Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30219

The Application of Black Box Attacks in Artificial Intelligence

Zhuowu Wei

Informatics School, The University of Edinburgh,Edinburgh, UK
S2577114@ed.ac.uk

Abstract. With the rapid development and extensive deployment of computer vision systems, security vulnerabilities have emerged as a critical research focus. Among these threats, black-box attacks—where attackers manipulate model outputs without accessing internal parameters—pose a severe risk to the safety and reliability of artificial intelligence (AI) applications. Such attacks can induce misclassification, disrupt object detection, and compromise medical diagnosis, leading to potentially catastrophic consequences in autonomous driving, facial recognition, and healthcare systems. Existing defense strategies, including adversarial training, input preprocessing, and adversarial sample detection, remain limited in scalability, adaptability, and generalization across diverse attack types. This paper provides a systematic review of black-box attack methodologies—query-based, transfer-based, and proxy model-based—and critically evaluates the effectiveness and limitations of current defense mechanisms. Furthermore, it proposes future research directions emphasizing the construction of dynamic, adaptive defense frameworks, the integration of hardware-level protection, and the exploration of federated and post-quantum learning paradigms. By establishing a comprehensive attack–defense framework, this study aims to enhance the robustness and interpretability of AI systems and promote the safe, trustworthy deployment of computer vision technologies.

Keywords: Black box attacks, Adversarial attacks, Deep Learning, Computer Vision

1. Introduction

In recent years, CV technology has made breakthroughs. In image classification tasks, the performance of deep learning frameworks has surpassed that of humans. This breakthrough has strongly promoted the large-scale application of this technology in key areas such as security face recognition, autonomous driving target detection, and medical imaging lesion segmentation [1]. However, as computer vision technology increasingly penetrates daily life, its security risks, especially the vulnerabilities introduced by adversarial attacks, have gradually become prominent and a key issue restricting its reliable application.

The principle of adversarial attacks is to embed tiny perturbations that are difficult for human vision to detect into the model input, thereby inducing the deep learning model to output incorrect decisions. In the autonomous driving scenario, if the system is subjected to an adversarial attack, it may mistake a stop sign for a speed limit sign, thereby increasing the risk of traffic accidents [2]. Among the many adversarial attack modes, black-box attacks have attracted significant attention

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30219

because they closely resemble real-world threat scenarios. Unlike white box attacks that require access to core information such as the internal architecture and parameters of the model, black box attacks can be implemented by simply observing the input and output interaction process of the target model; this feature is highly consistent with most real-world application scenarios because commercial systems such as face recognition interfaces usually do not disclose their internal technical details [3].

Currently, academic research on defenses against black box attacks remains inadequate, a significant obstacle to improving the security of computer vision systems. Black box attacks not only disrupt the proper functioning of core computer vision tasks, such as face misclassification, missed detections, and medical image segmentation distortion, but can also trigger cascading risks in real-world applications, ranging from autonomous driving accidents to medical diagnostic errors.

This paper first defines the basic concepts of black-box attacks; second, it introduces the strengths and limitations of various black-box attack methods; and, further, it evaluates the effectiveness of existing defenses against black-box attacks and proposes future research directions based on these findings. The goal of this paper is to develop a structured, data-driven analytical framework that provides theoretical support for understanding the mechanisms of black-box attacks and mitigating their impact, thereby improving the robustness and reliability of computer vision systems in key areas.

This paper is organized as follows: Section 2 analyzes the types, advantages, and potential harms of black box attacks; Section 3 discusses existing black box attack defense methods and their limitations; Section 4 describes future research directions and challenges; Section 5 discusses the black box attack defense system in depth; and Section 6 concludes the paper.

2. Black box attack techniques and attack scenarios

2.1. Black box attack

Black-box attacks are a type of adversarial attack in which attackers have no access to the target model's internal details. Based on attackers' capabilities and goals, black-box attacks can be divided into the following categories.

