Saliency Attack: Towards Imperceptible Black-box Adversarial Attack

ZEYU DAI, Southern University of Science and Technology, China and The Hong Kong Polytechnic University
SHENGCAI LIU, Southern University of Science and Technology
QING LI, The Hong Kong Polytechnic University
KE TANG, Southern University of Science and Technology

Deep neural networks are vulnerable to adversarial examples, even in the black-box setting where the attacker is only accessible to the model output. Recent studies have devised effective black-box attacks with high query efficiency. However, such performance is often accompanied by compromises in attack imperceptibility, hindering the practical use of these approaches. In this article, we propose to restrict the perturbations to a small salient region to generate adversarial examples that can hardly be perceived. This approach is readily compatible with many existing black-box attacks and can significantly improve their imperceptibility with little degradation in attack success rates. Furthermore, we propose the SaliencyAttack, a new black-box attack aiming to refine the perturbations in the salient region to achieve even better imperceptibility. Extensive experiments show that compared to the state-of-the-art black-box attacks, our approach achieves much better imperceptibility scores, including most apparent distortion (MAD), $L_0$ and $L_2$ distances, and also obtains significantly better true success rate and effective query number judged by a human-like threshold on MAD. Importantly, the perturbations generated by our approach are interpretable to some extent. Finally, it is also demonstrated to be robust to different detection-based defenses.

CCS Concepts: • Security and privacy; • Computing methodologies → Computer vision;

Additional Key Words and Phrases: Adversarial example, black-box adversarial attack, saliency map, deep neural networks

ACM Reference format:
Zeyu Dai, Shengcai Liu, Qing Li, and Ke Tang. 2023. Saliency Attack: Towards Imperceptible Black-box Adversarial Attack. *ACM Trans. Intell. Syst. Technol.* 14, 3, Article 45 (April 2023), 20 pages.

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1004102, in part by the National Natural Science Foundation of China under Grant 62250710682, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X386, and in part by the Hong Kong Research Grants Council under the General Research Fund (Project No. 15200021).

Authors’ addresses: Z. Dai, Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China and The Hong Kong Polytechnic University, Hong Kong, China; email: daizy2019@mail.sustech.edu.cn; S. Liu (corresponding author), Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China; email: liusc3@sustech.edu.cn; Q. Li, Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China; email: qing-prof.li@polyu.edu.hk; K. Tang, Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China; email: tangk3@sustech.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2157-6904/2023/04-ART45 $15.00

https://doi.org/10.1145/3582563
1 INTRODUCTION

Deep neural networks (DNNs) have achieved significant progress in wide applications, such as image classification [10], face recognition [30], object detection [31], speech recognition [16], and machine translation [3]. Despite their success, deep learning models have exhibited vulnerability to adversarial attacks [13, 23, 24, 39]. Crafted by adding some small perturbations to benign inputs, adversarial examples (AEs) can fool DNNs into making wrong predictions, which is a critical threat especially for some security-sensitive scenarios such as autonomous driving [37].

Existing adversarial attacks can be divided into white-box and black-box attacks according to the accessibility to the target model. White-box attacks [6, 13, 26, 29, 39] have full access to the architecture and parameters of the target model, and can generate successful AEs easily via back-propagation. However, in practice, the model internals are often unavailable to attackers. This gives rise to the more realistic and challenging black-box attacks [2, 17, 18, 22, 25, 27] that only require the output of the target model.

Motivated by the fact that many real-world online application programming interfaces (APIs) often pose mandatory time or monetary limits to user queries [17], most recent research on black-box attacks concerns improving query efficiency, which indeed has achieved notable progress. For example, the state-of-the-art (SOTA) Square Attack [2] can succeed in an untargeted attack on the ImageNet dataset [10] with only tens of queries on average. On the other hand, such performance is often achieved by applying large-region or even global perturbations (e.g., random vertical stripes (RVS) in Square Attack) to the original input, eventually resulting in unnatural AEs with significant visual differences from the original (see Figure 1 for some examples). In effect, imperceptibility is essential for attackers. Current online APIs usually integrate detectors into their services to detect anomalous inputs [22]. Those AEs with too visible perturbation are difficult to pass these detectors, not to mention human judgment, hindering the practical use of these black-box attacks.

This article aims to address the above issue. The key insight is that we can lead black-box attacks to search in a small salient region that has the most significant influences on the model output, to achieve successful attacks with imperceptible perturbations. Concretely, it has been shown in previous DNNs visualization work [35] that one can generate the so-called BP-saliency map to illustrate which features can influence the model prediction by calculating derivatives of the model output with respect to the input. As shown in the first column of Figure 2, the brighter pixels in the BP-saliency map have greater impacts on the model output, and therefore perturbing them is more likely to alter the model prediction. Actually, the classical white-box attack JSMA [29] constructs exactly such a map and iteratively selects pixels from it to perturb. Despite the fact that the BP-saliency map cannot be used for black-box attacks due to inaccessibility to model gradients, one can observe from the second column of Figure 2 that the region of bright pixels roughly represents the position of the main object in the image. That is to say, in black-box settings, we can leverage the existing salient object detection model [47] that requires no information other than the input image to approximately obtain the salient region, and then restrict the perturbations to it. This approach is appealing because it is readily compatible with most existing black-box attacks. By integrating it into SOTA black-box attacks, we find the modified attacks enjoy significant improvement in attack imperceptibility, with little degradation in success rate (SR) (see the results in Table 1). Nevertheless, they still suffer from the strategy of perturbing globally, eventually generating complex and irregular perturbations that span almost the entire salient region (see Figure 5 for some examples).

To achieve more imperceptible attacks, we then seek to further restrict the perturbations to even smaller regions. The intuition is that even in the salient region, some sub-regions are more critical. For example, it has been shown that the dog face region is the brightest in the salient region of a
Saliency Attack: Towards Imperceptible Black-box Adversarial Attack

Fig. 1. Some AEs generated by the SOTA black-box attacks.

Fig. 2. An illustration of our Saliency Attack.

