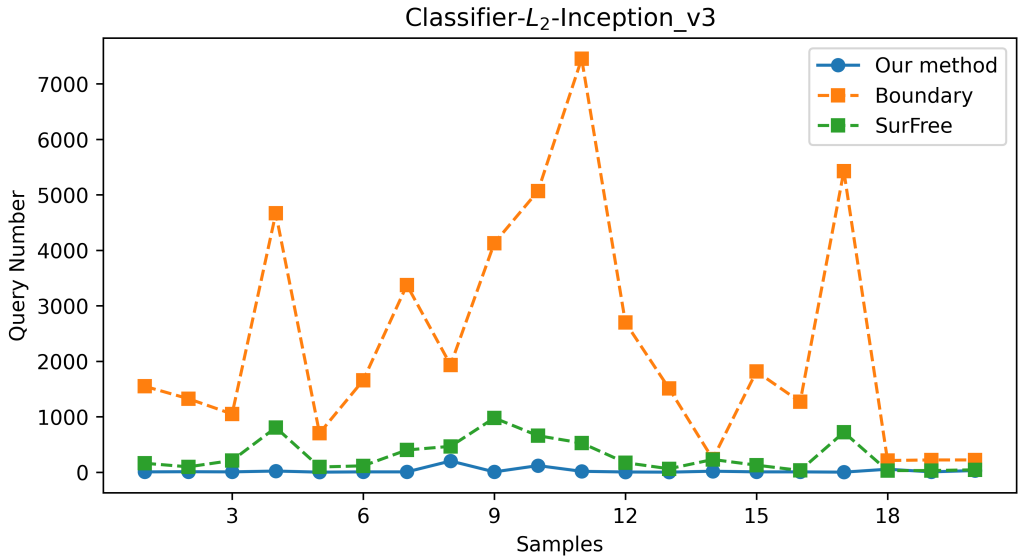

Figure 9: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

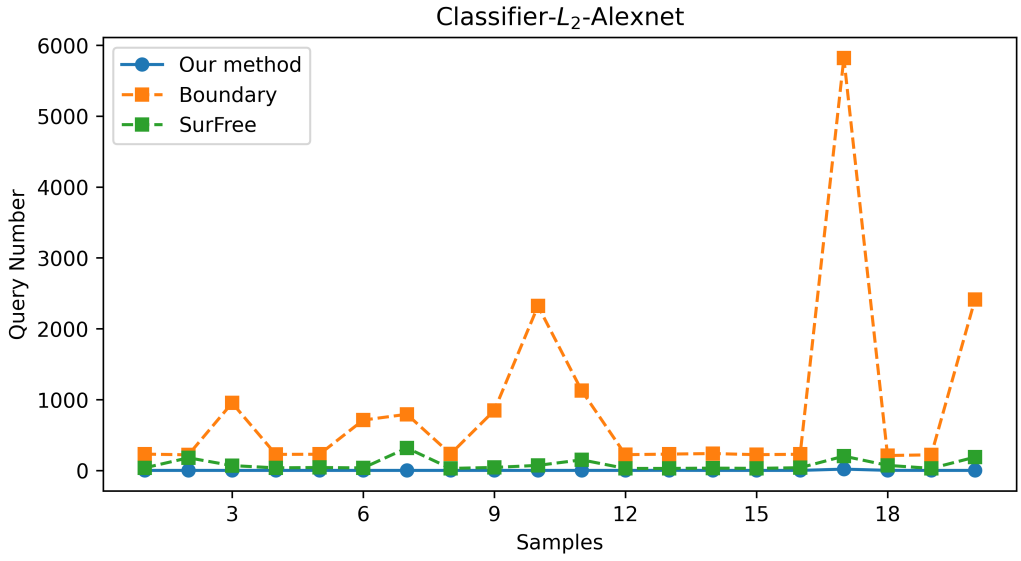

Figure 10: In the image classification task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 30.42