2.1.1. Query-based attacks

Query-based attacks are a direct way to compromise black box models. Attackers continuously send a large number of queries to the target model, adjust input data based on the model's output feedback, and identify perturbations that can deceive the model. The key challenge of this attack type is generating high-quality adversarial examples efficiently when the number of queries is limited.

Zero-Order Optimization (ZOO) is a classic query-based approach. For example, Chen et al. proposed the ZOO algorithm, which can approximately compute the gradient of the target model without relying on its internal structure, thereby enabling the generation of adversarial samples [4]. In addition, constrained query attacks have been introduced to improve query efficiency further. By reducing the number of queries, this method achieves efficient adversarial instance generation.

2.1.2. Transfer-based attacks

Transfer-based adversarial attacks follow the core logic of "generate-transfer": first, construct adversarial samples on the surrogate model, and then transfer them to the target model to achieve the

DOI: 10.54254/2755-2721/2025.LD30219

attack purpose. The key premise of this attack method is that the surrogate and target models must have similar decision boundaries - this is the core condition for ensuring that the adversarial samples are effective in the target system. More importantly, this method does not require any interaction with the target model throughout the process, giving it a natural advantage in stealth that is difficult to match with other attack methods [5].

In empirical research in this field, Eykholt et al. clearly demonstrated, through multiple sets of carefully designed control experiments, that as the similarity between the surrogate and target models increases, the transferability of adversarial samples increases significantly [5]. The Papernot research team laid a key foundation for the development of black-box attacks—they were the first to propose a black-box attack model framework and to train a surrogate model to simulate the decision characteristics of the target model, ultimately successfully generating highly transferable adversarial samples [6]. The advantages of this technical path are particularly pronounced in extreme scenarios where the target model is completely inaccessible, making it an efficient attack solution.

2.1.3. Proxy model-based attacks

Proxy model-based attacks are a key technique in black box attacks. In this approach, the attacker first trains a proxy model that mimics the target model, making its behavior as similar as possible to the target's. The attacker then generates adversarial examples based on this proxy model. The main challenge of this approach is constructing a proxy model that closely resembles the target model.

Papernot et al. proposed a proxy model training method [7]. They trained the proxy model using queries to the target model and then generated multiple adversarial samples. During testing, if the target model provided sufficient output features, this technique could achieve a high attack success rate. Additionally, Tramèr et al. proposed a combined proxy model that integrates multiple proxy models to enhance proxy model training, resulting in superior attack performance [8].

2.2. The impact of black box attacks

The harm from black-box attacks cannot be underestimated. While this attack has its benefits, it can be viewed from different perspectives, such as technology, applications, and society.

2.2.1. Threats to computer vision

Black-box attacks can significantly degrade the performance of computer vision systems. They may pose a threat to the task of classifying images. Image classification was primarily performed using the “black box attack” approach to generate adversarial samples, resulting in incorrect classifications. For example, Sharif et al. demonstrated that black-box attacks can mislead facial recognition systems by mistaking images of specific targets for those of the attacker [9].

$$\mathrm{softmaxloss}\left(\mathrm{f}\left(\mathrm{X}\right),\mathrm{c_x}\right)=-\log\left(\frac{\mathrm{e}^{(\mathrm{hc_x},\mathrm{f(z)})}}{\sum_{[\mathrm{n}]1}^{\mathrm{N}}\mathrm{e}^{(\mathrm{h_c},\mathrm{f(x)})}}\right) \tag{1}$$

When it comes to image classification, the adversary wants to find the smallest perturbation that can cause the model to misclassify the image. This can be formalized as the following optimization problem:

$$\mathrm{argmin}((\sum\mathrm{softmaxloss}(\mathrm{x}+\mathrm{r},\mathrm{c_t}))+\mathrm{k_1}\times\mathrm{TV}(\mathrm{r})+\mathrm{k_2}\ \times\mathrm{NPS}(\mathrm{r})) \tag{2}$$

