electronics

MDPI

Article

Black-Box Boundary Attack Based on Gradient Optimization

Yuli Yang, Zishuo Liu, Zhen Lei, Shuhong Wu and Yongle Chen *

College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China; yangyuli@tyut.edu.cn (Y.Y.); liuzishuo0651@link.tyut.edu.cn (Z.L.); leizhen0667@link.tyut.edu.cn (Z.L.); wushuhong@tyut.edu.cn (S.W.)
* Correspondence: chenyongle@tyut.edu.cn

Abstract: Deep neural networks have gained extensive applications in computer vision, demonstrating significant success in fundamental research tasks such as image classification. However, the robustness of these networks faces severe challenges in the presence of adversarial attacks. In real-world scenarios, addressing hard-label attacks often requires the execution of tens of thousands of queries. To combat these challenges, the Black-Box Boundary Attack leveraging Gradient Optimization (GOBA) has been introduced. This method employs a binary search strategy to acquire an initial adversarial example with significant perturbation. The Monte Carlo algorithm is utilized to estimate the gradient of the sample, facilitating iterative movement along the estimated gradient and the direction of the malicious label. Moreover, query vectors positively correlated with the gradient are extracted to construct a sampling space with an optimal scale, thereby enhancing the efficiency of the Monte Carlo algorithm. Experimental evaluations were conducted using the HSJA, QEBA, and NLBA attack methodologies on the ImageNet, CelebA, and MNIST datasets, respectively. The results indicate that, under the constraint of 3 k query times, the GOBA, compared to other methods, can, on average, reduce perturbation (L_2 distance) by 55.74% and simultaneously increase the attack success rate by an average of 13.78%.

Keywords: deep learning network; image classification; hard-label attack; adversarial samples

Citation: Yang, Y.; Liu, Z.; Lei, Z.; Wu, S.; Chen, Y. Black-Box Boundary Attack Based on Gradient Optimization. *Electronics* **2024**, *13*, 1009. https://doi.org/10.3390/electronics13061009

Academic Editor: Hung-Yu Chien

Received: 5 February 2024
Revised: 5 March 2024
Accepted: 5 March 2024
Published: 7 March 2024

1. Introduction

In recent years, significant strides have been made in the field of computer vision through the application of deep learning, particularly in tasks such as image classification [1–5], object detection [6–10], and image segmentation [11]. Object detection algorithms, in particular, have seen widespread use in critical areas such as autonomous driving [12,13] and industrial inspection [14]. Given the paramount importance of these applications to security, there is an escalated demand for algorithms to exhibit heightened robustness and security.

However, the susceptibility of deep learning to adversarial attacks [15], where even subtle perturbations to input images can drastically alter model outputs, has become a prominent concern. Consequently, there has been a growing emphasis on research regarding adversarial attacks and defenses within the context of deep learning [16–19]. Depending on the accessibility of information about the targeted model, adversarial attacks can be broadly categorized into two types. White-box attacks assume complete transparency of the internal structure and parameters of the targeted model, empowering attackers to obtain this information and construct precise adversarial samples. However, white-box scenarios become more challenging when attackers lack access to the model's structure and parameters, leading to a significant reduction in the success rate of attacks. On the other hand, black-box attacks, where attackers lack specific information about the model or training dataset, align more closely with real-world scenarios and play a crucial role in evaluating model robustness in practical applications. Black-box attacks are further divided into non-query attacks [20] and query-based attacks. Query-based attacks can be subcategorized into score-based and hard-label attacks [21]. Score-based attacks involve

attackers having access to complete data labels and corresponding probabilities, while hard-label attacks limit attackers to obtaining single-label model outputs, adding substantial difficulty to the attack by restricting access to rich information. This paper focuses on hard-label attacks, with the aim of enhancing the success rate of hard-label black-box attacks while adhering to a constrained query budget.

Presently, adversarial attacks in image classification have received considerable attention, resulting in the emergence of numerous noteworthy attack algorithms. However, existing decision-based approaches have not effectively tackled the challenges associated with reducing query counts and improving attack success rates. Consequently, it becomes relatively straightforward to detect and reject queries near the boundary. Analysis suggests that the most formidable aspect of minimizing query counts lies in gradient computation during the iterative image modification process. The Monte Carlo algorithm introduces probability errors when calculating gradient directions, emphasizing the importance of selecting optimal random variables to minimize their variance or increasing simulation counts under fixed variance conditions. Fewer attack attempts are more aligned with the practical scenario of hard-label problems.

Addressing these challenges, this paper introduces the Black-Box Boundary Attack based on Gradient Optimization (GOBA), a novel algorithm designed for black-box attacks. The GOBA introduces a low-dimensional noise history space to effectively approximate decision boundaries, capitalizing on the inherent smoothness of these boundaries. By constructing a vector distribution sampling space, the GOBA outperforms independent sampling directly in the original space, showcasing superior performance in enhancing gradient accuracy. Therefore, the GOBA achieves higher attack success rates for adversarial samples without the need for external information.

Major Contributions

The main contributions of this paper are as follows:

1. The introduction of an innovative Black-Box Boundary Attack based on Gradient Optimization (GOBA) is presented. By exploiting the flatness characteristics of classification boundaries, perturbation vectors positively correlated with them are extracted to construct a random vector sampling space. This minimizes the necessity for independent sampling, effectively reducing the query budget required for boundary attacks.
2. An optimal dimension subspace is employed to enhance the precision of gradients in high-dimensional spaces, and an optimized traditional binary search boundary method is introduced. This ensures the accurate calculation of the sample's movement step, leading to adversarial samples adhering more closely to the adversarial boundary, consequently increasing the success rate of sample attacks.
3. The proposed method's effectiveness and generality are validated through extensive comparative experiments conducted on the Imagenet, CelebA, and MNIST datasets. Experimental results demonstrate that, compared to existing attack methods, the GOBA not only exhibits robust generality but also demonstrates outstanding performance in black-box attack scenarios.

