DIFFERENTIABLE BOUNDARY PRIORS FOR BAYESIAN RAY SEARCH IN HARD-LABEL ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hard-label attacks, one of the most challenging settings in black-box adversarial attacks, rely solely on top-1 label feedback and thus suffer from extremely high query complexity. Existing approaches, such as OPT, formulate hard-label attacks as a zeroth-order ray search problem. Subsequent methods build on this formulation and seek better query efficiency through refined gradient estimation or surrogate priors, yet their improvements remain limited under the extremely sparse feedback of the hard-label setting. In this paper, we propose B-OPT, a Bayesian optimization framework that improves hard-label ray search by optimizing a reparameterized objective under extremely limited feedback. We further propose PB-OPT, which goes beyond local gradient priors by incorporating a global functional prior into the optimization process. To enable this in the hard-label setting, we design a decoupled forward–backward mechanism: the forward pass estimates the surrogate boundary distance via parallel queries, while the backward pass approximates its gradient using a substitute loss. Based on this design, PB-OPT incorporates the surrogate objective as a functional prior in the Gaussian-process mean, allowing Bayesian optimization to refine its posterior estimate using both the surrogate prior and observed target evaluations. We derive a regret bound showing that informative functional priors tighten optimization bounds. Extensive experiments on image classifiers and vision-language models demonstrate that our method significantly improves query efficiency and outperforms 16 state-of-the-art hard-label attacks.

1 INTRODUCTION

Adversarial attacks expose the vulnerability of deep neural networks (DNNs) by generating imperceptible perturbations that cause misclassification. Among different attack settings, hard-label attacks represent the most restrictive black-box scenario, where the adversary only observes the final predicted label without access to confidence scores, gradients, or model internals. Such discrete feedback provides extremely limited information about the decision boundary, making efficient optimization a fundamental challenge.

A representative family of hard-label attacks follows the ray-search paradigm, which searches for the optimal direction θ leading to the closest decision boundary. Specifically, OPT-based methods (Cheng et al., 2019; 2020; Chen & Gu, 2020; Ma et al., 2025; Xu et al., 2026a) optimize the boundary distance function $g(\theta)$, where the distance is estimated via binary search along the ray direction θ. However, this binary-search-based evaluation is inherently non-differentiable and query-intensive, which limits their query efficiency in practice.

Bayesian Optimization (BO) provides a promising framework for query-efficient black-box optimization by learning a surrogate model of the objective function. Recent studies have explored BO for hard-label attacks. Bayes-Attack (Shukla et al., 2021) applies BO in structured low-dimensional subspaces to alleviate the challenge of high-dimensional optimization. QE-DBA (Zhang et al., 2024) further integrates BO with the ray-search framework by modeling the decision boundary radius using Gaussian processes in a generator-induced low-dimensional space. However, existing BO-based hard-label attacks mainly rely on manually designed search spaces, such as predefined basis functions or perturbation generators, rather than exploiting transferable functional knowledge from surrogate models to guide the search process. Furthermore, the boundary distance objective of QE-DBA

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is obtained through binary search over discrete decision feedback, making it difficult to construct differentiable boundary priors for effective Bayesian ray search.

To address these challenges, we propose Bayesian OPT (B-OPT), a hard-label black-box Bayesian ray-search framework that reformulates the OPT objective as maximizing the reciprocal boundary objective $f(\theta) \coloneqq 1/g(\theta)$. B-OPT models this objective with a Gaussian Process surrogate and performs acquisition-guided optimization over normalized directions in a reduced-dimensional search space, with a Matérn kernel and dimension-scaled length-scale initialization for scalable optimization. To further improve query efficiency, we extend B-OPT with surrogate-derived functional priors and propose Prior-guided Bayesian OPT (PB-OPT). However, the boundary objective estimated through binary search over discrete labels is inherently non-differentiable, preventing surrogate models from providing gradient guidance for Bayesian ray search. To address this limitation, PB-OPT introduces a gradient-enabled boundary prior via a decoupled forward–backward scheme, estimating $g(\theta)$ with hard-label queries in the forward pass, and backpropagating implicit surrogate C&W gradients at the surrogate decision boundary. The resulting prior serves as a calibrated mean function for GP modeling: PB-OPT learns a residual GP from target-model observations and adds the prior back to the GP-UCB acquisition. Consequently, acquisition optimization combines gradients from the residual GP posterior and the surrogate-derived prior to select the next search direction. Furthermore, we introduce an adaptive prior weighting strategy with a learnable coefficient jointly optimized with GP hyperparameters, allowing the model to adjust prior influence and remain robust to inaccurate surrogate guidance. We conduct extensive experiments on ImageNet and CIFAR-10, showing that PB-OPT consistently outperforms previous state-of-the-art hard-label attacks, achieving the minimum distortion with only $2{,}000$ queries. We further validate its effectiveness on CLIP and robust defense models (MIMIR & AT). Our main contributions are summarized as follows.

1. **Bayesian optimization for hard-label attacks with functional priors.** We propose B-OPT, a Bayesian ray-search framework for hard-label attacks, and further develop PB-OPT by leveraging surrogate-derived functional priors. PB-OPT adopts a decoupled forward–backward design, where hard-label queries provide boundary-aware objective values and surrogate gradients supply optimization guidance for efficient acquisition maximization.
2. **Theoretical analysis.** We derive a regret bound for a simplified PB-OPT, showing that informative priors reduce the discrepancy between the target objective and prior, leading to tighter optimization bounds.
3. **SOTA performance.** Comprehensive experiments on the ImageNet and CIFAR-10 datasets, as well as attacks against CLIP and defense models, demonstrate that our approach significantly outperforms 16 state-of-the-art methods.

2 Related Work

Boundary-search methods. These methods approach the decision boundary from an adversarial image. Boundary Attack (BA) (Brendel et al., 2018) uses random walks, while HopSkipJumpAttack (HSJA) (Chen et al., 2020) improves query efficiency via binary search and label-based gradient estimation. Later works include curvature exploration in CGBA (Reza et al., 2023), tangent optimization in Tangent Attack (TA) (Ma et al., 2021), and gradient-free DCT searches in SurFree (Maho et al., 2021) and Triangle Attack (Wang et al., 2022). Evolutionary (Dong et al., 2019) and AHA (Li et al., 2021) further improve sampling efficiency.

Ray-search methods. Ray-search approaches optimize a ray direction θ to find minimum perturbation along θ. OPT (Cheng et al., 2019) and RayS (Chen & Gu, 2020) formulate this problem by minimizing the distance function $g(\theta)$, but RayS only supports ℓ_∞-norm attacks. Sign-OPT (Cheng et al., 2020) and Adam-OPT (Ding et al., 2026) reduce queries by estimating directional derivative signs, while Prior-OPT (Ma et al., 2025) and PARS-OPT (Xu et al., 2026a) improve efficiency by incorporating transfer-based priors. Bayes-Attack (Shukla et al., 2021) performs BO in low-dimensional subspaces, while QE-DBA (Zhang et al., 2024) combines BO with ray search by modeling boundary distances in generator-induced spaces. P-BO (Cheng et al., 2024) further demonstrates that surrogate-derived functional priors can improve BO efficiency in soft-label attacks. However, such priors cannot be directly applied to hard-label attacks, where only discrete predicted labels are available and differentiable priors are unavailable.

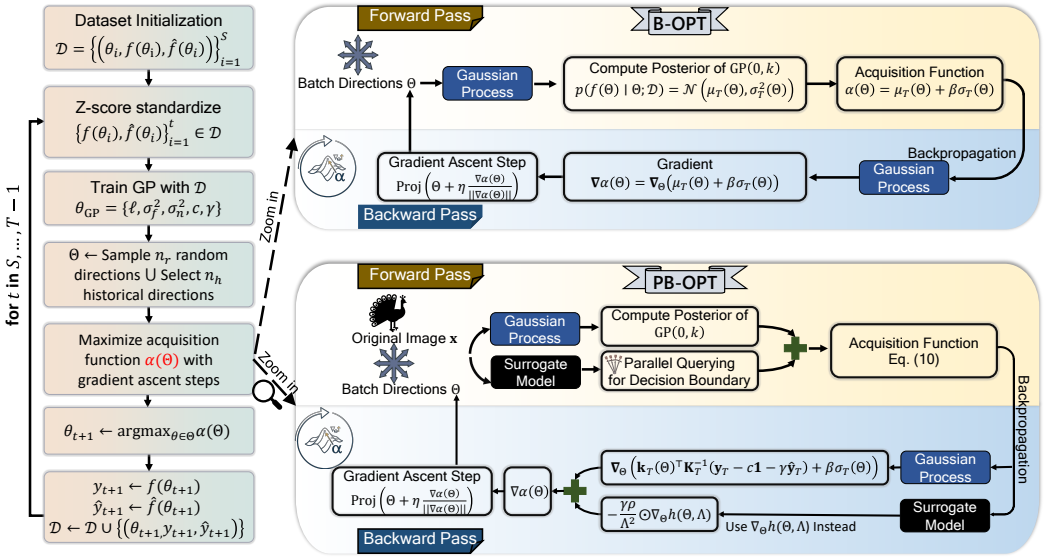


Figure 1: Illustration of the algorithmic pipeline of B-OPT and PB-OPT. B-OPT initializes a dataset $\mathcal{D}$ from S random directions and trains a Gaussian process on standardized observations. It then iteratively updates a batch of ray directions Θ via gradient ascent on the acquisition function. PB-OPT extends this pipeline by integrating a surrogate objective as a functional prior: the forward pass performs parallel querying on the surrogate model to rapidly locate the decision boundary, while the backward pass replaces gradients with $\nabla_\Theta h(\Theta, \Lambda)$, enabling efficient optimization despite the non-differentiability of the original objective.

3 The Proposed Approach

3.1 Problem Definition of Hard-label Attacks

Given a k-class classifier $\psi : \mathbb{R}^d \to \mathbb{R}^k$ and a correctly classified input $\mathbf{x} \in [0,1]^d$, the goal is to find an adversarial example $\mathbf{x}_{\text{adv}}$ with minimal perturbation such that $\psi(\mathbf{x}_{\text{adv}})$ is misclassified. This objective can be formulated as

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

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

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

where $\hat{y} = \arg\max_{i\in\{1,\ldots,k\}} \psi(\mathbf{x}_{\text{adv}})_i$ denotes the predicted label of ψ, $y \in \mathbb{R}$ being the ground-truth label of $\mathbf{x}$, and $y_{\text{adv}} \in \mathbb{R}$ is a specified target label. Following the ray optimization outlined by Cheng et al. (2019; 2020), we transform problem (1) into the task of identifying a ray direction θ that minimizes the distance from $\mathbf{x}$ to the adversarial region:

$$\min_{\theta\in\mathbb{R}^d\setminus\{\mathbf{0}\}} g(\theta) \text{ where } g(\theta) \coloneqq \inf\Big\{\lambda : \lambda > 0, \Phi\big(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|_p}\big) = 1\Big\}, \quad p \in \{2, \infty\}. \tag{3}$$

Note that $g(\theta) = +\infty$ when the set is empty, since $\inf \emptyset = +\infty$ by convention. Finally, the adversarial example is $\mathbf{x}^* = \mathbf{x} + g(\theta^*)\frac{\theta^*}{\|\theta^*\|_p}$, where θ^* denotes the optimal solution.

3.2 BO for Ray Direction Search

Bayesian Optimization (BO) (Jones et al., 1998) provides a query-efficient framework for optimizing expensive black-box functions by iteratively updating a Gaussian Process (GP) model and selecting

new candidates through an acquisition function. Given observations $\mathcal{D} = \{(\theta_i, f(\theta_i))\}_{i=1}^T$, BO alternates between GP posterior updating, acquisition optimization, and objective evaluation.

Based on this framework, we propose Bayesian OPT (B-OPT), which reformulates OPT-style hard-label attacks as Bayesian optimization over ray directions. Specifically, B-OPT maximizes the reciprocal objective $f(\theta) \coloneqq 1/g(\theta)$, where $g(\theta)$ denotes the minimum perturbation distance required to cross the decision boundary along direction θ from $\mathbf{x}$. This transformation converts the original minimization problem into maximization and enlarges objective variations, improving GP modeling stability (Table 4). Furthermore, B-OPT optimizes normalized ray directions with $\|\theta\|_p = 1$, where $p \in \{2, \infty\}$ denotes the attack norm, allowing Bayesian optimization to explore perturbation directions rather than the magnitudes. B-OPT models $f(\theta)$ using a zero-mean Gaussian Process with kernel $k(\cdot, \cdot)$. Given observations $\mathcal{D}$, the GP posterior follows a Gaussian distribution $\mathcal{N}(\mu_T(\theta), \sigma_T^2(\theta))$, where

$$\mu_T(\theta) = \mathbf{k}_T(\theta)^\top \mathbf{K}_T^{-1} \mathbf{y}_T, \tag{4}$$

$$\sigma_T^2(\theta) = k(\theta, \theta) - \mathbf{k}_T(\theta)^\top \mathbf{K}_T^{-1} \mathbf{k}_T(\theta), \tag{5}$$

with $\mathbf{k}_T(\theta) = [k(\theta_1, \theta), k(\theta_2, \theta), \ldots, k(\theta_T, \theta)]^\top$, $\mathbf{K}_T \in \mathbb{R}^{T \times T}$ denoting the kernel matrix whose (i, j)-th entry is $k(\theta_i, \theta_j)$, and $\mathbf{y}_T = [y_1, \ldots, y_T]^\top$. Based on the posterior distribution, B-OPT adopts the Upper Confidence Bound (UCB) acquisition function $\alpha(\theta) = \mu_T(\theta) + \beta\sigma_T(\theta)$ and selects the next ray direction by maximizing $\alpha(\theta)$. To enable Bayesian optimization in high-dimensional image spaces, B-OPT optimizes ray directions in a reduced-dimensional space and maps them to the image space via bilinear upsampling $\mathcal{U}(\cdot)$, followed by p-norm normalization as $\mathcal{U}(\theta)/\|\mathcal{U}(\theta)\|_p$ for boundary evaluation. Specifically, the attacks on ImageNet optimize directions in a $56 \times 56 \times 3$ space. Moreover, we employ a Matérn kernel with dimension-aware length-scale initialization $\ell_0 = 0.02\sqrt{d_r}$, where d_r denotes the reduced dimension. It follows the scaling property of high-dimensional distances and improves GP scalability during optimization (Xu et al., 2025).

3.3 Prior-Guided BO for Ray Direction Search

Next, we propose Prior-guided Bayesian OPT (PB-OPT), which improves query efficiency in hard-label attacks by incorporating a global functional prior derived from a surrogate model into B-OPT. Unlike B-OPT, which learns the optimization landscape solely from target-model observations, PB-OPT exploits transferable information from a surrogate model to guide Bayesian ray search. To isolate the effect of a functional prior, we analyze a simplified GP-UCB setting with a generic objective f, its finite-evaluation approximation $\widetilde{f}_L$, and a fixed functional prior $\hat{f}$. The result is intended to characterize how prior quality affects the optimization bound. We have

Theorem 3.1 (Regret with a functional prior)**.** *Let f, $\widetilde{f}_L$, and $\hat{f}$ be bounded real-valued functions on a compact domain Θ. Assume that the approximation error satisfies $\|\widetilde{f}_L - f\|_\infty \le \delta_L$. Further assume that $r_L \coloneqq \widetilde{f}_L - \hat{f} \in \mathcal{H}_k$, $B_L \coloneqq \|r_L\|_k$, and $k(\theta, \theta) \le 1$. Let $(\mu_{t-1}, \sigma_{t-1})$ be the regularized GP posterior with fixed mean $\hat{f}$, kernel k, and regularization parameter $\sigma_n^2 > 0$, constructed from $\{(\theta_i, \widetilde{f}_L(\theta_i))\}_{i<t}$. If $\theta_t \in \arg\max_{\theta\in\Theta}\big[\mu_{t-1}(\theta) + B_L\sigma_{t-1}(\theta)\big]$, then*

$$R_T \coloneqq \sum_{t=1}^{T} \left(\sup_{\theta\in\Theta} f(\theta) - f(\theta_t) \right) \le B_L \sqrt{\frac{8T\Gamma_T}{\log(1 + \sigma_n^{-2})}} + 2T\delta_L, \tag{6}$$

where $\Gamma_T \coloneqq \frac{1}{2}\sup_{\theta_{1:T}\in\Theta^T} \log\det(\mathbf{I} + \sigma_n^{-2}\mathbf{K}_T)$ and $\mathbf{K}_T = [k(\theta_i, \theta_j)]_{i,j=1}^T$.