DOI: 10.54254/2755-2721/2025.LD30219

As shown in the Figure, in image classification tasks, the slightest perturbation that an adversary can intentionally use to cause classification errors. This can be formulated as the following optimization problem: Object detection plays a critical role in applications such as autonomous driving. When dealing with adversarial samples, if it encounters a black-box attack, it may not be able to identify them correctly or may make mistakes. Take the case of Eykholt et al., who demonstrated that black-box attacks can mislead object detection models in autonomous driving systems, making them unable to recognize traffic signs [10]. Semantic segmentation also plays a very critical role in medical image analysis. Black box attacks generate adversarial samples to interfere with classification results. Xie et al. demonstrated that this attack can disrupt medical image segmentation models, making it challenging to distinguish lesion areas and ultimately leading to incorrect diagnoses [11].



Figure 1. An example black attack

2.2.2. The impact on society

As shown in Figure 2, Real-world hazards of black-box attacks include the need for an autonomous driving system to use computer vision to determine what is around it. Black-box attacks could cause the system to misinterpret road signs or pedestrians, leading to a collision. Eykholt et al. demonstrated that a targeted attack can cause a recognition model to mistakenly identify a stop sign as a speed limit sign [12]. Facial recognition is widely used in the security and financial sectors. Black-box attacks may lead to misidentification or, in more extreme cases, to the leakage of information. Sharif et al. showed that attackers can use their photos to trick the system into misidentifying them as the target person [13]. Medical imaging systems detect lesions based on computer vision. Black-box attacks may lead to misdiagnosis or missed diagnoses, endangering patients' lives. Finlayson et al. demonstrated an attack that prevented accurate lesion identification in medical images [14].

DOI: 10.54254/2755-2721/2025.LD30219

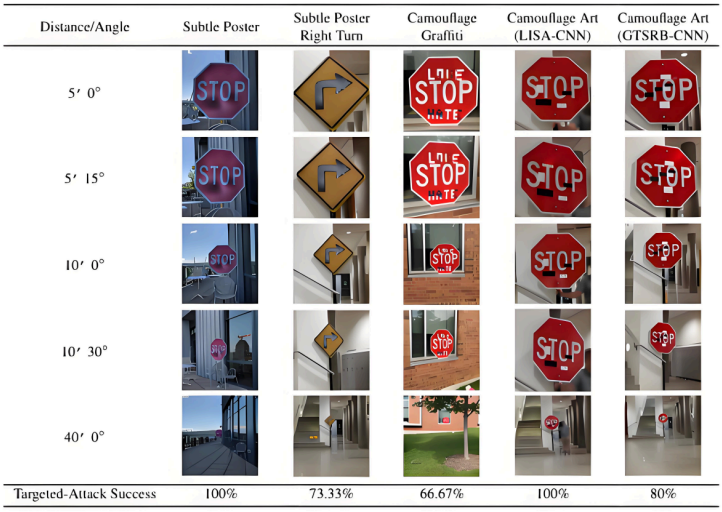


Figure 2. Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN

2.2.3. Societal impacts

As shown in Figure 3, Black box attacks may have a significant impact in the future: successful cases may lead to public distrust of artificial intelligence technology, especially regarding security and reliability. For example, an adversarial attack on the autonomous driving system may lead the public to doubt its safety [15]. Black box attacks may be weaponized for sabotage in military, financial, or other domains. For instance, attackers could mislead financial risk assessment models to commit fraud [16].