2. Related Work

Over the past few years, significant efforts have been dedicated to countering adversarial attacks in machine learning models.

Brendel et al. [22] introduced the pioneering hard-label-based attack method, which starts with a substantial adversarial perturbation while simultaneously diminishing perturbations in source and sphere directions. While their approach effectively addresses hard-label attacks, it is limited in efficiency due to reliance on a standard normal distribution. Cheng et al. [23] framed hard-label attacks as an optimization problem involving direction, distance, and gradient estimation of the decision boundary. However, in high-dimensional scenarios, the distance calculation and gradient estimation in their method

require a considerable number of queries. Evolutionary methods [24] have improved variance updates by replacing the normal distribution with custom-modeled weights for variance and pixels post successful sampling. Nevertheless, the use of symbol-agnostic variance introduces instability during the sampling process. Shi et al. [25] introduced the Customized Adversarial Boundary (CAB), utilizing the square of the difference between adversarial samples and source images as variance and cumulative direction in case of mean failure. However, this method did not significantly enhance the attack success rate of the samples.

Cheng et al. [26] introduced the sign function to approximate the direction of ascent or descent (sign-opt), but it still requires a significant number of queries. Liu et al. [27] iteratively extracted noise from a normal distribution to estimate the gradient direction of the decision boundary. Rahmati et al. [28] developed an algorithm to determine the optimal query distribution, yet both methods use random sampling within constrained spaces without a substantial reduction in query count. Guo et al. [29] leveraged the gradient of reference models to construct a search subspace, with the aim of enhancing query efficiency. However, this approach overly relies on external information and model portability, leading to a decrease in attack success rate despite some reduction in query count.

Chen et al. [30] proposed the HopSkipJump attack (HSJA), which directly computes gradient estimation on the decision boundary. While the algorithm achieves accurate gradient calculation through iterative updates of adversarial examples along the decision boundary, the query counts of the HSJA method remain high. The Query-Efficient Boundary-Based Black-box Attack (QEBA) [31] builds on the HSJA by performing gradient calculations through subspace sampling of low-dimensional vectors. Li et al. introduced Nonlinear Gradient Estimation for the Query-Efficient Black-box Attack (NLBA) [32], highlighting the existence of non-linear projections that achieve higher cosine similarity lower bounds. Zhang et al. [33] demonstrated the presence of an optimal scale in the projection space. However, these methods primarily focus on random vector dimension transformations and sampling space sizes' impact on the gradient without a substantial improvement in query efficiency. Maho et al. [34] explored moving in different directions based on the geometric properties of the decision boundary but observed a noticeable decrease in the success rate of the SurFree attack with a limited number of queries. The ongoing challenge of reducing query budget while increasing attack success rate remains a critical challenge in research on hard-label black-box attacks.

3. System Attack Model

3.1. Adversarial Attack

In the black-box attack scenario on neural networks used for image classification, the target model is denoted as $F(X)$, where X represents a specific image.

Adversarial samples are images that have been perturbed with imperceptible noise. This noise is sufficiently small, yet it causes the image to be misclassified into a malicious label y_{tgt} by the target model $F(X)$, as illustrated in Equation (1), where ρ represents the perturbation noise to be added, and X_{src} is the source image.

$$\begin{gathered} min\|\rho\|_2 \; s.t. \; F(X_{src}+\rho) = y_{tgt} \\ X_{src}+\rho \in [0-255]^n, \; \rho \to 0 \end{gathered} \tag{1}$$

Assuming that X^* represents the currently discovered adversarial example with the smallest noise amplitude and X' represents the adversarial sample obtained by adding new noise to the image X^*, then the objective of the adversarial attack is defined by Equation (2). The objective is to maximize the difference between the L_2 distance of image X' relative to the image X^* under the condition that images X^* and X' are consistently misclassified.

$$\begin{gathered} \underset{X'}{max}\|X^* - X_{src}\|_2 - \|X' - X_{src}\|_2 \\ F(X') \neq F(X_{src}) \end{gathered} \tag{2}$$

3.2. Black-Box Boundary Attack Based on Gradient Optimization

The boundary attack, which is a decision-based adversarial attack, aims to generate an adversarial sample that closely approaches the decision boundary. The adversary initiates the attack by starting with the source image X_{src} and perturbing it within the pixel space towards the direction of X_{tgt}; X_{tgt} is the target image when classified correctly.

An image situated on the decision boundary between two labels and classified as y_{tgt} ($F(X_{tgt}) = y_{tgt}$) is referred to as a boundary image. The primary objective of the attack is to discover an adversarial image X_{adv} such that $F(X_{adv}) = y_{tgt}$, while simultaneously minimizing the distance metric $D(X_{src}, X_{adv})$, which is typically calculated using the L_2 norm or L_∞ norm. Consequently, the generated adversarial image X_{adv} is a boundary image with an optimized, minimized distance from the source image.

The iterative process of generating adversarial examples is depicted in Figure 1, where the source image X_{src} and the target image X_{tgt}, labeled with a malicious category y_{tgt}, are selected. Initially, a binary search is performed on both images to derive the initial image X_0, which is misclassified as malicious, following the procedure described in Equation (3):

$$\begin{gathered} X_0 = \alpha X_{tgt} + \beta X_{src} \\ (\alpha + \beta) = 1 \end{gathered} \tag{3}$$

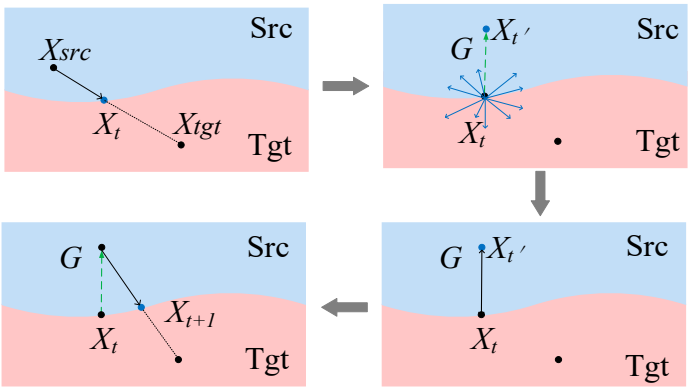

Figure 1. Iterative updates of adversarial examples.