Theorem 3.1 abstracts away the practical calibration and acquisition-optimization details to highlight this prior-quality effect, which shows that the regret bound depends on the residual complexity $B_L = \|\widetilde{f}_L - \hat{f}\|_k$, motivating adaptive prior calibration with learnable γ and c. In practice, PB-OPT uses $f(\theta) = 1/g(\theta)$ and $\hat{f}(\theta) = 1/\hat{g}(\theta)$. Although Bayesian optimization can model the non-differentiable objective $\hat{f}(\theta)$, our acquisition function is optimized through gradient-based updates of candidate directions. Since $\hat{f}(\theta)$ is obtained through discrete boundary search and has no direct differentiable path, we introduce a decoupled forward–backward scheme that estimates the boundary objective in the forward pass and provides surrogate boundary gradients in the backward pass.

Forward. The forward pass focuses on efficient estimation of the boundary distance. Given a direction θ_0, let $\lambda_0 = \hat{g}(\theta_0)$ denote the exact boundary distance on the surrogate model. In practice,

instead of sequential binary search, PB-OPT estimates λ_0 by querying multiple points along the ray to locate a boundary interval, followed by refinement between the closest non-adversarial and adversarial points. The corresponding functional prior is $\hat{f}(\theta_0) = 1/\lambda_0$, whose value is approximated using the estimated boundary distance.

Backward. Since $\hat{f}(\theta_0)$ is obtained through a discrete boundary-search operation without a differentiable computational path, we use surrogate boundary gradients, i.e., $\nabla_\theta h(\theta_0, \lambda_0)$, to provide guidance, where

$$h(\theta, \lambda) = \begin{cases} \hat{\psi}_y - \max_{i \neq y} \hat{\psi}_i, & \text{untargeted attack,} \\ \max_{i \neq y_{\text{adv}}} \hat{\psi}_i - \hat{\psi}_{y_{\text{adv}}}, & \text{targeted attack.} \end{cases} \tag{7}$$

with $\hat{\psi}_i = \hat{\psi}\left(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\right)_i$. The exact surrogate boundary is implicitly characterized by $h(\theta, \hat{g}(\theta)) = 0$. At points where h is locally differentiable and $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$, implicit differentiation at $\lambda_0 = \hat{g}(\theta_0)$ gives $\nabla_\theta \hat{g}(\theta_0) = \rho \nabla_\theta h(\theta_0, \lambda_0)$, where $\rho \coloneqq -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)}$. Since $\hat{f}(\theta) = 1/\hat{g}(\theta)$, we further obtain $\nabla_\theta \hat{f}(\theta_0) = -\frac{\rho}{\lambda_0^2} \nabla_\theta h(\theta_0, \lambda_0)$ as a substitute gradient.

With the differentiable functional prior, PB-OPT models the objective using a Gaussian Process: $f \sim \mathrm{GP}(\gamma \hat{f} + c, k)$, where γ and c are learnable scalars for controlling the prior contribution. The posterior mean and variance are[1]:

$$\mu_T(\theta) = \mathbf{k}_T(\theta)^\top \mathbf{K}_T^{-1}(\mathbf{y}_T - c\mathbf{1} - \gamma \hat{\mathbf{y}}_T) + \gamma \hat{f}(\theta) + c, \tag{8}$$

$$\sigma_T^2(\theta) = k(\theta, \theta) - \mathbf{k}_T(\theta)^\top \mathbf{K}_T^{-1} \mathbf{k}_T(\theta), \tag{9}$$

where $\hat{\mathbf{y}}_T = [\hat{y}_1, \ldots, \hat{y}_T]^\top$, and the remaining notation follows Eqs. (4)–(5). The acquisition function $\alpha(\theta) = \mu_T(\theta) + \beta \sigma_T(\theta)$ combines a GP term and a prior term:

$$\alpha(\theta) = \underbrace{\mathbf{k}_T(\theta)^\top \mathbf{K}_T^{-1}(\mathbf{y}_T - c\mathbf{1} - \gamma \hat{\mathbf{y}}_T) + \beta \sigma_T(\theta)}_{\text{GP term}} + \underbrace{\gamma \hat{f}(\theta) + c}_{\text{prior term}}. \tag{10}$$

Now, we further introduce a learnable nonnegative coefficient γ and a bias term c to adaptively calibrate the surrogate-derived prior, which reduces the discrepancy between the target objective and the surrogate prior. Ideally, the residual complexity between the target objective and the scaled prior is minimized as $\|f - (\gamma \hat{f} + c)\|_k^2$. Since f is unknown over the search space and only observed at queried directions, we learn c, γ, and the GP hyperparameters from the observed target and surrogate values. After standardization, we model $\mathbf{y}_T \mid \Theta_T, \gamma, c, \ell$. Let $\mathbf{r}_T = \mathbf{y}_T - c\mathbf{1} - \gamma \hat{\mathbf{y}}_T$. We jointly optimize γ and c by minimizing

$$-\log p(\mathbf{y}_T \mid \Theta_T, \gamma, c, \ell) = \frac{1}{2} \mathbf{r}_T^\top \mathbf{K}_T^{-1} \mathbf{r}_T + \frac{1}{2} \log |\mathbf{K}_T| + \frac{T}{2} \log(2\pi). \tag{11}$$

For fixed $\mathbf{K}_T$, this is equivalent to minimizing the kernel-weighted residual with respect to c and γ.

Algorithm 1 summarizes the procedure of ℓ_2-norm PB-OPT. We first sample S random unit directions and evaluate both the target boundary objective and the surrogate-derived prior. After standardizing the observation dataset, PB-OPT models the residual $f - (\gamma \hat{f} + c)$ with a GP and optimizes γ together with GP hyperparameters via marginal likelihood maximization. At each iteration, PB-OPT initializes a candidate pool using both newly sampled random directions and recently evaluated directions. It then performs N steps of gradient ascent on the prior-augmented acquisition function. During each step, the residual GP provides the UCB term, while parallel boundary probing on the surrogate model evaluates the prior. The backward pass implicitly differentiates $h(\theta, \lambda)$ at the estimated boundary, allowing the acquisition gradient to combine the GP-UCB gradient with surrogate-derived boundary guidance. After acquisition optimization, the candidate with the largest acquisition value is queried on the target model. Its target and surrogate observations are added to the dataset, and the residual GP is updated periodically. This process continues until the query budget is exhausted. Finally, the best observed direction is upsampled and normalized to construct the adversarial example.

[1]For simplicity, we reuse f, $\hat{f}$, $\mathbf{y}_T$, and $\hat{\mathbf{y}}_T$ after standardization.

Algorithm 1 ℓ_2 Norm of Prior-guided Bayesian OPT (PB-OPT)

Input: benign image $\mathbf{x} \in [0,1]^d$, GP with its length-scale ℓ, calibration parameters γ, c, UCB coefficient β, initial sample size S, number of iterations T, update steps N for acquisition, reduced search dimension d_r, learning rate η, random and historical candidate sizes n_r and n_h, target objective $f(\theta) = 1/g(\theta)$ with boundary-distance function g, and surrogate objective $\hat{f}$.
Output: adversarial example $\mathbf{x}^*$ that satisfies $\Phi(\mathbf{x}^*) = 1$.
Initialize $\mathcal{D} \leftarrow \{(\theta_i, f(\theta_i), \hat{f}(\theta_i))\}_{i=1}^{S}$ by querying f, $\hat{f}$
for $t = S, \ldots, T-1$ **do**
  Compute (μ_f, σ_f) and $(\mu_{\hat{f}}, \sigma_{\hat{f}})$ from $\mathcal{D}$ ▷ Mean and standard deviation.
  Standardize $\{f(\theta_i), \hat{f}(\theta_i)\}_{i=1}^{t}$ using the above statistics ▷ Z-score normalization.
  Update GP by maximizing Eq. (11) ▷ Fitting GP periodically.
  Sample and normalize random directions $\Theta_r \leftarrow \mathrm{Normalize}(\{\tilde{\theta}_i\}_{i=1}^{n_r}), \tilde{\theta}_i \sim U([-1,1]^{d_r})$
  Select recent directions $\Theta_h \leftarrow \{\theta_{t-n_h+1}, \ldots, \theta_t\}$
  $\Theta \leftarrow \Theta_r \cup \Theta_h$ ▷ Candidate directions.
  for $k = 1, \ldots, N$ **do**
    Compute posterior mean $\mu_t(\Theta)$ and variance $\sigma_t^2(\Theta)$
    Obtain surrogate boundary distances $\Lambda \leftarrow \hat{g}(\Theta)$
    Compute acquisition values $\alpha(\Theta) \leftarrow \mathbf{k}_t(\Theta)^\top \mathbf{K}_t^{-1}\big(\mathbf{y}_t - c\mathbf{1} - \gamma\hat{\mathbf{y}}_t\big) + \beta\sigma_t(\Theta) + \gamma\hat{f}(\Theta) + c$
    $\nabla_\Theta \alpha(\Theta) \leftarrow \nabla_\Theta\Big(\mathbf{k}_t(\Theta)^\top \mathbf{K}_t^{-1}\big(\mathbf{y}_t - c\mathbf{1} - \gamma\hat{\mathbf{y}}_t\big) + \beta\sigma_t(\Theta)\Big) - \frac{\gamma\rho}{\Lambda^2} \odot \nabla_\Theta h(\Theta, \Lambda)$
    Update candidates $\Theta \leftarrow \mathrm{Proj}\Big(\Theta + \eta \cdot \frac{\nabla_\Theta \alpha(\Theta)}{\|\nabla_\Theta \alpha(\Theta)\|}\Big)$ ▷ Clamp to $[-1,1]^{d_r}$ and normalize each $\theta_i \in \Theta$.
  end for
  Recompute acquisition values $\alpha(\Theta)$ for candidates
  $\theta_{t+1} \leftarrow \arg\max_{\theta \in \Theta} \alpha(\theta)$
  Evaluate $y_{t+1} \leftarrow f(\theta_{t+1})$ and $\hat{y}_{t+1} \leftarrow \hat{f}(\theta_{t+1})$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\theta_{t+1}, y_{t+1}, \hat{y}_{t+1})\}$
end for
$\theta^* \leftarrow \arg\max_{(\theta_i, y_i) \in \mathcal{D}} y_i$ ▷ Best observed direction.
Obtain boundary distance $\lambda^* \leftarrow g(\mathcal{U}(\theta^*))$ ▷ $\mathcal{U}(\theta)$ upsamples reduced-dimensional θ.
Compute normalized direction $\mathbf{v}_{\theta^*} \leftarrow \mathcal{U}(\theta^*)/\|\mathcal{U}(\theta^*)\|_2$
return $\mathbf{x}^* \leftarrow \mathbf{x} + \lambda^* \mathbf{v}_{\theta^*}$.

4 Experiment

4.1 Experimental Setting

Datasets. All experiments are conducted on two datasets, i.e., CIFAR-10 (Krizhevsky & Hinton, 2009) and ImageNet (Deng et al., 2009). The image sizes are $32 \times 32 \times 3$ for CIFAR-10, and $224 \times 224 \times 3$ for ImageNet. We randomly select 1,000 images from the validation sets for experiments.

Compared Methods. To provide a comprehensive comparison, we select 16 state-of-the-art hard-label attacks, including Sign-OPT (Cheng et al., 2020), HSJA (Chen et al., 2020), CGBA-H (Reza et al., 2023), QEBA (Li et al., 2020), AHA (Li et al., 2021), TA (Ma et al., 2021), G-TA (Ma et al., 2021), SurFree (Maho et al., 2021), GeoDA (Rahmati et al., 2020), Evolutionary (Dong et al., 2019), SQBA (Park et al., 2024), QE-DBA (Zhang et al., 2024), ARS-OPT, PARS-OPT (Xu et al., 2026a), Prior-Sign-OPT and Prior-OPT (Ma et al., 2025).

Target and Surrogate Models. In the ImageNet dataset, the target model is selected as Swin Transformer (Liu et al., 2021), and the surrogate model is adopted as ResNet-50 (He et al., 2016).

Method Setting. We initialize $\mathcal{D}$ with 10 random unit directions and standardize target/surrogate observations. The GP is refitted every 10 BO iterations using Adam for 50 steps with a learning rate of 0.1. Each acquisition optimization performs 10-step projected gradient ascent over 20 candidates (10 random and 10 recent directions). For ImageNet, we optimize directions in reduced spaces and upsample them to the original image resolution via bilinear interpolation. We use $56 \times 56 \times 3$ as

the reduced dimension for B-OPT and PB-OPT in all main experiments. The Matérn-$5/2$ kernel length-scale is initialized as $0.02\sqrt{d_r}$ for ℓ_2 attacks, where d_r is the reduced dimension.

Evaluation Metric. Each target-model label evaluation is counted as one query. Evaluating $f(\theta) = 1/g(\theta)$ may require multiple target-model queries during boundary search, all of which are included in the reported query budget. All methods are evaluated using the mean distortion over $1{,}000$ images as $\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}(\|\mathbf{x}_{\text{adv}} - \mathbf{x}\|_p)$ under different query budgets, where $\mathbf{X}$ is the test set and $p \in \{2, \infty\}$ is the attack norm. We also report the attack success rate (ASR), defined as the percentage of samples with distortion below a threshold ϵ, which is set to $\epsilon = \sqrt{0.0015 \times d}$ in ℓ_2-norm attacks.

4.2 Comparison with State-of-the-Art Attacks

Table 1: Mean ℓ_2 distortions vs. query budget for ℓ_2-norm attacks against WRN-28-10-drop and PyramidNet-272 on CIFAR-10.

Method	WRN-28-10-drop						PyramidNet-272					
	400	800	1000	1200	1600	2000	400	800	1000	1200	1600	2000
HSJA	1.89	0.99	0.80	0.80	0.57	0.51	3.14	1.67	1.29	1.29	0.85	0.71
GeoDA	2.70	1.90	1.57	1.52	1.35	1.23	4.10	2.90	2.26	2.23	1.85	1.59
TA	1.89	0.98	0.78	0.78	0.56	0.50	3.01	1.66	1.28	1.28	0.84	0.71
G-TA	1.87	0.97	0.78	0.78	0.56	0.50	3.14	1.65	1.26	1.26	0.83	0.70
Sign-OPT	3.03	1.79	1.50	1.36	1.04	0.83	4.41	2.73	2.35	2.08	1.56	1.19
QEBA	2.24	1.25	1.04	0.91	0.81	0.74	3.39	2.04	1.67	1.40	1.20	1.06
QE-DBA	3.02	2.50	2.37	2.28	2.15	2.08	4.19	3.61	3.49	3.38	3.24	3.16
Evolutionary	2.78	1.86	1.59	1.38	1.09	0.90	3.43	2.26	1.92	1.66	1.29	1.06
SurFree	2.20	1.57	1.39	1.25	1.06	0.94	2.95	1.99	1.73	1.55	1.30	1.13
AHA	3.82	2.92	2.67	2.49	2.26	2.14	5.00	4.10	3.79	3.56	3.25	3.06
CGBA-H	1.65	0.94	0.77	0.72	0.57	0.50	2.75	1.53	1.20	1.17	0.82	0.70
$\text{SQBA}_{\text{ResNet-110}}$	0.95	0.70	0.63	0.58	0.51	0.46	1.68	1.04	0.89	0.79	0.64	0.56
$\text{Prior-Sign-OPT}_{\text{ResNet-110}}$	3.01	1.74	1.43	1.32	1.01	0.78	4.38	2.65	2.30	2.07	1.52	1.15
$\text{Prior-OPT}_{\text{ResNet-110}}$	3.03	1.38	1.14	1.01	0.73	0.59	4.37	2.36	2.05	1.85	1.37	1.04
ARS-OPT	3.08	2.00	1.54	1.26	0.99	0.75	4.49	3.00	2.22	1.76	1.39	1.01
$\text{PARS-OPT}_{\text{ResNet-110}}$	3.09	1.84	1.32	1.19	0.78	0.63	4.49	2.91	2.09	1.99	1.28	1.03
B-OPT (Ours)	1.81	1.16	1.00	0.90	0.76	0.67	2.84	1.81	1.56	1.39	1.14	0.98
$\text{PB-OPT}_{\text{ResNet-110}}$ (Ours)	**0.48**	**0.37**	**0.34**	**0.33**	**0.30**	**0.29**	**1.00**	**0.63**	**0.56**	**0.52**	**0.46**	**0.41**

Results of Attacks against Undefended Models. We compared the results of B-OPT and PB-OPT under different datasets, models, and norm constraints, as shown in Tables 1, 2, and 3.

(1) In ℓ_2-Norm attacks (Tables 1 and 3), our approach achieves state-of-the-art performance on both datasets. Notably, PB-OPT achieves over 60% lower mean ℓ_2 distortion than Prior-OPT on ImageNet.

(2) In ℓ_∞-norm attacks (Table 3), PB-OPT achieves the best performance in untargeted settings, even attaining lower distortion at 400 queries than most other methods at $2{,}000$.