1st column: BP-saliency map, specific to a given image and the corresponding class; 2nd column: salient region, generated by salient object detection and roughly accords with the region of salient pixels in BP-saliency map; 3rd column: perturbation, generated by Saliency Attack and restricted in salient regions; 4th column: final adversarial example; 5th column: original image.

For crafting AEs in the black-box setting, we are the first to restrict perturbations to a salient region. This approach is readily compatible with many

dog image [49], and obscuring it will dramatically change the model output [45]. Furthermore, by combining internal feature visualization with output prediction, it has been revealed in [28] that within the dog face region, the ears and eyes seem to be more important than others. Therefore, it is reasonable to assume the salient region of an image is progressive with respect to its impact on model output. If we could find a smaller but more salient sub-region, the perturbations will be more efficient, leading to more imperceptible attacks (see the third column of Figure 2). Thus, we propose the Saliency Attack, a novel black-box attack that recursively refines the perturbations in the salient region.

It is worth mentioning that except white-box attack JSMA, the idea of restricting perturbations to a small region has also been implemented in transfer-based attacks [11, 43], where class activation mapping (CAM) [49] and Grad-CAM [33] are adopted to generate the saliency maps. However, transfer-based attacks assume the data distribution for training the target model is available and thus could build a substitute model to approximate it, which actually belong to the grey-box setting where partial knowledge of the target model is known. As a result, they cannot be applied to the more strict black-box setting where only the model output is available (see Section 2.3 for details).

In summary, we make the following contributions in this work.

— To the best of our knowledge, for crafting AEs in the black-box setting, we are the first to restrict perturbations to a salient region. This approach is readily compatible with many
existing black-box attacks and significantly improves their imperceptibility with little degradation in SR.

— We propose the Saliency Attack, a novel gradient-free black-box attack that iteratively refines perturbations in salient regions to keep them minimal and essential. Compared with the SOTA black-box attacks, our approach achieves much better attack imperceptibility in terms of most apparent distortion (MAD), $L_0$ and $L_2$ distances, and also obtains significantly better SRs and effective query number judged by a human-like threshold on MAD.

— We demonstrate that the perturbations generated by Saliency Attack are more robust against detection-based defenses, including Feature Squeezing and binary classifier detection.

The rest of this article is organized as follows: Section 2 presents a literature review on black-box attacks. Next, in Section 3, we define the optimization problem of our attack and detail the proposed Saliency Attack. Section 4 shows the experimental results and analysis, followed by a conclusion in Section 5.

2 RELATED WORK

In this part, we first overview the recent work of black-box attacks. Based on the generation method of AEs, these attacks can be divided into gradient estimation attacks and gradient-free attacks. Besides, we also introduce some work related to the imperceptibility in adversarial attacks. Finally, we discuss some methods that extract salient regions in an image.

2.1 Black-box Attacks

2.1.1 Gradient Estimation Attacks. Gradient estimation attacks first estimate the gradients by querying the target model and then use them to run white-box attacks. ZOO attack [8] first adopts symmetric difference quotient to approximate the gradients and then performs white-box Carlini-Wagner (CW) attack [6]. AutoZOOM [40], a variant of ZOO, uses a random vector based gradient estimation to reduce the query number per iteration from $2D$ in ZOO to $N+1$ ($D$ is the dimensionality and $N$ is the sample size). To further enhance query efficiency, Ilyas et al. [18] propose the “tiling trick” that updates a square of pixels simultaneously instead of a single pixel. This dramatically decreases the dimensionality by a factor of $k^2$ ($k$ is tile length).

2.1.2 Gradient-free Attacks. Gradient-free attacks do not estimate gradients and directly generate AEs with search heuristics according to the query result. Su et al. [36] propose One-pixel attack that adopts differential evolution algorithm to perturb the most important pixel in the image. Alzantot et al. [1] propose GenAttack, which uses genetic algorithm to generate AEs. To improve query efficiency, Moon et al. [25] consider a discrete surrogate optimization problem that transforms the original constraint of a continuous range $[-\epsilon, +\epsilon]$ to a discrete set $\{-\epsilon, +\epsilon\}$, achieving a massive reduction in the search space. This is motivated by linear program (LP) where the optimal solution is attained at an extreme point of the feasible set [32]. Combining tiling trick [18] and discrete optimization [25], Square Attack [2] has obtained the best result on SR and query performance so far with random search.

2.2 Attack Imperceptibility

The imperceptibility of AEs is essential for practical attackers, which has been investigated by previous studies from different perspectives. Guo et al. [14] consider low frequency perturbations as imperceptible, thus searching for AEs in frequency domain. But Zhang et al. [46] regard imperceptibility as visual smoothness in an image and integrate Laplacian smoothing into optimization. Croce and Hein [9] use a combination of $L_0$ and $L_{\infty}$ norms to generate sparse and imperceptible

ACM Transactions on Intelligent Systems and Technology, Vol. 14, No. 3, Article 45. Publication date: April 2023.
perturbations. Besides, some studies also leverage color distance [48] and image quality assessment (IQA) [21, 41] to improve attack imperceptibility.

On the other hand, determining how to assess the imperceptibility of AEs is still an open question. Most existing adversarial attacks use $L_p$ norms ($L_0, L_2$, and $L_{\infty}$) to measure the human perceptual distance between the perturbed image and the original one. Nonetheless, it has been shown that $L_p$ norms are not suitable enough for human vision system [34]. A recent study in [12] systematically examines various IQA metrics including $L_p$ norms through large-scale human evaluation on the imperceptibility of different AEs, and finds that among all the metrics the MAD [19] metric is closest to subjective scores (see Appendix A.1 for details of MAD). Hence, in this research, we adopt MAD as our main metric to assess attack imperceptibility.

