Improving the Convergence Rate of Ray Search Optimization for Query-Efficient Hard-Label Attacks

Xinjie Xu[1], **Shuyu Cheng**[2], **Dongwei Xu**[1], **Qi Xuan**[1,3], **Chen Ma**[1,3*]

[1]Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
[2]JQ Investments, Shanghai 200122, China
[3]Binjiang Institute of Artificial Intelligence, ZJUT, Hangzhou 310056, China
xxj1018@foxmail.com, csy530216@126.com, {dongweixu, xuanqi, machen}@zjut.edu.cn

Abstract

In hard-label black-box adversarial attacks, where only the top-1 predicted label is accessible, the prohibitive query complexity poses a major obstacle to practical deployment. In this paper, we focus on optimizing a representative class of attacks that search for the optimal ray direction yielding the minimum ℓ_2-norm perturbation required to move a benign image into the adversarial region. Inspired by Nesterov's Accelerated Gradient (NAG), we propose a momentum-based algorithm, ARS-OPT, which proactively estimates the gradient with respect to a future ray direction inferred from accumulated momentum. We provide a theoretical analysis of its convergence behavior, showing that ARS-OPT enables more accurate directional updates and achieves faster, more stable optimization. To further accelerate convergence, we incorporate surrogate-model priors into ARS-OPT's gradient estimation, resulting in PARS-OPT with enhanced performance. The superiority of our approach is supported by theoretical guarantees under standard assumptions. Extensive experiments on ImageNet and CIFAR-10 demonstrate that our method surpasses 13 state-of-the-art approaches in query efficiency.

Code — https://github.com/machanic/hard_label_attacks
Extended version — https://arxiv.org/abs/2512.21241

1 Introduction

We focus on hard-label adversarial attacks. Considered among the most practical and challenging black-box attacks, hard-label attacks operate under strict information constraints. While white-box attacks (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018) leverage model parameters and gradients, and score-based attacks (Ma, Chen, and Yong 2021) exploit confidence scores, hard-label attacks rely solely on top-1 predicted labels. This makes the efficient generation of adversarial examples substantially more difficult while enhancing their practical applicability.

Why study query-based black-box adversarial attacks under the hard-label setting? Real-world machine-learning services such as cloud vision APIs and biometric recognizers often reveal nothing more than the final predicted decision (i.e., the top-1 label) to external users. With gradients and confidence scores stripped away, an attacker is forced to treat the model as a hard-label black box to probe its decision boundary. This stringent setting accurately reflects the limited feedback of deployed services and raises three key challenges. (1) *Minimal feedback:* Each query yields only a hard-label response, demanding efficient exploration strategies. (2) *Practical relevance:* It closely mirrors restricted commercial platforms where probability scores and internal details are deliberately hidden. (3) *Security-critical:* Hard-label attacks reveal vulnerabilities in "security-through-obscurity" systems and underscore the urgent need for defenses against adversaries with minimal information. Consequently, designing query-efficient attacks based solely on hard-label feedback is essential for vulnerability assessment and robust defenses.

Why are hard-label attacks challenging? Because a model's predicted label typically changes only when an input moves across or near its decision boundary, hard-label attacks must restrict their search to this narrow region, making the optimization especially challenging. Early hard-label attacks like Boundary Attack (BA) (Brendel, Rauber, and Bethge 2018) and Biased BA (Brunner et al. 2019) initialize from a sample already in the adversarial region and progressively reduce the perturbation by stepping toward the original image while exploring directions on the decision boundary via randomly sampled spherical vectors. However, these approaches remain highly inefficient in terms of query cost: they rely almost entirely on random sampling and neglect valuable information from past queries, which impedes effective perturbation reduction. To address this challenge, recent studies have adopted zeroth-order (ZO) optimization techniques, which leverage boundary information more effectively to identify adversarial examples. Existing ZO-based attacks—such as HopSkipJumpAttack (HSJA) (Chen, Jordan, and Wainwright 2020), OPT (Cheng et al. 2019), Sign-OPT (Cheng et al. 2020), and Prior-OPT (Ma et al. 2025)—primarily focus on improving gradient estimation through finite differences. However, their optimization strategies rely on vanilla gradient descent, overlooking well-established acceleration methods such as momentum and Nesterov's accelerated gradient, which can enhance convergence rates even when the gradient estimation quality remains unchanged. To address these limitations, we propose ARS-OPT, a novel ZO optimization algorithm incorporating accelerated random search (ARS) (Nesterov and Spokoiny 2017). Our theoretical

*Corresponding author.

analysis demonstrates that the update of ARS-OPT can be interpreted as implicitly incorporating second-order curvature information without explicit Hessian estimation, and establishes a bound on the expected gap between the objective value at iteration T and the optimum value. Building on this, we introduce PARS-OPT, which integrates transfer-based priors to improve gradient estimation. PARS-OPT further extends to combine priors from multiple surrogate models, delivering additional gains in attack performance. Extensive experiments on ImageNet, CIFAR-10, and a CLIP-based model demonstrate that our framework, consisting of ARS-OPT and its prior-enhanced variant PARS-OPT, outperforms 13 state-of-the-art methods with superior query efficiency.

Our main contributions are summarized as follows.

- **Novelty in hard-label attacks.** We present ARS-OPT, a novel hard-label attack that accelerates convergence by estimating gradients along an interpolated "lookahead" direction, combining the search trajectory with accumulated momentum. We further introduce PARS-OPT, which integrates transfer-based priors from surrogate models to improve gradient estimation and enhance attack efficiency.
- **Novelty in theoretical analysis.** We establish an $\mathcal{O}(1/T^2)$ convergence rate under standard assumptions, supported by the construction of an unbiased estimator of the true gradient that is essential for ensuring this rate. The theoretical analysis provides a principled explanation for the acceleration behavior of our approach and clarifies its underlying optimization dynamics.
- **SOTA performance.** Experimental results show our approach outperforms 13 state-of-the-art attacks on ImageNet and CIFAR-10 across classifiers, including CLIP.