Following this, an iterative algorithm comprising three steps was implemented: (1) sampling vectors in the image space, (2) introducing them as perturbations to the image, and (3) estimating the current boundary gradient by evaluating the classification outcomes of the target model. In the case of score-based attacks, the exact prediction function was determined using the confidence scores derived from the target model's output, as illustrated in Equation (4):

$$S(X) = \left[F(X)_{y_{tgt}} - \max_{y \neq y_{tgt}} [F(X)]_y \right] \tag{4}$$

In the context of boundary-based attacks that yield only hard labels, a significant challenge inherent to decision-based attack methodologies is the inability to ascertain precise label confidence scores. This limitation necessitates a transformation of the prediction function into a computable indicator function $\varphi(X)$, as delineated in Equation (5):

$$\varphi(X) = sign(S(X)) = \begin{cases} 1 & S(X) \geq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

Upon acquiring the batch output values of the indicator function from the target model, we estimate the gradient at boundary points utilizing the Monte Carlo algorithm. This

estimation is formalized in Equation (6), where U_{rnd} denotes a batch of random vectors of size B and η is introduced as a minor weighting constant to refine the estimation accuracy.

$$G = \frac{1}{B}\sum_{i}^{B} \varphi\left[(X_q)^i\right] \cdot \eta(U_{rnd})^i \tag{6}$$

To refine the adversarial example's trajectory towards a correct classification boundary, we systematically adjust its position along the estimated gradient direction towards an image $X_{t'}$. The step size for each movement is meticulously calculated based on the proximity of the current adversarial example $X_{t'}$ and X_{src}, with a deliberate reduction in step magnitude as the iteration count escalates, mitigating the risk of undue distortion. Following this, we implement a strategic binary search technique to meticulously adjust the adversarial example X_{t+1}, ensuring its precise positioning near the decision boundary that delineates classes X_{src} and X_{tgt}.

This process ensures that the adversarial example undergoes a continuous cycle of iterative refinement, consistently edging closer to the decision boundary. The iteration step size, coupled with the gradient, dictates the directional adjustment for each boundary point, serving as a pivotal factor in accelerating the adversarial example's convergence towards the intended target.

In Figure 2, we delineate the procedural framework of the Black-Box Boundary Attack based on Gradient Optimization (GOBA) method. The process initiates with the selection of two distinct images, X_{src} and X_{tgt}, each accurately classified under their respective labels. Upon this initial setup, perturbations derived from random sampling are integrated into image X_t. These perturbed images are then fed into the target model to ascertain their classification labels, facilitating the computation of gradient values via the evaluation of the indicator function.

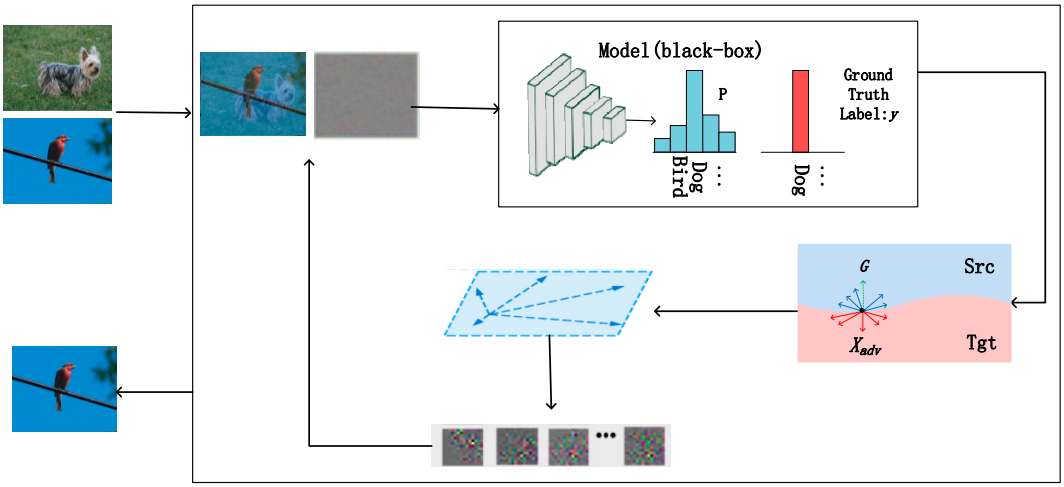

Figure 2. Attack frame.

Leveraging the computed gradient values alongside the initial perturbation vectors, we establish a novel sampling space, setting the stage for the subsequent iteration phase. This step is crucial for augmenting the synergy between the noise vector and the gradient values, thereby enabling the generation of more effective perturbations and the refinement of gradient value estimates. The ultimate goal is to iteratively update the adversarial examples, with the aim of crafting an image that, while visually resembling the X_{src} (bird), is misclassified as the malicious tag y_{tgt} (dog) by the target model $F(X)$.

4. Proposed System Design

Algorithm 1 provides a detailed overview of the implementation steps for the Gradient Optimization-based Black-Box Boundary Attack. In this algorithm, B denotes the size of

the random vector set, R^s represents the sampling space, τ signifies the threshold for the distance between the selected vector and the gradient, ξ stands for the step size for moving the boundary points in the direction of the gradient, ε represents the maximum threshold of the distance between the image moved by a certain step size and the current adversarial sample, and I indicates the number of iterations, which is typically set to 100. The specific implementation steps are outlined as follows:

In step 1, suitable coefficients are selected for X_{src} and X_{tgt} during initialization to guarantee that, when misclassified as y_{tgt}, the image X_t ($t = 0$) exhibits minimal deviation from X_{src}. The resulting image X_0 is then utilized as the original image for the subsequent boundary attack.