2.3 Extracting Salient Region

As aforementioned, there have been white-box attacks or transfer-based attacks that restrict the perturbations to a small salient region. Specifically, white-box attack JSMA [29] constructs a BP-saliency map by calculating derivatives of the model output w.r.t input pixels [35], while the two transfer-based attacks [11, 43] utilize CAM and Grad-CAM to extract salient regions, respectively. CAM [49] replaces the final fully connected layers with convolutional layers and global average pooling of a CNN, and localizes class-specific salient regions through forward propagation. Grad-CAM [33] improves CAM with on need to modify the network architecture, but it still requires access to the inner parameters of the model to calculate the gradients. Thus, CAM and Grad-CAM can only be applied to white-box attacks or transfer-based attacks, which first construct a transparent substitute model, craft AEs with gradient-based white-box attacks, and then transfer the generated AEs to the target model. It is conceivable that for transfer-based attacks, the transferability of both AEs and saliency maps highly depends on the similarity between the substitute model and the target model, which further depends on the prior knowledge of the training data distribution. Unfortunately, neither such knowledge nor model gradients are available in black-box settings. Considering the limitations of the above methods, we adopt a salient object detection model [47] to directly generate the saliency map given an input image with no need to access the target model’s architecture or parameters.

3 PROPOSED METHOD

In this section, we first formulate the problem of crafting AEs for image classification models, and then detail our approach. Figure 3 illustrates the overall flowchart of Saliency Attack.

3.1 Preliminary

Given a well-trained DNN classifier $h : [0, 1]^d \rightarrow \mathbb{R}^K$, where $d$ is the dimension of the input $x$, and $K$ is the number of classes. Let $h_k(x)$ denote the predicted score that $x$ belongs to class $k$. The classifier assigns the class that maximizes $h_k(x)$ to the input $x$. The goal of an untargeted attack is to find an AE $x_{adv}$, that results in the model misclassification from the ground-truth class $y_{gt}$, and
meanwhile keeps the distance between $x_{adv}$ and the benign input $x$ smaller than a threshold $\epsilon$:

$$\arg\max_{k=1,\ldots,K} h_k(x_{adv}) \neq y_{gt}, \quad \text{s.t. } ||x_{adv} - x||_p \leq \epsilon, \ x_{adv} \in [0, 1]^d.$$  

(1)

Note the last constraint indicates that $x_{adv}$ is a valid image. In this study, we focus on $L_\infty$ norm, following [2, 25]. Conventionally, the task of finding $x_{adv}$ can be rephrased as solving a constrained continuous problem:

$$\max_{x_{adv} \in [0, 1]^d} f(x_{adv}), \quad \text{s.t. } ||x_{adv} - x||_\infty \leq \epsilon,$$

(2)

where $f(x) = L(x, y_{gt})$ is the loss function. Similar to [25], we transform the continuous problem into a discrete surrogate problem, where the perturbation $x_{adv} - x = \delta \in \{-\epsilon, +\epsilon, 0\}^d$. Note unlike [25], where all pixels are perturbed either by $-\epsilon$ or $+\epsilon$, we allow the pixels to remain unperturbed to refrain from global perturbation. Besides, only the pixels in salient regions can be perturbed. The final problem is defined as the following set maximization problem.

$$\max_{S^+, S^- \subseteq S \ \ S^+ \cap S^- = \emptyset} \left\{ F(S^+, S^-) \triangleq f \left( x + \epsilon \sum_{i \in S^+} e_i - \epsilon \sum_{i \in S^-} e_i \right) \right\},$$

(3)

where $S$ denotes the set of pixels in salient regions, $S^+$ and $S^-$ denote the set of selected pixels with $+\epsilon$ and $-\epsilon$ perturbations, respectively, and $e_i$ is the $i$th standard basis vector. Let $\mathcal{V}$ denotes the ground set which is the set of all pixel locations ($|\mathcal{V}| = d$). Note that $S \subseteq \mathcal{V}$, and $S \setminus (S^+ \cup S^-)$ is the set of pixels in salient regions that are unperturbed. The goal of Equation (3) is to find $S^+$ and $S^-$ that will maximize the objective set function $F$.

### 3.2 Salient Object Detection

Salient object detection aims to automatically and accurately extract salient object(s) in an image. Compared with other salient region extraction methods like BP-saliency map [35], CAM [49], and Grad-CAM [33], this type of models do not require any information other than the input image, which is very suitable for the black-box setting. We adopt the Pyramid Feature Attention (PFA) network\(^1\) [47] that achieves the SOTA performance on multiple datasets via capturing high-level context features and low-level spatial structural features simultaneously (see Appendix A.2 for details). Specifically, given an input image, PFA can generate a saliency score between 0 and 1 for each pixel. The higher value denotes higher visual saliency. We then use a threshold $\phi$ to transform the saliency scores into a binary saliency mask, which determines the salient region. The binarization can be expressed as

$$s_i^* = \begin{cases} 0 & s_i < \phi \\ 1 & s_i \geq \phi \end{cases},$$

(4)

where $s_i$ and $s_i^*$ are the saliency score and the binary mask at the $i$th pixel position, respectively. Thus, the salient region is the set of the pixels masked by 1.

$$\mathcal{S} \triangleq \{ i \mid s_i^* = 1, \ i \in \mathcal{V} \}.$$  

(5)

\(^1\)Implementation of PFA [https://github.com/sairajk/PyTorch-Pyramid-Feature-Attention-Network-for-Saliency-Detection].
3.3 Refining Perturbations in Salient Region

The proposed Saliency Attack is outlined in Algorithm 1. The search process (lines 6–35) recursively refines perturbations based on a tree structure for an image, which is illustrated in Figure 4. Concretely, an input image is first split into some initial blocks (a coarse grid in Figure 4), and only the blocks in salient regions are kept (lines 7–11). Then we try to add $+\epsilon$ or $-\epsilon$ perturbations on each initial block individually (lines 12–19) and sort them according to their influence ($F$) on model output (line 26). After that, we choose the block with most influence for further refinement if its $F$ is better than the current best $\hat{F}$ (lines 27–35). We then recursively refine the current best block (line 32), split this block into smaller blocks (a finer grid in Figure 4), and again try to add a perturbation on each block individually (at this time we just flip the perturbation (lines 21–24) for convenience, e.g., $+\epsilon$ to $-\epsilon$) to find the best smaller block. The refinement process recurs until the minimal block (e.g., 1 pixel, line 30) or no smaller block has a better $F$ (line 28). Afterwards, we backtrack to the last level of split blocks and use the second-best block for further perturbation. In this way, the most important block will be explored first, and the perturbations could be as small as possible, with no need to initialize a global perturbation like Parsimonious Attack [25].