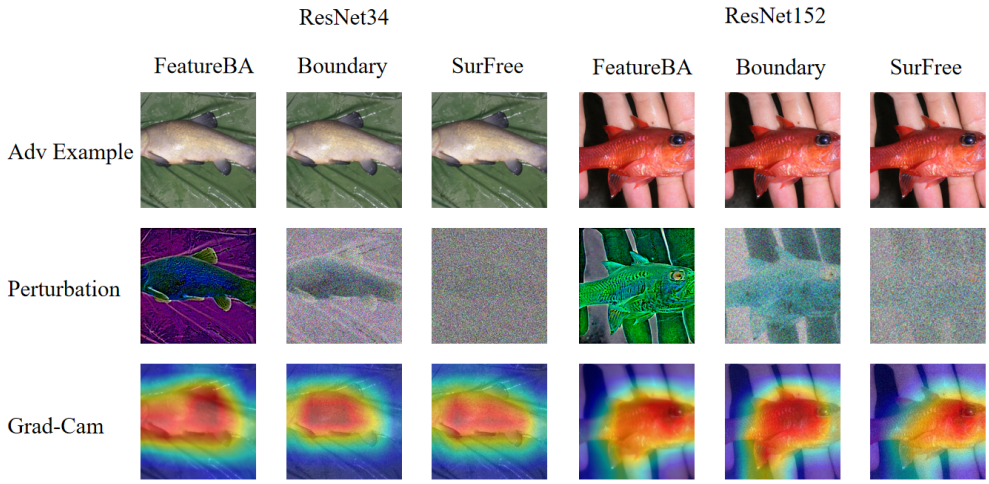

Figure 11: In the image classification task, adversarial examples, noise and attention maps of our algorithm and the comparison algorithm when attacking the black-box model. All methods are under the constraint of L_2 norm bound 30.42.

in Context) is a widely used computer vision dataset designed to provide rich contextual information, including a large number of multi-category objects, dense annotations, and multiple task scenarios. This dataset was created by Microsoft and is widely used for tasks such as object detection, segmentation, and key point detection. The COCO dataset contains about 330,000 images, of which about 200,000 images are annotated, including objects in 80 categories, and 360,000 object instances for training models. We evaluate the attack performance on the Pascal VOC 2007 (Everingham et al., 2010) dataset. The pre-trained model we use is trained on COCO because COCO contains 80 object categories (a superset of the 20 categories of the VOC dataset). When testing on the VOC dataset, we only return objects that exist in VOC. The VOC dataset contains 20 categories. We follow the settings in (Cai et al., 2023, 2022) and randomly select 500 images containing multiple (2-6) objects from the VOC 2007 test set. And these images can be successfully recognized by the pre-trained model. We choose the widely used YOLO series model as the victim model. These victim models are pre-trained on the VOC dataset. We choose YOLOv4 as the victim model. In order to create a strict black-box environment, we choose YOLOv3, which is different from the victim model, as a surrogate model.

We use the perturbation L_2 norm and the number of queries as evaluation

indicators. We give the same L_2-norm perturbation budgets to the attack methods, and the number of queries for these attack methods determines the effectiveness of the attack. The goal of all attack algorithms is to minimize the average number of queries required to obtain 100 successful adversarial examples. The perturbation budgets we give for the L_2 norm are 28.25, 56.51, and 84.76. The image size is [416, 416, 3].