2 Related Work

Hard-label attacks, also known as decision-based black-box attacks, are among the most challenging adversarial scenarios. They rely solely on the target model's top-1 predicted label without access to internal structure or confidence scores, and craft perturbations by querying and exploiting information near the decision boundary. Boundary Attack (BA) (Brendel, Rauber, and Bethge 2018) was one of the earliest methods, performing random walks on the boundary to minimize perturbations, but suffers from low query efficiency. Biased BA (BBA) (Brunner et al. 2019) improves BA via three biases: (1) low-frequency Perlin noise, (2) regional masking, and (3) surrogate-model gradients. The Evolutionary Attack (abbreviated as Evolutionary) (Dong et al. 2019) adopts random sampling with adaptive covariance, while AHA (Li et al. 2021) exploits historical queries to guide the search. HopSkipJumpAttack (HSJA) (Chen, Jordan, and Wainwright 2020) refines adversarial examples via (1) gradient approximation at the boundary and (2) binary search projection onto the boundary toward the benign image. SQBA (Park, Miller, and McLaughlin 2024) combines surrogate-model gradients with HSJA's gradient estimation to improve query efficiency. QEBA (Li et al. 2020) lowers HSJA's query cost using subspaces derived from spatial transformations, low-frequency components, and intrinsic features. GeoDA (Rahmati et al. 2020) leverages the boundary's low curvature via local linearization to estimate gradients and reduce queries. Triangle Attack (Wang et al. 2022) applies the law of sines in a low-frequency subspace, removing boundary projections and gradient estimation. Tangent Attack (TA) (Ma et al. 2021) locates an optimal tangent point to minimize perturbations, while SurFree (Maho, Furon, and Le Merrer 2021) uses geometry-driven directional trials without gradient estimation. CGBA and its variant CGBA-H (Reza et al. 2023) search along a semicircular path on a restricted 2D plane to find boundary points. Another direction formulates hard-label attacks as continuous optimization problems. OPT (Cheng et al. 2019) employs zeroth-order (ZO) optimization based on random-direction finite differences. Sign-OPT (Cheng et al. 2020) reduces queries by using directional derivative signs but sacrifices gradient precision. Prior-OPT (Ma et al. 2025) integrates transfer-based priors into the ray-search optimization, while RayS (Chen and Gu 2020) removes gradient estimation entirely by using hierarchical search, but is limited to untargeted ℓ_∞-norm attacks. QE-DBA (Zhang, Ahmed, and Yu 2024) applies Bayesian optimization to explore the perturbation space, effectively addressing hard-label ZO optimization problems. However, existing methods overlook established acceleration strategies—such as momentum and Nesterov's accelerated gradient—that can greatly improve convergence rates without requiring better gradient estimates. In this work, we address this gap by integrating acceleration techniques to enhance query efficiency. Moreover, our framework can further boost efficiency by incorporating transfer-based priors.

3 Problem Statement of Hard-Label Attacks

Given a classifier $\psi : \mathbb{R}^d \to \mathbb{R}^C$ designed for a C-class classification task, and a correctly classified input image $\mathbf{x} \in [0,1]^d$, where d is the dimension of the input image, the adversary seeks to generate an adversarial example $\mathbf{x}_{\text{adv}}$ by crafting a minimal perturbation such that the classifier's prediction for $\mathbf{x}_{\text{adv}}$ becomes incorrect. This adversarial objective can be formally expressed as:

$$\min_{\mathbf{x}_{\text{adv}}} \|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_p \quad \text{s.t.} \quad \Phi(\mathbf{x}_{\text{adv}}) = 1, \tag{1}$$

where $\|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_p$ is the p-norm distortion, and the constraint $\Phi(\mathbf{x}_{\text{adv}})$ is defined as an attack success indicator:

$$\Phi(\mathbf{x}_{\text{adv}}) \coloneqq \begin{cases} 1 & \text{if } \hat{y} = y_{\text{adv}} \text{ in a targeted attack,} \\ & \quad \text{or } \hat{y} \neq y \text{ in an untargeted attack,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\hat{y} = \arg\max_{i \in \{1,\dots,C\}} \psi(\mathbf{x}_{\text{adv}})_i$ denotes the top-1 predicted label by classifier ψ, y is the true label of $\mathbf{x}$, and y_{adv} is the target label in a targeted attack scenario.

Following the ray-search methods (Cheng et al. 2019, 2020; Ma et al. 2025), we reformulate the optimization problem in Eq. (1) as finding the optimal ray direction θ^* from $\mathbf{x}$ that yields the minimal distance $f(\theta)$ to the boundary of the adversarial region. This can be formulated as:

$$\begin{gathered} \min_{\theta \in \mathbb{R}^d \setminus \{\mathbf{0}\}} f(\theta), \\ \text{where} \quad f(\theta) \coloneqq \inf\Big\{\lambda > 0 : \Phi\big(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\big) = 1\Big\}. \end{gathered} \tag{3}$$

By convention, $f(\theta) = +\infty$ if the set is empty. Consequently, the resulting adversarial example is constructed as

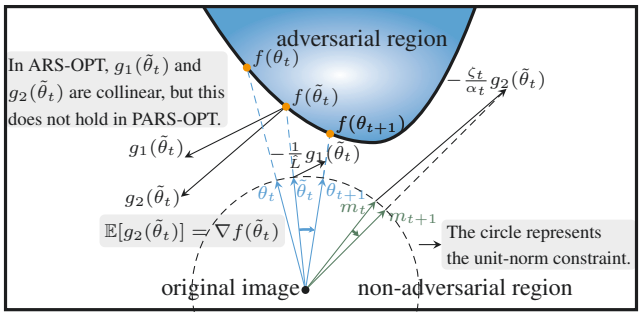


Figure 1: Illustration of a three-step update: first, compute the perturbation direction $\tilde{\theta}_t = (1-\alpha_t)\theta_t + \alpha_t m_t$; then estimate gradients at $\tilde{\theta}_t$ using a biased $g_1(\tilde{\theta}_t)$ and an unbiased $g_2(\tilde{\theta}_t)$; finally, update θ_{t+1} and m_{t+1} via a gradient descent step.

$\mathbf{x}^* = \mathbf{x} + f(\theta^*)\frac{\theta^*}{\|\theta^*\|}$, where θ^* is the optimal solution obtained from the minimization problem defined in Eq. (3).

4 The Proposed Approach

Previous works (Cheng et al. 2019, 2020; Ma et al. 2025) focus on efficient gradient estimation to optimize the direction θ, with step size typically determined by line search. However, they do not explore any optimization acceleration techniques beyond gradient estimation. Next, we present an overview of ARS-OPT and its prior-enhanced variant PARS-OPT, both equipped with theoretical convergence guarantees.

Conceptual Sketch and Overview

Nesterov and Spokoiny (2017) propose an Accelerated Random Search (ARS) method for ZO optimization, which rigorously establishes explicit non-asymptotic convergence rates under various convexity and smoothness assumptions by introducing an accelerated ZO framework. In *the score-based setting*, Cheng et al. (2021) extend ARS to score-based attacks and provide an analysis of the convergence rate. However, in *hard-label attacks*, obtaining function values requires extensive binary searches, significantly reducing the query efficiency of gradient estimation based on finite differences.

To address these limitations, we introduce ARS-OPT, a novel ZO optimization framework that can be seamlessly augmented with transfer-based priors to further boost query efficiency. The primary challenge is accelerating convergence in gradient descent when only poorly estimated gradients are available. At iteration t, we employ the following three-step update process for θ_t (Fig. 1):

1 Compute the perturbation direction $\tilde{\theta}_t \leftarrow (1-\alpha_t)\theta_t + \alpha_t m_t$, where m_0 is initialized to θ_0.
2 At $\tilde{\theta}_t$, we use multiple queries to estimate gradients $g_1(\tilde{\theta}_t)$ (biased estimator, e.g., Sign-OPT or Prior-OPT method) and $g_2(\tilde{\theta}_t)$ (unbiased estimator of $\nabla f(\tilde{\theta}_t)$).
3 Update both parameters by gradient descent: $\theta_{t+1} \leftarrow \tilde{\theta}_t - \frac{1}{\hat{L}} g_1(\tilde{\theta}_t)$, $m_{t+1} \leftarrow m_t - \frac{\zeta_t}{\alpha_t} g_2(\tilde{\theta}_t)$.