To make full use of query budget and combine the advantages of different initial block sizes $k_{int}$ (large $k_{int}$ can quickly lead to successful attacks with large perturbations, while small $k_{int}$ enables the refinement for small perturbations in finer grids), we leverage an outer iteration to run the aforementioned Refine search with decreasing $k_{int}$ (lines 2–5). During the iterations, the algorithm will stop if the generated $x_{adv}$ succeeds to fool the model ($F > 0$) or the termination condition is reached (line 2).

We use the loss function from CW attack [6] of untargeted attack for Equation (2):

$$L(x, y_{gt}) = -\max(Z(x_{adv})_{y_{gt}} - \max_{i \neq y_{gt}}(Z(x_{adv})_i), -\kappa), \quad (6)$$

where $Z(x_{adv})_{y_{gt}}$ is the logit of the AE with respect to the ground-truth class of the original image. In this way, the AE tends to be misclassified into the class $i$, with a margin $\kappa$ of the loss function between the ground-truth class and other classes (We set $\kappa = 0$ in our work).
ALGORITHM 1: Saliency Attack

1: **Input**: Objective set function $F$, Ground set $\mathcal{V}$, salient region $\mathcal{S}$, initial block size $k_{int}$, query budget $\hat{F}$

2: **Output**: $S^+, S^-$ set of pixels with $+\epsilon$ and $-\epsilon$ perturbations

3: while $k_{int} > 1$ and not exceeding query budget do

4: Refine($\mathcal{V}, k_{int}, 1$);

5: $k_{int} \leftarrow k_{int}/2$;

6: end

7: Procedure Refine($B, k, \text{split\_level}$)

8: /* Split the current block into finer blocks in the salient region. */

9: $k' \leftarrow \sqrt{|B|}$ is the block size of $B$; $n \leftarrow (k'/k)^2$ is the number of split blocks;

10: $(B_1, B_2, \ldots, B_n) \leftarrow$ split $B$ into $n$ finer blocks;

11: for $i \leftarrow 1$ to $n$ do

12: $B_i \leftarrow B_i \cap \mathcal{S}$;

13: end

14: /* Sort the finer blocks according to their influence on model output. */

15: if split\_level = 1 then

16: for $j \leftarrow 1$ to $n$ do

17: if $F(S^+ \cup B_j, S^- \setminus B_j) \geq F(S^+ \setminus B_j, S^- \cup B_j)$ then

18: $S^+_j \leftarrow S^+ \cup B_j; S^-_j \leftarrow S^- \setminus B_j; F_j \leftarrow F(S^+ \cup B_j, S^- \setminus B_j)$;

19: else

20: $S^+_j \leftarrow S^+ \setminus B_j; S^-_j \leftarrow S^- \setminus B_j; F_j \leftarrow F(S^+ \setminus B_j, S^- \cup B_j)$;

21: end

22: end

23: else

24: for $j \leftarrow 1$ to $n$ do

25: $S^+_j \leftarrow (S^+ \setminus B_j) \cup (S^- \cap B_j); S^-_j \leftarrow (S^- \setminus B_j) \cup (S^+ \cap B_j)$;

26: $F_j \leftarrow F(S^+, S^-)$;

27: end

28: end

29: $(F_{\pi(1)}, F_{\pi(2)}, \ldots, F_{\pi(n)}) \leftarrow$ sort $\{F_1, F_2, \ldots, F_n\}$ in descending order;

30: /* Recursively select the finer block with most influence for perturbation. */

31: for $j \leftarrow 1$ to $n$ do

32: if $F_{\pi(j)} > \hat{F}$ then

33: $S^+ \leftarrow S^+_{\pi(j)}; S^- \leftarrow S^-_{\pi(j)}; \hat{F} \leftarrow F_{\pi(j)}$;

34: if $k > 1$ then

35: $k \leftarrow k/2$;

36: Recursively call Refine($B_{\pi(j)}, k, \text{split\_level} + 1$);

37: end

38: end

39: end

4 EXPERIMENTS

The main goal of the experiments is to validate: (1) whether salient regions could improve the imperceptibility of existing black-box attacks; (2) whether our Saliency Attack could further enhance the imperceptibility performance. Hence, we first compare our Saliency Attack with the baselines, including SOTA black-box attacks and their modified version restricted in salient regions. Then, we
perform ablation studies to verify the effectiveness of salient regions and Refine search separately. Besides, we measure the hyperparameter sensitivity of our approach. Finally, we test different detection-based defenses to evaluate the imperceptibility from the defensive perspective.

4.1 Settings

We compare our Saliency Attack against SOTA black-box attacks including Boundary Attack [5], HSJA [7], TVDBA [21], Parsimonious attack [25], and Square attack [2]. Among them, Boundary Attack and HSJA start from an already misclassified noise and gradually approach the original image, TVDBA tries to minimize the perturbation via integrating Structural SIMilarity (SSIM) [42] into the loss, Parsimonious Attack proposes discrete optimization to improve query efficiency, and Square Attack can achieve the current SOTA query efficiency and SR in the black-box attack setting. In addition, we design Parsimonious-sal Attack and Square-sal Attack as baselines that restrict their perturbations in salient regions. For our Saliency Attack, the threshold $\phi$ to produce binary saliency masks is chosen to be 0.1. For splitting, original images are resized to $256 \times 256$, and we set the initial block size $k_{int}$ to 16. All parameters of the compared attacks remain consistent with those recommended in their papers.