(3) Table 2 reports the mean ℓ_2 distortion and attack success rate (ASR) on CLIP (ViT-L/14) under untargeted attacks. CLIP is a vision-language model that enables zero-shot classification by aligning images and text in a shared embedding space. PB-OPT consistently achieves the lowest distortion and highest ASR across all query budgets, demonstrating its effectiveness on vision-language models.

Results of Attacks against Defense Models. We conduct untargeted attacks against two defense models, i.e., adversarial training (AT) (Madry et al., 2018) and MIMIR (Xu et al., 2026b), where MIMIR is a state-of-the-art self-supervised adversarial defense that leverages masked image modeling with mutual information regularization to learn robust representations. We use an independently trained AT ResNet-50 ($\epsilon_{\ell_\infty} = 4/255$) as the surrogate model, while the target models are independently trained AT (ResNet-50, $\epsilon_{\ell_\infty} = 8/255$) and MIMIR (ViT) with different parameters and decision boundaries. Fig. 2 shows that PB-OPT achieves the lowest distortion and highest ASR across all query budgets for both defenses, which maintains clear advantages in the low-query regime.

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

Method	Mean ℓ_2 Distortions ↓						Attack Success Rate ↑					
	400	800	1000	1200	1600	2000	400	800	1000	1200	1600	2000
HSJA	63.06	61.38	61.38	60.65	59.59	58.95	7.1%	8.1%	8.1%	8.4%	9.2%	9.8%
GeoDA	46.55	44.59	43.16	43.15	41.61	40.11	11.8%	13.2%	14.6%	14.6%	16.5%	18.6%
TA	61.20	59.45	59.45	58.65	57.75	56.89	7.7%	9.1%	9.1%	9.4%	10.2%	10.6%
G-TA	60.41	58.42	58.42	57.64	56.81	55.91	7.8%	9.0%	9.0%	9.2%	9.5%	9.6%
Sign-OPT	65.85	60.06	58.18	55.93	52.02	49.44	14.9%	16.0%	16.1%	16.7%	17.6%	17.9%
QEBA	35.09	21.92	17.90	17.90	12.72	10.89	22.9%	45.8%	54.3%	54.3%	69.8%	75.0%
QE-DBA	22.40	19.94	19.30	18.76	18.13	17.64	49.4%	53.7%	55.0%	56.0%	57.4%	59.1%
Evolutionary	50.54	37.41	32.90	29.05	23.26	19.21	9.9%	17.9%	22.9%	28.3%	39.5%	48.1%
SurFree	30.59	22.89	20.33	18.36	15.33	13.20	31.0%	44.1%	50.3%	55.4%	62.2%	68.9%
AHA	54.33	41.68	37.05	33.06	27.30	22.93	7.3%	12.2%	16.3%	20.9%	29.7%	37.5%
CGBA-H	37.10	25.93	22.47	22.22	17.01	15.03	23.2%	39.1%	46.3%	46.3%	58.0%	62.7%
$\text{SQBA}_{\text{ResNet-50}}$	20.20	16.48	15.03	13.92	12.13	10.77	47.8%	57.1%	61.1%	65.0%	71.7%	76.0%
$\text{Prior-Sign-OPT}_{\text{ResNet-50}}$	66.41	59.51	56.93	54.14	50.38	47.23	13.9%	15.7%	16.4%	17.3%	17.8%	19.3%
$\text{Prior-OPT}_{\text{ResNet-50}}$	66.32	44.59	38.93	35.43	30.76	27.38	14.2%	25.1%	28.9%	32.1%	38.1%	41.5%
ARS-OPT	66.61	59.72	57.12	54.03	49.90	47.41	14.4%	15.6%	16.3%	17.1%	18.5%	19.7%
$\text{PARS-OPT}_{\text{ResNet-50}}$	66.53	49.71	40.83	37.65	29.69	26.17	14.3%	20.0%	25.1%	28.7%	37.8%	43.2%
B-OPT (Ours)	27.23	18.10	15.62	13.84	11.66	10.20	43.1%	59.6%	65.3%	69.6%	75.3%	79.7%
$\text{PB-OPT}_{\text{ResNet-50}}$ (Ours)	**20.10**	**14.14**	**12.50**	**11.34**	**9.80**	**8.77**	**55.2%**	**68.0%**	**72.3%**	**75.7%**	**80.6%**	**82.2%**

Table 3: Mean distortion vs. query budget for ℓ_2- and ℓ_∞-norm attacks against Swin Transformer on ImageNet. "–" indicates that the corresponding attack method does not support the specified norm.

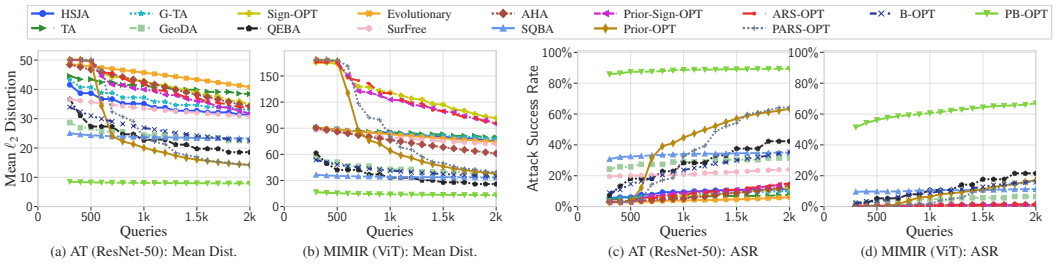
Method	Mean ℓ_2 Distortions ↓						Mean ℓ_∞ Distortions ↓					
	400	800	1000	1200	1600	2000	400	800	1000	1200	1600	2000
HSJA	94.60	74.64	74.64	65.95	51.50	45.82	0.35	0.31	0.31	0.29	0.25	0.23
GeoDA	65.28	56.69	48.94	48.94	42.22	36.72	**0.23**	0.21	0.19	0.19	0.17	0.15
TA	91.37	73.24	73.24	65.47	52.06	46.73	0.55	0.54	0.54	0.53	0.52	0.52
G-TA	90.72	72.26	72.26	64.14	50.54	44.88	0.35	0.31	0.31	0.29	0.25	0.23
Sign-OPT	97.95	81.37	75.49	68.23	56.89	49.15	0.90	0.83	0.79	0.74	0.65	0.58
QEBA	67.54	51.78	45.48	45.48	35.67	31.76	0.31	0.25	0.23	0.22	0.19	0.17
QE-DBA	39.46	34.62	33.51	32.71	31.53	30.56	–	–	–	–	–	–
Evolutionary	89.62	74.90	69.27	64.18	55.83	49.24	–	–	–	–	–	–
SurFree	64.72	51.77	47.44	43.78	38.22	34.28	–	–	–	–	–	–
AHA	89.96	72.96	66.66	61.33	53.02	46.76	–	–	–	–	–	–
CGBA-H	63.04	44.37	39.40	38.38	29.86	26.69	–	–	–	–	–	–
$\text{SQBA}_{\text{ResNet-50}}$	39.49	31.35	28.49	26.30	22.89	20.38	–	–	–	–	–	–
RayS	–	–	–	–	–	–	0.34	0.25	0.23	0.21	0.19	0.17
$\text{Prior-Sign-OPT}_{\text{ResNet-50}}$	98.26	80.87	74.84	69.41	59.06	49.99	0.90	0.84	0.83	0.79	0.74	0.69
$\text{Prior-OPT}_{\text{ResNet-50}}$	98.18	71.15	62.44	56.16	44.28	36.74	0.90	0.80	0.74	0.70	0.59	0.52
ARS-OPT	118.21	93.44	81.31	70.69	59.69	49.46	0.91	0.86	0.82	0.81	0.74	0.68
$\text{PARS-OPT}_{\text{ResNet-50}}$	98.30	77.26	63.62	58.87	42.83	35.34	0.91	0.84	0.84	0.76	0.69	0.63
B-OPT (Ours)	57.36	42.39	38.42	35.46	31.42	28.78	0.29	0.23	0.20	0.19	0.16	0.15
$\text{PB-OPT}_{\text{ResNet-50}}$ (Ours)	**34.54**	**23.95**	**21.51**	**19.78**	**17.23**	**15.75**	0.24	**0.18**	**0.16**	**0.15**	**0.13**	**0.12**

4.3 Comprehensive Understanding of (P)B-OPT

We conduct comprehensive ablation studies on ImageNet with the Swin Transformer to examine the key components of B-OPT and PB-OPT. Table 4 isolates the effects of the objective transformation, functional prior, GP acquisition gradient, prior gradient, and GP hyperparameter fitting.

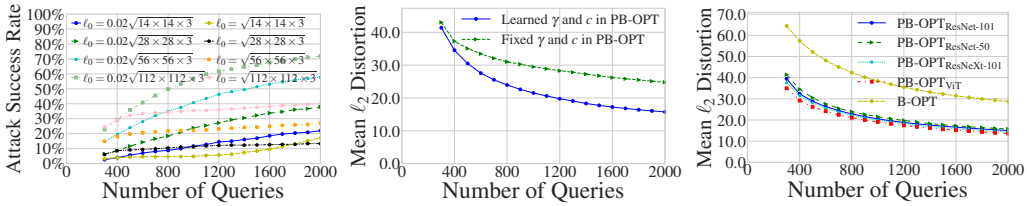
For B-OPT, the reciprocal objective $1/g(\theta)$ achieves lower distortion and higher ASR than directly optimizing $-g(\theta)$, validating the reciprocal transformation. This result confirms that the reciprocal mapping provides a more suitable objective.

For PB-OPT, the complete method consistently outperforms all ablated variants. At $2{,}000$ queries, it achieves a mean ℓ_2 distortion of 15.75 and an ASR of 57.79%. Removing either gradient component or GP fitting degrades performance, while a random Gaussian prior or the P-BO-style C&W margin prior (Cheng et al., 2024) also leads to inferior results, showing the importance of informative priors.



(a) AT (ResNet-50): Mean Dist. (b) MIMIR (ViT): Mean Dist. (c) AT (ResNet-50): ASR (d) MIMIR (ViT): ASR

Figure 2: Performance of ℓ_2-norm untargeted attacks against AT (ResNet-50) and MIMIR (ViT) defense models on ImageNet. (a)–(b) Mean ℓ_2 distortions. (c)–(d) Attack success rates. The surrogate model is independently trained AT with $\epsilon_\infty = 4/255$.



(a) Effect of ℓ_0 and d in PB-OPT. (b) Learned vs. Fixed γ and c. (c) Effect of surrogate models.

Figure 3: Ablation studies of PB-OPT on ImageNet (Swin Transformer). (a) Reduced dimension and initial length-scale ℓ_0. (b) Learned vs. fixed γ and c. (c) Comparison of different surrogate models.

Table 4: Ablation study of different components in B-OPT and PB-OPT with attacks against Swin Transformer. We evaluate the effects of prior functions, objective mappings, GP acquisition gradients, prior gradients, and GP hyperparameter fitting.

Method	Prior Func.	Obj. Trans.	GP Grad.	Prior Grad.	GP Fit.	Mean ℓ_2 Distortion ↓ 400	800	1200	1600	2000	ASR (%) ↑ 400	800	1200	1600	2000
B-OPT	–	$1/g(\theta)$	✓	×	×	58.79	45.21	37.22	31.59	27.54	3.92	10.85	19.40	26.73	33.47
	–	$-g(\theta)$	✓	×	✓	59.61	45.74	38.55	34.03	31.01	3.02	6.03	10.95	14.97	18.09
	–	$1/g(\theta)$	✓	×	✓	57.36	42.39	35.46	31.42	28.78	4.42	12.76	20.00	27.24	31.06
PB-OPT	Random	$1/g(\theta)$	✓	✓	✓	58.11	42.71	35.61	31.43	28.69	4.02	11.76	18.69	26.13	29.95
	C&W Margin	$-g(\theta)$	✓	✓	✓	44.15	31.72	26.31	23.01	20.72	7.04	14.27	21.91	30.25	37.39
	Standard	$-g(\theta)$	✓	✓	✓	35.14	25.56	21.71	19.28	17.70	17.19	28.24	36.58	42.61	49.25
	Standard	$1/g(\theta)$	×	✓	✓	37.09	31.68	29.68	28.48	27.50	16.28	21.41	23.12	25.53	26.23
	Standard	$1/g(\theta)$	✓	×	✓	57.97	42.44	35.72	31.68	28.92	4.02	12.16	19.60	24.62	29.75
	Standard	$1/g(\theta)$	✓	✓	×	43.74	37.78	35.60	34.02	32.80	10.85	13.57	14.97	15.88	16.68
	Standard	$1/g(\theta)$	✓	✓	✓	**34.54**	**23.95**	**19.78**	**17.23**	**15.75**	**19.60**	**35.08**	**46.33**	**53.07**	**57.79**

Fig. 3 further analyzes the effects of key components. Fig. 3a shows that PB-OPT achieves the best ASR with $d_r = 112 \times 112 \times 3$ and $\ell_0 = 0.02\sqrt{d_r}$, while B-OPT favors $56 \times 56 \times 3$ (Appendix). For fair comparison and efficiency, we use $56 \times 56 \times 3$ in main experiments. Learning γ and c consistently outperforms fixed weights in Fig. 3b. Fig. 3c shows that PB-OPT achieves similar performance across different surrogate models and consistently surpasses B-OPT.

5 Conclusion

In this work, we propose a Bayesian optimization framework for query-efficient hard-label attacks, including B-OPT and PB-OPT. PB-OPT leverages a functional prior with a decoupled forward–backward design for acquisition optimization under non-differentiable hard label. Extensive experiments on CLIP, AT, and MIMIR show state-of-the-art performance.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used to assist with language editing and writing refinement, such as improving grammar, clarity, and presentation, and to assist in the derivation and presentation of the proof of Theorem 3.1. They were not used to generate the research ideas, formulate the proposed methods, conduct experiments, or analyze experimental results. All AI-assisted text and mathematical reasoning were carefully reviewed, verified, and revised by the authors. The authors take full responsibility for the scientific content, correctness, and integrity of this work.

ETHICS STATEMENT

This work studies query-efficient hard-label black-box adversarial attacks with the goal of improving the understanding and evaluation of the robustness of machine learning systems. Although adversarial attack techniques may be misused to evade deployed classifiers, they also provide important tools for identifying model vulnerabilities and developing more robust learning systems. Our experiments are conducted on research datasets and models in controlled experimental settings and do not involve attacks on real-world deployed systems or unauthorized access to external services. We encourage the use of the proposed methods for authorized robustness evaluation, security research, and the development of stronger defenses.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate the reproducibility of this work. The proposed B-OPT and PB-OPT methods are described in detail in Section 3, including the optimization objectives, Gaussian-process modeling, surrogate functional prior, and the forward–backward mechanism. The experimental settings, datasets, target and surrogate models, evaluation metrics, and major hyperparameters are provided in Section 4, with additional implementation details, ablation studies, and theoretical derivations included in the appendix. We plan to publicly release the implementation of B-OPT and PB-OPT, together with the configurations necessary to reproduce the main experiments, upon acceptance of the paper.

REFERENCES

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, pp. 1–12, 2018.

Jianbo Chen, Jordan Michael I., and Wainwright Martin J. HopSkipJumpAttack: a query-efficient decision-based adversarial attack. In *IEEE Symposium on Security and Privacy*, pp. 1277–1294, 2020.

Jinghui Chen and Quanquan Gu. RayS: a ray searching method for hard-label adversarial attack. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1739–1747, 2020.

Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: an optimization-based approach. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlk6iRqKX.

Minhao Cheng, Simranjit Singh, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: a query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SklTQCNtvS.

Shuyu Cheng, Yibo Miao, Yinpeng Dong, Xiao Yang, Xiao-Shan Gao, and Jun Zhu. Efficient black-box adversarial attacks via bayesian optimization guided by a function prior. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Tianyi Ding, Xinjie Xu, Qi Xuan, Hanzhe Yu, and Chen Ma. Query-efficient hard-label attack: A prior-guided adam ray search optimization. *Sensors*, 26(10), 2026. ISSN 1424-8220. doi: 10.3390/s26103272. URL https://www.mdpi.com/1424-8220/26/10/3272.

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5927–5935, 2017.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global Context Vision Transformers. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12633–12646, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. ISSN 1573-2916. doi: 10.1023/A:1008306431147. URL https://doi.org/10.1023/A:1008306431147.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: Query-efficient boundary-based blackbox attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1221–1230, 2020.

Jie Li, Rongrong Ji, Peixian Chen, Baochang Zhang, Xiaopeng Hong, Ruixin Zhang, Shaoxin Li, Jilin Li, Feiyue Huang, and Yongjian Wu. Aha! Adaptive history-driven attack for decision-based black-box models. In *IEEE/CVF International Conference on Computer Vision*, pp. 16168–16177, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Chen Ma, Xiangyu Guo, Li Chen, Jun-Hai Yong, and Yisen Wang. Finding optimal tangent points for reducing distortions of hard-label attacks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19288–19300, 2021.