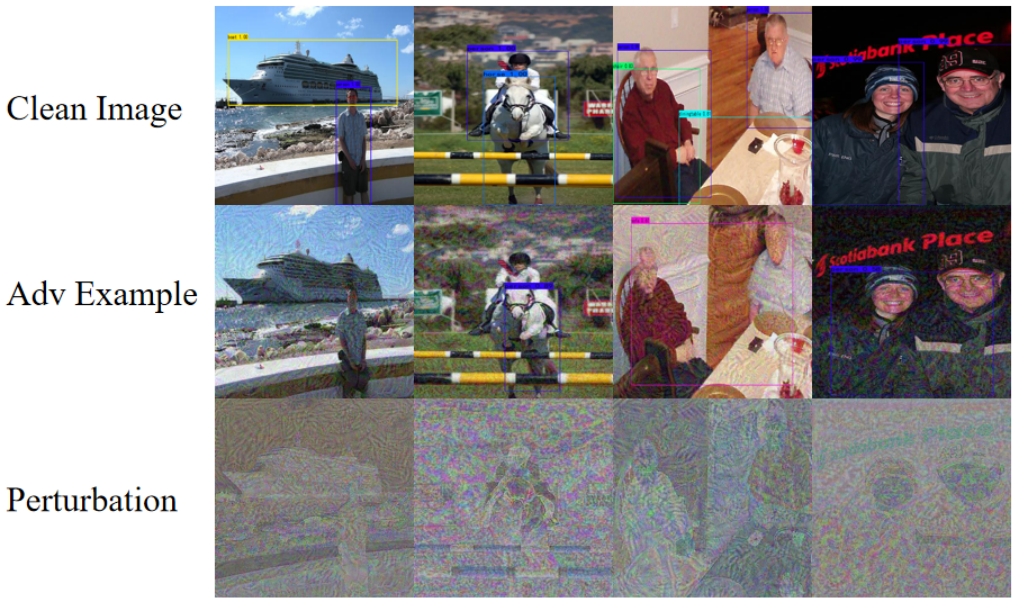

Figure 12: Clean image, adversarial examples and perturbation noise generated by our algorithm for object detection.

5.2. Face Recognition Dataset and Victim Model

We use the LFW (Huang et al., 2008): Labeled Faces in the Wild dataset to verify the effectiveness of our attack method. LFW is a face recognition dataset containing 13,233 images collected from 5,749 websites of different topics. Each image is annotated with the real identity of the face. The LFW dataset is mainly used to evaluate the accuracy of facial recognition systems and is often used to benchmark the performance of models on real-world datasets. We randomly sampled 1,000 pairs of faces of different identities to evaluate the attack performance.

The FaceNet (Schroff et al., 2015) face recognition model is a face recognition algorithm published by Google in 2015. It uses the characteristics that

the same face has high cohesion in photos of different angles and postures, and different faces have low coupling. It also proposes the use of CNN (Convolutional Neural Network is a class of deep learning models designed for processing structured grid data, such as images, by leveraging convolutional layers to automatically extract hierarchical features.) + Triplet Mining to achieve face recognition tasks, and the accuracy rate on the LFW dataset reached 99.63%.

We chose the FaceNet face recognition model with MobileNet V1 as the backbone network as the surrogate model, and downloaded its weight obtained by training on the CASIA-WebFace face dataset. We selected the FaceNet face recognition model with the backbone network of Inception-ResNet V1 as the victim model for testing, and downloaded the weight obtained by training it on the CASIA-WebFace face dataset.

For the trained face recognition model, we need to determine its threshold for judging whether two face images belong to the same person. The discrimination thresholds of the FaceNet face recognition models with the backbone networks of MobileNet V1 and Inception-ResNet V1 were measured, and the LFW dataset was selected as the test dataset. Under the condition of FAR<0.001, the threshold T with the largest TAR was used as the discrimination threshold of our model. After the measurement, the discrimination threshold result of the FaceNet model with the backbone network of MobileNet V1 was 1.150, and the success rate of its model discrimination was 98.28%; the discrimination threshold result of the FaceNet model with the backbone network of Inception-ResNet V1 was 1.170, and the success rate of its model discrimination was 98.73%. The performance of the two backbone network models is basically similar.

We use the perturbation L_2 norm and the number of queries as evaluation indicators. We give the same L_2-norm perturbation budgets to the attack methods, and the number of queries for these attack methods determines the effectiveness of the attack. The goal of all attack algorithms is to minimize the average number of queries required to obtain 100 successful adversarial samples.

The perturbation budgets we give are L_2 norms of 10.86, 21.73, and 32.60. When the image size is [160, 160, 3], these L_2 norm perturbation budgets correspond to the perturbation levels used for evaluating attack methods.

Table 3: The average query count of our algorithm and the comparison algorithm in object detection and face recognition respectively

Task	L_2 bound	Boundary	Ours
Object Detection	28.25	6153.74	**9.66**
	56.51	3356.76	**3.47**
	84.76	2117.56	**3.48**
Face Recognition	10.86	15572.4	**348.44**
	21.73	9236.7	**341.3**
	32.60	3002	**328.93**

5.3. Results

We have expanded the application scenarios of the attack method. Since the accuracy of target detection is also highly correlated with the classification model, we apply our attack method to the target detection model. We attack the target detection model YOLO v4 and measure the number of queries to the black-box model when 100 adversarial samples are successfully generated under the conditions of L_2 norm of 28.25, 56.51, and 84.76. The experimental results are shown in Table 3. Whether under the L_2 norm conditions, for the selected target detection model, the order of magnitude of our attack method's queries is significantly smaller than that of Boundary Attack; under the condition of L_2 norm of 56.51, our improved Boundary Attack for target detection needs 3356.76 queries, and our attack method only needs 3.47 queries, and the number of queries has dropped by 99.90%. Visualization of adversarial examples is shown in Figure 12.

