Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU32310

Black-Box Adversarial Attacks: A Threat Modeling Perspective on Score-Based, Decision-Based, and Transfer-Based Paradigms

Hongyue Fang

Department of Cyber Science and Engineering, Nanjing University of Science and Technology, Wuxi, China
fanghongyue@njust.edu.cn

Abstract. Due to the growing use of deep learning models in contexts where security is very important, these models are usually made unreachable to internal parameters and architecture to maintain intellectual property by the model providers. This restricts the attackers to utilizing just a few external interfaces, and so black-box adversarial attacks present the most realistic threat scenario and a central point of security testing. In this paper, black-box adversarial attacks have been reviewed using the lens of threat modeling and the three paradigms include score-based, decision-based, and transfer-based. It highlights the main concepts and the characteristic works of each of the paradigms and presents the recent significant developments of each. The paper shows that the association between these paradigms can not merely be that of strength and weakness; their effectiveness and cost structures are co-determined by the nature of the feedback and interaction constraints of the target model. Moreover, it is indeed becoming more and more common in recent years that hybrid strategies are obtained that combine attack practices of the customs of other paradigms. This paper gives a brief source of knowledge about the application requirements, evaluation aspects, and technological development of every black-box attack paradigm to provide the foundation for future defensive design and risk determination.

Keywords: Black-box Adversarial Attacks, Score-based Attacks, Decision-based Attacks, Transfer-based Attacks

1. Introduction

With large-scale deep learning models being used on cloud services, mobile applications and commercial APIs, there are usually strong protection measures both on the parameters and structure of that architecture and on the training specifics. The interfaces that external attackers have access to are usually the inputs submitted and their output is limited. On these practical assumptions, the application of adversarial analysis according to the traditional white-box assumptions becomes challenging to implement. Therefore, more relevant to real-life circumstances, black-box adversarial attacks have become a crucial subject of adversarial machine learning studies and security analysis [1].

© 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: 10.54254/2755-2721/2026.GU32310

Black-box adversarial attacks are attack techniques in which the attackers do not have access to information about the internals of the model. They depend on feedback of the model to the query inputs to create adversarial samples and induce the model to make wrong predictions or generate desired results [2]. Black-box adversarial attacks are further hampered by other challenges like lack of feedback, limited queries and costly queries compared to the white-box attacks [3]. Thus, improved efficiency of queries is an important line of research in the black-box adversarial attacks to accomplish successful attacks with a limited query budget. Various researchers [3-5] have suggested different approaches in this direction, such as effective query-based attacks.

The paper sorts black-box adversarial attacks using the threat model, which is in line with existing classification techniques that are used in black-box attacks [6]. Upon defining the main principles and development pattern of classical methods, the paper is structured to align and compare the most representative techniques of the offered methods suggested in the last several years. This paper will attempt to present a perspective of relative conciseness of reference in understanding the differences and applicability conditions between the various black-box attack paradigms.

2. Fundamentals of black-box adversarial attacks

This paper discusses black-box adversarial attacks in the context of adversarial examples. Black-box adversarial attacks do not provide the attacker with the opportunity to access the internal information of the target model. Instead, they treat the target model as a query-only black-box function, generating adversarial examples by querying the model with inputs and using output-feedback to create suitable perturbations [2]. The aim of the attack is to make the outputs erroneous while satisfying the perturbation constraints. These limitations can be separated in terms of the $_{p}$ norms or using structural limitations. There are also black-box adversarial attacks, which can be classified based on the type of target: untargeted and targeted. The former simply requires that the output must be different from the actual class, whereas the latter expects that the output must be within a given target class, as it may be more expensive in terms of queries.

The important aspect of black-box attacks is the feedback information available to the attacker. Attackers are not aware of the internal structure and parameters of the model; hence, they are required to query the model in order to receive the measurements such as confidence scores relative to input samples. It is known as a query-based attack because this type of attack involves retrieving the information by queries. According to the response received from queries, the attackers choose suitable techniques to produce the adversarial examples and run the attacks against the model. Judging by the nature of the feedback, black-box adversarial attacks are divided into score-based attacks and decision-based attacks. Also, there exist scenarios in which the attack may not require or may only slightly require queries to the target model. In the transfer-based attacks, attackers make surrogate models on the basis of obtained information and produce adversarial examples on surrogate models with the application of white-box methods.