Chen Ma, Xinjie Xu, Shuyu Cheng, and Qi Xuan. Boosting ray search procedure of hard-label attacks with transfer-based priors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tIBAOcAvn4.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Thibault Maho, Teddy Furon, and Erwan Le Merrer. SurFree: a fast surrogate-free black-box attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10430–10439, 2021.

Jeonghwan Park, Paul Miller, and Niall McLaughlin. Hard-label based small query black-box adversarial attack. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3986–3995, 2024.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: a geometric framework for black-box adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8446–8455, 2020.

Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, and Huaiyu Dai. CGBA: curvature-aware geometric black-box attack. In *IEEE/CVF International Conference on Computer Vision*, pp. 124–133, 2023.

Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467386. URL https://doi.org/10.1145/3447548.3467386.

Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle Attack: a query-efficient decision-based adversarial attack. In *European Conference on Computer Vision*, pp. 156–174, 2022.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Xinjie Xu, Shuyu Cheng, Dongwei Xu, Qi Xuan, and Chen Ma. Improving the convergence rate of ray search optimization for query-efficient hard-label attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11451–11459, Mar. 2026a. doi: 10.1609/aaai.v40i14.38127. URL https://ojs.aaai.org/index.php/AAAI/article/view/38127.

Xinjie Xu, Shuyu Yu, Ziyu Liu, and Stjepan Picek. Mimir: Masked image modeling for mutual information-based adversarial robustness. In *Proceedings of the Network and Distributed System Security Symposium*, San Diego, CA, USA, 2026b. The Internet Society. doi: 10.14722/ndss.2026.241813.

Zhitong Xu, Haitao Wang, Jeff M. Phillips, and Shandian Zhe. Standard gaussian process is all you need for high-dimensional bayesian optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kX8h23UG6v.

Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. Shakedrop regularization for deep residual learning. *IEEE Access*, 7:186126–186136, 2019. doi: 10.1109/ACCESS.2019.2960566.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pp. 87.1–87.12, 2016. doi: 10.5244/C.30.87.

Zhuosheng Zhang, Noor Ahmed, and Shucheng Yu. QE-DBA: Query-efficient decision-based adversarial attacks via bayesian optimization. In *2024 International Conference on Computing, Networking and Communications*, pp. 783–788, 2024. doi: 10.1109/ICNC59896.2024.10555954.

APPENDIX

A THEORETICAL ANALYSIS OF SIMPLIFIED PB-OPT

A.1 PROOF OF THEOREM 3.1

Proof. Define the deterministic approximation error

$$e_L(\theta) \coloneqq \widetilde{f}_L(\theta) - f(\theta). \tag{12}$$

By the assumption in Theorem 3.1,

$$\|e_L\|_\infty = \|\widetilde{f}_L - f\|_\infty \le \delta_L. \tag{13}$$

This error is deterministic; no probabilistic noise model is imposed on e_L. The parameter $\sigma_n^2 > 0$ below is used only to regularize the GP posterior.

We first establish a deterministic confidence bound for the approximated objective $\widetilde{f}_L$. For $t \ge 1$, define

$$\mathbf{k}_{t-1}(\theta) \coloneqq \big(k(\theta_1,\theta),\dots,k(\theta_{t-1},\theta)\big)^\top, \tag{14}$$

$$\mathbf{K}_{t-1} \coloneqq \big[k(\theta_i,\theta_j)\big]_{i,j=1}^{t-1}, \tag{15}$$

$$\mathbf{A}_{t-1} \coloneqq \big(\mathbf{K}_{t-1} + \sigma_n^2\mathbf{I}\big)^{-1}. \tag{16}$$

Let

$$\mathbf{r}_{t-1} \coloneqq \big(r_L(\theta_1),\dots,r_L(\theta_{t-1})\big)^\top. \tag{17}$$

The regularized posterior mean and latent posterior variance can then be written as

$$\mu_{t-1}(\theta) = \hat{f}(\theta) + \mathbf{k}_{t-1}(\theta)^\top \mathbf{A}_{t-1}\mathbf{r}_{t-1}, \tag{18}$$

$$\sigma_{t-1}^2(\theta) = k(\theta,\theta) - \mathbf{k}_{t-1}(\theta)^\top \mathbf{A}_{t-1}\mathbf{k}_{t-1}(\theta). \tag{19}$$

For $t = 1$, these expressions are interpreted using empty vectors and matrices.

Let $\mathcal{S}_{t-1} : \mathcal{H}_k \to \mathbb{R}^{t-1}$ be the evaluation operator

$$\mathcal{S}_{t-1}h = \big(h(\theta_1),\dots,h(\theta_{t-1})\big)^\top. \tag{20}$$

Then $\mathcal{S}_{t-1}r_L = \mathbf{r}_{t-1}$, $\mathcal{S}_{t-1}\mathcal{S}_{t-1}^* = \mathbf{K}_{t-1}$, and

$$\mathcal{S}_{t-1}^*\mathbf{a} = \sum_{i=1}^{t-1} a_i k(\theta_i,\cdot). \tag{21}$$

For any $\theta \in \Theta$, define

$$q_{t-1,\theta} \coloneqq k(\theta,\cdot) - \mathcal{S}_{t-1}^*\mathbf{A}_{t-1}\mathbf{k}_{t-1}(\theta). \tag{22}$$

Using the reproducing property and Eq. (18), we obtain

$$\widetilde{f}_L(\theta) - \mu_{t-1}(\theta) = r_L(\theta) - \mathbf{k}_{t-1}(\theta)^\top \mathbf{A}_{t-1}\mathbf{r}_{t-1} \tag{23}$$

$$= \langle r_L, q_{t-1,\theta}\rangle_k. \tag{24}$$

Moreover, since $\mathbf{A}_{t-1}\mathbf{K}_{t-1} = \mathbf{I} - \sigma_n^2\mathbf{A}_{t-1}$,

$$\|q_{t-1,\theta}\|_k^2 = k(\theta,\theta) - 2\mathbf{k}_{t-1}(\theta)^\top \mathbf{A}_{t-1}\mathbf{k}_{t-1}(\theta) + \mathbf{k}_{t-1}(\theta)^\top \mathbf{A}_{t-1}\mathbf{K}_{t-1}\mathbf{A}_{t-1}\mathbf{k}_{t-1}(\theta) \tag{25}$$

$$= \sigma_{t-1}^2(\theta) - \sigma_n^2\big\|\mathbf{A}_{t-1}\mathbf{k}_{t-1}(\theta)\big\|_2^2 \tag{26}$$

$$\le \sigma_{t-1}^2(\theta). \tag{27}$$

Therefore, Cauchy–Schwarz in $\mathcal{H}_k$ yields the deterministic confidence bound

$$\Big|\widetilde{f}_L(\theta) - \mu_{t-1}(\theta)\Big| \le B_L\sigma_{t-1}(\theta), \qquad \forall\theta \in \Theta. \tag{28}$$

Next, define

$$U_{t-1}(\theta) \coloneqq \mu_{t-1}(\theta) + B_L \sigma_{t-1}(\theta). \tag{29}$$

For any $\varepsilon > 0$, choose $\theta^\varepsilon \in \Theta$ such that

$$\widetilde{f}_L(\theta^\varepsilon) \geq \sup_{\theta \in \Theta} \widetilde{f}_L(\theta) - \varepsilon. \tag{30}$$

By Eq. (28) and the choice of θ_t,

$$\sup_{\theta \in \Theta} \widetilde{f}_L(\theta) - \varepsilon \leq \widetilde{f}_L(\theta^\varepsilon) \tag{31}$$

$$\leq U_{t-1}(\theta^\varepsilon) \tag{32}$$

$$\leq U_{t-1}(\theta_t). \tag{33}$$

Applying Eq. (28) once more at θ_t gives

$$\sup_{\theta \in \Theta} \widetilde{f}_L(\theta) - \widetilde{f}_L(\theta_t) \leq U_{t-1}(\theta_t) - \widetilde{f}_L(\theta_t) + \varepsilon \tag{34}$$

$$\leq 2B_L \sigma_{t-1}(\theta_t) + \varepsilon. \tag{35}$$

Letting $\varepsilon \downarrow 0$, we obtain

$$\sup_{\theta \in \Theta} \widetilde{f}_L(\theta) - \widetilde{f}_L(\theta_t) \leq 2B_L \sigma_{t-1}(\theta_t). \tag{36}$$

We now transfer this bound from the finite-search objective $\widetilde{f}_L$ to the true objective f. By Eq. (13),

$$\sup_{\theta \in \Theta} f(\theta) \leq \sup_{\theta \in \Theta} \widetilde{f}_L(\theta) + \delta_L, \tag{37}$$

$$f(\theta_t) \geq \widetilde{f}_L(\theta_t) - \delta_L. \tag{38}$$

Thus, the instantaneous regret with respect to f satisfies

$$\sup_{\theta \in \Theta} f(\theta) - f(\theta_t) \leq \sup_{\theta \in \Theta} \widetilde{f}_L(\theta) - \widetilde{f}_L(\theta_t) + 2\delta_L \tag{39}$$

$$\leq 2B_L \sigma_{t-1}(\theta_t) + 2\delta_L. \tag{40}$$

It remains to bound the accumulated posterior variances. Since $k(\theta, \theta) \leq 1$, Eq. (19) implies $0 \leq \sigma^2_{t-1}(\theta_t) \leq 1$. For $a \coloneqq \sigma_n^{-2}$, concavity of $s \mapsto \log(1 + as)$ on $[0, 1]$ gives

$$\log(1 + as) \geq s \log(1 + a), \qquad 0 \leq s \leq 1. \tag{41}$$

Consequently,

$$\sigma^2_{t-1}(\theta_t) \leq \frac{\log\big(1 + \sigma_n^{-2}\sigma^2_{t-1}(\theta_t)\big)}{\log(1 + \sigma_n^{-2})}. \tag{42}$$

By the Schur-complement determinant identity,

$$\det\left(\mathbf{I} + \sigma_n^{-2}\mathbf{K}_T\right) = \prod_{t=1}^{T} \left(1 + \sigma_n^{-2}\sigma^2_{t-1}(\theta_t)\right). \tag{43}$$

and hence

$$\frac{1}{2}\log\det\left(\mathbf{I} + \sigma_n^{-2}\mathbf{K}_T\right) = \frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \sigma_n^{-2}\sigma^2_{t-1}(\theta_t)\right). \tag{44}$$

Combining Eqs. (42) and (44), and using the definition of Γ_T, yields

$$\sum_{t=1}^{T} \sigma^2_{t-1}(\theta_t) \leq \frac{1}{\log(1 + \sigma_n^{-2})} \sum_{t=1}^{T} \log\left(1 + \sigma_n^{-2}\sigma^2_{t-1}(\theta_t)\right) \tag{45}$$

$$= \frac{2}{\log(1 + \sigma_n^{-2})} \cdot \frac{1}{2}\log\det\left(\mathbf{I} + \sigma_n^{-2}\mathbf{K}_T\right) \tag{46}$$

$$\leq \frac{2\Gamma_T}{\log(1 + \sigma_n^{-2})}. \tag{47}$$

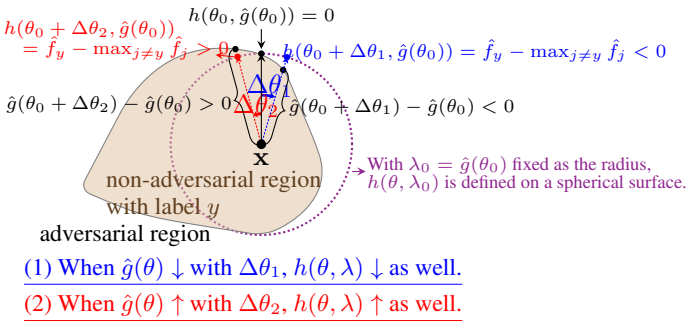


Figure 4: Geometrical explanation of $\nabla\hat{g}(\theta_0) \propto \nabla_\theta h(\theta_0, \lambda_0)$ for an untargeted attack. When $\hat{g}(\theta)$ increases/reduces with a small $\Delta\theta$, $h(\theta, \lambda)$ changes proportionally.

Finally, summing Eq. (40) over t and applying Cauchy–Schwarz gives

$$R_T \leq 2B_L \sum_{t=1}^{T} \sigma_{t-1}(\theta_t) + 2T\delta_L \tag{48}$$

$$\leq 2B_L \sqrt{T \sum_{t=1}^{T} \sigma_{t-1}^2(\theta_t)} + 2T\delta_L \tag{49}$$

$$\leq B_L \sqrt{\frac{8T\Gamma_T}{\log(1 + \sigma_n^{-2})}} + 2T\delta_L. \tag{50}$$

This proves the claimed regret bound. □

A.2 Proof for Surrogate Gradient Estimation

A.2.1 Geometrical Explanation

The first challenge of PB-OPT is how to obtain the gradient $\nabla\hat{g}(\theta)$ from a surrogate model $\hat{\psi}$, where $\hat{g}(\theta)$ represents the shortest distance along the direction θ to the adversarial region of $\hat{\psi}$. This is challenging because $\hat{g}(\theta)$ is obtained through discrete boundary search, which provides no direct differentiable computational path. To address this, for any non-zero vector $\theta_0 \in \mathbb{R}^d$ such that $\hat{g}(\theta_0) < +\infty$, we define a surrogate function $h(\theta, \lambda)$ such that $\nabla\hat{g}(\theta_0) = c \cdot \nabla_\theta h(\theta_0, \lambda_0)$, where $\lambda_0 = \hat{g}(\theta_0)$ is treated as a constant during differentiation. Here, λ is a scalar, and c is a non-zero constant. The surrogate function $h(\theta, \lambda)$ is defined as the negative Carlini & Wagner (C&W) loss function:

$$h(\theta, \lambda) := \begin{cases} \hat{\psi}_y - \max_{j \neq y} \hat{\psi}_j, & \text{if untargeted attack,} \\ \max_{j \neq \hat{y}_{\text{adv}}} \hat{\psi}_j - \hat{\psi}_{\hat{y}_{\text{adv}}}, & \text{if targeted attack,} \end{cases} \tag{51}$$

where $\hat{\psi}_i := \hat{\psi}\big(\mathbf{x} + \lambda \cdot \frac{\theta}{\|\theta\|}\big)_i$ is an abbreviation for the i-th element of the output of $\hat{\psi}$, and $\mathbf{x}$ is the original image. Although $h(\theta, \lambda)$ is defined for arbitrary λ, the implicit-gradient relation is evaluated at the exact surrogate boundary $\lambda_0 = \hat{g}(\theta_0)$, where $h(\theta_0, \lambda_0) = 0$. In practice, PB-OPT uses a finite-search approximation $\widetilde{\lambda}_0$ obtained by parallel boundary probing in the forward pass, and evaluates the substitute gradient at $\widetilde{\lambda}_0$. The geometric explanation of the relation $\nabla\hat{g}(\theta_0) = c \cdot \nabla_\theta h(\theta_0, \lambda_0)$ is presented in Fig. 4. Next, we present the formal proof.[2]

[2]For notational simplicity, we reuse the symbols f, $\hat{f}$, and g in this subsection; their meanings are local to Appendix A.2 and differ from those in Appendix A.1 and the main text of the paper.

A.2.2 PROOF

For a surrogate model $\hat{f}$, we define f to be the negative C&W loss function as follows.

$$f(\mathbf{x}') \coloneqq \begin{cases} \hat{f}(\mathbf{x}')_y - \max_{j \neq y} \hat{f}(\mathbf{x}')_j, & \text{if untargeted attack,} \\ \max_{j \neq \hat{y}_{\text{adv}}} \hat{f}(\mathbf{x}')_j - \hat{f}(\mathbf{x}')_{\hat{y}_{\text{adv}}}, & \text{if targeted attack,} \end{cases} \tag{52}$$

where $\hat{f}(\mathbf{x}')_i$ denotes the i-th element of the output of $\hat{f}(\mathbf{x}')$. Note that $h(\theta, \lambda)$ defined in Eq. (51) equals $f\left(\mathbf{x} + \lambda \cdot \frac{\theta}{\|\theta\|}\right)$. Now we consider $g(\theta)$, the distance from the benign image $\mathbf{x}$ to the adversarial region along the ray direction θ, as defined in Eq. (3) of the main paper. Ignoring exact logit ties and their classifier-dependent tie-breaking, the closure of the adversarial region is characterized by $f(\mathbf{x}') \leq 0$. Hence, the boundary distance can be written as

$$g(\theta) = \inf \left\{ \lambda : \lambda > 0, f\big(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\big) \leq 0 \right\}, \tag{53}$$

where $\inf(\cdot)$ denotes the infimum of a subset of $\mathbb{R}$. We define $g(\theta) = +\infty$ when no valid λ exists, since $\inf \emptyset = +\infty$ by convention. Now we can prove the following proposition.