In steps 2 to 7, at $t = 1$, the initial sampling space is randomly sampled to acquire perturbations. Subsequent boundary point sampling is then carried out in the optimal dimensional subspace R^θ corresponding to the guidance vectors. The low-dimensional samples obtained are subsequently projected back into the original input space. To address the potential loss of vector information resulting from nonlinear transformations, a linear projection method is utilized to derive the perturbation vector for the next boundary point. To maintain consistent guidance from historical information throughout the gradient estimation process, the sampling subspace constructed by historical data is dynamically updated. The iteration of historical information does not necessitate the design of an adaptive factor, thereby effectively reducing the algorithm's complexity. Following the sampling process, the outcomes are projected back into the original space, resulting in B perturbation vectors.

Algorithm 1: GOBA

input: Model $F(X)$, X_{src}, X_{tgt}, indicator function φ_x, τ, B, I, ξ, ε
output: X_{adv}
1: X_0 = Initial((X_{src}, X_{tgt}), F)
2: **for** t in 1, 2, …, $I - 1$ **do:**
3: **if** $t = 1$ **then:**
4: get $U_{rnd} \in R^S$.
5: **else:**
6: get $V_{rnd} \in R^\theta$.
7: U_{rnd} = Bil_Interp (V_{rnd}).
8: $X_q[i] = X_t + U_{rnd}[i]$
9: Generate G
10: **if** $\|U_{rnd}[i] - G\|_p < \tau$ **then:**
11: resize $U_{rnd}[i]$, get R^θ
12: $\xi_t = \|X_t - X_{src}\|_p / \sqrt{t}$
13: **while** $\varphi_x(X_t + \xi_t \cdot G_t) = 0$ **do:**
14: $\xi_t \leftarrow \xi_t / 2$.
15: **end while:**
16: $X_{t'} = X_t + \xi_t \cdot G_t$
17: X_{t+1} = Binary Search($X_{t'}$, φ_x, X_{tgt}, ε)
18: $d_{t+1} = \|X_{t+1} - X_{src}\|_p$
19: **return** $X_{adv} = X_{t+1}$

In steps 8 to 9, a random perturbation is added to the current image to create a new input, denoted as X_q. This input X_q is then fed into the target model, and the resulting output is transformed to derive the value of the indicator function $\varphi(X_q)$, thus completing the gradient estimation.

In steps 10 to 11, the process involves generating the guidance vector set. The GOBA utilizes historical information obtained through sampling to construct the guidance vector set. The selection of information is illustrated in Figure 3.

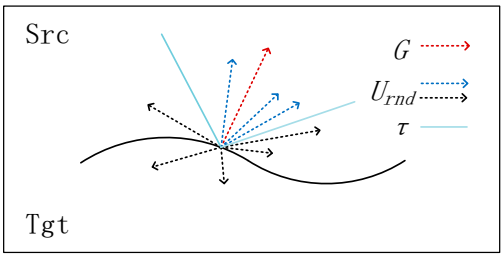

Figure 3. Selection the guide vector.

Each random vector generated near a boundary point is represented by U_{rnd}. Following the calculation of the gradient vector G, historical vectors are either expanded or removed from the guidance vectors based on a predefined threshold τ. This adjustment is made to ensure the condition $\|U_{rnd} - G\|_p < \tau$ is met, while continuously updating τ to maintain the number of vectors within a fixed range. Subsequently, a subset of vectors is chosen to form the guidance vector set, which in turn guides the update of the next boundary point.

When dealing with a large sample space, the effectiveness of random sampling attacks in high-dimensional spaces heavily relies on the scale of the sampling subspace. Enhancing efficiency significantly can be achieved by conducting attacks with an optimal scale. To fine-tune the subspace scale, PGAN (Progressive GAN) [31] is leveraged. Several dimensionality reduction techniques have been proposed, encompassing methods such as dimensionality reduction [22,35] and low-frequency constraints [29,36], which serve to expedite the attack process.

Upon acquiring the historical vector set, normalization and dimensionality reduction are performed to adjust to the optimal dimension corresponding to the current dataset. In comparison to alternative dimensionality reduction approaches, bilinear interpolation stands out for its simplicity and speed. Consequently, the vectors within the set undergo bilinear interpolation, reducing them to the optimal dimension space, thus forming the sampling space R^{θ}.

The set of random vectors, represented as $U_{rnd} = [\omega_1, ..., \omega_n] \in R^{m \times n}$, comprises n orthogonal basis vectors in R^m. Here, the query space $R^s \subseteq R^m$ denotes the sampling within the original space. Upon acquiring the guidance vectors in the reduced dimension, they are utilized to construct the sampling space R^{θ} of random vectors. The objective is to sample random perturbations from the optimal dimension rather than directly from the original space R^m. A vector $V_{rnd} \in R^{\theta}$ is randomly sampled from the unit ball in R^{θ}. If the $span\ (\theta) = R^m$, then this sampling process is equivalent to sampling in the original space R^m.

In steps 12 to 14, the step size ξ for movement in the gradient direction is calculated. The image classification result is continuously updated based on the adjusted image to ensure the image label remains y_{tgt}, with ξ_t representing the step size at the t-th step. Consequently, the prediction score for the adversarial class is expected to increase.

In step 16, within the iterative loop of boundary points, the adversarial example X_t is advanced by a distance of ξ_t along the gradient direction G_t.