Inspired by Nesterov's accelerated gradient method, our approach dynamically tracks two sequences, i.e., the direction θ_t and the momentum vector m_t, and then computes a lookahead vector $\tilde{\theta}_t$ by linearly interpolating between θ_t and m_t, controlled by an interpolation coefficient α_t. At $\tilde{\theta}_t$, we estimate two gradients, $g_1(\tilde{\theta}_t)$ and $g_2(\tilde{\theta}_t)$, to compute the updates of θ_{t+1} and m_{t+1}, respectively. Although we adopt the same estimation procedure for $g_1(\tilde{\theta}_t)$ as in Prior-OPT, our algorithm converges substantially faster, as demonstrated by our experiments. The convergence guarantee of our approach relies on two technical assumptions: (1) $g_2(\tilde{\theta}_t)$ serves as an unbiased estimator of $\nabla f(\tilde{\theta}_t)$, and (2) $\zeta_t \leq \mathbb{E}_t\left[(\nabla f(\tilde{\theta}_t)^\top \mathbf{v}_t)^2\right] / \left(\hat{L} \cdot \mathbb{E}_t\left[\|g_2(\tilde{\theta}_t)\|^2\right]\right)$, with the full derivation given in the Appendix. We also note that our framework can incorporate various gradient estimation techniques, such as prior-guided estimation, to further improve performance. Our approach can be intuitively understood through the analogy of a walker descending a valley: rather than relying solely on the current slope, the walker looks ahead to anticipate the upcoming terrain and adjust the direction of motion accordingly, thereby achieving smoother and faster progress toward the minimum.

ARS-OPT

Our framework, spanning from Step 1 to Step 3, is compatible with various gradient estimation techniques, enabling flexible algorithmic implementations. In this section, we provide a detailed introduction to the fundamental algorithm, ARS-OPT. In Step 1, unlike standard gradient descent, the gradient is not computed at the current direction θ_t. Instead, the algorithm predicts a candidate ray direction $\tilde{\theta}_t$ by interpolating between the momentum vector m_t and the current direction θ_t. The sequences of θ_t and m_t are referred to as *the main sequence* and *the auxiliary sequence*, respectively. $\tilde{\theta}_t$ is referred to as the *lookahead position* of θ_t, and is computed via interpolation: $\tilde{\theta}_t \leftarrow (1-\alpha_t)\theta_t + \alpha_t m_t$, where $\alpha_t \in [0,1]$ is the interpolation coefficient. The value of α_t is defined as the positive root of the equation $\alpha_t^2 = \zeta_t \gamma_t (1-\alpha_t)$, where γ_t is a scalar determined in Algorithm 1, and $\zeta_t = \left(\frac{2(q-1)+\pi}{d\pi}\right) / \left(\hat{L}\left(\frac{d\pi}{2(q-1)+\pi}\right)\right)$. This expression is derived from the convergence analysis of ARS-OPT. This choice of α_t is critical to establishing the algorithm's theoretical convergence guarantees. For detailed derivations, we refer readers to Appendix A. To maintain two sequences—the optimization variable θ_t and the auxiliary variable m_t (which accumulates historical momentum to capture global optimization trends)—we employ two gradient estimates, $g_1(\tilde{\theta}_t)$ and $g_2(\tilde{\theta}_t)$, to update θ_t and m_t, respectively:

$$g_1(\tilde{\theta}_t) \coloneqq \nabla f(\tilde{\theta}_t)^\top \mathbf{v}_t \cdot \mathbf{v}_t \approx \frac{f(\tilde{\theta}_t + \epsilon \mathbf{v}_t) - f(\tilde{\theta}_t)}{\epsilon} \cdot \mathbf{v}_t, \quad (4)$$

$$g_2(\tilde{\theta}_t) \coloneqq \frac{d}{\frac{2}{\pi}(q-1)+1} \nabla f(\tilde{\theta}_t)^\top \mathbf{v}_t \cdot \mathbf{v}_t \quad (5)$$

$$\approx \frac{d\left(f(\tilde{\theta}_t + \epsilon \mathbf{v}_t) - f(\tilde{\theta}_t)\right)}{\frac{2\epsilon}{\pi}(q-1)+\epsilon} \cdot \mathbf{v}_t, \quad (6)$$

where d is the dimension of the input image, q is the number of vectors in gradient estimation, and $\mathbf{v}_t$ is the sign-based gradient estimate (Cheng et al. 2020) as $\mathbf{v}_t \coloneqq \frac{1}{\sqrt{q}} \sum_{i=1}^{q} \text{sign}(f(\tilde{\theta}_t + \epsilon \mathbf{u}_i) - f(\tilde{\theta}_t))\mathbf{u}_i$, which calculates the

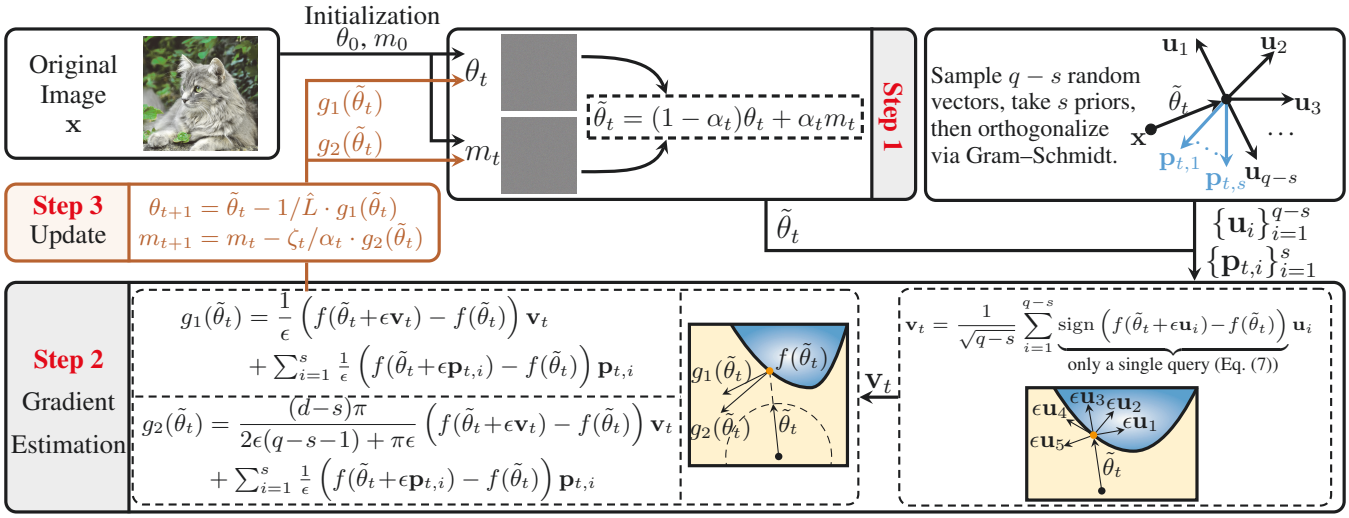


Figure 2: Illustration of one iteration in PARS-OPT. We first form a lookahead point $\tilde{\theta}_t$ by linearly interpolating between the current direction θ_t and the momentum term m_t (with $m_0 = \theta_0$). Next, we estimate $\mathbf{v}_t$ via a sign-based procedure over a set of randomly sampled orthonormal basis vectors. Finally, we use $\mathbf{v}_t$ to compute the biased gradient estimate $g_1(\tilde{\theta}_t)$ and the unbiased estimate $g_2(\tilde{\theta}_t)$, which are then used to update θ_t and m_t, yielding θ_{t+1} and m_{t+1} for the next iteration.

sign of the directional derivative with a single query:

$$\text{sign}(f(\theta+\epsilon\mathbf{u}) - f(\theta)) = \begin{cases} +1, & \Phi\left(\mathbf{x} + f(\theta)\frac{\theta+\epsilon\mathbf{u}}{\|\theta+\epsilon\mathbf{u}\|}\right) \neq 1, \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

Eq. (4) can be regarded as the projection of the true gradient onto $\mathbf{v}_t$. Eq. (5) is an unbiased estimator of $\nabla f(\tilde{\theta}_t)$, derived from Theorem 4.1[1].