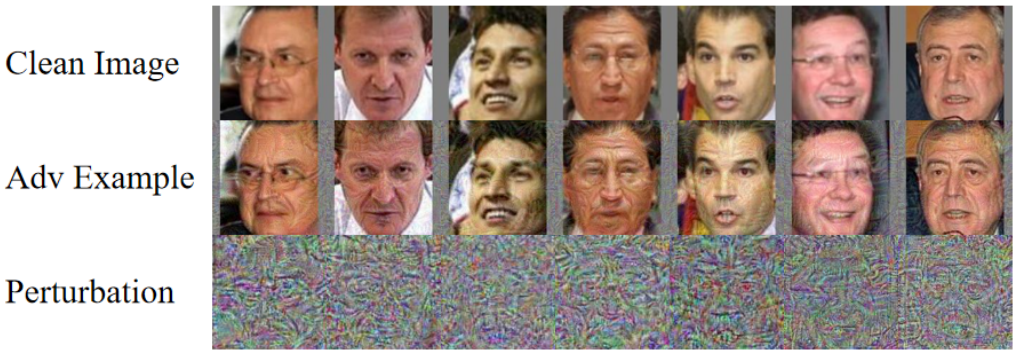

Figure 13: Clean image, adversarial examples and perturbation noise generated by our algorithm for face recognition.

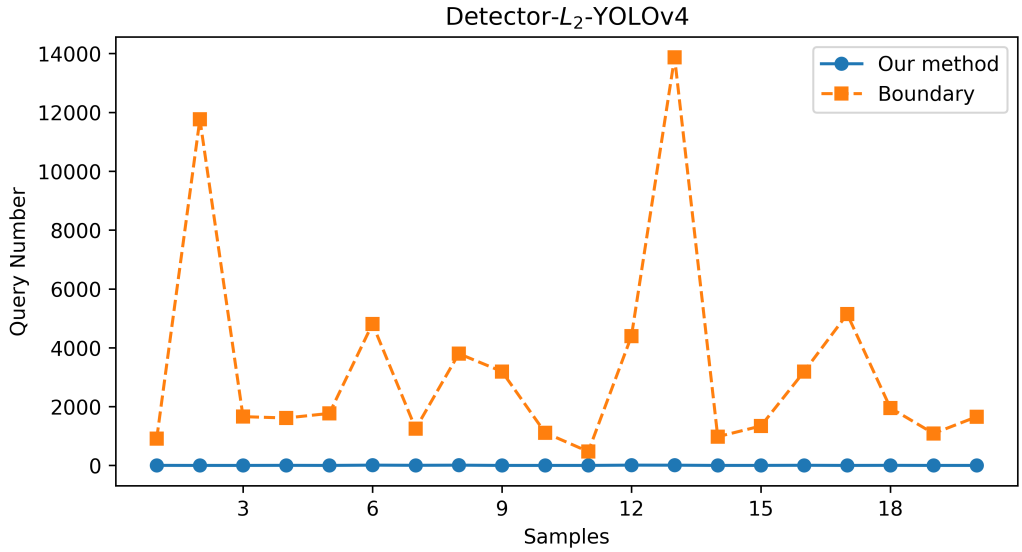

Figure 14: In the object detection task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 56.51