Proceeding to step 17, within the iterative loop of boundary points, the adversarial example advanced along the gradient direction undergoes further refinement using a binary search, as optimized in Algorithm 2. The decision boundary range is progressively narrowed in the initial 6 steps, followed by the last 4 steps within a specific small perturbation range where the misclassified example $X_{t'}$ is sought.

Algorithm 2: Binary Search

input: $X_{t'}$, indicator function φ_x, ε
output: X_{t+1}
1: $X_{tem} = X_{tgt}$
2: **while** $\|X_{tem} - X_{t'}\|_p > \varepsilon$ **do:**
3: $X_{t'} = X_{tem}/2 + X_{t'}/2$
4: **if** $\varphi_x(X_{t'}) = 1$:
5: $X_{tem} = X_{tem}/2$
6: **else**: break
7: **while** $\varphi_x(X_{t'}) = 0$ **do:**
8: $X_{tem} = 2X_{tem}$
9: $X_{t'} = X_{tem}/2 + X_{t'}/2$
10: $X_{t+1} = X_{t'}$
11: **return** X_{t+1}

In step 18, the distance between the current adversarial example X_{t+1} and the original image X_{src} is computed to evaluate the magnitude of the perturbation introduced to the existing example.

5. Results

5.1. Experimental Setup

The experiments were conducted on three datasets: MNIST, Imagenet, and CelebA. MNIST comprises 70,000 28 × 28 grayscale images of handwritten digits ranging from 0 to 9. Imagenet consists of 1000 classes, with images resized to 224 × 224 × 3 dimensions. The CelebA dataset includes 202,599 178 × 218 face images representing 10,177 celebrities.

Fifty pairs of source and target images were randomly selected from the validation set, with the target model predicting different classes for these pairs. A pre-trained ResNet-18 model was utilized as the target model for the experiment. The evaluation primarily focused on six methods, HSJA, QEBA-S, QEBA-F, NLBA-AE, NLBA-VAE, and GOBA, comparing the perturbation sizes (measured by L_2 distance) of generated samples under consistent constraints and the attack success rates under varying perturbation thresholds across the datasets.

The L_2 distance was employed as the criterion for evaluating perturbations, with the attack methods' superiority determined based on their attack success rates. Each attack method exhibited distinct effects on individual images, posing challenges in accurately assessing their efficiency. Thus, the overall attack success rates across the dataset were considered a more comprehensive indicator of attack efficiency, as calculated in Equation (7). Perturbation thresholds were set at 1×10^{-3} for the Imagenet dataset, 1×10^{-4} for CelebA, and 5×10^{-3} for MNIST.

$$ASR = \frac{N_{adv}}{N} \tag{7}$$

Here, N represents the total number of samples, while N_{adv} denotes the number of samples for which the L_2 distance of the generated adversarial samples, obtained after a limited number of queries, falls below a specified L_2 threshold.

5.2. Comparative Experiments

5.2.1. Noise Similarity Analysis

To assess the efficacy of utilizing a constructed sampling space for random vector sampling, this study calculated the similarity between random vectors (noise) and gradient vectors across six methodologies as the number of queries increased, as illustrated in Figure 4. This evaluation was crucial for validating the effectiveness of employing the bilinear interpolation method to compile historical information for sampling space construction. As shown in Figure 4a, on the MNIST dataset, a substantial enhancement in the correlation between gradient and random vectors was observed, with the similarity consistently exceeding 0.2 despite minor fluctuations as the number of queries escalated. Similarly, as depicted in Figure 4b,c for the Imagenet and CelebA datasets, respectively, a

significant increase in vector similarity was noted, sustaining levels above 0.04 and 0.05, correspondingly, alongside an upward trend in query volumes.

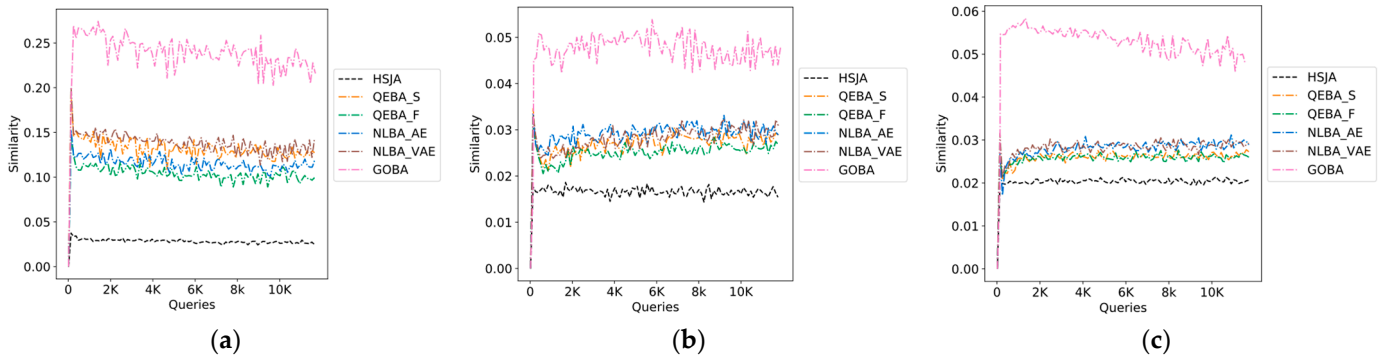

Figure 4. Similarity between random vectors and estimated gradients in different datasets. (**a**) MNIST, (**b**) Imagenet, (**c**) CelebA.

These findings underscore the benefits of dynamically modifying the sampling space and fine-tuning its dimensional parameters, which markedly bolsters the Monte Carlo algorithm's efficiency in gradient estimation compared to static-value approaches. This dynamic strategy enhances the alignment between noise and gradient directions, thereby optimizing the generation of adversarial examples with greater precision and reduced computational overhead.