Proposition A.1. *If f is continuous, $f(\mathbf{x}) > 0$, then given any $\theta \neq \mathbf{0}$ s.t. $g(\theta) < +\infty$, we have $g(\theta) > 0$ and $f\left(\mathbf{x} + g(\theta)\frac{\theta}{\|\theta\|}\right) = 0$.*

Proof. We first prove that $g(\theta) > 0$. We begin by defining the function $f_\theta(\lambda) \coloneqq f\left(\mathbf{x} + \lambda \cdot \frac{\theta}{\|\theta\|}\right)$. Note that f_θ is continuous (w.r.t. λ) because f is continuous. Since $f_\theta(0) = f(\mathbf{x}) > 0$, there exists $\delta > 0$ such that $\forall 0 < \lambda < \delta$, $f_\theta(\lambda) > 0$. Since $g(\theta) < +\infty$, by the definition we have $g(\theta) = \inf(\{\lambda : \lambda > 0, f_\theta(\lambda) \leq 0\}) \geq \delta > 0$.

Next, we want to prove $f\left(\mathbf{x} + g(\theta)\frac{\theta}{\|\theta\|}\right) = 0$, i.e., $f_\theta(g(\theta)) = 0$. To simplify notation in the proof, let us denote $A_\theta \coloneqq \{\lambda : \lambda > 0, f_\theta(\lambda) \leq 0\}$.

If $f_\theta(g(\theta)) < 0$, then since f_θ is continuous and $g(\theta) > 0$, there exists $\epsilon > 0$ such that $f_\theta(g(\theta) - \epsilon) < 0$ and $g(\theta) - \epsilon > 0$. Therefore, we have $g(\theta) - \epsilon \in A_\theta$, which implies that $g(\theta) > \inf(A_\theta)$. This contradicts the definition $g(\theta) = \inf(A_\theta)$.

If $f_\theta(g(\theta)) > 0$, then there exists $\epsilon > 0$ such that $f_\theta(\lambda) > 0$ holds for all $g(\theta) \leq \lambda \leq g(\theta) + \epsilon$. This means that $[g(\theta), g(\theta) + \epsilon] \cap A_\theta = \emptyset$. Noting that $g(\theta)$ is a lower bound of A_θ, this implies that $g(\theta) + \epsilon$ is also a lower bound of A_θ, which contradicts the definition $g(\theta) = \inf(A_\theta)$.

Therefore, $f_\theta(g(\theta)) = 0$. □

Next, we show how to calculate $\nabla g(\theta)$ based on some weak assumptions.

Theorem A.2. *Suppose that f is continuous and $f(\mathbf{x}) > 0$. Define $h(\theta, \lambda) = f\left(\mathbf{x} + \lambda \frac{\theta}{\|\theta\|}\right)$. For any $\theta_0 \neq \mathbf{0}$ such that $g(\theta_0) < +\infty$, let $\lambda_0 = g(\theta_0)$. Assume that h is continuously differentiable in a neighborhood of (θ_0, λ_0) and that $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$. Then g is differentiable at θ_0, and*

$$\nabla g(\theta_0) = -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)} \nabla_\theta h(\theta_0, \lambda_0). \tag{54}$$

Remark A.3. The assumptions above combine standard feasibility conditions with local regularity conditions around the decision-boundary point (θ_0, λ_0). The condition $f(\mathbf{x}) > 0$ means that the original input lies strictly inside the correctly classified region, while $g(\theta_0) < +\infty$ ensures that the ray along θ_0 reaches the adversarial region.

For a piecewise-smooth neural network, the required local differentiability holds when the boundary point lies in the interior of a smooth activation region and the active competing class remains unique in a neighborhood. The condition $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$ requires a nondegenerate, transverse crossing of the decision boundary along the ray. It excludes exceptional cases in which the ray is tangent to the boundary or the margin has a higher-order zero at the crossing point.

Proof. By Proposition A.1, we have $h(\theta_0, \lambda_0) = 0$. By assumption, h is continuously differentiable in a neighborhood of (θ_0, λ_0) and satisfies $\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0) \neq 0$. Therefore, by the Implicit Function Theorem, there exist a neighborhood Θ of θ_0, an open interval $\Lambda = (\lambda_0 - \eta, \lambda_0 + \eta)$, and a unique continuously differentiable function $\widetilde{g} : \Theta \to \Lambda$ such that $\widetilde{g}(\theta_0) = \lambda_0$ and $h(\theta, \widetilde{g}(\theta)) = 0$ for all $\theta \in \Theta$. By shrinking the neighborhoods if necessary, we may assume that $0 < \eta < \lambda_0$. Moreover, the Implicit Function Theorem gives

$$\nabla \widetilde{g}(\theta_0) = -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)} \nabla_\theta h(\theta_0, \lambda_0). \tag{55}$$

It remains to show that g and $\widetilde{g}$ coincide in a neighborhood of θ_0. Since $\widetilde{g}(\theta) \in \Lambda$ and $\lambda_0 - \eta > 0$, we have $\widetilde{g}(\theta) > 0$. Consequently, $h(\theta, \widetilde{g}(\theta)) = 0$ and the definition of g imply that $g(\theta) \leq \widetilde{g}(\theta) < +\infty$ for all $\theta \in \Theta$. By Proposition A.1, we therefore have $h(\theta, g(\theta)) = 0$. The local uniqueness guaranteed by the Implicit Function Theorem implies that if $g(\theta) \in \Lambda$, then $g(\theta) = \widetilde{g}(\theta)$. Since $g(\theta) \leq \widetilde{g}(\theta) < \lambda_0 + \eta$, it remains to establish $g(\theta) > \lambda_0 - \eta$ for all θ sufficiently close to θ_0.

For every $\lambda \in [0, \lambda_0 - \eta]$, we have $h(\theta_0, \lambda) > 0$ because $g(\theta_0) = \lambda_0$. Since $\lambda \mapsto h(\theta_0, \lambda)$ is continuous on the compact interval $[0, \lambda_0 - \eta]$, it attains a strictly positive minimum. Hence, there exists $\epsilon > 0$ such that $h(\theta_0, \lambda) \geq \epsilon$ for all $\lambda \in [0, \lambda_0 - \eta]$.

Choose $r > 0$ sufficiently small such that $\overline{B}(\theta_0, r) \subset \mathbb{R}^d \setminus \{\mathbf{0}\}$. Since f is continuous and $\theta \mapsto \theta / \|\theta\|_2$ is continuous away from $\mathbf{0}$, h is continuous on the compact set $\overline{B}(\theta_0, r) \times [0, \lambda_0 - \eta]$. Therefore, by the Heine–Cantor theorem, h is uniformly continuous on this set. It follows that there exists $\delta \in (0, r)$ such that, whenever $\|\theta - \theta_0\|_2 < \delta$, we have $|h(\theta, \lambda) - h(\theta_0, \lambda)| < \epsilon$ for every $\lambda \in [0, \lambda_0 - \eta]$. Consequently, $h(\theta, \lambda) > 0$ for every $\lambda \in [0, \lambda_0 - \eta]$, which implies $g(\theta) > \lambda_0 - \eta$.

Therefore, for every $\theta \in \Theta \cap B(\theta_0, \delta)$, we have $\lambda_0 - \eta < g(\theta) \leq \widetilde{g}(\theta) < \lambda_0 + \eta$. Hence, $g(\theta) \in \Lambda$, and the local uniqueness guaranteed by the Implicit Function Theorem yields $g(\theta) = \widetilde{g}(\theta)$. Thus, g and $\widetilde{g}$ coincide in a neighborhood of θ_0. Since $\widetilde{g}$ is continuously differentiable, g is differentiable at θ_0. Finally, Eq. (55) yields

$$\nabla g(\theta_0) = -\frac{1}{\frac{\partial h}{\partial \lambda}(\theta_0, \lambda_0)} \nabla_\theta h(\theta_0, \lambda_0). \tag{56}$$

□

B Experiment

B.1 Practical Algorithm Details

In practical implementation, the boundary evaluation (originally defined in Eq. (3)) is performed as

$$\min_{\theta \in \mathbb{R}^d \setminus \{\mathbf{0}\}} g(\theta) \quad \text{where} \quad g(\theta) \coloneqq \inf \left\{ \lambda : \lambda > 0,\ \Phi\left(\Pi_{[0,1]^d} \left(\mathbf{x} + \lambda \frac{\mathcal{U}(\theta)}{\|\mathcal{U}(\theta)\|_p} \right) \right) = 1 \right\}, \quad p \in \{2, \infty\}, \tag{57}$$

where $\mathcal{U}(\theta)$ maps the reduced-dimensional direction θ to the image space via bilinear interpolation, and $\Pi_{[0,1]^d}(\cdot)$ projects the resulting image onto the valid pixel range $[0, 1]^d$. The ℓ_∞-norm version of PB-OPT is shown in Algorithm 2.

B.2 Experimental Setting

In this section, we provide the hyperparameter settings of our approach and compared methods, which include HSJA, TA, G-TA, GeoDA, Sign-OPT, QEBA, QE-DBA, Evolutionary, SurFree, AHA, CGBA-H, SQBA, Prior-Sign-OPT, Prior-OPT, ARS-OPT, and PARS-OPT. All methods are evaluated on the same images and under identical query budgets, while their method-specific hyperparameters follow the corresponding official implementations.

Experimental Equipment. All experiments are carried out with PyTorch 2.0.1 on NVIDIA GPUs, including the RTX 3090, RTX 4090, V100, and A100. While the RTX 4090 and A100 support TensorFloat-32 (TF32) on its Tensor Cores for improved throughput, TF32 can

Algorithm 2 ℓ_∞ Norm of Prior-guided Bayesian OPT (PB-OPT$_{\ell_\infty}$)

Input: benign image $\mathbf{x} \in [0,1]^d$, GP with its length-scale ℓ, calibration parameters γ, c, UCB coefficient β, initial sample size S, number of iterations T, update steps N for acquisition, reduced search dimension d_r, learning rate η, random and historical candidate sizes n_r and n_h, target objective $f(\theta) = 1/g(\theta)$ with boundary-distance function g, and surrogate objective $\hat{f}$.
Output: adversarial example $\mathbf{x}^*$ that satisfies $\Phi(\mathbf{x}^*) = 1$.
Initialize $\mathcal{D} \leftarrow \{(\theta_i, f(\theta_i), \hat{f}(\theta_i))\}_{i=1}^{S}$ by querying f, $\hat{f}$
for $t = S, \ldots, T-1$ **do**
  Compute (μ_f, σ_f) and $(\mu_{\hat{f}}, \sigma_{\hat{f}})$ from $\mathcal{D}$
  Standardize $\{f(\theta_i), \hat{f}(\theta_i)\}_{i=1}^{t}$ using the above statistics
  Update GP by maximizing Eq. (11) ▷ Fitting GP periodically.
  Sample and normalize random directions $\Theta_r \leftarrow \{\tilde{\theta}_i\}_{i=1}^{n_r}, \tilde{\theta}_i \sim U([-1,1]^{d_r})$
  Select recent directions $\Theta_h \leftarrow \{\theta_{t-n_h+1}, \ldots, \theta_t\}$
  $\Theta \leftarrow \Theta_r \cup \Theta_h$ ▷ Candidate directions.
  for $k = 1, \ldots, N$ **do**
    Compute posterior mean $\mu_t(\Theta)$ and variance $\sigma_t^2(\Theta)$
    Obtain surrogate boundary distances $\Lambda \leftarrow \hat{g}(\Theta)$
    Compute acquisition values $\alpha(\Theta) \leftarrow \mathbf{k}_t(\Theta)^\top \mathbf{K}_t^{-1}\big(\mathbf{y}_t - c\mathbf{1} - \gamma\hat{\mathbf{y}}_t\big) + \beta\sigma_t(\Theta) + \gamma\hat{f}(\Theta) + c$
    $\nabla_\Theta \alpha(\Theta) \leftarrow \nabla_\Theta\Big(\mathbf{k}_t(\Theta)^\top \mathbf{K}_t^{-1}\big(\mathbf{y}_t - c\mathbf{1} - \gamma\hat{\mathbf{y}}_t\big) + \beta\sigma_t(\Theta)\Big) - \frac{\gamma\rho}{\Lambda^2} \odot \nabla_\Theta h(\Theta, \Lambda)$
    Update candidates $\Theta \leftarrow \mathrm{Proj}(\Theta + \eta \cdot \mathrm{sign}\,(\nabla_\Theta \alpha(\Theta)))$ ▷ Clamp to $[-1,1]^{d_r}$.
  end for
  Recompute acquisition values $\alpha(\Theta)$ for candidates
  $\theta_{t+1} \leftarrow \arg\max_{\theta \in \Theta} \alpha(\theta)$
  Evaluate $y_{t+1} \leftarrow f(\theta_{t+1})$ and $\hat{y}_{t+1} \leftarrow \hat{f}(\theta_{t+1})$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\theta_{t+1}, y_{t+1}, \hat{y}_{t+1})\}$
end for
$\theta^* \leftarrow \arg\max_{(\theta_i, y_i) \in \mathcal{D}} y_i$ ▷ Best observed direction.
Obtain boundary distance $\lambda^* \leftarrow g(\mathcal{U}(\theta^*))$. ▷ $\mathcal{U}(\theta)$ upsamples reduced-dimensional θ.
Compute normalized direction $\mathbf{v}_{\theta^*} \leftarrow \mathcal{U}(\theta^*)$
return $\mathbf{x}^* \leftarrow \mathbf{x} + \lambda^* \mathbf{v}_{\theta^*}$.

introduce larger numerical errors compared to standard FP32, which may impact gradient-sensitive tasks such as adversarial attacks on ViTs. To ensure consistent numerical accuracy across all hardware platforms, we disable TF32 by setting the following parameters in all experiments: `torch.backends.cuda.matmul.allow_tf32 = False` and `torch.backends.cudnn.allow_tf32 = False`. We adopt GPyTorch version 1.11 for implementing B-OPT and PB-OPT.

CIFAR-10 dataset. For CIFAR-10, we use two wide residual networks with 28 and 40 layers (WRN-28 and WRN-40) (Zagoruyko & Komodakis, 2016) and a 272-layer PyramidNet+ShakeDrop network (PyramidNet-272) (Han et al., 2017; Yamada et al., 2019), and DenseNet-BC ($L = 190, k = 40$) as the target models. ResNet-110 is used as the surrogate model for all surrogate-assisted methods.

ImageNet dataset. For ImageNet, we adopt ResNet-101 (He et al., 2016), ResNeXt-101 (64×4d) (Xie et al., 2017), SENet-154 (Hu et al., 2018), and Vision Transformer (ViT) (Dosovitskiy et al., 2021), Global Context Vision Transformer (GC ViT) (Hatamizadeh et al., 2023), and Swin Transformer (Liu et al., 2021) as target models. ResNet-50 is used as the surrogate model for all surrogate-assisted methods.

Prior-OPT and Prior-Sign-OPT. The hyperparameters of Prior-OPT and Prior-Sign-OPT are summarized in Table 5. For both datasets, we use $q = 200$ directions for gradient estimation and run the optimization for at most 1,000 iterations. The boundary-search tolerance and gradient-clipping threshold are set separately for CIFAR-10 and ImageNet. For a fair comparison, PB-OPT uses the same surrogate architectures as the surrogate-assisted baselines, namely ResNet-110 on CIFAR-10 and ResNet-50 on ImageNet.

Table 5: The hyperparameters of Prior-OPT and Prior-Sign-OPT.

Dataset	Hyperparameter	Value
CIFAR-10	q, total number of vectors for estimating a gradient, including priors and random vectors	200
	the binary search's stopping threshold	$\frac{\beta}{500}$
	the number of iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm for the gradient clipping operation	0.1
ImageNet	q, total number of vectors for estimating a gradient, including priors and random vectors	200
	the binary search's stopping threshold	1×10^{-4}
	the number of iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm for the gradient clipping operation	1.0

Table 6: The hyperparameters of ARS-OPT and PARS-OPT for untargeted ℓ_2-norm attacks.