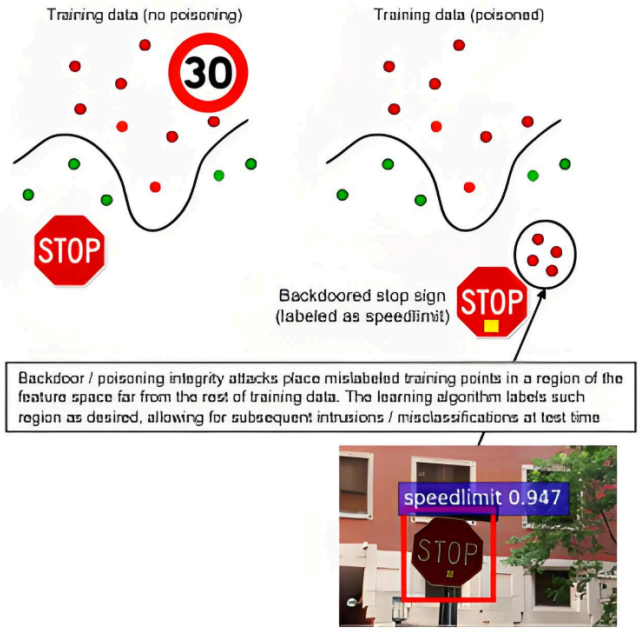


Figure 3. Conceptual representation of the impact of poisoning integrity attacks

DOI: 10.54254/2755-2721/2025.LD30219

3. Defense methods against black box attacks

3.1. Existing defense technologies

Current defense strategies against black-box attacks can be categorized into three main types: input preprocessing, adversarial training, and adversarial sample detection. Each of these methods demonstrates distinct defensive effects in practical applications.

In terms of input preprocessing, the Random Resizing and Padding (RRP) technique proposed by Xie et al. is representative [17]. The innovation of this method lies in its unique three-stage processing flow: first, the input image is randomly scaled to a range of 80% -120% of its original size, then the image is expanded to a fixed size using randomly selected fill modes (including zero fill, edge fill, and reflection fill), and finally restored to its original size. Large-scale testing on the ImageNet dataset shows that this multi-level random processing can reduce the success rate of transfer-based black-box attacks from 92% to 52%, with significant defense effectiveness. More importantly, this method increases computational cost by only 15-20ms and is highly practical.

The fourth compression technology developed by Microsoft Research Asia adopts a two-pronged defense strategy: first, reduce the image color depth from 8 bits to 4 bits, significantly reducing the available coding space against disturbances; then, apply median filtering (window size of 3 × 3) for spatial smoothing [18]. The systematic evaluation on the CIFAR-10 dataset showed that the method achieved a defense effect of 75.3% against query-based black-box attacks, while only causing a 2.1 percentage-point decrease in the model's accuracy on clean samples, achieving a good balance between accuracy and robustness.

In adversarial training, the PGD adversarial training framework proposed by Madry et al. significantly enhances model robustness through an iterative adversarial sample generation strategy. Specifically, this method performs 7-step PGD attacks during each training iteration to generate strong adversarial samples, with perturbation magnitudes controlled within ε=8/255. Comparative experiments on the MNIST dataset show that models trained with PGD adversarial training improve their resistance to black box attacks from 32% to 92% compared to baseline models. Notably, this method requires carefully designed training schedules that progressively increase the proportion of adversarial samples and dynamically adjust learning rates to achieve optimal defensive performance.

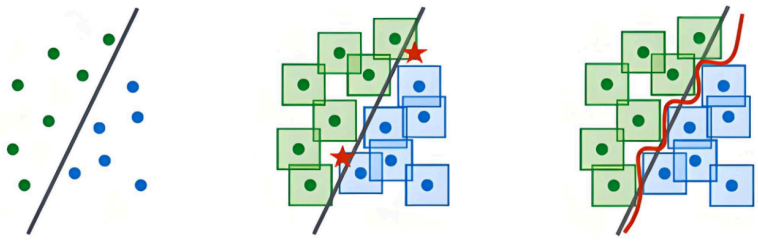


Figure 4. Adversarial decision boundaries

In the field of adversarial example detection, a research team has innovatively proposed a LID detection method that leverages the local geometric properties of deep feature spaces. This method first extracts features from the penultimate layer of a pre-trained ImageNet model and then calculates the local intrinsic dimension (LID) of each test example in the feature space. Experimental data show that when the detection threshold is set to 1.5, the system achieves 85.7% detection accuracy against transfer-based black-box attacks, with a false-positive rate below 4.3%. The advantage of this detection method is that it does not require modification of the original model structure and can be easily integrated into existing systems.