The most basic part of black-box attacks is the top-1 misclassification, where only the highest confidence category is to be misclassified. Recent studies have made the top-K attack goals stricter and require the top-K prediction set in the model or the ranking structure to be adapted to correspond to realistic multi-label retrieval applications [7,8].

Black-box adversarial attacks are normally evaluated on the effectiveness of the attack, query cost, perturbation and generalization ability. Recent knowledge [6] developed a universal assessment threshold of black-box adversarial attacks. Attack Success Rate (ASR) is the fundamental universal

DOI: 10.54254/2755-2721/2026.GU32310

data of attack effectiveness. For a sample set $\{(x_i, y_i)\}_{i=1}^{N}$, let the adversarial sample be $x_i' = x_i + \delta_i$, and the target model output be $f(\cdot)$, A untargeted attack can be written as

$$\mathrm{ASR}_{unt} = \frac{1}{N}\sum_{i=1}^{N} 1\left[f\left(x_i^{'}\right) \neq y_i\right] \tag{1}$$

Targeted attacks for class t can be expressed as

$$\mathrm{ASR}_{tar} = \frac{1}{N}\sum_{i=1}^{N} 1\left[f\left(x_i^{'}\right) = t\right] \tag{2}$$

In the transfer-based attack, ASR often further aggregates the results by averaging over the target model set $\mathcal{M}$, which facilitates the comparison of different transfer strategies.

$$\mathrm{ASR}_{trans} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathrm{ASR}^{(m)} \tag{3}$$

The magnitude of perturbation is used to quantify the difference between the adversarial sample and the original input, and is an important indicator for measuring the concealment of the attack. In visual tasks, related works usually report the size of the adversarial perturbation under $_2$ or $_\infty$ constraints, or compare the attack success rates of different methods under a fixed norm budget.

For query-based attacks, the query cost is usually represented by the average and median of the number of queries per sample (qi), and the combined use of both can reduce the interference from extreme values and better approximate the actual cost.

$$Q_{\mathrm{mean}} = \frac{1}{N}\sum_{i=1}^{N} q_i, Q_{med} = \mathrm{median}\left(\{q_i\}_{i=1}^{N}\right) \tag{4}$$

Ideal attacks tend to achieve both high ASR and low $\mathrm{Q_{mean}}$, $\mathrm{Q_{med}}$.

For decision-based attacks, the evaluation also often focuses on the process where the perturbations gradually shrink with each query iteration under fixed success conditions.This can be represented by either $_2$ or $_\infty$. Some benchmark works also use MSE to characterize the perturbation magnitude.

$$\mathrm{MSE}\left(x^{'}, x\right) = \frac{1}{d}\left\|x^{'} - x\right\|_2^2 = \frac{1}{d}\left\|\delta\right\|_2^2 \tag{5}$$

where d denotes the input dimension.

Zheng et al. innovatively proposed the metric of adversarial divergence [6]. They constructed a conditional adversarial distribution between the surrogate model and the target model, and calculated the distribution difference between the two to evaluate the transferability. The lower the adversarial divergence value is, the stronger the transferability is.

3. Classification of black-box adversarial attacks

This paper classifies black-box adversarial attacks based on the threat model of the attack. This classification consists of three categories: score-based attack, decision-based attack, and transfer-based attack.

DOI: 10.54254/2755-2721/2026.GU32310

3.1. Score-based attack

3.1.1. Core concepts and representative work

In score-based attacks, the target model returns continuous confidence scores or logits upon query [2]. Attackers iteratively adjust perturbations based on score changes from queries to decrease the score of the true class or increase the score of the target class, while satisfying the $_{p}$ norm constraint.