Theorem 4.1. *Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q\}$ be an orthonormal set obtained by orthogonalizing q vectors independently and uniformly sampled from the unit sphere in $\mathbb{R}^d$. Suppose $\mathbf{g}$ is a fixed vector in $\mathbb{R}^d$ (for example, it is the true gradient to be estimated). Let $\mathbf{v} := \sum_{i=1}^{q}\text{sign}(\mathbf{g}^\top\mathbf{u}_i)\mathbf{u}_i$, and $\hat{\mathbf{g}} := \mathbf{g}^\top\overline{\mathbf{v}}\cdot\overline{\mathbf{v}}$. Then we have*

$$\mathbb{E}[\hat{\mathbf{g}}] = \mathbb{E}[(\overline{\mathbf{g}}^\top\overline{\mathbf{v}})^2]\cdot\mathbf{g}. \tag{8}$$

The proof of Theorem 4.1 is shown in Appendix A. In Eq. (8), $\hat{\mathbf{g}}$ is equal to $g_1(\tilde{\theta}_t)$, and $\mathbb{E}[(\overline{\mathbf{g}}^\top\overline{\mathbf{v}})^2] = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right)$ based on Lemma A.5 (see Appendix A). Thus we have $\mathbb{E}[g_1(\tilde{\theta}_t)] = \frac{1}{d}\left(\frac{2}{\pi}(q-1)+1\right)\cdot\mathbf{g}$. Consequently, the true gradient can be recovered as $\mathbf{g} = \frac{d}{\frac{2}{\pi}(q-1)+1}\mathbb{E}[g_1(\tilde{\theta}_t)] = \frac{d}{\frac{2}{\pi}(q-1)+1}\mathbb{E}[\nabla f(\tilde{\theta}_t)^\top\mathbf{v}_t\cdot\mathbf{v}_t]$, which shows that $g_2(\tilde{\theta}_t)$ is an unbiased estimator of $\nabla f(\tilde{\theta}_t)$.

PARS-OPT

ARS-OPT relies exclusively on random orthonormal vectors to estimate the gradient, which leads to inaccurate gradient approximation and poor query efficiency. To further enhance the efficiency of the algorithm, we propose a variant algorithm named Prior-guided ARS-OPT (PARS-OPT) within our framework. An ideal prior would be the gradient of $\hat{f}(\theta)$ derived from a surrogate model. However, since $\hat{f}(\theta)$ is non-differentiable due to the binary search process, this gradient cannot be directly computed. To overcome this challenge, we employ a differentiable surrogate function $h(\theta,\lambda)$ in Eq. (9), following Ma et al. (2025), which ensures the gradient relationship: $\nabla\hat{f}(\theta_0) = c\cdot\nabla_\theta h(\theta_0,\lambda_0)$ for any non-zero vector $\theta_0 \in \mathbb{R}^d$ with $\hat{f}(\theta_0) < +\infty$. Here, $\hat{f}(\cdot)$ is defined on the surrogate model $\hat{\psi}$, $\lambda_0 = \hat{f}(\theta_0)$ is treated as a constant scalar during differentiation, and c is a non-zero constant.

$$h(\theta,\lambda) := \begin{cases} \hat{\psi}_y - \max_{j\neq y}\hat{\psi}_j, & \text{if untargeted attack,} \\ \max_{j\neq y_{\text{adv}}}\hat{\psi}_j - \hat{\psi}_{y_{\text{adv}}}, & \text{if targeted attack,} \end{cases} \tag{9}$$

where $\hat{\psi}_i := \hat{\psi}\left(\mathbf{x}+\lambda\cdot\frac{\theta}{\|\theta\|}\right)_i$ is an abbreviation for the i-th element of the output of the surrogate model $\hat{\psi}$, and $\mathbf{x}$ is the original image. Given s non-zero vectors $\mathbf{k}_{t,1},\ldots,\mathbf{k}_{t,s}$ computed as $\nabla_\theta h(\theta_0,\lambda_0)$ from s surrogate models and $q-s$ randomly sampled vectors $\mathbf{r}_1,\ldots,\mathbf{r}_{q-s}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, we apply Gram–Schmidt orthogonalization to these q vectors to obtain an orthonormal set $\mathbf{p}_{t,1},\ldots,\mathbf{p}_{t,s},\mathbf{u}_1,\ldots,\mathbf{u}_{q-s}$, which are used by the gradient estimation formulas:

$$g_1(\tilde{\theta}_t) = \nabla f(\tilde{\theta}_t)^\top\mathbf{v}_t\cdot\mathbf{v}_t + \sum_{i=1}^{s}\nabla f(\tilde{\theta}_t)^\top\mathbf{p}_{t,i}\cdot\mathbf{p}_{t,i} \tag{10}$$

$$\approx \frac{f(\tilde{\theta}_t+\epsilon\mathbf{v}_t)-f(\tilde{\theta}_t)}{\epsilon}\mathbf{v}_t + \sum_{i=1}^{s}\frac{f(\tilde{\theta}_t+\epsilon\mathbf{p}_{t,i})-f(\tilde{\theta}_t)}{\epsilon}\mathbf{p}_{t,i}. \tag{11}$$

$$g_2(\tilde{\theta}_t) = \frac{d-s}{\frac{2}{\pi}(q-s-1)+1}\nabla f(\tilde{\theta}_t)^\top\mathbf{v}_t\cdot\mathbf{v}_t + \sum_{i=1}^{s}\nabla f(\tilde{\theta}_t)^\top\mathbf{p}_{t,i}\cdot\mathbf{p}_{t,i} \tag{12}$$

$$\approx \frac{(d-s)\left(f(\tilde{\theta}_t+\epsilon\mathbf{v}_t)-f(\tilde{\theta}_t)\right)}{\frac{2\epsilon}{\pi}(q-s-1)+\epsilon}\mathbf{v}_t + \sum_{i=1}^{s}\frac{f(\tilde{\theta}_t+\epsilon\mathbf{p}_{t,i})-f(\tilde{\theta}_t)}{\epsilon}\mathbf{p}_{t,i}, \tag{13}$$

where $\mathbf{v}_t := \frac{1}{\sqrt{q-s}}\sum_{i=1}^{q-s}\text{sign}(f(\tilde{\theta}_t+\epsilon\mathbf{u}_i)-f(\tilde{\theta}_t))\mathbf{u}_i$. To ensure the convergence of PARS-OPT, we still require $g_2(\tilde{\theta}_t)$

[1]Throughout this paper, for any vector $\mathbf{v}$, we denote $\overline{\mathbf{v}}$ as its ℓ_2-normalized vector, where $\overline{\mathbf{v}} := \frac{\mathbf{v}}{\|\mathbf{v}\|}$.