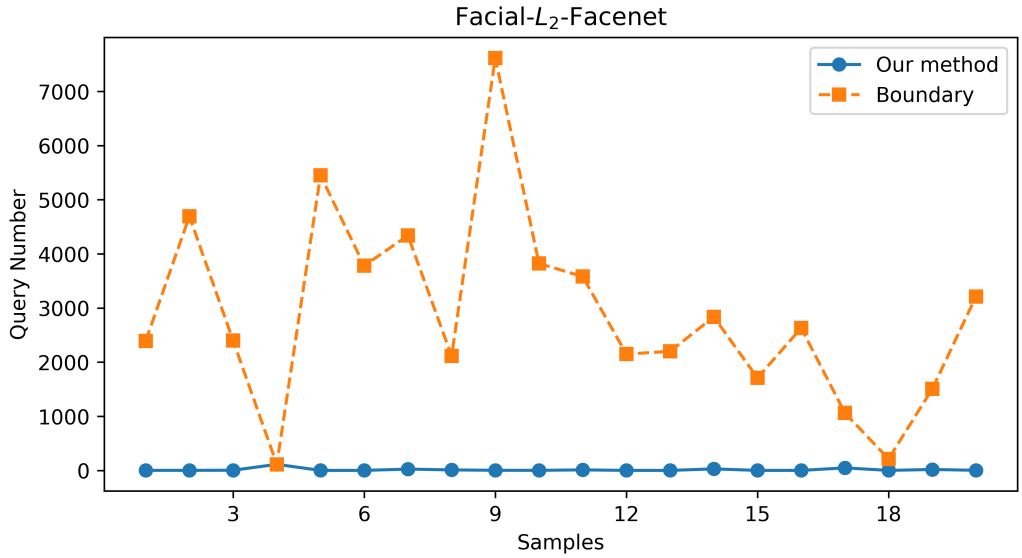

Figure 15: In the face recognition task, the number of queries per sample of our method and the comparison method under the constraint of L_2 norm bound 21.73

In addition, we also applied this method to the field of face recognition, attacking the face recognition model facenet, and measured the number of queries to the black-box model when the L_2 norm is 10.86, 21.73, and 32.60, and 100 adversarial samples are successfully generated. The experimental results are shown in Table 3 and visualization in Figure 13. Whether under the L_2 norm conditions, for the selected face recognition model, the number of queries of our attack method is significantly smaller than that of Boundary Attack. Under the condition of L_2 norm of 10.86, our improved Boundary Attack for face recognition needs 15572.4 queries, while our attack method only needs 348.41 queries, a decrease of 97.76%. However, since the face recognition task is a vector distance comparison, it requires more queries than image classification and object detection tasks. In object detection and face recognition, we compare with BoundaryAttack algorithm. we show in Figure 14 and 15, the number of queries for each of the 20 adversarial examples that were successfully attacked. For the selected victim model, the order of magnitude of queries of our attack method is significantly smaller than that of Boundary Attack and Surfree.

6. Ablation Study

The key of the proposed FeatureBA is to perform random masking and aggregate gradients for the intermediate layer features Z, which significantly improves the transferability, thus reducing the number of queries. As shown in the above results. To highlight the contribution of the algorithm, we conducted an ablation study to compare the performance of the algorithm without aggregated gradients and without considering gradients. We constructed three objective functions as shown below, where L_1 is equivalent to the loss we proposed (Equation 7), $L_2 = \Delta_i \odot Z_i$ is the use of non-aggregated gradients, and $L_3 = Z_i$ uses the feature map directly as the loss function.

$$L_1 = (\frac{1}{C}\sum_{n=1}^{N} \Delta_i^{Z_i \odot M_p^n}) \odot Z_i \tag{7}$$

As shown in Figure 16, in all cases, the proposed loss L_1 outperforms other losses, indicating the effectiveness of the proposed FeatureBA. Recall the ablation experiment in the main text, we conducted an ablation study to compare the performance of the algorithm without aggregated gradients and without considering gradients. We present the average number of queries for

Table 4: Comparison of the average query times of the three losses in the ablation study.

L_∞ 20/255 average query count							
Method	ResNet34	ResNet152	VGG16	DenseNet201	Efficientnet-b0	Inception-v3	Alexnet
L_1	**12.31**	**45.05**	**7.21**	**70.94**	**38.24**	**57.75**	**12.68**
L_2	14.12	52.21	8.23	75.67	46.67	58.48	14.71
L_3	13.01	53.18	8.74	71.68	44.55	60.19	14.99

attacking each black-box model using Resnet18 as a surrogate model with L_1, L_2, and L_3 losses in Table 4. The goal of all attack algorithms is to minimize the average number of queries required to obtain 100 successfully attacked adversarial examples. The perturbation limit for ablation study is the L_∞ norm bound 20/255.