The early research on score-based attacks mainly relied on the zeroth-order optimization concept. Zeroth Order Optimization based Black-box Attacks (ZOO) approximates the gradient through finite differences or random estimation to guide the perturbation update, but it faces problems of low query efficiency and poor scalability to high-dimensional input [4]. Ilyas et al. proposed a method based on NES that replaces the element-wise difference with the expected gradient estimation in random directions, improving stability and efficiency, but still being constrained by the query budget [3]. Ilyas et al. regarded the attack as a Bandit optimization problem and used the time-dependent priors to adaptively update the search direction, significantly reducing the query complexity under a limited query setting [9]. Subsequently, a representative method emerged that does not explicitly construct the gradient but only uses score feedback for direction or random search [5, 8]. In recent years, most research has focused on the above paradigm, further optimizing in terms of query efficiency and perturbation stealthiness.

Unlike the common black-box attacks based on $_{2}$ or $_{\infty}$ constraints that involve dense perturbations, sparse black-box attacks are based on $_{0}$ constraints and only allow the modification of a few key pixels, which is closer to real threat models such as physical attacks and local tampering. In recent years, in Score-based attacks, related work on such attacks has also been increasing [10-13].

3.1.2. A query-efficient score-based black-box sparse adversarial attack

A query-efficient score-based black-box sparse adversarial attack (BruSLeAttack) belongs to the gradient-free random search paradigm and is a sparse black-box attack method for $_{0}$ constraints [13]. This method mitigates the shortcomings of the versatile framework for query-efficient sparse black-box adversarial attacks (Sparse-RS), which has low random search efficiency and redundant search space [10]. Under the $_{0}$ constraint, Sparse-RS conducts random local searches on fixed cardinality pixel subsets [10], iteratively replacing a part of the current pixel subset with randomly sampled random perturbation pixel groups, and updates based on the scores.

BruSLeAttack fixes the perturbation color by introducing synthetic color images and only searches for pixel positions, thereby reducing the dimensionality of the search space [13]. On this basis, BruSLeAttack also adds a dual search guidance mechanism. BruSLeAttack models the selection probability of pixels during the search process using a Dirichlet–Categorical distribution, gradually updating the weights of each pixel based on historical query feedback, and retaining high-impact pixels and replacing low-impact pixels during the search to achieve adaptive search with memory. BruSLeAttack integrates pixel difference maps, prioritizing the processing of pixels with large differences between the original image and the synthetic image, thereby accelerating the crossing of decision boundaries. In addition, BruSLeAttack further introduces an adaptive scheduling mechanism, dynamically adjusting the pixel replacement ratio. In the early iterations, it replaces more pixels for exploration, and in the later stages, it makes more use of the advantageous pixels.

DOI: 10.54254/2755-2721/2026.GU32310

3.1.3. Top-K Geometric Score-based Black-box Attack

Top-K Geometric Score-based Black-box Attack ($GSBAK^K$) is one of the earlier systematic studies on score-based black-box attacks in the top-K scenario, applicable to both untargeted and targeted top-K attack scenarios [7]. Most of the existing score-based black-box attack methods focus on the single highest confidence category and construct top-1 adversarial samples. The existing top-K attack methods are mostly white-box attack methods [14], while top-K attacks in the black-box domain are relatively limited.

$GSBAK^K$ introduced the geometric optimization idea of decision-based attack into the score-based framework and designed a two-level gradient estimation for the top-K scenario. In the top-K scenario, it is difficult for the attacker to find the initial boundary point that satisfies the top-K objective constraint. Therefore, $GSBAK^K$ designed a non-adversarial zone gradient estimation, starting from the source samples and iteratively approximating the decision boundary in a directed manner to find high-quality initial boundary points that satisfy the top-K target requirements. The gradient estimation on the decision boundary is susceptible to abnormal queries, and in the top-K scenario, the attacker needs to consider the joint influence of multiple labels. To solve this problem, $GSBAK^K$ designed decision boundary gradient estimation, filtering abnormal queries, and designed a dual-weight gradient formula for the top-K scenario, taking into account the number of target classes affected by noise and the total change in confidence of these target classes, to improve the accuracy of the gradient direction.