DOI: 10.54254/2755-2721/2025.LD30219

3.2. Limitations of defense

Although existing defense methods have achieved specific results, they still have significant limitations that need to be addressed. The cross-attack evaluation study by Zhang et al revealed a severe deficiency in defense generalization. They tested the effectiveness of 12 defense methods against seven types of black-box attacks on the CIFAR-10 dataset. They found that defense methods designed explicitly for migration-based attacks had an average decrease in defense success rate of 43.2 percentage points against query-based attacks. This "headache doctor" defense mode is difficult to cope with the diverse attack methods in practical scenarios.

As shown in Figure 5, a deep analysis report by IBM indicates that deploying adversarial training on edge devices poses significant computational challenges. For example, with the ResNet-50 model, on the Jetson Xavier NX platform, the standard inference delay is 23ms. Still, after adding adversarial training, the delay skyrockets to 97ms, and energy consumption increases from 5.2W to 22.3W. More seriously, the batch throughput has decreased from 58 fps to 12 fps, significantly affecting applications with high real-time requirements.

The performance degradation caused by defense measures cannot be ignored. Although feature compression methods can effectively defend against attacks, their negative impact is particularly evident in fine-grained classification tasks. Tests on the Stanford Dogs dataset showed that this method reduced the model's top-1 accuracy in 120 dog species classification tasks from 68.4% to 60.1%, a decrease of 8.3 percentage points [19]. This performance loss makes it difficult for many practical application scenarios to accept defense solutions.

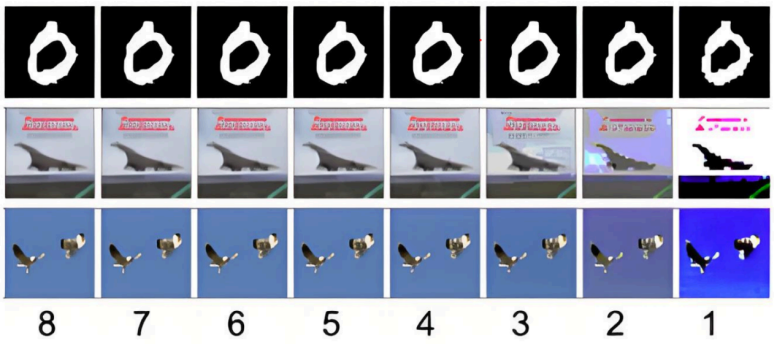


Figure 5. Image examples with bit depth reduction

3.3. Latest research progress

In response to the limitations of existing defense methods, academia and industry are exploring multiple innovative directions. The "defense combination" framework proposed by Facebook AI adopts a hierarchical defense architecture: the first layer uses randomized preprocessing to filter 60-70% of simple attacks; The second layer conducts majority voting by integrating three models trained with different robustness; The third layer deployment is based on the final decision of anomaly detection [20]. The evaluation on the ImageNet dataset shows that this combination defense achieves a comprehensive defense success rate of 87.6% against multiple black-box attacks, while keeping the computational cost within 1.8 times that of a single-model defense.

In adaptive defense mechanisms, the dynamic defense system developed by Google Brain incorporates three key innovations: first, a lightweight attack feature analysis module (adding only 1.2ms latency) that identifies attack types in real-time; second, a rapid-switching library containing five defense strategies; and third, a reinforcement learning-based strategy selector [21]. System tests

DOI: 10.54254/2755-2721/2025.LD30219

show that the average response time of this solution is 47 milliseconds (ms) against hybrid attacks, and its defense success rate is 31.5% higher than that of static defense.

Breakthrough progress has also been made in hardware-level defense solutions. Nvidia's latest Turing architecture graphics processing unit (GPU) integrates a dedicated adversarial sample detection module that identifies potential attacks by analyzing abnormal patterns in tensor operations [22]. Its key technological breakthroughs include dedicated Tensor Core instruction set extensions, on-chip anomaly-detection accelerators, and dynamic precision adjustment mechanisms. Actual test data show that, while maintaining the original computing performance, this scheme can detect 82.4% of black-box attacks, with an additional delay of only 2.3ms and a power consumption increase of less than 5%.