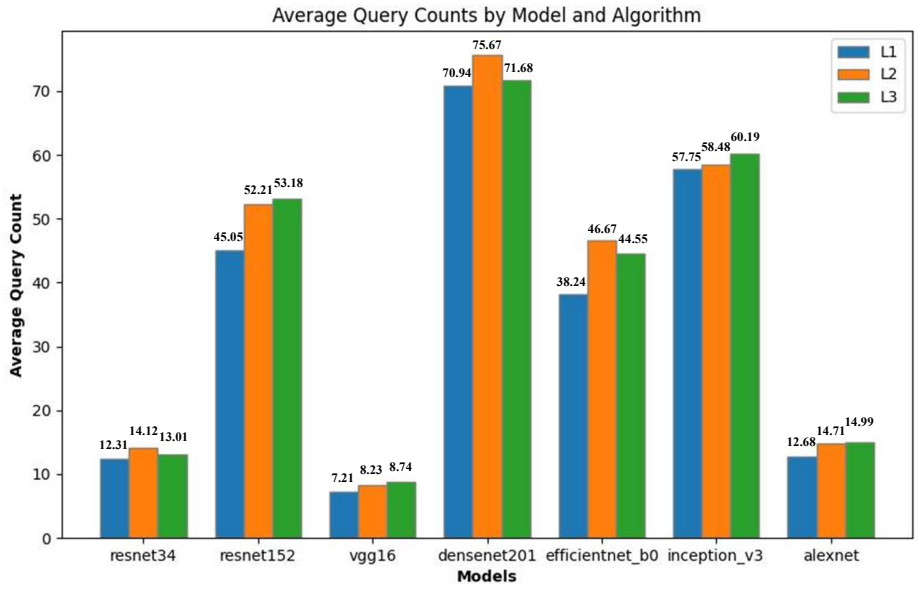

Figure 16: The impact of the aggregated gradient of the intermediate layer features on the number of queries. L_1 is equivalent to the loss we proposed, L_2 uses non-aggregated gradients, and L_3 does not use gradients.

7. Discussion

7.1. Theoretical and Practical Implications

The proposed method introduces a novel integration of surrogate model features into decision-based attacks, providing theoretical insights into transferability and robustness. Practically, the method's applicability to diverse tasks enhances its value for real-world adversarial testing.

7.2. Limitations

Despite its advantages, the method's reliance on surrogate models may limit its effectiveness against highly dissimilar architectures. Future work should explore adaptive mechanisms for cross-architecture transferability.

8. Conclusion

In this study, we proposed a novel decision-based black-box adversarial attack method, FeatureBA, which leverages intermediate layer features of surrogate models to achieve competitive results without requiring additional priors or heuristics. Our method not only outperforms existing algorithms in image classification but also generalizes to object detection and face recognition tasks. This work lays a new foundation for combining decision-based (hard label) attacks with transfer-based methods, offering practical implications for real-world adversarial attack scenarios and theoretical contributions to the field of adversarial machine learning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. U2336201).

References

Brendel, W., Rauber, J., Bethge, M., 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, in: International Conference on Learning Representations. doi:`https://doi.org/10.15760/honors.1581`.

Brunner, T., Diehl, F., Le, M.T., Knoll, A., 2019. Guessing smart: Biased sampling for efficient black-box adversarial attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4958–4966. doi:https://doi.org/10.1109/iccv.2019.00506.

Cai, Z., Tan, Y., Asif, M.S., 2023. Ensemble-based blackbox attacks on dense prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4045–4055. doi:https://doi.org/10.1109/cvpr52729.2023.00394.

Cai, Z., Xie, X., Li, S., Yin, M., Song, C., Krishnamurthy, S.V., Roy-Chowdhury, A.K., Asif, M.S., 2022. Context-aware transfer attacks for object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 149–157. doi:https://doi.org/10.1609/aaai.v36i1.19889.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks, in: 2017 Ieee Symposium on Security and Privacy (Sp), pp. 39–57. doi:https://doi.org/10.1109/sp.2017.49.