Dataset	Hyperparameter	Value
CIFAR-10	q, total number of directions used for gradient estimation, including surrogate-prior and random directions	200
	α, the learning rate for updating the ray direction	2.0
	β, the finite-difference step size	1×10^{-3}
	the target boundary-search tolerance during optimization	$\frac{\beta}{500} = 2 \times 10^{-6}$
	the initial boundary-search tolerance	1×10^{-3}
	the surrogate boundary-search tolerance	1×10^{-3}
	the maximum number of optimization iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm used for gradient clipping	0.1
	ϵ, the internal distortion threshold for determining attack success	1.0
ImageNet	q, total number of directions used for gradient estimation, including surrogate-prior and random directions	200
	α, the learning rate for updating the ray direction	0.4
	β, the finite-difference step size	1×10^{-3}
	the target boundary-search tolerance during optimization	1×10^{-4}
	the initial boundary-search tolerance	1×10^{-3}
	the surrogate boundary-search tolerance	1×10^{-3}
	the maximum number of optimization iterations	1,000
	$\mathbf{g}_{\max}$, the maximum gradient norm used for gradient clipping	1.0
	ϵ, the internal distortion threshold for determining attack success	10.0

Table 7: The hyperparameters of Sign-OPT.

Hyperparameter	Value
q, number of queries for estimating an approximate gradient	200
the number of iterations	1,000
the binary search's stopping threshold of the CIFAR-10 dataset	$\frac{\beta}{500}$
the binary search's stopping threshold of the ImageNet dataset	1×10^{-4}

Table 8: The hyperparameters of HSJA, TA and G-TA.

Hyperparameter	Value
γ, threshold of the binary search	1.0
B_0, the initial batch size for gradient estimation	100
$B_{\max}$, the maximum batch size for gradient estimation	10,000
the search method for step size	geometric progression
the number of iterations	64
radius ratio r for attacking ImageNet's models in G-TA	1.1
radius ratio r for attacking CIFAR-10's models in G-TA	1.5

Table 9: The hyperparameters of QEBA.

Hyperparameter	Value
γ, it controls the threshold of the binary search	0.01
`pgen`, the dimension reduction method	resize
B_0, the initial batch size for gradient estimation	100
B_{max}, the maximum batch size of gradient estimation	10,000
the search method for step size	geometric progression
the number of iterations in the CIFAR-10 dataset	150
the number of iterations in the ImageNet dataset	100

Table 10: The hyperparameters of QE-DBA for untargeted ℓ_2-norm attacks.

Setting	Hyperparameter	Value
Both	the structured noise family used to parameterize perturbations	Perlin noise
	the dimensionality of the Bayesian optimization search space	3
	the number of randomly sampled initial BO points	5
	the acquisition function	expected improvement
	the acquisition evaluator	sequential
	the acquisition batch size	1
	the Gaussian-process model	GPyOpt `GPModel`
	whether the objective evaluations are treated as exact	false
	the number of restarts for GP hyperparameter optimization	5
	τ, the stopping tolerance for boundary binary search	0.01
	the initial Perlin perturbation scale	$16/255$
	the search interval of the Perlin wavelength	$[10, 200]$
	the candidate set of the number of Perlin octaves	$\{1, 2, 3, 4\}$
	the search interval of the sinusoidal color frequency	$[4, 32]$
CIFAR-10	ϵ, the internal distortion threshold for determining attack success	1.0
	the maximum target-model query budget	2,000
	the random seed	0
ImageNet	ϵ, the internal distortion threshold for determining attack success	10.0
	the maximum target-model query budget	2,000
	the random seed	0

Table 11: The hyperparameters of Evolutionary.

Hyperparameter	Value
c_{cov}, the hyperparameter of updating the diagonal covariance matrix $\mathbf{C}$	0.001
σ, the deviation for bias	0.03
μ, the mean for bias, and it is a critical hyper-parameter controlling the strength of going towards the original image	0.01
`maxlen`, the maximum length of successful attacks for calculating μ	30

Table 12: The hyperparameters of GeoDA.

Dataset	Hyperparameter	Value
CIFAR-10	subspace dimension, the dimension of 2D DCT basis's subspace	10
	ϵ, the step size of searching the decision boundary	0.5
ImageNet	subspace dimension, the dimension of 2D DCT basis's subspace	75
	ϵ, the step size of searching the decision boundary	5

ARS-OPT and PARS-OPT. The hyperparameters of ARS-OPT and PARS-OPT for untargeted ℓ_2-norm attacks are reported in Table 6. Both methods use $q = 200$ directions for gradient estimation and a finite-difference step size of $\beta = 10^{-3}$. The maximum number of optimization iterations is 1,000, while the initial and surrogate boundary-search tolerances are both set to 10^{-3}. On CIFAR-10,

Table 13: The hyperparameters of SurFree.

Hyperparameter	Value
`BS_gamma`, the stopping threshold in the binary search of α	0.01
`BS_max_iteration`, the maximum iterations in the binary search for α	10
ρ, the parameter for determining θ_{max}	0.98
`T`, the parameter for determining the range of α and the best θ	3
θ_{max}, the parameter for determining the range of α	30
`n_ortho`, the parameter for finding the direction of the lowest ϵ in `_get_candidates`	100
the binary search's stopping threshold of the ImageNet dataset	1×10^{-4}
`frequence_range`, the parameter used in constructing `dct_mask`	$0 \sim 0.5$
`with_distance_line_search`, the parameter used in `_get_candidates`	false
`with_interpolation`, the parameter used in `_get_candidates`	false
`with_alpha_line_search`, the parameter used in `_get_best_theta`	true

Table 14: The hyperparameters of SQBA.

Hyperparameter	Value
`threshold`, the stopping threshold in the binary search	0.001
`min_randoms`, the value indirectly determines the number of queries in each gradient estimation	10

Table 15: The hyperparameters of AHA.

Hyperparameter	Value
`dim_reduce_factor`, the spatial reduction factor for random-direction sampling	4
α, the step size for the source direction	0.01
β, the step size for the spherical direction	0.01

the learning rate, target boundary-search tolerance, and gradient-clipping threshold are set to 2.0, $\beta/500 = 2 \times 10^{-6}$, and 0.1, respectively. On ImageNet, the corresponding values are 0.4, 10^{-4}, and 1.0. The internal distortion thresholds are set to 1.0 for CIFAR-10 and 10.0 for ImageNet. ARS-OPT uses only random directions, whereas PARS-OPT additionally incorporates surrogate-prior directions within the same total direction budget.

Sign-OPT. The hyperparameters of Sign-OPT are listed in Table 7. We use $q = 200$ queries for approximate-gradient estimation and run the optimization for at most 1,000 iterations. The boundary-search tolerance is set to $\beta/500$ on CIFAR-10 and 10^{-4} on ImageNet.

HSJA, TA, and G-TA. The hyperparameters of HSJA, TA, and G-TA are summarized in Table 8. We set the binary-search parameter γ to 1.0, the initial and maximum batch sizes for gradient estimation to $B_0 = 100$ and $B_{\max} = 10{,}000$, respectively, and use geometric progression for step-size search. All three methods are run for at most 64 iterations. TA follows the same basic hyperparameter settings as HSJA and introduces no additional tunable hyperparameters in our implementation. G-TA additionally uses a dataset-specific radius ratio r, which is set to 1.1 for ImageNet and 1.5 for CIFAR-10.

QEBA. The hyperparameters of QEBA are reported in Table 9. QEBA uses the same initial and maximum gradient-estimation batch sizes as HSJA, namely $B_0 = 100$ and $B_{\max} = 10{,}000$, and also adopts geometric progression for step-size search. We set the binary-search parameter γ to 0.01 and use resizing as the dimension-reduction method through `pgen`. The maximum number of iterations is set to 150 on CIFAR-10 and 100 on ImageNet.

Evolutionary. We follow the official implementation of Evolutionary to configure its hyperparameters, as summarized in Table 11. Specifically, the covariance-update coefficient c_{cov}, the standard deviation σ, and the mean parameter μ are set to 0.001, 0.03, and 0.01, respectively. The maximum history length used to update μ is set to 30.

GeoDA. GeoDA is designed for untargeted attacks, and its convergence analysis is established for the ℓ_2-norm setting. We therefore evaluate GeoDA only for untargeted ℓ_2-norm attacks. As shown in

Table 16: The hyperparameters of CGBA-H for untargeted ℓ_2-norm attacks.

Dataset	Hyperparameter	Value
CIFAR-10	ϵ, the distortion threshold used to determine attack success	1.0
	τ, the stopping tolerance for iterative boundary search	0.002
	σ, the perturbation step size used to estimate the decision-boundary normal vector	0.003
	r, the spatial reduction factor defining the retained low-frequency DCT subspace	3
	the retained low-frequency DCT region	$10 \times 10 \times 3$
	N_0, the initial number of directions for normal-vector estimation, with $q_t = \lfloor N_0\sqrt{t} \rfloor$	100
	the maximum number of attack iterations	200
	the maximum query budget	10,000
ImageNet	ϵ, the distortion threshold used to determine attack success	10.0
	τ, the stopping tolerance for iterative boundary search	0.002
	σ, the perturbation step size used to estimate the decision-boundary normal vector	0.003
	r, the spatial reduction factor defining the retained low-frequency DCT subspace	6
	the retained low-frequency DCT region	$37 \times 37 \times 3$
	N_0, the initial number of directions for normal-vector estimation, with $q_t = \lfloor N_0\sqrt{t} \rfloor$	100
	the maximum number of attack iterations	200
	the maximum query budget	10,000

Table 17: The common hyperparameters of B-OPT and PB-OPT.

Norm	Hyperparameter	Value
ℓ_2	the GP covariance kernel	Matérn-$5/2$
	the number of initial random directions	10
	the number of random candidate initializations per BO iteration	10
	the number of recent historical directions per BO iteration	10
	the number of selected directions per BO iteration	1
	d_r, the search-space dimension on CIFAR-10	$32 \times 32 \times 3 = 3{,}072$
	d_r, the reduced search-space dimension on ImageNet	$56 \times 56 \times 3 = 9{,}408$
	the reconstruction method on ImageNet	bilinear interpolation
	the initial GP length-scale	$0.02\sqrt{d_r}$
	the optimizer for GP hyperparameter fitting	Adam
	the learning rate for GP hyperparameter fitting	0.1
	the number of optimization steps for GP hyperparameter fitting	50
	the frequency of GP hyperparameter fitting	every 10 BO iterations
	the learning rate and number of gradient-ascent steps for acquisition maximization	0.1 and 10
ℓ_∞	the GP covariance kernel	Matérn-$5/2$
	the number of initial random directions	10
	the number of random candidate initializations per BO iteration	10
	the number of recent historical directions per BO iteration	10
	the number of selected directions per BO iteration	1
	d_r, the search-space dimension on CIFAR-10	$32 \times 32 \times 3 = 3{,}072$
	d_r, the reduced search-space dimension on ImageNet	$56 \times 56 \times 3 = 9{,}408$
	the reconstruction method on ImageNet	bilinear interpolation
	the initial GP length-scale	$\sqrt{d_r}$
	the optimizer for GP hyperparameter fitting	Adam
	the learning rate for GP hyperparameter fitting	0.1
	the number of optimization steps for GP hyperparameter fitting	50
	the frequency of GP hyperparameter fitting	every 10 BO iterations
	the learning rate and number of gradient-ascent steps for acquisition maximization	0.1 and 30

Table 12, the dimensions of the two-dimensional DCT subspace are set to 10 and 75 on CIFAR-10 and ImageNet, respectively, while the corresponding boundary-search step sizes are set to 0.5 and 5.

QE-DBA. The hyperparameters of QE-DBA are summarized in Table 10. We use Perlin noise to parameterize the perturbation, resulting in a three-dimensional Bayesian optimization space consisting of the wavelength, number of octaves, and sinusoidal color frequency. Their search domains are set to $[10, 200]$, $\{1, 2, 3, 4\}$, and $[4, 32]$, respectively. The optimization is initialized with five randomly sampled points and uses a Gaussian-process model with five hyperparameter-optimization restarts.

Expected improvement is adopted as the acquisition function and is evaluated sequentially with a batch size of one. The boundary-search tolerance is set to 0.01, and the initial Perlin perturbation scale is set to $16/255$. The maximum target-model query budget is 2,000 for both datasets. The internal distortion thresholds used to determine attack success are set to 1.0 on CIFAR-10 and 10.0 on ImageNet.

SurFree. SurFree supports untargeted ℓ_2-norm attacks. We translate its official implementation into PyTorch while preserving the original algorithmic settings. The complete hyperparameters are provided in Table 13, including the binary-search parameters, angular-search parameters, DCT frequency range, and line-search options.

SQBA. SQBA supports untargeted ℓ_2-norm attacks. As shown in Table 14, the stopping tolerance for boundary search is set to 0.001, and `min_randoms`, which indirectly controls the number of queries used in each gradient-estimation step, is set to 10.

CGBA-H. CGBA-H supports both targeted and untargeted attacks, although we use it for untargeted ℓ_2-norm attacks in the main comparison. Its hyperparameters are summarized in Table 16. The boundary-search tolerance and perturbation step size for normal-vector estimation are set to 0.002 and 0.003, respectively. CGBA-H samples random directions from a low-frequency DCT subspace. The spatial reduction factor is set to 3 on CIFAR-10 and 6 on ImageNet, corresponding to retained low-frequency regions of $10 \times 10 \times 3$ and $37 \times 37 \times 3$, respectively. The number of directions used at iteration t is given by $q_t = \lfloor N_0\sqrt{t} \rfloor$, where $N_0 = 100$. The attack is run for at most 200 iterations.

AHA. The hyperparameters of AHA are listed in Table 15. AHA employs dimension reduction to improve the quality and efficiency of gradient estimation. We set `dim_reduce_factor` to 4, while the step sizes for the source and spherical directions, denoted by α and β, are both set to 0.01.

B-OPT and PB-OPT. The common hyperparameters of B-OPT and PB-OPT are summarized in Table 17. Both methods employ a Gaussian process with a Matérn-$5/2$ covariance kernel. At initialization, we randomly generate 10 directions to construct the initial observation set $\mathcal{D}$. At each subsequent BO iteration, acquisition maximization is initialized with 10 newly sampled random directions and 10 directions selected from recent optimization history, and the direction with the largest acquisition value is evaluated on the target model.

The GP hyperparameters are refitted every 10 BO iterations using Adam for 50 optimization steps with a learning rate of 0.1. Acquisition maximization is performed by projected gradient ascent with the same learning rate. We use 10 gradient-ascent steps for ℓ_2-norm attacks and 30 steps for ℓ_∞-norm attacks.

On CIFAR-10, the ray direction is optimized in the original $32 \times 32 \times 3$ search space, corresponding to $d_r = 3{,}072$. On ImageNet, it is optimized in a reduced $56 \times 56 \times 3$ search space, corresponding to $d_r = 9{,}408$, and subsequently reconstructed to the original input resolution using bilinear interpolation. The initial GP length-scale is set to $\ell_0 = 0.02\sqrt{d_r}$ for ℓ_2-norm attacks and $\ell_0 = \sqrt{d_r}$ for ℓ_∞-norm attacks. B-OPT models the target objective using the GP alone, whereas PB-OPT additionally incorporates a differentiable functional prior derived from the surrogate model.

B.3 Additional Experimental Results

The ℓ_2 distortion thresholds used to compute the attack success rate (ASR) are 15.026 for ImageNet and 2.147 for CIFAR-10. They are calculated as $\sqrt{0.0015 \times d}$, where $d = 224 \times 224 \times 3$ for ImageNet and $d = 32 \times 32 \times 3$ for CIFAR-10. A sample is considered successfully attacked if its best distortion within the specified query budget does not exceed the corresponding threshold. In the summary tables, "Mean ℓ_2" and "ASR" denote the final mean distortion and attack success rate at $2{,}000$ queries, respectively. "AUC" denotes the area under the mean ℓ_2 distortion–query curve over 400–2,000 queries. Hence, lower mean ℓ_2 distortion and AUC indicate better performance, whereas a higher ASR is preferred.

Ablation study of the reduced dimension and initial length-scale on B-OPT. Fig. 5 investigates the effects of the reduced search dimension and the initial GP length-scale ℓ_0 on B-OPT. We evaluate four reduced dimensions, namely $14 \times 14 \times 3$, $28 \times 28 \times 3$, $56 \times 56 \times 3$, and $112 \times 112 \times 3$, with two length-scale settings, $\ell_0 = 0.02\sqrt{d_r}$ and $\ell_0 = \sqrt{d_r}$, where d_r denotes the reduced dimension. The results show that the performance of B-OPT is sensitive to both settings. Among the evaluated

configurations, reducing the perturbation direction to $56 \times 56 \times 3$ and setting $\ell_0 = 0.02\sqrt{d_r}$ achieves the highest ASR throughout most of the query range and at 2,000 queries. The $112 \times 112 \times 3$ configuration with the same normalized length-scale also performs competitively, but requires optimization in a substantially larger search space. In contrast, using $\ell_0 = \sqrt{d_r}$ consistently results in substantially lower attack success rates. These observations motivate the use of a moderately reduced search space and the smaller initialization $\ell_0 = 0.02\sqrt{d_r}$ in the main ℓ_2-norm experiments.