5.2.2. Attack Performance Analysis

The experiment was designed to monitor the variations in the L_2 norm distance throughout the adversarial attack process across the MNIST, Imagenet, and CelebA datasets, incorporating varying quantities of queries. This study conducted a comparative analysis of six distinct methods, focusing on the magnitude of perturbations generated as the number of queries escalated. This analysis was quantified by measuring the L_2 distance between the adversarial and original samples, with the results depicted in Figure 5. Notably, the L_2 distance on the MNIST dataset was observed to be generally larger compared to that on Imagenet and CelebA, aligning with findings from prior research. This phenomenon is likely attributable to the increased challenge of deceiving models tasked with simpler problems, where the models' heightened sensitivity to perturbations demands larger modifications to the input images to successfully induce misclassification.

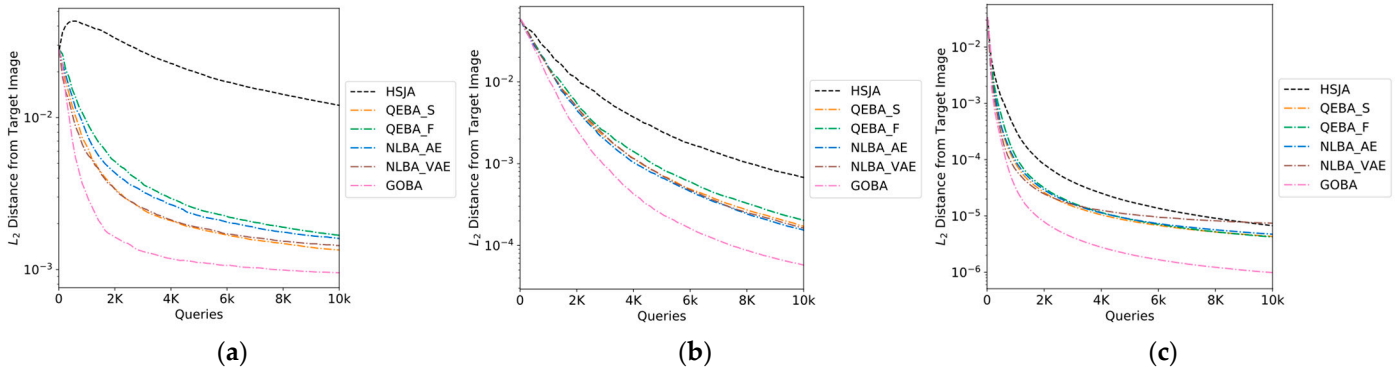

Figure 5. L_2 distance between the adversarial example and the target image in different datasets. (**a**) MNIST, (**b**) Imagenet, (**c**) CelebA.

1. Adversarial examples crafted using the GOBA method demonstrate superior visual quality. Remarkably, when the L_2 norm distance is minimized to a negligible level, these adversarial examples become virtually indistinguishable to the human eye from their original counterparts, yet they significantly impair the classification accuracy

of machine learning models. In scenarios involving up to 10,000 queries, the GOBA achieved a reduction in the L_2 norm distance between the adversarial and original images of 5.38×10^{-4} for the MNIST dataset, 1.03×10^{-5} for the Imagenet dataset, and an impressive 4.88×10^{-7} for the CelebA dataset. This performance markedly outstrips that of the previous most effective method, the QEBA-S, by achieving an average reduction in perturbation magnitude (measured by the L_2 distance) of 54.31%. Thus, the GOBA's efficacy not only surpasses that of the QEBA-S but also exceeds the capabilities of the other five evaluated methods, highlighting its effective application in generating adversarial examples with minimal perturbation deviations yet a maximal misclassification impact.

The proposed methodology for constructing a sampling space, predicated on historical data, is designed to augment the efficacy of queries involving random vectors. By dynamically modulating the sampling space, this approach endeavors to identify the optimal perturbation vector for a given image. It accomplishes this with a finite number of queries, thereby computing gradients with greater accuracy and minimizing the overall image perturbation, significantly bolstering the success rate of attacks.

In stark contrast, alternative methods relying on independent sampling fail to efficiently leverage the quantity of model queries to enhance algorithmic accuracy, often resulting in inefficiencies.

This nuanced approach not only improves the strategic utilization of query capacities but also underscores the importance of adaptively adjusting the sampling space to achieve more precise gradient estimations. The outcome is a marked advancement in the generation of adversarial examples, characterized by reduced perturbations while maintaining high levels of misclassification effectiveness.

2. Superior attack performance of the GOBA: Figure 6 illustrates the attack success rate curves of the six methods for generating adversarial samples on the three datasets. With a query budget of 1 k, the GOBA's attack success rate on the MNIST dataset is 2.67 times that of the HSJA and 2.28 times that of NLBA-VAE. Under a 1.5 k query budget on the Imagenet dataset, the GOBA's attack success rate is at least 2% higher than the HSJA and QEBA-S. On the CelebA dataset, it is at least 9% higher. Evidently, the GOBA's attack success rate surpasses the HSJA, QEBA, and NLBA.