After identifying the initial boundary point and the gradient direction on the boundary, $GSBAK^K$ draws on the geometric concept of the decision-based attack, Curvature-Aware Geometric Black-Box Attack (CGBA) [15]. It iteratively searches for the next boundary point along a semicircular trajectory on the decision boundary, and gradually reduces the perturbation by leveraging the geometric features of the decision boundary to achieve a high success rate with small perturbations.

3.2. Decision-based attack

3.2.1. Core concepts and representative work

In decision-based attacks, the attacker can only obtain the final prediction label of the model. The attacker continuously queries and searches and approximates near the decision boundary based on the predicted labels output by the model, thereby constructing adversarial samples.

Boundary Attack laid the basic framework for decision-based attacks, which starts with a feasible adversarial sample as the initial point, locates the decision boundary and performs geometric approximation, gradually reducing the perturbation [16]. Subsequently, other techniques like HopSkipJumpAttack [17] and Query-Efficient Hard-Label Adversarial Attack (Sign-OPT) [18] started implementing the ideas of direction estimation and zeroth-order optimization. This shifted the boundary search toward a more targeted continuous optimization process that achieved much better query efficiency. Query-Efficient Boundary-Based Blackbox Attack (QEBA) [19], A Geometric Framework of Black-Box Adversarial Attack (GeoDA) [20], and A Ray Searching Method of Hard-Label Adversarial Attack (RayS) [21] have provided further improvements in the efficiency of decision boundary approximation from perspectives such as dimensionality reduction, geometric modeling, and structured search. In recent years, methods such as Fast Surrogate-Free Black-Box Attack (SurFree) [22], Triangle Attack [23], and CGBA have enhanced search efficiency and convergence stability by improving direction generation strategies, introducing subspace

DOI: 10.54254/2755-2721/2026.GU32310

constraints, designing geometric rules, or utilizing curvature-aware updates to improve the classical boundary search model.

3.2.2. Decision-Based Query Efficient Adversarial Attack via Adaptive Boundary Learning

The Decision-Based Query Efficient Adversarial Attack via Adaptive Boundary Learning (DEAL) is a decision-based black-box attack method based on adaptive boundary learning [24]. The core idea is to replace a large number of boundary exploration queries with local boundary predictions by constructing a boundary learner that can gradually fit the decision boundaries of the target model. This reduces the direct queries to the target model. Unlike traditional boundary approximation methods that rely on random exploration or direction estimation, this method introduces a binary classifier specifically for predicting decision boundaries, transferring some boundary judgment tasks to the local model for completion, thereby reducing the overhead of invalid queries.

DEAL is divided into two main stages: offline initialization and online optimization [24]. In the offline stage, an auxiliary dataset is constructed and combined with a meta-learning mechanism to pre-train the boundary learner, enabling it to quickly adapt to the decision boundaries of different models. In the online stage, the boundary learner is warmed up and fine-tuned based on a small number of real queries, and candidate adversarial samples are generated under the guidance of the local model. Verification and correction are only performed through necessary queries. By alternately executing the process of generating adversarial samples and updating the boundaries, the algorithm gradually narrows the gap between the learner and the target model's decision boundaries, thereby significantly reducing the number of queries while maintaining a high attack success rate.

3.2.3. Boosting the Ray Search Procedure of Hard-label Attacks with Transfer-based Priors

The Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors [25] is a recent enhancement of the ray-search methods OPT [26] and Sign-OPT [18] in decision-based black-box attacks. These methods rely solely on top-1 label feedback and formulate the attack as a direction-space optimization problem: given a direction, the shortest distance along that direction to reach the decision boundary is estimated through ray search and binary queries, and the direction is iteratively updated to minimize the perturbation. However, the evaluation of boundary distance requires a large number of binary queries, and the zeroth-order direction estimation has high noise, high query cost, and low estimation accuracy.