Algorithm 1: (P)ARS-OPT Attack

Input: L-smooth function f, $\hat{L} \geq L$, the original image $\mathbf{x}$, the success indicator function $\Phi(\cdot)$, initial ray direction θ_0, number of estimation vectors q, finite-difference step size ϵ, input dimension d, number of iterations T, maximum gradient norm $g_{\max}$, $\gamma_0 > 0$, surrogate model set $\mathbb{S} = \{\hat{\psi}^{(1)}, \ldots, \hat{\psi}^{(s)}\}$ with $s > 0$ for PARS-OPT, and $\mathbb{S} = \emptyset$ for ARS-OPT.
Output: Adversarial example $\mathbf{x}_{\text{adv}}$.
$m_0 \leftarrow \theta_0, \quad \|\hat{\nabla} f_{-1}\|^2 \leftarrow +\infty$;
for $t = 0$ to $T-1$ **do**
  for $\hat{\psi}^{(i)}$ **in** $\mathbb{S}$ **do**
    $\lambda_t \leftarrow \text{BinarySearch}(\mathbf{x}, \theta_t, \hat{\psi}^{(i)}, \Phi)$;
    $\mathbf{k}_{t,i} \leftarrow \nabla_\theta h(\theta_t, \lambda_t)$ on $\hat{\psi}^{(i)}$ with λ_t treated as a constant in differentiation; ▷ obtain s priors.
  end for
  $\mathbf{r}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \ldots, q-s$;
  $\mathbf{p}_{t,1}, \ldots, \mathbf{p}_{t,s}, \mathbf{u}_1, \ldots, \mathbf{u}_{q-s} \leftarrow$ Gram–Schmidt–Orthogonalize$(\{\mathbf{k}_{t,1}, \ldots, \mathbf{k}_{t,s}, \mathbf{r}_1, \ldots, \mathbf{r}_{q-s}\})$;
  $\hat{D}_t \leftarrow \frac{\sum_{i=1}^{s} (\nabla f(\theta_t)^\top \mathbf{p}_{t,i})^2}{\|\hat{\nabla} f_{t-1}\|^2}$; ▷ It requires extra queries.
  $\zeta_t \leftarrow \frac{\hat{D}_t + \frac{(2(q-s-1)+\pi)}{(d-s)\pi}(1-\hat{D}_t)}{\hat{L}\left(\hat{D}_t + \frac{(d-s)\pi}{(2(q-s-1)+\pi)}(1-\hat{D}_t)\right)}$;
  $\tilde{\theta}_t \leftarrow (1-\alpha_t)\theta_t + \alpha_t m_t$, where $\alpha_t \geq 0$ is a positive root of the equation $\alpha_t^2 = \zeta_t \gamma_t (1-\alpha_t)$;
  $\gamma_{t+1} \leftarrow (1-\alpha_t)\gamma_t$;
  $\mathbf{v}_t \leftarrow \frac{1}{\sqrt{q-s}} \sum_{i=1}^{q-s} \text{sign}(f(\tilde{\theta}_t + \epsilon \mathbf{u}_i) - f(\tilde{\theta}_t))\mathbf{u}_i$;
  $\nabla f(\tilde{\theta}_t)^\top \mathbf{v}_t \leftarrow \frac{f(\tilde{\theta}_t + \epsilon \mathbf{v}_t) - f(\tilde{\theta}_t)}{\epsilon}$; ▷ Directional derivative approximation by finite differences.
  $\nabla f(\tilde{\theta}_t)^\top \mathbf{p}_{t,i} \leftarrow \frac{f(\tilde{\theta}_t + \epsilon \mathbf{p}_{t,i}) - f(\tilde{\theta}_t)}{\epsilon}, \quad \forall i = 1, \ldots, s$;
  Estimate $g_1(\tilde{\theta}_t)$, $g_2(\tilde{\theta}_t)$ by using Eq. (11) and Eq. (13);
  $g_1(\tilde{\theta}_t) \leftarrow \text{ClipGradNorm}\big(g_1(\tilde{\theta}_t), g_{\max}\big)$;
  $\|\hat{\nabla} f_t\|^2 \leftarrow \sum_{i=1}^{s} \left(\nabla f(\tilde{\theta}_t)^\top \mathbf{p}_{t,i}\right)^2 + \frac{(d-s)\pi}{2(q-s-1)+\pi} \left(\nabla f(\tilde{\theta}_t)^\top \mathbf{v}_t\right)^2$; ▷ This line is used only in PARS-OPT.
  $\theta_{t+1} \leftarrow \tilde{\theta}_t - \frac{1}{\hat{L}} g_1(\tilde{\theta}_t), \quad m_{t+1} \leftarrow m_t - \frac{\zeta_t}{\alpha_t} g_2(\tilde{\theta}_t)$;
end for
return $\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x} + f(\theta_T)\frac{\theta_T}{\|\theta_T\|}$.

to be an unbiased estimator of $\nabla f(\tilde{\theta}_t)$, whose proof is more involved than in ARS-OPT; see the Appendix for details.

Algorithm 1 presents a unified framework covering both ARS-OPT and PARS-OPT, and Fig. 2 offers an overview of the PARS-OPT procedure. In targeted attacks, we initialize θ_0 with the direction to an image $\tilde{\mathbf{x}}$ from the target class in the training set. The momentum term m_0 is initialized as θ_0 in the first iteration. Specifically, setting $s = 0$ reduces Eq. (11) and Eq. (13) to their counterparts in ARS-OPT, namely Eq. (4) and Eq. (6). Note that $\hat{D}_t$ and $\|\hat{\nabla} f_t\|^2$ are estimators rather than exact values, and $\{\nabla f(\theta_t)^\top \mathbf{p}_{t,i}\}_{i=1}^{s}$ in $\hat{D}_t$ require additional finite-difference approximations. Details are provided in Remark A.12 of Appendix A. Algorithm 1 is a practical approximation of an idealized version presented in Appendix A. Theorem 4.2 establishes the convergence guarantee for this idealized algorithm, giving an $\mathcal{O}(1/T^2)$ rate under smooth convex assumptions, whereas a variant of Sign-OPT only attains an $\mathcal{O}((\ln T)/T)$ rate (Theorem A.10), implying faster convergence for the idealized PARS-OPT.

Theorem 4.2. *Let θ^* denote the optimal solution of Problem* (3)*, and let θ_0, θ_T, γ_0, and ζ_t denote the corresponding quantities in the idealized version of Algorithm 1. Assuming that $f(\cdot)$ is smooth and convex, we have*

$$\mathbb{E}\left[(f(\theta_T) - f(\theta^*))\left(1 + \frac{\sqrt{\gamma_0}}{2}\sum_{t=0}^{T-1}\sqrt{\zeta_t}\right)^2\right] \leq f(\theta_0) - f(\theta^*) + \frac{\gamma_0}{2}\|\theta_0 - \theta^*\|^2. \quad (14)$$

The proof is given in Appendix A (Theorem A.11).

5 Experiments

Experimental Setting

Dataset. We evaluate the proposed method on two publicly available datasets, CIFAR-10 (Krizhevsky and Hinton 2009) and ImageNet (Deng et al. 2009), with images resized to $3 \times 32 \times 32$ and $3 \times 299 \times 299$, respectively. For all experiments, 1,000 images are randomly selected from each dataset as test samples for evaluation. In the case of targeted attacks, the target class is defined as $y_{\text{adv}} = (y+1) \mod C$, where y denotes the true class. For the same target class, we use the same image $\tilde{\mathbf{x}}$ as the initialization for all methods.