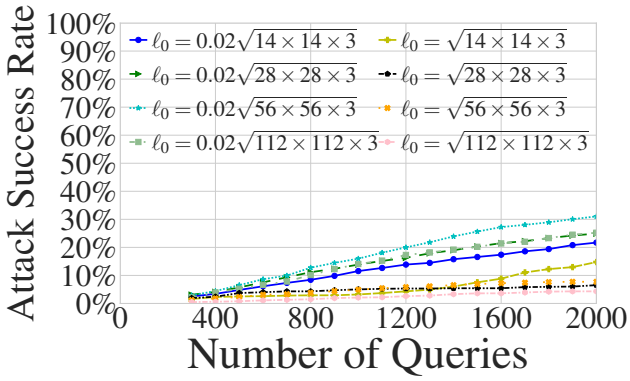


Figure 5: Ablation study of initial length scale and reduced dimension in B-OPT.

Ablation study of different mappings from the reduced-dimensional search space to the image space. Fig. 6 compares different mappings from the reduced-dimensional search space to the image space. Bilinear interpolation and DCT achieve comparable performance for both B-OPT and PB-OPT, while nearest-neighbor interpolation consistently results in larger distortion across query budgets. Moreover, PB-OPT maintains a clear advantage over B-OPT under all three strategies, indicating that its improvement is not tied to a particular upsampling scheme. We therefore adopt bilinear interpolation in the main experiments due to its simplicity and effectiveness.

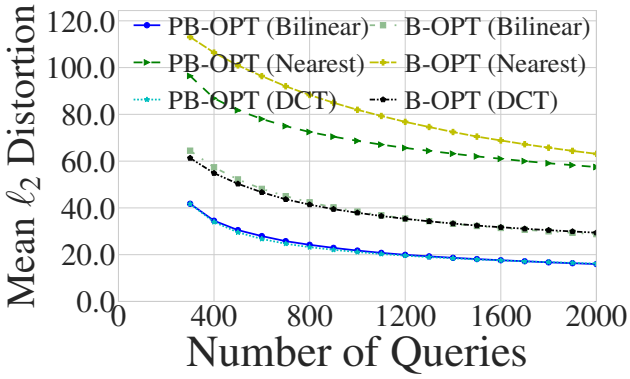


Figure 6: Ablation study of different mappings from the reduced-dimensional search space to the image space for B-OPT and PB-OPT.

Results on ImageNet under different query budgets. Table 18 reports the mean ℓ_2 distortions and ASRs of untargeted attacks under query budgets ranging from 400 to 2,000. The experiments cover ResNet-101, ViT, and GC ViT, thereby including both convolutional and transformer-based target models. B-OPT consistently improves substantially over the original Sign-OPT baseline, demonstrating that Bayesian optimization provides a more query-efficient mechanism for optimizing the ray direction. PB-OPT further improves B-OPT by incorporating the surrogate-derived boundary prior and achieves the strongest overall performance across all three target models.

On ResNet-101, PB-OPT already obtains a mean ℓ_2 distortion of 6.26 and an ASR of 92.8% with only 400 queries. At 2,000 queries, its distortion decreases to 3.19 and its ASR increases to 98.9%,

outperforming strong surrogate-assisted competitors such as SQBA, Prior-OPT, and PARS-OPT. In comparison, B-OPT obtains a distortion of 10.78 and an ASR of 78.1% at the same query budget, indicating that the functional boundary prior provides substantial additional guidance beyond the GP-based search alone.

A similar improvement is observed on ViT. PB-OPT reduces the final mean distortion to 6.00 and achieves a 90.9% ASR, compared with 9.10 and 82.2% for B-OPT. It also slightly improves upon the strongest competing attacks, including CGBA-H and QEBA, whose final distortions are 7.42 and 7.66, respectively. The advantage is particularly pronounced at low query budgets: at 400 queries, PB-OPT achieves a distortion of 13.76 and an ASR of 64.6%, whereas B-OPT reaches only a 33.1% ASR.

GC ViT constitutes a more challenging target, as reflected by the larger distortions and lower success rates of all methods. Nevertheless, PB-OPT maintains a clear advantage. At $2{,}000$ queries, it achieves a mean distortion of 11.48 and an ASR of 75.8%, compared with 19.28 and 51.0% for B-OPT. The improvements over SQBA, Prior-OPT, and PARS-OPT further demonstrate that the proposed boundary prior remains effective when transferred to a target architecture that differs considerably from the surrogate model.

Results on CIFAR-10 under different query budgets. Table 19 reports the budget-dependent results on WRN-40-10 and PyramidNet-272. PB-OPT exhibits particularly strong low-budget performance on both models. On WRN-40-10, it achieves a mean distortion of 0.43 and an ASR of 99.7% with only 400 queries, and reaches 0.28 distortion with a 100% ASR at $2{,}000$ queries. These results are already comparable to, or better than, the final results of competing methods using the complete query budget. For example, SQBA obtains a final distortion of 0.45, while HSJA, TA, and G-TA obtain approximately 0.50.

On PyramidNet-272, PB-OPT achieves a distortion of 1.00 and an ASR of 92.4% at 400 queries. At $2{,}000$ queries, it reaches a distortion of 0.41 and an ASR of 99.9%, outperforming SQBA, which obtains 0.56 and 99.8%. The consistent advantage from 400 to $2{,}000$ queries shows that PB-OPT improves both the early-stage search efficiency and the final perturbation quality.

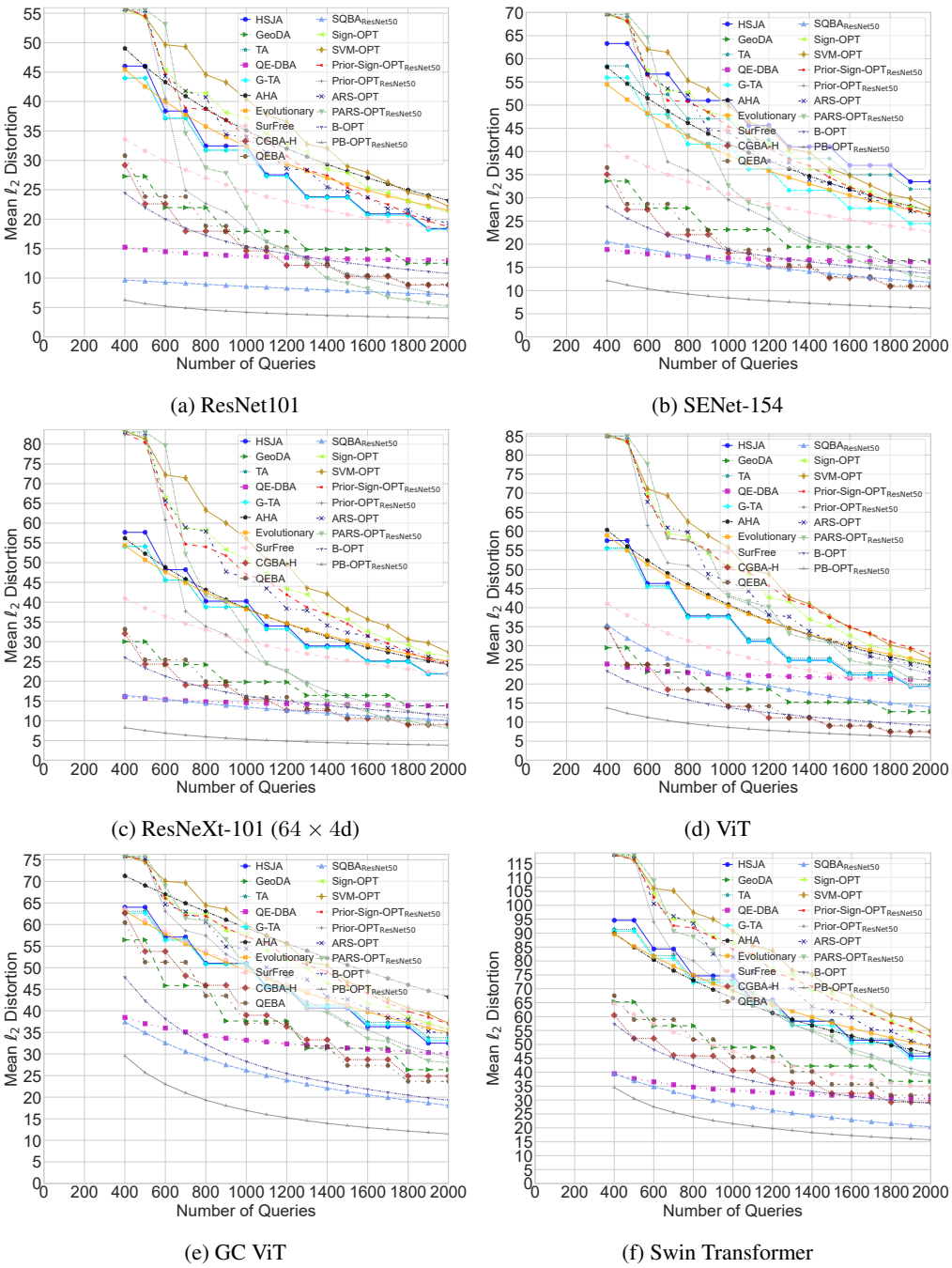
Aggregate CIFAR-10 comparison. Table 20 extends the comparison to PyramidNet-272, WRN-28-10, WRN-40-10, and DenseNet-BC with $L = 190$ and $k = 40$. PB-OPT obtains the lowest final mean distortion and AUC on PyramidNet-272, WRN-28-10, and WRN-40-10. Its final distortions on these three models are 0.414, 0.290, and 0.276, respectively, with ASRs of 99.9%, 100.0%, and 100.0%. The corresponding AUC values are also substantially lower than those of B-OPT and the competing attacks, confirming that PB-OPT converges faster over the entire query range.

DenseNet-BC presents a notable exception. On this target model, SQBA achieves the best results, with a final distortion of 0.386, an AUC of 852.4, and a 100.0% ASR, whereas PB-OPT obtains 1.156, 1982.4, and 97.0%, respectively. This result suggests that the effectiveness of the ResNet-110 boundary prior can depend on the architectural relationship between the surrogate and target models. Nevertheless, the results on the other three target models demonstrate that PB-OPT generally transfers effectively across different architectures. This result also reveals that the effectiveness of functional priors remains partially dependent on the surrogate-target architecture relationship, highlighting the importance of improving surrogate selection and prior transfer robustness.

Convergence curves across additional architectures. Figs. 7 and 8 provide the complete mean-distortion and ASR curves on six ImageNet target models: ResNet-101, SENet-154, ResNeXt-101 (64×4d), ViT, GC ViT, and Swin Transformer. These figures extend the tabulated evaluation to additional CNN architectures and provide a more detailed view of the optimization dynamics.

Across the six targets, PB-OPT generally follows the lowest mean-distortion trajectories and the highest ASR trajectories. Its advantage is already visible at the earliest reported budgets and typically increases as more observations are incorporated into the GP. In contrast, several conventional attacks improve slowly during the early stage and require substantially more queries to approach comparable distortions. The consistent trends across ResNet, SENet, ResNeXt, and multiple transformer architectures demonstrate that the benefit of the differentiable boundary prior is not restricted to a single target-model family.

Overall, the additional experiments show that B-OPT considerably improves the query efficiency of conventional OPT-style attacks, while PB-OPT provides further gains by exploiting transferable functional information from the surrogate decision boundary. PB-OPT achieves the best aggregate results on all evaluated ImageNet target models and on three of the four CIFAR-10 target models.



(a) ResNet101

(b) SENet-154

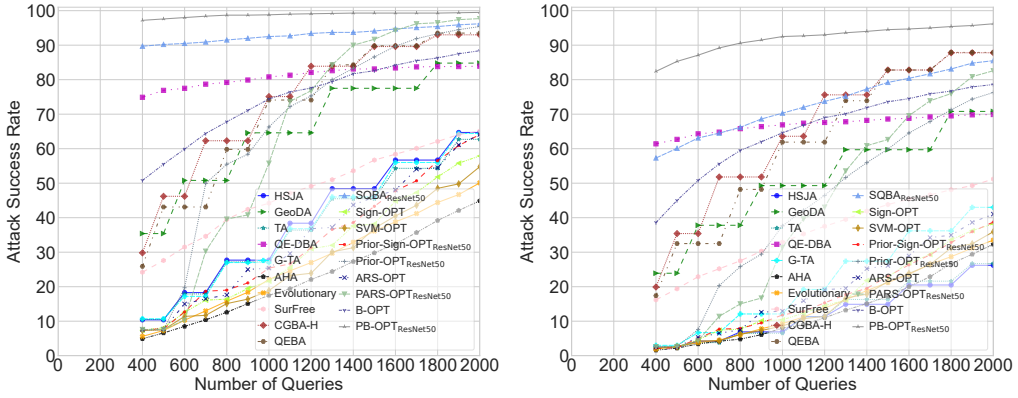
(c) ResNeXt-101 (64 × 4d)

(d) ViT

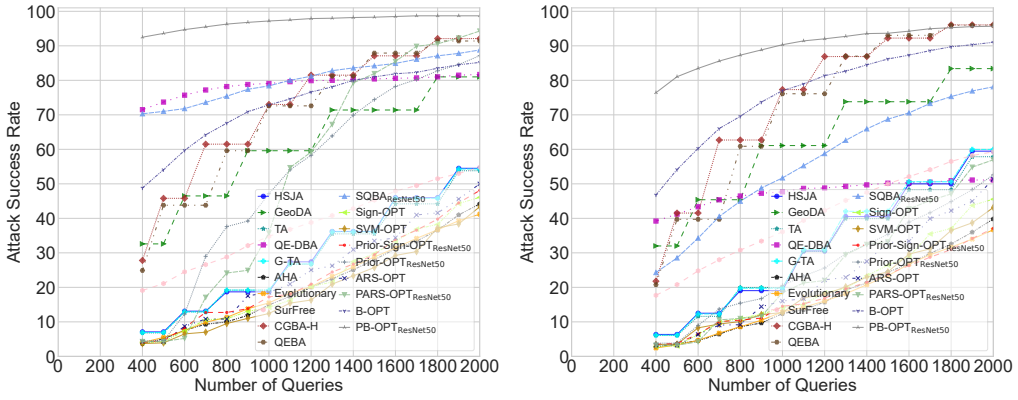
(e) GC ViT

(f) Swin Transformer

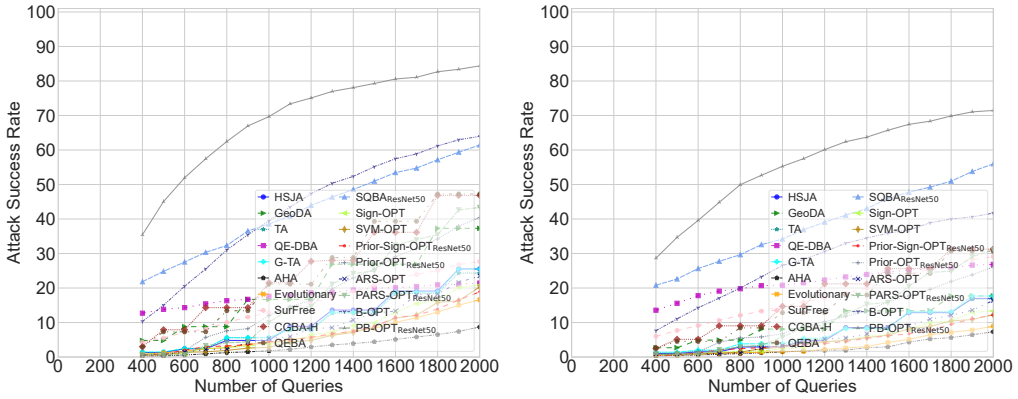
Figure 7: Mean distortions of ℓ_2 norm untargeted attacks under different query budgets on the ImageNet dataset.



(a) ResNet101 (b) SENet-154



(c) ResNeXt-101 (64×4d) (d) ViT



(e) GC ViT (f) Swin Transformer

Figure 8: Attack success rates of ℓ_2 norm untargeted attacks under different query budgets on the ImageNet dataset.

Table 18: Mean ℓ_2 distortions and attack success rates of different query budgets on the ImageNet dataset.