Therefore, the paper proposes to incorporate transfer priors into the ray search for hard-label attacks [25]. It utilizes a surrogate model to generate transferable prior directions, and together with random vectors, constructs a subspace. Then, the approximation of the true gradient is projected onto this subspace to achieve efficient estimation, balancing accuracy and query cost.

3.3. Transfer-based attack

3.3.1. Core concepts and representative work

Transfer-based black-box attacks achieve the attack by generating adversarial samples on a local surrogate model and relying on the transferability of these adversarial samples across different models. Such attack methods typically do not require querying the target model and are the most basic form of black-box attacks in scenarios with limited information.

The development of transfer-based attacks began with the systematic discovery of cross-model transferability of adversarial samples [27,28]. Subsequently, Papernot et al. proposed a training

DOI: 10.54254/2755-2721/2026.GU32310

framework for surrogate models [29], and Liu et al. further demonstrated that integrating substitute models could significantly increase the success rate of targeted transfer, making the ensemble-based transfer generation strategy a common design approach in subsequent work [30].

3.3.2. Frequency-Aware Contrastive Learning for Transferable Adversarial Attacks

Frequency-Aware Contrastive Learning for Transferable Adversarial Attacks (FACL-Attack) is one of the representative generative methods in recent transfer-based black-box attacks [31], specifically designed for strict black-box scenarios where both the target domain and target architecture are unknown. The work points out that the lack of transferability is related to the overfitting of the generator to the source domain and the surrogate model, and reconstructs the training mechanism from the frequency domain to enhance the cross-domain and cross-model generalization ability.

Its method introduces two types of modules during the training stage. One of them is Frequency-Aware Domain Randomization (FADR), which involves randomizing perturbations on components in the spectral domain, such as low and high frequencies, that are more likely to vary across domains, to simulate domain shifts and force the generator to learn more robust perturbation generation rules. The second is Frequency-Enhanced Contrastive Learning (FACL), which imposes contrastive constraints on intermediate frequency features based on frequency domain decomposition, making clean samples and adversarial samples more distinguishable in the intermediate frequency representation, thereby concentrating attacks on semantic representations that are more consistent across models. The experimental results show that this method exhibits stronger transfer attack performance under various cross-domain and cross-architecture settings.

3.3.3. Diffusion Models for Imperceptible and Transferable Adversarial Attacks

Diffusion Models for Imperceptible and Transferable Adversarial Attack (DiffAttack) is an unrestricted adversarial attack method based on diffusion models [32]. In response to the core pain points of traditional L_P norm pixel attacks, which are easily detectable by the human eye and have no L_P constraints, resulting in poor transferability for black-box models, this method relies on the generation prior and semantic alignment mechanism of the diffusion model to construct perturbations in its latent space, achieving high imperceptibility of adversarial samples and cross-model transferability.

This method is based on the Stable Diffusion model and uses Denoising Diffusion Implicit Models (DDIM) for inversion to map clean images to the latent space of the diffusion model [31]. It no longer directly applies L_P -limited perturbations in the pixel space or modifies the text embeddings. Instead, it optimizes on the latent space variables. This method designs a triple-weighted loss function for attack, transfer, and structure. The attack loss achieves a basic attack on the surrogate model through cross-entropy. The transfer loss minimizes the variance of the cross-attention map between the text and the image, dispersing the attention of the diffusion model on the target area and using it as an implicit recognition surrogate to enhance cross-model transferability. The structure loss retains the core semantics of the image based on self-attention consistency constraints, and at the same time, it achieves a balance between the editability in the latent space and the retention of the original content in the final image by controlling the inversion strength.

This framework demonstrates greater imperceptibility and transferability across various model architectures, datasets, and defense settings. However, it requires inversion and iterative optimization, and has relatively higher computational costs.

DOI: 10.54254/2755-2721/2026.GU32310

4. Comparative analysis and discussion

Both score-based and decision-based attack approaches belong to query-based attack methods, but the former relies on continuous scores while the latter utilizes only discrete labels. Transfer-based methods rely weakly or not at all on target model queries, instead focusing more on surrogate models and transferability. Different paradigms do not represent a simple hierarchy of strength; their applicability is jointly determined by the feedback mechanisms and interaction constraints of the target model.