In this study, we consider $L_\infty$ threat model on ImageNet dataset [10], and set the perturbation magnitude $\epsilon$ to 0.05 in $[0,1]$ scale. We use Inception v3 [38] and ResNet50 [15] as the target model, and the query budget is 10,000 for untargeted attack. We provide our implementation publicly.\footnote{https://github.com/Daizy97/SaliencyAttack.}

For performance metrics, we employ commonly used SR and average number of queries (AQ). Only the AEs that can fool the target model below the max query budget will be recorded as successful. And the number of queries used by successful AEs will be counted. To evaluate the imperceptibility of AEs, we consider $L_0$, $L_2$, and MAD. All these three imperceptibility metrics are the smaller the better. In practice, a successful AE with imperceptible perturbation is what we need. So besides SR, we use a new metric $\text{SR}_{\text{true}}$. It indicates the rate of successful AEs whose MAD values are below a human-like threshold, i.e., 30, that AEs with MAD $\leq$ 30 are basically imperceptible to human eyes (see Appendix A.3). We further calculate the effective number of queries $\text{EQ} = \frac{\text{AQ}}{\text{SR}_{\text{true}}}$ to denote the number of queries needed to generate a successful AE with imperceptible perturbations. EQ is a comprehensive metric, taking into account attack SR, query efficiency, and imperceptibility.

4.2 Results and Analysis

Some examples of different attacks are shown in Figure 5. We can easily find the perturbations of original Parsimonious Attack and Square Attack are very obvious in the entire image due to global perturbation, while their modified versions are relatively more imperceptible since no perturbation exists in background regions. However, their perturbations are still complex and irregular, taking up almost all salient regions. In comparison, even restricted in the same salient regions, the perturbations generated by Saliency Attack are smaller and more critical, roughly corresponding to the bright pixels in BP-saliency map. They further represent the positions of dogs’ noses or ears, which accord with our inspiration before. Besides, TVDBA produces global and irregular perturbations, and the AEs of Boundary Attack contain noticeable coarse textures due to inadequate query budget. We omit HSJA since it generates similar but blurrier AEs with Boundary Attack’s.

The quantitative results are reported in Table 1 and Table 2. In the experiments, we randomly choose 10,000 images from the ImageNet validation set and divide them into 10 groups. Then we calculate the mean and standard deviation (SD). The best results are recorded in bold based on Wilcoxon signed-rank test with significance level at 0.05. We can find in all query budgets, our
Saliency Attack statistically significantly outperforms all the baselines on different target models with a huge gap in terms of $\text{SR}_{\text{true}}$, $\text{EQ}$, and three imperceptibility metrics, which demonstrates the superiority of our method to refine the perturbations in salient regions. Specifically, for Boundary Attack and HSJA, although they gradually sample AEs to approach the original images while keeping adversarial, their perturbations are hard to decrease to the $L_\infty$ constraint under the max query budget. As an example, it takes at least hundreds of thousands of queries for Boundary Attack to converge \cite{5}, which is infeasible in practice. That is why they have very bad SR and imperceptibility performance (Boundary Attack and HSJA can exhaust all the query budget to reduce perturbations, so we do not count their query number). On the contrary, the remaining baselines and our Saliency attack start from original images and gradually add perturbations, constrained in the pre-defined $L_\infty$ distance. Among TVDBA, Parsimonious Attack, Square Attack, and Saliency

Fig. 5. Examples generated by different attacks. For each example, the upper row is AEs and original image. The lower row is perturbations and BP-saliency map.
Attack, they all adopt the “tiling trick” [18] to sample or perturb a square of pixels to improve efficiency, and Square Attack achieves the best SR and AQ due to the RVS [2] as initializations (a part of test images can directly fool the target model with RVS using only one query). However, this performance significantly sacrifices the imperceptibility of Square attack with worst SR_{true}, EQ, and three imperceptibility metrics among four attacks. For our Saliency Attack, despite slight degradation in SR, it can achieve much better SR_{true}, EQ, and three imperceptibility metrics, which means Saliency Attack can efficiently generate successful AEs with imperceptible perturbations, and are more likely to evade various defenses or human judgements in practical.

For Parsi-sal and Square-sal attack, they obtain better SR_{true} and three imperceptibility metrics than their original version due to restricting perturbations in small but critical regions, which demonstrates different black-box attack can also benefit from salient regions. Their SR and AQ also have some slight degradation because the search space is reduced and some test images need more queries to generate successful AEs. In terms of the comprehensive metric EQ, We can find Square-sal attack is improved while Parsi-sal Attack is degraded. We further study the change of their MAD scores in Appendix A.4 and find most samples’ MAD scores of Square-sal Attack are enhanced due to its random search scheme. But Parsimonious Attack adopts greedy local search, that plays a similar role with salient regions, so the benefit from salient regions is relatively limited.

We also plot the convergence curve of SR_{true} versus query number under different MAD thresholds in Figure 6. We omit Boundary Attack and HSJA since they have very low SR. It shows that Saliency Attack can lead other attacks stably most of the time. For some baselines such as Square Attack and Parsimonious Attack, they can generate some successful AEs with very few queries due to global perturbations and some tricks like RVS as initializations, but they lack the ability to further improve the imperceptibility. Instead, Saliency Attack conservatively selects small regions to perturb, hence its query efficiency is a little lower than some baselines at the beginning but soon exceeds them.
4.3 Ablation Study

We carry out ablation studies of Saliency Attack, including refining in salient region, in non-salient region, and without saliency (refining in the whole image). We also design a greedy search as a baseline to verify our \textit{Refine} search. We test multiple block sizes for greedy search and use 32 as the best choice (see Appendix A.5). The results on 1,000 randomly selected images and some examples are given in Table 3 and Figure 7. Note that refining in salient region and refining without saliency generate the same or almost the same perturbations, which means the salient regions indeed contain useful parts and enhance the query efficiency by limiting the search space. But for refining in non-salient region, its perturbation is more complex and visible with worse query efficiency and SR due to unuseful search space. Compared with greedy search, our \textit{Refine} search has much better query efficiency and \(\text{SR}_{\text{true}}\), which demonstrates its superiority. Therefore, we can conclude that both salient region and \textit{Refine} search facilitate Saliency Attack.