Model	Method	Mean ℓ_2 Distortions					Attack Success Rate				
		400	800	1,200	1,600	2,000	400	800	1,200	1,600	2,000
ResNet-101	HSJA	46.01	32.45	27.55	20.90	18.39	5.9%	17.7%	24.5%	39.5%	47.7%
	GeoDA	27.29	21.98	17.94	14.87	12.53	21.1%	32.8%	45.3%	58.3%	69.4%
	TA	43.97	31.72	27.36	20.92	18.49	5.5%	15.5%	22.9%	39.2%	46.2%
	G-TA	43.99	31.73	27.30	20.66	18.21	5.4%	15.9%	23.1%	40.1%	47.7%
	Sign-OPT	55.49	41.45	32.68	25.27	21.23	4.4%	9.9%	18.1%	29.8%	39.7%
	QEBA	30.81	18.84	15.23	10.47	8.94	15.5%	43.0%	56.4%	78.7%	85.7%
	QE-DBA	15.25	14.05	13.49	13.24	13.05	56.7%	62.0%	64.2%	64.9%	66.0%
	Evolutionary	45.46	35.76	29.42	24.91	21.51	3.0%	7.3%	15.3%	23.8%	32.4%
	SurFree	33.59	27.03	22.98	20.18	18.10	14.8%	24.8%	33.4%	41.5%	49.3%
	AHA	49.02	38.76	31.99	27.02	23.18	2.7%	6.4%	11.8%	18.5%	27.1%
	CGBA-H	29.19	17.97	12.17	10.27	8.80	16.5%	45.6%	69.4%	78.1%	85.1%
	$\text{SQBA}_{\text{resnet50}}$	9.67	8.95	8.27	7.68	7.15	83.2%	84.5%	86.6%	88.9%	90.8%
	$\text{Prior-Sign-OPT}_{\text{resnet50}}$	55.83	38.64	30.74	23.60	18.73	4.2%	11.2%	20.3%	32.0%	46.1%
	$\text{Prior-OPT}_{\text{resnet50}}$	55.44	22.62	14.81	9.88	7.10	4.4%	41.5%	62.6%	81.3%	91.5%
	ARS-OPT	55.64	40.75	28.61	23.73	19.33	4.5%	10.6%	22.1%	33.0%	45.9%
	$\text{PARS-OPT}_{\text{resnet50}}$	55.64	28.70	14.87	8.20	5.17	4.5%	25.2%	64.8%	88.1%	95.7%
	B-OPT (Ours)	24.47	17.24	14.13	12.18	10.78	33.0%	54.9%	64.1%	71.7%	78.1%
	$\text{PB-OPT}_{\text{resnet50}}$ (Ours)	**6.26**	**4.63**	**3.91**	**3.48**	**3.19**	**92.8%**	**96.7%**	**97.9%**	**98.7%**	**98.9%**
ViT	HSJA	57.58	37.79	31.19	22.36	19.38	3.5%	13.1%	18.9%	35.5%	41.7%
	GeoDA	29.48	23.14	18.62	15.24	12.73	20.3%	30.9%	43.2%	56.6%	66.7%
	TA	55.63	37.92	31.74	22.89	19.87	4.1%	12.1%	18.5%	34.2%	41.3%
	G-TA	55.49	37.42	31.10	22.36	19.35	4.1%	12.5%	18.9%	35.9%	44.1%
	Sign-OPT	84.92	58.68	42.66	32.62	25.29	1.9%	7.7%	11.8%	18.3%	31.3%
	QEBA	35.09	18.53	14.19	9.18	7.66	12.7%	43.5%	60.7%	83.7%	89.9%
	QE-DBA	25.21	22.97	22.12	21.55	21.19	23.7%	27.7%	30.1%	32.0%	33.4%
	Evolutionary	58.97	45.29	36.38	30.19	25.71	1.2%	4.4%	9.3%	15.5%	22.9%
	SurFree	41.06	31.34	25.63	21.74	19.02	11.7%	21.6%	29.5%	37.4%	46.2%
	AHA	60.39	46.03	36.46	29.73	24.80	1.1%	4.1%	9.1%	16.1%	24.0%
	CGBA-H	34.81	18.50	11.14	8.97	7.42	13.9%	45.5%	75.3%	84.5%	90.0%
	$\text{SQBA}_{\text{resnet50}}$	35.44	24.89	19.50	16.17	14.00	16.3%	29.7%	43.8%	53.6%	62.5%
	$\text{Prior-Sign-OPT}_{\text{resnet50}}$	85.17	57.61	45.72	34.99	27.95	1.7%	7.7%	11.2%	16.3%	23.9%
	$\text{Prior-OPT}_{\text{resnet50}}$	84.69	51.01	37.93	28.70	22.70	2.2%	10.3%	17.4%	26.5%	37.1%
	ARS-OPT	84.98	59.78	38.12	30.67	23.04	1.9%	5.7%	15.9%	21.5%	35.7%
	$\text{PARS-OPT}_{\text{resnet50}}$	84.98	57.51	40.16	26.15	20.82	1.9%	6.8%	15.6%	29.6%	41.6%
	B-OPT (Ours)	23.38	15.79	12.45	10.50	9.10	33.1%	55.4%	68.0%	76.5%	82.2%
	$\text{PB-OPT}_{\text{resnet50}}$ (Ours)	**13.76**	**9.68**	**7.84**	**6.77**	**6.00**	**64.6%**	**79.0%**	**85.1%**	**88.4%**	**90.9%**
GC ViT	HSJA	64.00	51.02	45.50	36.33	32.51	0.5%	2.6%	5.0%	11.1%	15.1%
	GeoDA	56.44	45.85	37.65	31.33	26.37	2.4%	5.1%	9.1%	15.1%	23.1%
	TA	63.07	50.95	45.89	37.34	33.78	0.3%	3.0%	5.6%	10.6%	15.0%
	G-TA	62.64	50.79	45.58	36.80	33.12	0.6%	2.9%	5.0%	10.9%	15.0%
	Sign-OPT	75.85	62.30	52.05	41.63	35.76	0.3%	1.3%	4.0%	7.0%	12.2%
	QEBA	60.47	43.49	37.11	27.34	23.68	1.8%	7.8%	13.2%	26.4%	33.6%
	QE-DBA	38.47	34.21	32.35	31.11	30.12	6.3%	8.4%	9.2%	10.3%	11.2%
	Evolutionary	63.14	53.35	45.78	39.78	34.96	0.4%	1.5%	3.6%	5.7%	7.8%
	SurFree	63.61	53.91	46.25	40.19	35.73	2.9%	5.5%	8.9%	13.1%	18.1%
	AHA	71.27	63.01	55.63	49.05	43.23	0.1%	0.4%	1.1%	2.7%	4.5%
	CGBA-H	62.63	45.93	36.35	28.71	24.87	1.8%	8.4%	19.4%	25.3%	31.5%
	$\text{SQBA}_{\text{resnet50}}$	37.43	28.97	23.96	20.58	18.08	15.2%	23.0%	30.9%	39.0%	45.5%
	$\text{Prior-Sign-OPT}_{\text{resnet50}}$	75.83	61.94	52.41	44.07	36.95	0.3%	2.2%	3.0%	6.2%	11.2%
	$\text{Prior-OPT}_{\text{resnet50}}$	76.14	57.41	44.24	35.61	29.09	0.3%	4.5%	9.9%	20.3%	28.9%
	ARS-OPT	75.84	60.60	48.00	40.48	34.61	0.3%	1.3%	4.8%	7.9%	14.5%
	$\text{PARS-OPT}_{\text{resnet50}}$	75.84	60.93	45.24	33.48	27.97	0.3%	2.2%	8.4%	22.3%	32.9%
	B-OPT (Ours)	47.78	32.18	25.47	21.84	19.28	4.8%	19.7%	33.5%	43.9%	51.0%
	$\text{PB-OPT}_{\text{resnet50}}$ (Ours)	**29.57**	**19.33**	**15.23**	**12.96**	**11.48**	**23.9%**	**48.3%**	**62.4%**	**71.1%**	**75.8%**

Table 19: Mean ℓ_2 distortions and attack success rates of different query budgets on the CIFAR-10 dataset.

Model	Method	Mean ℓ_2 Distortions					Attack Success Rate				
		400	800	1,200	1,600	2,000	400	800	1,200	1,600	2,000
WRN-40-10-drop	HSJA	1.83	0.97	0.78	0.56	0.50	67.6%	95.4%	98.7%	99.9%	100.0%
	GeoDA	2.71	1.97	1.60	1.43	1.27	51.4%	66.5%	77.5%	84.7%	88.2%
	TA	1.87	0.98	0.78	0.56	0.49	66.5%	94.7%	98.2%	**100.0%**	**100.0%**
	G-TA	1.86	0.97	0.78	0.56	0.50	69.4%	95.0%	98.9%	**100.0%**	**100.0%**
	Sign-OPT	4.67	3.49	2.95	2.34	1.83	38.7%	71.9%	82.5%	90.9%	94.5%
	QEBA	2.11	1.24	0.91	0.82	0.76	58.0%	89.2%	96.5%	98.1%	99.1%
	QE-DBA	2.81	2.38	2.21	2.10	2.02	49.1%	56.9%	60.9%	62.7%	64.3%
	Evolutionary	2.68	1.80	1.34	1.06	0.88	38.5%	68.7%	87.0%	95.0%	98.6%
	SurFree	2.14	1.56	1.25	1.06	0.94	55.6%	77.3%	88.3%	93.4%	96.3%
	AHA	3.70	2.85	2.43	2.21	2.09	20.5%	33.5%	44.1%	51.9%	57.6%
	CGBA-H	2.32	1.47	1.15	1.07	1.01	48.7%	81.0%	90.4%	93.4%	95.0%
	$\text{SQBA}_{\text{ResNet-110}}$	0.92	0.68	0.56	0.50	0.45	96.1%	99.3%	99.9%	100.0%	100.0%
	$\text{Prior-Sign-OPT}_{\text{ResNet-110}}$	4.77	2.92	1.80	1.12	0.80	38.6%	71.8%	83.2%	93.0%	96.7%
	$\text{Prior-OPT}_{\text{ResNet-110}}$	4.20	2.32	1.47	0.84	0.57	38.8%	82.6%	89.5%	96.5%	99.2%
	ARS-OPT	4.62	3.12	1.72	1.00	0.74	39.2%	64.1%	87.6%	93.9%	97.1%
	$\text{PARS-OPT}_{\text{ResNet-110}}$	4.71	3.16	2.10	0.86	0.63	39.2%	70.7%	86.2%	95.1%	99.0%
	B-OPT (Ours)	1.80	1.15	0.89	0.75	0.66	71.5%	90.9%	96.0%	98.4%	99.3%
	$\text{PB-OPT}_{\text{ResNet-110}}$ (Ours)	**0.43**	**0.34**	**0.31**	**0.29**	**0.28**	**99.7%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**
PyramidNet-272	HSJA	3.14	1.67	1.29	0.85	0.71	37.9%	75.7%	86.1%	95.7%	98.0%
	GeoDA	4.10	2.90	2.23	1.85	1.59	30.2%	45.9%	59.5%	70.5%	78.3%
	TA	3.01	1.66	1.28	0.84	0.71	38.2%	75.1%	85.3%	96.1%	98.2%
	G-TA	3.14	1.65	1.26	0.83	0.70	37.0%	74.2%	86.8%	95.7%	98.4%
	Sign-OPT	4.41	2.73	2.08	1.56	1.19	19.3%	46.5%	62.2%	79.1%	89.5%
	QEBA	3.39	2.04	1.40	1.20	1.06	28.1%	62.8%	83.1%	89.2%	93.1%
	QE-DBA	4.19	3.61	3.38	3.24	3.16	24.0%	31.4%	33.3%	35.7%	36.9%
	Evolutionary	3.43	2.26	1.66	1.29	1.06	24.4%	52.9%	76.2%	87.4%	94.0%
	SurFree	2.95	1.99	1.55	1.30	1.13	41.4%	64.9%	78.3%	87.2%	91.3%
	AHA	5.00	4.10	3.56	3.25	3.06	8.9%	14.9%	21.1%	26.0%	29.6%
	CGBA-H	3.64	2.37	1.79	1.61	1.47	23.7%	50.9%	67.7%	74.7%	79.4%
	$\text{SQBA}_{\text{ResNet-110}}$	1.68	1.04	0.79	0.64	0.56	73.8%	92.2%	97.2%	99.2%	99.8%
	$\text{Prior-Sign-OPT}_{\text{ResNet-110}}$	4.38	2.65	2.07	1.52	1.15	18.3%	47.0%	63.9%	79.4%	89.6%
	$\text{Prior-OPT}_{\text{ResNet-110}}$	4.37	2.36	1.85	1.37	1.04	19.8%	58.6%	72.3%	84.4%	92.3%
	ARS-OPT	4.49	3.00	1.76	1.39	1.01	17.3%	39.0%	72.9%	83.5%	93.7%
	$\text{PARS-OPT}_{\text{ResNet-110}}$	4.49	2.91	1.99	1.28	1.03	17.3%	41.5%	66.0%	87.0%	93.2%
	B-OPT (Ours)	2.84	1.81	1.39	1.14	0.98	47.3%	71.2%	83.1%	88.9%	92.5%
	$\text{PB-OPT}_{\text{ResNet-110}}$ (Ours)	**1.00**	**0.63**	**0.52**	**0.46**	**0.41**	**92.4%**	**98.9%**	**99.6%**	**99.8%**	**99.9%**

Table 20: Performance comparison of hard-label attacks on CIFAR-10 against PyramidNet-272, WRN-28-10, WRN-40-10, and DenseNet-BC ($L = 190, k = 40$), evaluated by the final mean ℓ_2 distortion and attack success rate (ASR) at 2,000 queries, together with the AUC of the ℓ_2 distortion–query curve over 400–2,000 queries.

Method	PyramidNet-272			WRN-28-10			WRN-40-10			DenseNet-BC$_{190,40}$		
	Mean $\ell_2 \downarrow$	AUC $\downarrow$	ASR $\uparrow$	Mean $\ell_2 \downarrow$	AUC $\downarrow$	ASR $\uparrow$	Mean $\ell_2 \downarrow$	AUC $\downarrow$	ASR $\uparrow$	Mean $\ell_2 \downarrow$	AUC $\downarrow$	ASR $\uparrow$
HSJA	0.715	2233.7	98.0%	0.508	1383.0	**100.0%**	0.499	1351.6	**100.0%**	0.438	1227.4	**100.0%**
GeoDA	1.593	3725.6	78.3%	1.225	2578.4	89.0%	1.270	2687.9	88.2%	1.148	2602.6	88.5%
TA	0.710	2236.0	98.2%	0.496	1388.8	**100.0%**	0.495	1385.1	**100.0%**	0.433	1251.8	99.9%
G-TA	0.701	2209.1	98.4%	0.498	1357.5	**100.0%**	0.496	1352.9	**100.0%**	0.439	1227.4	99.9%
Sign-OPT	1.194	3611.4	89.5%	0.826	2405.5	97.5%	1.827	4819.0	94.5%	0.743	2154.7	97.9%
QEBA	1.059	2720.3	93.1%	0.742	1742.2	99.1%	0.757	1730.9	99.1%	0.682	1771.6	98.8%
QE-DBA	3.164	5540.3	36.9%	2.077	3766.3	64.2%	2.021	3617.0	64.3%	2.180	3905.0	59.4%
Evolutionary	1.056	2938.6	94.0%	0.898	2425.3	98.7%	0.879	2360.3	98.6%	0.761	2099.5	99.5%
SurFree	1.134	2710.1	91.3%	0.937	2157.4	96.6%	0.938	2137.7	96.3%	0.820	1932.2	97.9%
AHA	3.058	5941.9	29.6%	2.143	4219.8	56.4%	2.088	4116.9	57.6%	1.985	3904.2	62.2%
CGBA-H	1.473	3288.0	79.4%	1.019	2141.0	94.6%	1.008	2100.8	95.0%	0.950	2089.7	95.8%
SQBA$_{\text{ResNet-110}}$	0.556	1403.8	99.8%	0.465	989.7	**100.0%**	0.448	959.6	**100.0%**	**0.386**	**852.4**	**100.0%**
Prior-Sign-OPT$_{\text{ResNet-110}}$	1.151	3547.2	89.6%	0.784	2334.7	98.0%	0.803	3384.3	96.7%	0.693	2089.1	98.8%
Prior-OPT$_{\text{ResNet-110}}$	1.045	3271.5	92.3%	0.589	1926.7	99.3%	0.572	2759.2	99.2%	0.576	1812.4	99.2%
ARS-OPT	1.011	3461.8	93.7%	0.753	2390.2	99.1%	0.744	3355.4	97.1%	0.652	2076.9	99.3%
PARS-OPT$_{\text{ResNet-110}}$	1.031	3520.5	93.2%	0.630	2220.1	99.3%	0.630	3380.8	99.0%	0.609	2008.2	99.0%
B-OPT (Ours)	0.980	2455.5	92.5%	0.669	1590.2	99.3%	0.657	1575.1	99.3%	1.509	2860.6	96.2%
PB-OPT$_{\text{ResNet-110}}$ (Ours)	**0.414**	**901.6**	**99.9%**	**0.290**	**544.6**	**100.0%**	**0.276**	**510.4**	**100.0%**	1.156	1982.4	97.0%