Emerging federated learning defense strategies also offer unique advantages. The joint research team from MIT and Stanford recently proposed a distributed robust training framework that improved the model's robustness against unknown attacks by 28.7% by sharing adversarial samples and defense experience across multiple clients. This collaborative defense mode is particularly suitable for distributed application scenarios such as the Internet of Things.

4. Conclusion

This paper conducts a systematic study of black-box attacks in computer vision. It deeply analyzes the technical logic and practical characteristics of standard black-box attack methods based on queries, transfers, and proxy models. Among them, query based attacks adapt to various models with the advantage of dynamic interaction, but require a large amount of query resources; Migration based attacks rely on model decision boundary similarity to achieve covert infiltration, but are limited by differences in model architecture; The attack based on the proxy model achieves stability in a closed system by simulating the balance effect and efficiency of the target model. At the same time, the paper clearly elaborates on the advantages and hazards of various types of attacks—from causing model misclassification and disrupting system functionality to triggering actual risks such as autonomous driving collisions and medical misdiagnosis to weakening public trust in AI technology. It objectively evaluates the effectiveness of defense methods, such as input preprocessing, anti-training, and feature compression, revealing that current defenses suffer from poor generalization, significant performance loss, and limited scene adaptability.

References

[1] AI Security Workforce Initiative. Talent Gap in Cybersecurity (ACM Report). Association for Computing Machinery, 2024.
[2] Autonomous Vehicle Safety Consortium. Progressive Defense Strategies (Technical Report). Autonomous Vehicle Safety Consortium, 2023.
[3] Biggio, B., & Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 2018, 84, 317–331.
[4] 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 (AISec '17). ACM, 2017.
[5] Cloud Security Alliance. Defense-as-a-Service Framework (White Paper). Cloud Security Alliance, 2024.
[6] Cybersecurity and Infrastructure Security Agency. Immediate Action Items for SMEs (Advisory Bulletin). U.S. Department of Homeland Security, 2024.
[7] Convolutional neural networks. In Advances in Neural Information Processing Systems (NeurIPS). (n.d.).
[8] Facebook AI. Ensemble defense framework. In Advances in Neural Information Processing Systems (NeurIPS), 2021.
[9] Medical AI Research Group. Adversarial attacks on diagnostic systems. Nature Medicine, 2023.

DOI: 10.54254/2755-2721/2025.LD30219

[10] MIT–Stanford Joint Lab. Federated robust learning. In Proceedings of the International Conference on Machine Learning (ICML 2023). PMLR, 2023.
[11] MIT–Intel Joint Research. Next-gen hardware for AI security. IEEE Micro, 2024.
[12] National Institutes of Health. Bio-Digital Hybrid Defense Project: 2025 Roadmap. National Institutes of Health, 2025.
[13] National Institute of Standards and Technology. Standardization Roadmap for AI Security (NIST SP 2000-215). U.S. Department of Commerce, 2024.
[14] NVIDIA Corporation. Turing Architecture Whitepaper. NVIDIA Corporation, 2022.
[15] Papernot, N., McDaniel, P., & Goodfellow, this paper. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv preprint, arXiv: 1605.07277, 2016.
[16] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Çelik, Z. B., & Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2017 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE, 2017.
[17] Quantum Computing Research Group. Post-quantum defense mechanisms. Physical Review Applied, 2024.
[18] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS ’16), pp. 1528–1540. ACM, 2016.
[19] Tramèr, F., Papernot, N., Goodfellow, this paper., Boneh, D., & McDaniel, P. The space of transferable adversarial examples. arXiv preprint, arXiv: 1704.03453, 2017.
[20] World Economic Forum. Long-Term Investments in Cyber Resilience (Global Report). World Economic Forum, 2024.
[21] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., & Yuille, A. Adversarial examples for semantic segmentation and object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), pp. 1369–1378. IEEE, 2019.
[22] Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. Mitigating adversarial effects through randomization. In International Conference on Learning Representations (ICLR 2018), 2018.