### Table 3. Ablation Study of Saliency Attack against Inception v3 Model

| Method                              | SR  | \(\text{SR}_{\text{true}}\) | AQ  | EQ  | \(L_2\) | \(L_0\) | MAD   |
|-------------------------------------|-----|-----------------------------|-----|-----|---------|---------|-------|
| Refining in salient region          | 93.6% | 86.2%                      | 1958| 2271| 3.71    | 3.8%    | 12.88 |
| Refining in non-salient region      | 78.2% | 57.0%                      | 3128| 5488| 4.94    | 6.5%    | 21.58 |
| Refining without saliency           | 95.5% | 79.6%                      | 2563| 3220| 3.84    | 4.2%    | 16.35 |
| Greedy search in salient region     | 56.0% | 50.7%                      | 2727| 3579| 4.37    | 4.7%    | 12.87 |

4.4 Hyperparameter Sensitivity

The hyperparameter \(k_{int}\) is the initial block size that determines the first level of split blocks. We test the effect of different \(k_{int}\) on SR, query number, and imperceptibility (\(L_2\) and MAD) over 1,000 randomly selected examples in Figure 8. As \(k_{int}\) decreases, SR and imperceptibility can be improved, and SR reaches the peak when \(k_{int}\) equals 16. This is because, with smaller initial blocks, Saliency Attack can search for perturbations more finely and accurately leading to higher SR and better imperceptibility. Meanwhile, inevitably more queries are needed, especially for sorting initial blocks. Under a limited query budget like 10,000, the remaining query budget for searching in finer blocks could be inadequate. That is why a turning point occurs in SR.

We also show some examples in Figure 9. It can be observed that as \(k_{int}\) decreases, the generated perturbations become smaller but more salient. For instance, in the first row, the perturbation with \(k = 128\) roughly covers the region of the dog face, which is also the brightest region in the BP-saliency map. While the perturbation with \(k = 32\) or \(k = 16\) focuses on smaller region of the dog ear. This indicates that our Saliency Attack could find smaller and more important perturbations progressively as the \(k_{int}\) decreases, which coincides with our assumption before.
Saliency Attack: Towards Imperceptible Black-box Adversarial Attack

Fig. 7. Examples generated by Saliency Attack with different strategies.

![Saliency Attack Examples](image)

(a) Salient region  (b) In salient region  (c) Without saliency  (d) In non-salient region

Fig. 8. The effect of initial block size on SR, average query number, and imperceptibility ($L_2$ and MAD).

![Initial Block Size Effect](image)

Fig. 9. Examples generated by Saliency Attack with different initial block sizes.

![Initial Block Size Examples](image)

(a) Salient region  (b) $k_{int} = 128$  (c) $k_{int} = 32$  (d) $k_{int} = 16$  (e) BP-saliency map

For another hyperparameter $\phi$, which is the threshold to produce binary saliency masks from the saliency maps generated by salient object detection models. We use SOTA salient object detection models PFA [47] and TRACER [20] to generate saliency masks with different thresholds {0.025, 0.05, 0.1, 0.2, 0.3}. The results are reported in Table 4 and some examples are displayed in Figure 10. We can find using different salient object detection models with different thresholds generate similar...
Table 4. Results of Saliency Attack against Inception v3 Model using Different Salient Object Detection Models with Different Thresholds

| Method | Threshold | SR   | SRtrue | AQ   | EQ   | L2   | L0   | MAD  |
|--------|-----------|------|--------|------|------|------|------|------|
| PFA    | 0.025     | 92.90% | 85.30% | 1891 | 2217 | 3.69 | 3.80%| 12.52|
|        | 0.05      | 92.00% | 85.60% | 1842 | 2152 | 3.67 | 3.70%| 12.39|
|        | 0.1       | 93.90% | 86.00% | 1979 | 2301 | 3.73 | 3.90%| 12.94|
|        | 0.2       | 89.90% | 84.30% | 1751 | 2077 | 3.58 | 3.50%| 11.89|
|        | 0.3       | 89.60% | 84.50% | 1721 | 2037 | 3.59 | 3.50%| 12.00|
| TRACER | 0.025     | 92.90% | 85.90% | 1891 | 2201 | 3.69 | 3.80%| 12.53|
|        | 0.05      | 92.00% | 85.60% | 1842 | 2152 | 3.67 | 3.70%| 12.51|
|        | 0.1       | 92.10% | 85.30% | 1785 | 2093 | 3.66 | 3.70%| 12.51|
|        | 0.2       | 92.10% | 85.80% | 1784 | 2079 | 3.66 | 3.70%| 12.45|
|        | 0.3       | 92.00% | 85.30% | 1772 | 2077 | 3.66 | 3.70%| 12.40|

Fig. 10. Binary saliency masks from the saliency maps generated by PFA and TRACER with different thresholds. For each example, the upper row is generated by PFA. The lower row is generated by TRACER.

salient masks, extracting the regions of main object(s) in images. Although PFA and TRACER may generate some distinct saliency masks (e.g., the first dog example in Figure 10), these saliency masks all cover the critical regions (e.g., the region of the dog’s face). That is why only slight differences exist in the quantitative results using PFA or TRACER with different thresholds. In our work, we select PFA with $\phi = 0.1$, which slightly outperform others in terms of SR and SRtrue.

4.5 Attacking Detection-based Defense

To further validate the imperceptibility of AEs, we attack against two detection-based defenses that attempt to detect AEs from clean images. The first is Feature Squeezing\(^3\) [44]. Its basic idea is to squeeze out unnecessary input features that may be utilized by an adversary to generate AEs. Comparing the model’s prediction on the original image with its prediction on the image after squeezing, the input is likely to be adversarial if the original and squeezed inputs produce sub-