Continuous scoring feedback provides more refined improvement signals for score-based attacks, facilitating more targeted searches by the attacker and reducing the number of ineffective attempts. Decision-based attacks lack continuous signal feedback and often require more queries during the boundary positioning and perturbation compression stages. Transfer-based attacks almost do not consume query budgets in scenarios where the target is non-interactable or has strong rate limiting, but their cost is transferred to the computational and engineering costs of local training, data acquisition, or generative optimization, and they are highly sensitive to transferability. BlackboxBench systematically compared a large number of query-based and transfer-based methods under unified code and evaluation standards [6], pointing out the overall progress trend of increasing attack success rates and decreasing query requirements in recent years, and also revealing the differentiation in query efficiency and stability among different technical routes.

Query-based attacks can obtain feedback from the target model through repeated queries, explicitly impose perturbation constraints during the iterative process, and adjust the search strategy according to different attack targets, thereby achieving more controllable construction of adversarial samples. However, this type of method is often directly constrained by the query budget. Transfer-based attacks, on the other hand, operate in the absence of feedback from the target model and utilize a set of perturbation patterns learned from a surrogate model to improve the transfer success rate. Their advantage lies in almost not consuming the query budget of the target model and having weaker online interaction traces. However, their effect highly depends on the consistency of feature representation and decision boundaries between the surrogate model and the target model, so their performance fluctuates significantly in cross-architecture, cross-domain, or cross-task settings. Currently, the boundaries of these three paradigms are being blurred. The geometric optimization idea of decision boundaries is introduced into score-based attacks [7], transfer priors assist query search [25], and a hybrid strategy that integrates local priors and a small amount of verification under a strict black-box setting [24] is increasingly common.

5. Conclusion

This paper systematically reviews the core paradigms and evaluation frameworks of black-box adversarial attacks. Starting from the threat model, the paper comprehensively summarizes the representative works and technological development trends of three types of methods: score-based, decision-based, and transfer-based, and introduces the important recent works. The paper also discusses the trend of extending the attack target from top-1 to top-K and the evaluation system centered on ASR, perturbation norms, MSE, average and median query numbers, and adversarial divergence.

According to the current benchmark performance, one can notice that over the past few years, the black-box attacks tended to increase their success rates and reduce their query costs. Nevertheless, the various paradigms have very marked variations in stability and the cost structure. Query-based attacks are explicitly limited by query budgets, whereas transfer-based attacks transfer the cost to

DOI: 10.54254/2755-2721/2026.GU32310

local training, generative optimization, and data acquisition and are extremely sensitive to transferability. Over time, over the past years, a convergent nature has been observed between the three paradigms due to the blurring of traditional technical boundaries. Others are experimenting with adding transfer priors to query search, or using small numbers of verification queries together with local priors to jointly optimize in very restrictive black-box environments, to seek more moderate tradeoffs between success rate and query cost. The present paper offers a brief reference point of the selection of attack techniques, integrated assessment, and design of the defense in the conditions of various interaction constraints.