Models. On the ImageNet dataset, we evaluate two target models: Inception-v4 (Szegedy et al. 2017) and Swin Transformer (Liu et al. 2021). For Inception-v4 (input resolution 299×299), we use Inception-ResNet-v2 (IncResV2) and Xception as surrogate models. For Swin Transformer (inputs resized to 224×224), the surrogate models are ResNet-50 and ConViT (D'Ascoli et al. 2021). See Appendix for details.

Baseline Methods. We compare ARS-OPT and PARS-OPT against baselines, including HSJA, TA, GeoDA, Evolutionary, SurFree, AHA, QEBA, CGBA-H, SQBA, BBA, Sign-OPT, Prior-Sign-OPT and Prior-OPT. In our methods, the suffix "-S" (e.g., ARS-OPT-S) means the random vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{q-s}$ for gradient estimation are drawn from a $3 \times 56 \times 56$-dimensional subspace. AHA, QEBA, and CGBA-H also adopt subspace sampling, while SQBA, BBA, Prior-Sign-OPT, Prior-OPT, and PARS-OPT leverage surrogate models, denoted by subscripts; e.g., PARS-OPT$_{\text{IncResV2}}$ uses Inception-ResNet-v2 as the surrogate model.

Metrics. We report the mean ℓ_2 distortion as $\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}} \|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_2$, where $\mathbf{X}$ denotes the test dataset. Additionally, we present the attack success rate (ASR), defined as the proportion of samples with distortions below a threshold $\sqrt{0.001 \times d}$ for a given query budget.

Experimental Results on the ImageNet Dataset

Results of Attacks against Undefended Models. Tables 1 and 2 report the results of attacks against undefended models on 1,000 ImageNet images. In summary:

(1) In Table 1, PARS-OPT performs the best in untargeted attacks, while ARS-OPT-S achieves state-of-the-art performance in targeted attacks due to its stabilized optimization via the lookahead direction, reducing the risk of local minima.

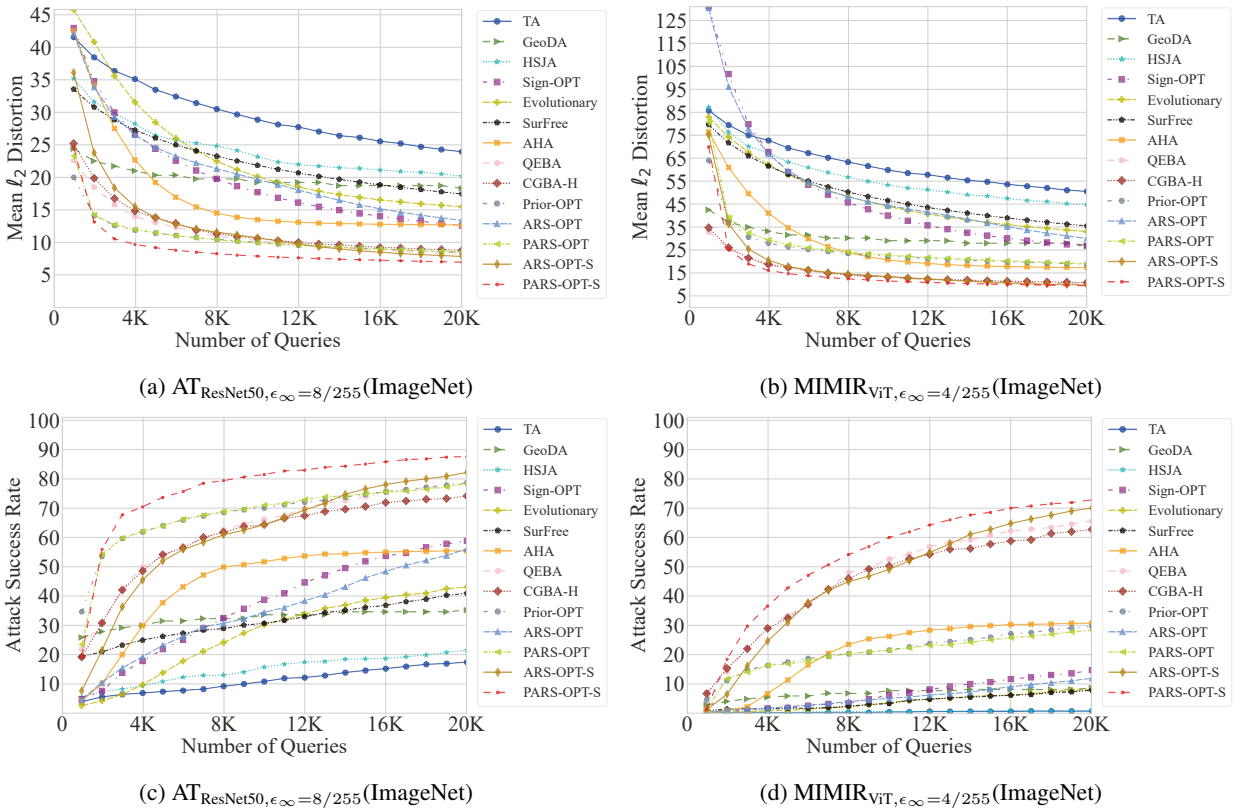
	Method	with D.R.[1]	Untargeted Attack 2K	4K	6K	8K	10K	Targeted Attack 2K	4K	6K	8K	10K	15K	20K
Inception-v4	HSJA	×	44.53	26.31	17.92	14.19	11.65	79.00	60.90	47.25	39.19	32.95	24.55	19.52
	TA	×	42.23	25.86	17.80	14.17	11.69	61.99	47.07	37.16	31.51	27.11	21.08	17.32
	Sign-OPT	×	48.23	23.27	14.97	11.07	8.79	65.20	48.33	38.49	32.10	27.53	20.39	16.28
	GeoDA	×	20.12	14.33	12.49	11.01	9.69	-	-	-	-	-	-	-
	Evolutionary	×	42.66	25.32	17.60	13.38	10.84	65.06	48.37	38.72	32.12	27.39	19.94	15.61
	SurFree	×	38.48	26.35	20.17	16.37	13.82	74.89	61.16	51.56	44.48	39.00	29.35	23.15
	AHA	✓	42.06	23.52	15.41	11.10	8.52	54.12	36.09	26.46	20.50	16.49	10.86	8.12
	QEBA	✓	16.54	8.08	5.82	4.26	3.66	58.31	37.68	28.56	21.74	18.00	12.07	9.25
	CGBA-H	✓	15.12	7.83	5.86	4.61	4.10	56.32	37.82	29.69	23.86	20.00	14.31	11.56
	$\text{SQBA}_{\text{IncResV2}}$	×	19.03	12.80	10.01	8.43	7.42	-	-	-	-	-	-	-
	$\text{BBA}_{\text{IncResV2}}$	×	28.44	20.74	17.37	15.47	14.19	56.28	44.98	38.43	34.07	30.94	25.76	22.63
	$\text{Prior-Sign-OPT}_{\text{IncResV2}}$	×	42.40	17.16	10.19	7.36	5.84	55.42	37.00	28.14	22.96	19.51	14.36	11.66
	$\text{Prior-Sign-OPT}_{\text{IncResV2\&Xception}}$	×	37.10	12.57	7.10	5.19	4.20	49.37	31.34	23.67	19.32	16.70	12.82	10.77
	$\text{Prior-OPT}_{\text{IncResV2}}$	×	18.13	6.80	5.15	4.45	4.03	49.84	36.80	31.04	27.60	25.28	21.84	19.80
	$\text{Prior-OPT}_{\text{IncResV2\&Xception}}$	×	13.42	4.49	3.64	3.32	3.12	**42.63**	30.32	25.60	23.01	21.44	19.19	17.98
	ARS-OPT	×	46.60	24.24	15.74	11.68	9.30	65.53	46.60	35.84	28.84	24.02	16.63	12.69
	$\text{PARS-OPT}_{\text{IncResV2}}$	×	14.02	6.31	4.93	4.24	3.82	49.37	33.88	26.91	22.72	19.94	16.06	13.67
	$\text{PARS-OPT}_{\text{IncResV2\&Xception}}$	×	**9.91**	**4.41**	**3.62**	**3.28**	**3.05**	43.91	**28.16**	**22.56**	19.36	17.23	14.13	12.32
	ARS-OPT-S	✓	25.02	10.38	6.46	4.85	3.92	59.15	37.52	26.37	19.62	15.37	**9.94**	**7.38**
	$\text{PARS-OPT-S}_{\text{IncResV2}}$	✓	19.55	7.82	5.36	4.23	3.54	55.18	34.12	24.22	18.73	**15.02**	10.18	7.84
	$\text{PARS-OPT-S}_{\text{IncResV2\&Xception}}$	✓	20.52	7.25	5.02	4.05	3.45	55.28	33.30	23.70	**18.64**	15.31	10.78	8.33
Swin Transformer	HSJA	×	45.86	27.32	17.92	13.50	10.64	50.96	39.26	30.66	25.64	21.73	16.19	12.75
	TA	×	46.73	27.85	18.02	13.38	10.51	40.72	31.92	25.88	22.25	19.45	15.52	12.89
	Sign-OPT	×	53.40	26.41	16.93	12.41	9.90	44.91	35.98	30.89	27.52	25.27	21.84	19.95
	GeoDA	×	36.92	28.03	24.54	21.59	19.12	-	-	-	-	-	-	-
	Evolutionary	×	49.24	31.19	23.04	18.60	15.74	51.71	38.29	31.23	26.85	23.76	19.28	16.56
	SurFree	×	34.28	23.58	18.37	15.18	13.06	61.31	47.67	39.39	33.84	29.73	22.96	18.73
	AHA	✓	46.76	30.37	23.35	19.39	17.02	36.11	28.04	23.68	20.78	18.76	15.51	13.72
	QEBA	✓	31.11	16.99	12.07	8.46	7.02	42.99	30.31	24.38	19.40	16.52	11.58	8.91
	CGBA-H	✓	29.24	17.01	12.60	9.26	7.81	37.81	27.83	23.19	19.67	17.17	13.10	10.83
	$\text{SQBA}_{\text{ResNet50}}$	×	20.40	13.40	10.33	8.62	7.56	-	-	-	-	-	-	-
	$\text{BBA}_{\text{ResNet50}}$	×	29.37	20.94	17.59	15.47	14.08	35.28	28.45	24.65	22.16	20.34	17.54	15.98
	$\text{Prior-Sign-OPT}_{\text{ResNet50}}$	×	52.88	26.19	16.45	11.88	9.25	43.88	34.32	29.23	26.06	23.86	20.52	18.66
	$\text{Prior-Sign-OPT}_{\text{ResNet50\&ConViT}}$	×	43.06	17.96	10.91	7.90	6.33	43.20	33.48	28.21	24.99	22.84	19.66	17.94
	$\text{Prior-OPT}_{\text{ResNet50}}$	×	39.45	20.26	14.13	11.24	9.62	42.96	33.51	28.64	25.67	23.72	20.86	19.45
	$\text{Prior-OPT}_{\text{ResNet50\&ConViT}}$	×	17.98	8.66	6.45	5.45	4.90	39.62	30.27	25.75	23.12	21.45	19.27	18.33
	ARS-OPT	×	41.91	20.04	12.76	9.26	7.21	38.85	26.14	19.70	15.67	12.99	9.10	6.96
	$\text{PARS-OPT}_{\text{ResNet50}}$	×	29.26	12.77	8.41	6.22	5.01	38.01	25.72	19.72	15.73	13.15	9.39	7.29
	$\text{PARS-OPT}_{\text{ResNet50\&ConViT}}$	×	**12.73**	**6.11**	**4.56**	**3.73**	**3.23**	36.53	23.60	17.98	14.50	12.20	8.61	6.90
	ARS-OPT-S	✓	23.04	10.61	6.88	5.06	3.99	**34.77**	**20.92**	**14.46**	**10.71**	**8.31**	**5.24**	**3.79**
	$\text{PARS-OPT-S}_{\text{ResNet50}}$	✓	23.91	10.96	7.23	5.40	4.31	37.10	22.92	15.85	11.88	9.42	6.03	4.30
	$\text{PARS-OPT-S}_{\text{ResNet50\&ConViT}}$	✓	19.84	8.74	6.09	4.68	3.85	37.06	22.56	16.26	12.26	9.83	6.54	4.80