\(^3\)Implementation of Feature Squeezing https://github.com/mzweilin/EvadeML-Zoo.
Table 5. Results for Attacking against Detection-based Defenses

| Method          | SR   | SR_true | Feature Squeezing DR | Classifier DR |
|-----------------|------|---------|----------------------|---------------|
| TVDBA           | 96.9%| 23.2%   | 26.5%                | 61.9%         |
| Parsimonious    | 98.2%| 14.1%   | 33.3%                | 89.7%         |
| Parsimonious-sal| 96.7%| 12.0%   | 29.7%                | 67.6%         |
| Square          | 99.7%| 1.9%    | 47.8%                | 85.4%         |
| Square-sal      | 97.1%| 22.5%   | 26.1%                | 63.1%         |
| Saliency (ours) | 93.6%| 86.2%   | 21.2%                | 50.4%         |

stantially different outputs. Feature Squeezing adopts color bits reduction of each pixel and spatial smoothing (local and non-local smoothing) as squeezers. We use the recommended parameters ($5 - bit, 2 \times 2, 11 - 3 - 4$) for ImageNet dataset. Another is a CNN-based binary classifier based on a Steganography detector \cite{4}, which is designed for detecting small noise specifically and outperforms classic classification models such as Inception-v3 or ResNet. We finetune the pre-trained model\footnote{\url{https://github.com/brijeshiitg/Pytorch-implementation-of-SRNet}.} with AEs generated by different attacks. Concretely, we randomly select 1,000 clean images and generate the corresponding AEs with Saliency Attack and six baselines, respectively. Thus, we use $2 \times 7,000$ pairs of clean images and AEs as the training set to finetune the model. The training is run for 100 epochs, and the learning rate is 0.001. We test another 1,000 samples and record the detection rate (DR), which is the lower the better.

From Table 5, we can find our Saliency attack can achieve lower DR against both Feature Squeezing and binary classifier detection than other baselines. For Feature Squeezing, although the area of perturbation generated by Saliency Attack is quite small and thus suffering a higher risk of being denoised by squeezers, the DR of Saliency Attack is still better than other attacks due to our effective perturbation. For the binary classifier detector, our Saliency Attack obtains nearly 50% accuracy/DR, which means the perturbation is invisible enough that the classifier will be difficult to converge and results in random guess. Therefore, our Saliency Attack is able to generate imperceptible AEs that evade different detection-based defenses.

5 CONCLUSION

In this article, we propose restricting black-box attacks to perturb in salient regions to improve the imperceptibility of AEs, and propose a novel black-box attack, Saliency Attack. Experiments show that SOTA black-box attacks restricted in salient regions can indeed achieve better imperceptibility performance, while Saliency Attack further enhances this by recursively refining perturbations in salient regions. Its perturbation is much smaller, imperceptible, and interpretable to some extent. Besides, we also find that the salient regions of some examples are indeed progressive with respect to their impact on model output, which is consistent with our assumption from the previous visualization studies. Finally, the evaluation demonstrates that the AEs generated by Saliency Attack are harder to be detected, which validates its imperceptibility from the defensive perspective.

Although our Refine search provides significant benefit to the imperceptibility, it inevitably consumes more queries compared with the black-box attacks with global perturbations. In the future, we will try to find better ways to balance the three objectives of query efficiency, imperceptibility, and attack SR for black-box attacks. Furthermore, we will apply the salient region to more black-box attacks to improve their imperceptibility. Our Saliency Attack could also be tested against other kinds of defenses, such as robustness-based defenses (adversarial training, network distillation, etc.), to validate its effectiveness.
APPENDIX

A.1 MAD Metric

MAD metric [19] is one of the SOTA full-reference IQA methods. MAD attempts to merge two separate strategies for two kinds of distorted images, respectively. For high-quality images with near-threshold distortion (just visible), MAD focuses on detection-based strategy to look for distortions in the presence of the image. While for low-quality images with suprathreshold distortion (clearly visible), MAD focuses on appearance-based strategy to look for image content in the presence of the distortions. MAD will control the weight of two strategies according to the type of distorted images. The calculation process of MAD can be summarized as following steps and we recommend interested readers to read the original literature.

- Compute locations of visible distortions based on luminance images.
- Combine the visibility map with local error image.
- Decompose both the distorted and original images into log-Gabor subbands.
- Calculate different statistics of each subband.
- Calculate the adaptive blending score.

A.2 PFA Network for Saliency Detection

PFA network [47] is a novel salient object detection method considering the different characteristics level features. Specifically, the saliency maps from low-level features contain many noises, while the saliency maps from high-level features only get an approximate area. Therefore, PFA network first devises a context-aware pyramid feature extraction (CPFE) module to get multi-scale multi-receptive-field high-level features, and then uses channel-wise attention (CA) to select appropriate scale and receptive-field for generating saliency regions. On the other hand, to refine the boundaries of saliency regions, PFA network uses spatial attention to better focus on the effective low-level features, and obtain clear saliency boundaries. After the processing of different attention mechanisms, the high-level features and low-level features are complementary-aware and suitable to generate saliency map. Besides, an edge preservation loss is proposed to guide the network to learn more detailed information in boundary localization. With these powerful feature extraction methods and attention mechanisms, PFA network can achieve robust and effective salient object detection, outperforming SOTA methods under different evaluation metrics.

A.3 Threshold for MAD Metric

Since Boundary Attack [5] can reduce the distortion gradually, we generate multiple AEs with different MAD scores in Figure A.1 to find a proper threshold for MAD metric.
Fig. A.1. Adversarial examples generated by boundary attack with different MAD scores. We can find the threshold MAD ≤ 30 is roughly enough to indicate an imperceptible adversarial example.

A.4 Scatter Plot of MAD Scores

As shown in Figure A.2, via restricting perturbations in salient region, 68.1% and 94.3% examples of Parsimonious-sal attack and Square-sal attack have better MAD scores compared with their original version, respectively. Since Parsimonious Attack uses a greedy local search that plays a similar role with salient regions, while Square attack adopts a random search for sampling perturbations, it is obvious that limiting the search in salient regions is more helpful to Square Attack than Parsimonious Attack.