Chen, J., Gu, Q., 2020. Rays: A ray searching method for hard-label adversarial attack, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1739–1747. doi:https://doi.org/10.1145/3394486.3403225.

Chen, J., Jordan, M.I., Wainwright, M.J., 2020. Hopskipjumpattack: A query-efficient decision-based attack, in: 2020 ieee symposium on security and privacy (sp), IEEE. pp. 1277–1294. doi:https://doi.org/10.1109/sp40000.2020.00045.

Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J., 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM workshop on artificial intelligence and security, pp. 15–26. doi:https://doi.org/10.1145/3128572.3140448.

Cheng, M., Le, T., Chen, P.Y., Zhang, H., Yi, J., Hsieh, C.J., 2018. Query-efficient hard-label black-box attack: An optimization-based approach, in: International Conference on Learning Representations. doi:https://doi.org/10.1109/wacv57701.2024.00394.

Cheng, S., Dong, Y., Pang, T., Su, H., Zhu, J., 2019. Improving black-box adversarial attacks with a transfer-based prior. Advances in neural information processing systems 32. doi:https://doi.org/10.1117/12.3013308.a4587f73-c566-ee11-a99c-005056914f1c.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255. doi:https://doi.org/10.1109/cvpr.2009.5206848.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J., 2018. Boosting adversarial attacks with momentum, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193. doi:https://doi.org/10.1109/cvpr.2018.00957.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J., 2019. Efficient decision-based black-box adversarial attacks on face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7714–7722. doi:https://doi.org/10.1109/cvpr.2019.00790.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88, 303–338. doi:https://doi.org/10.1007/s11263-009-0275-4.

Fang, J., Jiang, Y., Jiang, C., Jiang, Z.L., Liu, C., Yiu, S.M., 2024. State-of-the-art optical-based physical adversarial attacks for deep learning computer vision systems. Expert Systems with Applications 250, 123761. doi:https://doi.org/10.1016/j.eswa.2024.123761.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. Proc. Int. Conf. Learn. Representations doi:https://doi.org/10.5220/0006123702260234.

Guo, Y., Yan, Z., Zhang, C., 2019. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. Advances in Neural Information Processing Systems 32. doi:https://doi.org/10.5220/0012359900003654.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. doi:https://doi.org/10.1109/cvpr.2016.90.

Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E., 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, in: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. doi:https://doi.org/10.1117/12.2080393.

Li, H., Xu, X., Zhang, X., Yang, S., Li, B., 2020a. Qeba: Query-efficient boundary-based blackbox attack, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1221–1230. doi:https://doi.org/10.1109/cvpr42600.2020.00130.

Li, J., Ji, R., Chen, P., Zhang, B., Hong, X., Zhang, R., Li, S., Li, J., Huang, F., Wu, Y., 2021. Aha! adaptive history-driven attack for decision-based black-box models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16168–16177. doi:https://doi.org/10.1109/iccv48922.2021.01586.

Li, M., Deng, C., Li, T., Yan, J., Gao, X., Huang, H., 2020b. Towards transferable targeted attack, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 641–649. doi:https://doi.org/10.1109/cvpr42600.2020.00072.

Li, Q., Guo, Y., Zuo, W., Chen, H., 2024. Improving adversarial transferability via intermediate-level perturbation decay. Advances in Neural Information Processing Systems 36. doi:https://doi.org/10.2139/ssrn.4954123.

Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E., 2019. Nesterov accelerated gradient and scale invariance for adversarial attacks, in: International Conference on Learning Representations. doi:https://doi.org/10.1109/cac59555.2023.10452055.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755. doi:https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, J., Jin, H., Xu, G., Lin, M., Wu, T., Nour, M., Alenezi, F., Alhudhaif, A., Polat, K., 2023. Aliasing black box adversarial attack with joint self-attention distribution and confidence probability. Expert Systems with Applications 214, 119110. doi:https://doi.org/10.1016/j.eswa.2022.119110.

Liu, Y., Chen, M., Zhu, C., Liang, H., Chen, J., 2025. You cannot handle the weather: Progressive amplified adverse-weather-gradient projection adversarial attack. Expert Systems with Applications 266, 126143. doi:https://doi.org/10.1016/j.eswa.2024.126143.