[1] D.R. denotes the use of dimension reduction technique.

Table 1: Mean ℓ_2 distortions of different query budgets on the ImageNet dataset.

Method	Mean ℓ_2 Distortions 2K	4K	6K	8K	10K	Attack Success Rate 2K	4K	6K	8K	10K
Sign-OPT	49.44	42.29	38.93	37.02	35.71	15.2%	16.3%	18.0%	19.1%	19.4%
$\text{Prior-OPT}_{\text{ResNet50}}$	27.38	21.52	19.34	18.52	18.15	34.9%	44.9%	48.7%	50.6%	51.4%
$\text{Prior-OPT}_{\text{ConViT}}$	21.27	16.54	15.14	14.62	14.36	43.3%	54.7%	57.5%	58.8%	58.9%
$\text{Prior-OPT}_{\text{ResNet50\&ConViT}}$	**18.09**	12.66	11.22	10.72	10.43	**50.1%**	**65.6%**	70.2%	72.2%	73.4%
ARS-OPT	47.42	37.08	31.00	26.97	24.02	16.2%	20.2%	24.7%	28.2%	30.7%
$\text{PARS-OPT}_{\text{ResNet50}}$	26.18	17.65	14.45	12.69	11.55	36.2%	49.8%	56.8%	62.2%	65.8%
$\text{PARS-OPT}_{\text{ConViT}}$	20.80	14.89	12.68	11.47	10.61	43.3%	57.0%	62.0%	66.5%	69.1%
$\text{PARS-OPT}_{\text{ResNet50\&ConViT}}$	18.67	**12.25**	**10.16**	**9.07**	**8.38**	48.7%	65.1%	**72.6%**	**76.0%**	**78.8%**

Table 2: The experimental results of attacking against CLIP with the backbone of ViT-L/14.