Fig. A.2. MAD scores of Parsimonious Attack versus Parsimonious-sal attack and Square Attack versus Square-sal Attack.
A.5 Greedy Search in Salient Region with Different Block Sizes

We test different block sizes for greedy search in salient region in Table A.1, and choose the best one (block size equals 32) to be compared in ablation study.

| Block size | SR | SR<sub>TRUE</sub> | AQ | EQ | L₂ | L₀ | MAD |
|------------|----|-------------------|----|----|----|----|-----|
| 128        | 20.6% | 18.4% | 57 | 310 | 7.32 | 12.8% | 11.50 |
| 64         | 37.4% | 31.3% | 512 | 1636 | 6.37 | 10.2% | 15.18 |
| 32         | 56.0% | 50.7% | 2727 | 5379 | 4.37 | 4.7% | 12.87 |
| 16         | 35.4% | 35.2% | 4039 | 11474 | 1.79 | 0.8% | 4.84 |

REFERENCES

[1] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B. Srivastava. 2019. GenAttack: Practical black-box attacks with gradient-free optimization. In Proceedings of the Genetic and Evolutionary Computation Conference. 1111–1119.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square attack: A query-efficient black-box adversarial attack via random search. In Proceedings of the 16th European Conference on Computer Vision, Part XXII, Vol. 12368. 484–501.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations.

[4] Mehdi Boroumand, Mo Chen, and Jessica J. Fridrich. 2019. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security 14, 5 (2019), 1181–1193.

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In Proceedings of the 6th International Conference on Learning Representations.

[6] Nicholas Carlini and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy. CA, 39–57.

[7] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. 2020. HopSkipJumpAttack: A query-efficient decision-based attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy. 1277–1294.

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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. 15–26.

[9] Francesco Croce and Matthias Hein. 2019. Sparse and imperceivable adversarial attacks. In Proceedings of the 2019 IEEE International Conference on Computer Vision. 4723–4731.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 248–255.

[11] Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiaying Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, and Nenghai Yu. 2020. Robust Superpixel-guided attentional adversarial attack. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. 12892–12901.

[12] Sid Ahmed Fezza, Yassine Bakhti, Wassim Hamidouche, and Olivier Déforges. 2019. Perceptual evaluation of adversarial attacks for CNN-based image classification. In Proceedings of the 11th International Conference on Quality of Multimedia Experience. 1–6.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In Proceedings of the 3rd International Conference on Learning Representations.

[14] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. 2019. Low frequency adversarial perturbation. In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence, Vol. 115. 1127–1137.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 770–778.

[16] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29, 6 (2012), 82–97.
Saliency Attack: Towards Imperceptible Black-box Adversarial Attack

[17] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In Proceedings of the 35th International Conference on Machine Learning. Vol. 80. 2142–2151.

[18] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. 2019. Prior convictions: Black-box adversarial attacks with bandits and priors. In Proceedings of the 7th International Conference on Learning Representations.

[19] Eric C. Larson and Damon M. Chandler. 2010. Most apparent distortion: Full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging 19, 1 (2010), 011006.

[20] Min Seok Lee, WooSeok Shin, and Sung Won Han. 2022. TRACER: Extreme attention guided salient object tracing network (student abstract). In Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event, 12993–12994.

[21] Nannan Li and Zhenzhong Chen. 2021. Toward visual distortion in black-box attacks. IEEE Transactions on Image Processing 30 (2021), 6156–6167.

[22] Xurong Li, Shouling Ji, Meng Han, Juntao Ji, Zhenyu Ren, Yushan Liu, and Chunming Wu. 2021. Adversarial examples versus cloud-based detectors: A black-box empirical study. IEEE Transactions on Dependable and Secure Computing 18, 4 (2021), 1933–1949.

[23] Shengcai Liu, Ning Lu, Wenjing Hong, Chao Qian, and Ke Tang. 2021. Effective and imperceptible adversarial textual attack via multi-objectivization. arXiv:2111.01528. (2021). Retrieved from https://arxiv.org/abs/.

[24] Shengcai Liu, Ning Lu, Cheng Chen, and Ke Tang. 2022. Efficient combinatorial optimization for word-level adversarial textual attack. IEEE ACM Transactions on Audio, Speech, and Language Processing 30 (2022), 98–111.

[25] Seungyong Moon, Gaon An, and Hyun Oh Song. 2019. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In Proceedings of the 36th International Conference on Machine Learning, Vol. 97. 4636–4645.

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2574–2582.

[27] Nina Narodytska and Shiva Kasiviswanathan. 2017. Simple black-box adversarial attacks on deep neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 1310–1318.

[28] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. Distill 3, 3 (2018), e10.

[29] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In Proceedings of the IEEE European Symposium on Security and Privacy. 372–387.

[30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In Proceedings of the 26th British Machine Vision Conference. 41.1–41.12.

[31] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 779–788.

[32] Alexander Schrijver. 1999. Theory of Linear and Integer Programming. Wiley.

[33] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision. 618–626.

[34] Mahmoud Sharif, Lujo Bauer, and Michael K. Reiter. 2018. On the suitability of Lp-norms for creating and preventing adversarial examples. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 1605–1613.

[35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the 2nd International Conference on Learning Representations.

[36] Jiawei Su, Danilo Vasconcellos Vargas, and Kouchi Sakurai. 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23, 5 (2019), 828–841.

[37] Qi Sun, Arjun Ashok Rao, Xufeng Yao, Bei Yu, and Shiyuan Hu. 2020. Counteracting adversarial attacks in autonomous driving. In Proceedings of the IEEE/ACM International Conference On Computer Aided Design. 83:1–83:7.

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2818–2826.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations.

[40] Chun-Chen Tu, Pai-Shun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. 2019. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence. 742–749.
[41] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. 2021. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. In Proceedings of the 30th International Joint Conference on Artificial Intelligence. 3125–3133.

[42] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612.

[43] Tao Xiang, Hangcheng Liu, Shangwei Guo, Tianwei Zhang, and Xiaofeng Liao. 2021. Local black-box adversarial attacks: A query efficient approach. arXiv:2101.01032 (2021). Retrieved from https://arxiv.org/abs/2101.01032.

[44] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. In Proceedings of the 25th Annual Network and Distributed System Security Symposium.

[45] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In Proceedings of the 13th European Conference on Computer Vision, Part I, Vol. 8689. 818–833.

[46] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. 2020. Smooth adversarial examples. EURASIP Journal on Information Security 2020, 1 (2020), 15.

[47] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3085–3094.

[48] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. 2020. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, 1036–1045.

[49] Bolei Zhou, Aditya Khosla, Ágata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2921–2929.

Received 4 June 2022; revised 13 December 2022; accepted 9 January 2023