Liu, Y., Moosavi-Dezfooli, S.M., Frossard, P., 2019. A geometry-inspired decision-based attack, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4890–4898. doi:https://doi.org/10.1109/iccv.2019.00499.

Ma, C., Guo, X., Chen, L., Yong, J.H., Wang, Y., 2021. Finding optimal tangent points for reducing distortions of hard-label attacks. Advances in Neural Information Processing Systems 34, 19288–19300. doi:https://doi.org/10.1016/j.jco.2008.10.001.

Maho, T., Furon, T., Le Merrer, E., 2021. Surfree: a fast surrogate-free black-box attack, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10430–10439. doi:https://doi.org/10.1109/cvpr46437.2021.01029.

Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P., 2016. Deepfool: a simple and accurate method to fool deep neural networks, in: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 2574–2582. doi:https://doi.org/10.1109/cvpr.2016.282.

Muthalagu, R., Malik, J., Pawar, P.M., 2025. Detection and prevention of evasion attacks on machine learning models. Expert Systems with Applications 266, 126044. doi:https://doi.org/10.1016/j.eswa.2024.126044.

Rahmati, A., Moosavi-Dezfooli, S.M., Frossard, P., Dai, H., 2020. Geoda: a geometric framework for black-box adversarial attacks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8446–8455. doi:https://doi.org/10.1109/cvpr42600.2020.00847.

Ran, Y., Zhang, A.X., Li, M., Tang, W., Wang, Y.G., 2025. Black-box adversarial attacks against image quality assessment models. Expert Systems with Applications 260, 125415. doi:https://doi.org/10.1016/j.eswa.2024.125415.

Reza, M.F., Rahmati, A., Wu, T., Dai, H., 2023. CGBA: Curvature-aware Geometric Black-box Attack, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 124–133. doi:https://doi.org/10.1109/iccv51070.2023.00018.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823. doi:https://doi.org/10.1109/cvpr.2015.7298682.

Shen, Z., Li, Y., 2025. Temporal characteristics-based adversarial attacks on time series forecasting. Expert Systems with Applications 264, 125950. doi:https://doi.org/10.1016/j.eswa.2024.125950.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. Proceedings of International Conference on Learning Representations (ICLR) doi:https://doi.org/10.1109/cvpr.2016.182.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 doi:https://doi.org/10.1007/springerreference_64406.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826. doi:https://doi.org/10.1109/cvpr.2016.308.

Tashiro, Y., Song, Y., Ermon, S., 2020. Diversity can be transferred: Output diversification for white-and black-box attacks. Advances in neural information processing systems 33, 4536–4548. doi:https://doi.org/10.1109/ijcnn52387.2021.9533772.

Wang, X., He, K., 2021. Enhancing the transferability of adversarial attacks through variance tuning, in: Proceedings of the IEEE/CVF conference

on computer vision and pattern recognition, pp. 1924–1933. doi:https://doi.org/10.1109/cvpr46437.2021.00196.

Wang, X., He, X., Wang, J., He, K., 2021a. Admix: Enhancing the transferability of adversarial attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16158–16167. doi:https://doi.org/10.1109/iccv48922.2021.01585.

Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., Ren, K., 2021b. Feature Importance-aware Transferable Adversarial Attacks, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7619–7628. doi:https://doi.org/10.1109/iccv48922.2021.00754.

Wu, W., Su, Y., Lyu, M.R., King, I., 2021. Improving the transferability of adversarial samples with adversarial transformations, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9024–9033. doi:https://doi.org/10.1109/cvpr46437.2021.00891.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L., 2019. Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2730–2739. doi:https://doi.org/10.1109/cvpr.2019.00284.

Yang, J., Jiang, Y., Huang, X., Ni, B., Zhao, C., 2020. Learning black-box attackers with transferable priors and query feedback. Advances in Neural Information Processing Systems 33, 12288–12299. doi:https://doi.org/10.1007/s10994-022-06251-3.

Zhang, Y., Tan, Y.a., Chen, T., Liu, X., Zhang, Q., Li, Y., 2022. Enhancing the transferability of adversarial examples with random patch., in: IJCAI, pp. 1672–1678. doi:https://doi.org/10.24963/ijcai.2022/233.