References

[1] RCarlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., … Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv: 1902.06705.
[2] Vassilev, A., Oprea, A., Fordyce, A., & Andersen, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations.
[3] Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. arXiv: 1804.08598.
[4] Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017, November). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 15-26).
[5] Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square Attack: A query-efficient black-box adversarial attack via random search. In ECCV 2020. arXiv: 1912.00049.
[6] Zheng, M., Yan, X., Zhu, Z., Chen, H., & Wu, B. (2025). Blackboxbench: A comprehensive benchmark of black-box adversarial attacks. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[7] Reza, M. F., Jin, R., Wu, T., & Dai, H. (2025). GSBAK: top-K GEOMETRIC SCORE-BASED BLACKBOX ATTACK.
[8] Guo, C., Gardner, J., You, Y., Wilson, A. G., & Weinberger, K. (2019, May). Simple black-box adversarial attacks. In International conference on machine learning (pp. 2484-2493). PMLR.
[9] Ilyas, A., Engstrom, L., & Madry, A. (2018). Prior convictions: Black-box adversarial attacks with bandits and priors. arXiv preprint arXiv: 1807.07978.
[10] Croce, F., Andriushchenko, M., Singh, N. D., Flammarion, N., & Hein, M. (2022, June). Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 6, pp. 6437-6445).
[11] Williams, P. N., & Li, K. (2023). Black-box sparse adversarial attack via multi-objective optimisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12291-12301).
[12] Chen, J., Chen, T., Xu, X., Zhang, J., Yang, Y., & Shen, H. T. (2023). Coreset learning-based sparse black-box adversarial attack for video recognition. IEEE Transactions on Information Forensics and Security, 19, 1547-1560.
[13] Vo, V. Q., Abbasnejad, E., & Ranasinghe, D. C. (2024). Brusleattack: A query-efficient score-based black-box sparse adversarial attack. arXiv preprint arXiv: 2404.05311.
[14] Hu, S., Ke, L., Wang, X., & Lyu, S. (2021). Tkml-ap: Adversarial attacks to top-k multi-label learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7649-7657).
[15] Reza, M. F., Rahmati, A., Wu, T., & Dai, H. (2023). CGBA: Curvature-aware geometric black-box attack. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 124-133).
[16] Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv: 1712.04248.
[17] Chen, J., Jordan, M. I., & Wainwright, M. J. (2020, May). Hopskipjumpattack: A query-efficient decision-based attack. In 2020 ieee symposium on security and privacy (sp) (pp. 1277-1294). IEEE.
[18] Cheng, M., Singh, S., Chen, P., Chen, P. Y., Liu, S., & Hsieh, C. J. (2019). Sign-opt: A query-efficient hard-label adversarial attack. arXiv preprint arXiv: 1909.10773.
[19] Li, H., Xu, X., Zhang, X., Yang, S., & Li, B. (2020). Qeba: Query-efficient boundary-based blackbox attack. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1221-1230).
[20] Rahmati, A., Moosavi-Dezfooli, S. M., Frossard, P., & Dai, H. (2020). Geoda: a geometric framework for black-box adversarial attacks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8446-8455).

DOI: 10.54254/2755-2721/2026.GU32310

[21] Chen, J., & Gu, Q. (2020, August). Rays: A ray searching method for hard-label adversarial attack. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1739-1747).
[22] Maho, T., Furon, T., & Le Merrer, E. (2021). Surfree: a fast surrogate-free black-box attack. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10430-10439).
[23] Wang, X., Zhang, Z., Tong, K., Gong, D., He, K., Li, Z., & Liu, W. (2022, October). Triangle attack: A query-efficient decision-based adversarial attack. In European conference on computer vision (pp. 156-174). Cham: Springer Nature Switzerland.
[24] Shen, M., Li, C., Yu, H., Li, Q., Zhu, L., & Xu, K. (2023). Decision-based query efficient adversarial attack via adaptive boundary learning. IEEE Transactions on Dependable and Secure Computing, 21(4), 1740-1753.
[25] Ma, C., Xu, X., Cheng, S., & Xuan, Q. (2025). Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors. arXiv preprint arXiv: 2507.17577.
[26] Cheng, M., Le, T., Chen, P. Y., Yi, J., Zhang, H., & Hsieh, C. J. (2018). Query-efficient hard-label black-box attack: An optimization-based approach. arXiv preprint arXiv: 1807.04457.
[27] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv: 1312.6199.
[28] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv: 1412.6572.
[29] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia conference on computer and communications security (pp. 506-519).
[30] Liu, Y., Chen, X., Liu, C., & Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv: 1611.02770.
[31] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv: 2010.02502.
[32] Chen, J., Chen, H., Chen, K., Zhang, Y., Zou, Z., & Shi, Z. (2024). Diffusion models for imperceptible and transferable adversarial attack. IEEE Transactions on Pattern Analysis and Machine Intelligence.