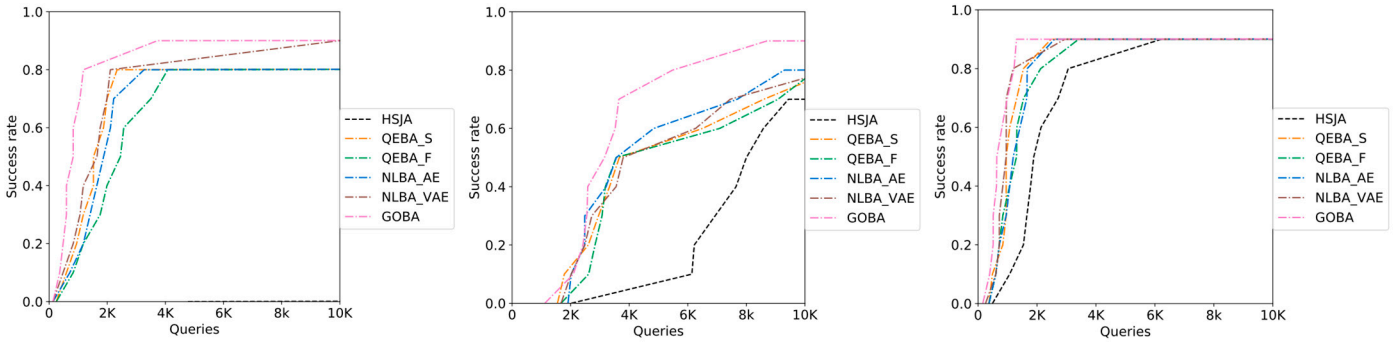

Figure 6. The success rate of adversarial examples against the target model in different datasets. (**a**) MNIST, (**b**) Imagenet, (**c**) CelebA.

3. Efficient convergence in attack success rate: The GOBA exhibits faster convergence in attack success rate. Under the constraint of a 90% success rate, the GOBA requires 1 k queries on the CelebA dataset; a reduction of 52.3% compared to the QEBA-S. On the MNIST dataset, the GOBA reduces the required query quantity by 60% compared to the NLBA-VAE. On the Imagenet dataset, when the attack success rate threshold for the GOBA is set to 80%, it reduces the query quantity by 47% compared to the NLBA-AE, accelerating the convergence speed of adversarial samples.

In the experiments, QEBA, NLBA, and GOBA all leverage low-frequency noise for attacks, whereas HSJA employs Gaussian noise. The smooth characteristics of low-frequency noise allow images to exhibit texture features closely resembling real images. The classifier captures features introduced by low-frequency noise, incorporating them with the original features, thereby diminishing confidence in the correct class. Conversely, Gaussian noise is sharper, less likely to form features resembling real images, and is easily filtered out by linear filters, rendering the attack less effective. This contributes to the suboptimal performance of the HSJA, particularly with a low query count. Although the QEBA and NLBA attacks utilize low-frequency noise in experiments, the QEBA directly applies linear transformations based on uniform noise during adversarial sample initialization. Similarly, the NLBA involves nonlinear transformations, substantially increasing the time cost of the attack, and the attack effect is not stable.

Table 1 provides a summary of the datasets and models employed for the HSJA, QEBA-S, QEBA-F, NLBA-AE, NLBA-VAE, and GOBA, presenting the attack success rate (ASR) and the perturbation size (L_2 distance) between adversarial samples and original images with a fixed query number of 3 k. The data highlights a substantial improvement in the GOBA attacks on adversarial samples after 3 k queries, showcasing a reduction in perturbation size (L_2 distance) of at least 55.74% and an average increase in ASR of 13.78%. This underscores the GOBA as an efficient and rapidly converging, query-efficient attack.

Table 1. L_2 distance and success rate of different attacks in the queries of 3 K.

Method	**HSJA**		**QEBA-S**		**QEBA-F**		**NLBA-AE**		**NLBA-VAE**		**GOBA**	
Dataset	L_2	**ASR**	L_2	**ASR**	L_2	**ASR**	L_2	**ASR**	L_2	**ASR**	L_2	**ASR**
MNIST	0.0273	0	0.0025	0.80	0.0037	0.64	0.0033	0.80	0.0025	0.81	0.00132	0.90
Imagenet	0.0061	0.02	0.0021	0.31	0.0024	0.27	0.0019	0.37	0.0022	0.33	0.00099	0.47
CelebA	4.15×10^{-5}	0.78	1.52×10^{-5}	0.90	1.73×10^{-5}	0.86	1.66×10^{-5}	0.90	1.62×10^{-5}	0.90	4.26×10^{-6}	0.90

6. Security Analysis

The results presented by the Black-Box Boundary Attack based on Gradient Optimization (GOBA) are encouraging in terms of attack success rates and query efficiency. The following provides an in-depth scrutiny of the security aspects of the GOBA.

In adversarial attacks, a primary concern is the robustness of the proposed method against various defense mechanisms. Evaluating the GOBA's performance when confronted with common defense strategies employed by machine learning models, such as adversarial training, input preprocessing, and gradient masking, is essential. Such an evaluation facilitates the identification of more targeted defensive strategies. As adversarial attacks grow increasingly sophisticated, analyzing the detectability of adversarial samples generated by the GOBA using state-of-the-art methods enables an understanding of current detection technique limitations and facilitates ongoing efforts to advance adversarially robust models.

In conclusion, the proposed GOBA method is crucial for ensuring the effectiveness and reliability of existing defense methods in real-world scenarios, contributing to the development of more robust and secure machine learning models and enhancing their resilience against adversarial attacks.

7. Conclusions

This paper addresses the issue of hard-label attacks and proposes an efficient query attack based on gradient direction optimization (GOBA). In the presence of gradients, this method utilizes historical query information as prior knowledge for optimizing random vectors, dynamically constructing an optimal dimensional sampling space based on historical information. The evaluation results of this method on three natural image datasets indicate that, under the same attack success rate threshold, a reduction of over 40% in query budget can be achieved by the GOBA, accompanied by an improvement of 13.78% in attack success rate with the minimum possible number of queries. However, this study

still has limitations: Existing hard-label attacks can be defended against by restricting queries near the boundary or by inserting additional 'unknown' classes for low-confidence inputs to expand decision boundaries. This renders existing methods ineffective. Therefore, exploring how to introduce perturbations that are not limited to the vicinity of the boundary (global perturbations) or obtaining decision boundaries with higher accuracy during attacks should be further investigated. Subsequent research can focus on these two directions to generate higher-quality adversarial samples.