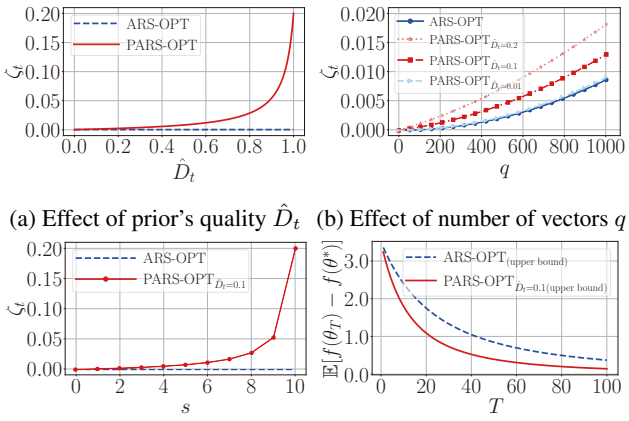
(a) $\text{AT}_{\text{ResNet50},\epsilon_\infty=8/255}$(ImageNet)

(b) $\text{MIMIR}_{\text{ViT},\epsilon_\infty=4/255}$(ImageNet)

(c) $\text{AT}_{\text{ResNet50},\epsilon_\infty=8/255}$(ImageNet)

(d) $\text{MIMIR}_{\text{ViT},\epsilon_\infty=4/255}$(ImageNet)

Figure 3: Mean distortions and attack success rates of untargeted attacks with ℓ_2 norm constraint against defense models. The surrogate model of PARS-OPT and Prior-OPT is the adversarially trained ResNet-50 model (PGD, $\epsilon_{\ell_\infty} = 4/255$).



(a) Effect of prior's quality $\hat{D}_t$ (b) Effect of number of vectors q

(c) Effect of number of priors s (d) Convergence of (P)ARS-OPT

Figure 4: Experimental results of ablation studies.

(2) Table 2 reports untargeted attack results on CLIP (ViT-L/14). Our methods outperform the baselines (Sign-OPT and Prior-OPT) in mean ℓ_2 distortion and attack success rate.

Results of Attacks against Defense Models. We evaluate untargeted attacks against two types of defense models, i.e., adversarial training (AT) (Madry et al. 2018) and MIMIR (Xu et al. 2025). MIMIR achieves state-of-the-art performance on RobustBench (Croce et al. 2021). Fig. 3 shows that our methods achieve the best performance on ImageNet.

Comprehensive Understanding of (P)ARS-OPT

In our ablation studies, we perform controlled experiments based on our theoretical analysis with dimensionality $d = 3{,}072$. Fig. 4a shows the relationship between $\hat{D}_t$ and ζ_t. As $\hat{D}_t$ increases, ζ_t increases accordingly, which in turn improves the convergence rate of PARS-OPT (Eq. (14)). Fig. 4b illustrates that increasing the number of vectors q used for gradient estimation leads to larger ζ_t and improved performance. Fig. 4c shows that when each prior has the same quality, defined as $\bar{D}_t := \hat{D}_t/s$, increasing the number of priors yields a larger ζ_t and higher attack efficiency. Fig. 4d shows that when the prior is effective, even with a small $\hat{D}_t$, PARS-OPT achieves a lower convergence bound than ARS-OPT, indicating better potential performance.

6 Conclusion

We propose a novel hard-label attack approach, comprising two algorithms—ARS-OPT and PARS-OPT—that accelerate convergence and improve attack success rates by leveraging Nesterov-style acceleration and transfer-based priors. We provide convergence guarantees through theoretical analysis and validate our methods with extensive experiments, demonstrating improvements over 13 state-of-the-art approaches.

Acknowledgments

This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LMS25F020005, and by the Key R&D Program of Zhejiang Province under Grant No. 2024C01164.

References

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*.

Brunner, T.; Diehl, F.; Le, M. T.; and Knoll, A. 2019. Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks. In *IEEE/CVF International Conference on Computer Vision*, 4957–4965. IEEE Computer Society.

Chen, J.; and Gu, Q. 2020. RayS: A Ray Searching Method for Hard-label Adversarial Attack. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, 1739–1747. Association for Computing Machinery. ISBN 9781450379984.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *IEEE Symposium on Security and Privacy*, 1277–1294.

Cheng, M.; Le, T.; Chen, P.-Y.; Zhang, H.; Yi, J.; and Hsieh, C.-J. 2019. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In *International Conference on Learning Representations*.

Cheng, M.; Singh, S.; Chen, P. H.; Chen, P.-Y.; Liu, S.; and Hsieh, C.-J. 2020. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. In *International Conference on Learning Representations*.

Cheng, S.; Wu, G.; and Zhu, J. 2021. On the Convergence of Prior-Guided Zeroth-Order Optimization Algorithms. In *Advances in Neural Information Processing Systems*, volume 34, 14620–14631. Curran Associates, Inc.

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; and Hein, M. 2021. RobustBench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

D'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2286–2296. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; and Zhu, J. 2019. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7706–7714.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report 0, University of Toronto.

Li, H.; Xu, X.; Zhang, X.; Yang, S.; and Li, B. 2020. QEBA: Query-Efficient Boundary-Based Blackbox Attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1218–1227. IEEE Computer Society.

Li, J.; Ji, R.; Chen, P.; Zhang, B.; Hong, X.; Zhang, R.; Li, S.; Li, J.; Huang, F.; and Wu, Y. 2021. Aha! Adaptive History-driven Attack for Decision-based Black-box Models. In *IEEE/CVF International Conference on Computer Vision*, 16148–16157. IEEE Computer Society.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision*, 9992–10002. IEEE Computer Society.

Ma, C.; Chen, L.; and Yong, J.-H. 2021. Simulating Unknown Target Models for Query-Efficient Black-box Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11830–11839.

Ma, C.; Guo, X.; Chen, L.; Yong, J.-H.; and Wang, Y. 2021. Finding Optimal Tangent Points for Reducing Distortions of Hard-label Attacks. In *Advances in Neural Information Processing Systems*, volume 34, 19288–19300.

Ma, C.; Xu, X.; Cheng, S.; and Xuan, Q. 2025. Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors. In *International Conference on Learning Representations*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Maho, T.; Furon, T.; and Le Merrer, E. 2021. SurFree: a fast surrogate-free black-box attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10425–10434. IEEE Computer Society.

Nesterov, Y.; and Spokoiny, V. 2017. Random Gradient-Free Minimization of Convex Functions. *Foundations of Computational Mathematics*, 17(2): 527–566.

Park, J.; Miller, P.; and McLaughlin, N. 2024. Hard-label based Small Query Black-box Adversarial Attack. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3974–3983.

Rahmati, A.; Moosavi-Dezfooli, S.-M.; Frossard, P.; and Dai, H. 2020. GeoDA: A Geometric Framework for Black-Box Adversarial Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8443–8452.

Reza, M. F.; Rahmati, A.; Wu, T.; and Dai, H. 2023. CGBA: Curvature-aware Geometric Black-box Attack. In *IEEE/CVF International Conference on Computer Vision*, 124–133.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI Conference on Artificial Intelligence*, AAAI'17, 4278–4284. AAAI Press.

Wang, X.; Zhang, Z.; Tong, K.; Gong, D.; He, K.; Li, Z.; and Liu, W. 2022. Triangle Attack: A Query-Efficient Decision-Based Adversarial Attack. In *European Conference on Computer Vision*, 156–174. Springer-Verlag. ISBN 978-3-031-20064-9.

Xu, X.; Yu, S.; Liu, Z.; and Picek, S. 2025. MIMIR: Masked Image Modeling for Mutual Information-based Adversarial Robustness. arXiv:2312.04960.

Zhang, Z.; Ahmed, N.; and Yu, S. 2024. QE-DBA: Query-Efficient Decision-Based Adversarial Attacks via Bayesian Optimization. In *International Conference on Computing, Networking and Communications*, 783–788.