Author Contributions: Conceptualization, Y.Y. and Z.L. (Zishuo Liu); formal analysis, S.W.; funding acquisition, Y.C.; investigation, Y.Y. and Z.L. (Zishuo Liu); methodology, Z.L. (Zishuo Liu) and Z.L. (Zhen Lei); project administration, Y.C.; software, Z.L. (Zishuo Liu) and Z.L. (Zhen Lei); supervision, Y.C.; validation, Y.Y. and Z.L. (Zishuo Liu); visualization, Z.L. (Zhen Lei); writing—original draft, Z.L. (Zishuo Liu); writing—review and editing, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Natural Science Foundation of Shanxi Province, grant numbers 20210302123131, 20210302124395, and 202303021221017. The APC was funded by Taiyuan University of Technology.

Data Availability Statement: The MNIST dataset is available at http://yann.lecun.com/exdb/mnist/ (accessed on 5 February 2024). The Imagenet dataset is available at https://image-net.org/ (accessed on 5 February 2024). The CelebA dataset is available at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 5 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, J.; Su, H.; Zhu, J.; Wang, S.; Zhang, B. Textbook question answering under instructor guidance with memory networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3655–3663.
2. Gong, Z.; Zhong, P.; Yu, Y.; Hu, W.; Li, S. A CNN with multiscale convolution and diversified metric for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3599–3618. [CrossRef]
3. Gong, Z.; Zhong, P.; Hu, W. Statistical loss and analysis for deep learning in hyperspectral image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 322–333. [CrossRef] [PubMed]
4. Albert, A.; Kaur, J.; Gonzalez, M.C. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1357–1366.
5. Pritt, M.; Chern, G. Satellite image classification with deep learning. In Proceedings of the 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 10–12 October 2017; IEEE: Piscataway, NC, USA, 2017; pp. 1–7.
6. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
7. Joseph, K.J.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5830–5840.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
12. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1625–1634.
13. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [CrossRef]
14. Hu, Y.; Yang, A.; Li, H.; Sun, Y.; Sun, L. A survey of intrusion detection on industrial control systems. *Int. J. Distrib. Sens. Netw.* **2018**, *14*, 1550147718794615. [CrossRef]

15. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. 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, Dallas, TX, USA, 3 November 2017; pp. 15–26.
16. Jia, X.; Zhang, Y.; Wu, B.; Wang, J.; Cao, X. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Trans. Image Process.* **2022**, *31*, 4417–4430. [CrossRef] [PubMed]
17. Bai, J.; Chen, B.; Li, Y.; Wu, D.; Guo, W.; Xia, S.-T.; Yang, E.-H. Targeted attack for deep hashing based retrieval. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 618–634.
18. Jia, X.; Zhang, Y.; Wu, B.; Ma, K.; Wang, J.; Cao, X. LAS-AT: Adversarial training with learnable attack strategy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13398–13408.
19. Gu, Z.; Hu, W.; Zhang, C.; Lu, H.; Yin, L.; Wang, L. Gradient shielding: Towards understanding vulnerability of deep neural networks. *IEEE Trans. Netw. Sci. Eng.* **2020**, *8*, 921–932. [CrossRef]
20. Yu, M.; Sun, S. FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks. *Comput. Secur.* **2022**, *113*, 102555. [CrossRef]
21. Brunner, T.; Diehl, F.; Le, M.T.; Knoll, A. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4958–4966.
22. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv* **2017**, arXiv:1712.04248.
23. Cheng, M.; Le, T.; Chen, P.Y.; Yi, J.; Zhang, H.; Hsieh, C.J. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv* **2018**, arXiv:1807.04457.
24. Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; Zhu, J. Efficient decision-based black-box adversarial attacks on face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7714–7722.
25. Shi, Y.; Han, Y.; Tian, Q. Polishing decision-based adversarial noise with a customized sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1030–1038.
26. Cheng, M.; Singh, S.; Chen, P.; Chen, P.Y.; Liu, S.; Hsieh, C.J. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv* **2019**, arXiv:1909.10773.
27. Liu, Y.; Moosavi-Dezfooli, S.M.; Frossard, P. A geometry-inspired decision-based attack. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4890–4898.
28. Rahmati, A.; Moosavi-Dezfooli, S.M.; Frossard, P.; Dai, H. Geoda: A geometric framework for black-box adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8446–8455.
29. Guo, Y.; Yan, Z.; Zhang, C. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *arXiv* **2019**, arXiv:1906.04392.
30. Chen, J.; Jordan, M.I.; Wainwright, M.J. Hopskipjumpattack: A query-efficient decision-based attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (sp), Francisco, CA, USA, 18–20 May 2020; IEEE: Piscataway, NC, USA, 2020; pp. 1277–1294.
31. Li, H.; Xu, X.; Zhang, X.; Yang, S.; Li, B. Qeba: Query-efficient boundary-based blackbox attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1221–1230.
32. Li, H.; Li, L.; Xu, X.; Zhang, X.; Yang, S.; Li, B. Nonlinear gradient estimation for query efficient blackbox attack. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS 2021), Proceedings of Machine Learning Research, PMLR, Virtual, 13–15 April 2021; pp. 13–15.
33. Zhang, J.; Li, L.; Li, H.; Zhang, X.; Yang, S.; Li, B. Progressive-scale boundary blackbox attack via projective gradient estimation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12479–12490.
34. Maho, T.; Furon, T.; Le Merrer, E. SurFree: A fast surrogate-free black-box attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10430–10439.
35. Serban, A.; Poll, E.; Visser, J. Adversarial examples on object recognition: A comprehensive survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 66. [CrossRef]
36. Liu, J.; Jin, H.; Xu, G.; Lin, M.; Wu, T.; Nour, M.; Alenezi, F.; Alhudhaif, A.; Polat, K. Aliasing black box adversarial attack with joint self-attention distribution and confidence probability. *Expert Syst. Appl.* **2023**, *214*, 119110. [CrossRef]