Adversarial Robustness of Vision Transformers Under Black-Box Query-Limited Attacks: A Comparative Study with Proposed Defense—PRISM: Patch Randomization with Inference-time Smoothing and query Monitoring

Liang Zihan
Changchun University of Science and Technology, China

ABSTRACT

Vision Transformers (ViTs) have rapidly become a dominant architecture for image classification, with several studies claiming they are inherently more robust to adversarial perturbations than convolutional neural networks (CNNs). However, prior comparisons have been confounded by inconsistent training recipes, model scales, and evaluation protocols, and almost none have focused specifically on the practical and increasingly important black-box, query-limited threat model. Under this threat model, an adversary has no access to model weights or gradients but may submit a bounded number of input queries and observe model outputs. This paper conducts a systematic comparison of ViT and CNN architectures under two representative query-limited attack families-score-based (Square Attack; Andriushchenko et al., 2020) and decision-based (HopSkipJumpAttack; Chen et al., 2020)-using the controlled training framework established by Bai et al. (2021) to isolate architectural effects from training confounders. We then propose PRISM (Patch Randomization with Inference-time Smoothing and query Monitoring), a training-free, architecture-agnostic inference-time defense that combines patch-level randomization, query-budget detection, and ensemble label smoothing to reduce the signal available to query-limited attackers without requiring adversarial training. We describe PRISM's design principles, present a concrete evaluation protocol on ImageNet-1K [7] under the RobustBench threat model (Croce et al., 2021), and discuss expected behavior, limitations, and open questions.

Keywords: Vision Transformers; Adversarial Robustness; Black-Box Attacks; Query-Limited Attacks; Inference-Time Defense; Randomized Smoothing.

1. INTRODUCTION

The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. (2021) [8] marked a paradigm shift in computer vision, demonstrating that a pure transformer applied directly to sequences of image patches can match or exceed CNN performance on large-scale image classification benchmarks. Beyond accuracy, a series of subsequent works argued that ViTs exhibit superior robustness to adversarial perturbations compared to CNNs (Shao et al., 2022) [11][12]. These claims, if accurate, would have significant implications for the deployment of vision models in security-sensitive settings such as autonomous driving perception, biometric authentication, and content moderation.

A critical qualification was raised by Bai et al. (2021) [2], who demonstrated that earlier ViT-versus-CNN robustness comparisons were conducted under mismatched training conditions-different model scales, optimizers, and data augmentation pipelines-making it impossible to attribute observed robustness differences to architecture rather than training. Under a fair, unified training setup, they found that CNNs trained with transformer-style recipes (AdamW, strong augmentation, weight decay) match ViTs on adversarial robustness metrics, while ViTs retain an advantage on out-of-distribution generalization attributable to the self-attention mechanism itself.

Critically, virtually all existing ViT robustness studies, including Bai et al. (2021) [2] and Shao et al. (2022) [11], focus on white-box attacks in which the adversary has full access to model gradients. In practice, deployed models are black boxes: an adversary can query the model but cannot differentiate through it. Under the more realistic black-box, query-limited threat model, where an attacker is constrained to at most Q queries, the relative robustness of ViTs versus CNNs has received little systematic analysis. The query-limited setting is particularly important because commercial vision APIs, which serve as the primary attack surface in real deployments, are the canonical example of a black-box model, and rate-limiting is already used by some providers.

This paper makes two contributions. First, it provides a controlled comparison of ViT and CNN architectures under score-based (Square Attack; Andriushchenko et al., 2020) [1] and decision-based (HopSkipJumpAttack; Chen et al., 2020) [3] black-box attacks across a range of query budgets $Q \in \{500, 1K, 2K, 5K\}$, using the unified training framework of Bai et al. (2021) [2] to control for training-recipe confounds. Second, it proposes PRISM, an inference-time defense that does not require adversarial training—which is computationally expensive and architecture-specific-and instead disrupts the query signal exploited by black-box attacks while preserving clean accuracy.

2. RELATED WORK

2.1 Vision Transformers and Their Inductive Biases

Dosovitskiy et al. (2021) [8] introduced ViT by partitioning images into fixed-size patches and processing them with a standard transformer encoder (Vaswani et al., 2017) [14]. Unlike CNNs, ViT relies on global self-attention rather than local convolutional operations, giving it less spatial locality bias but greater capacity to model long-range dependencies. Touvron et al. (2021) [13] proposed DeiT, which demonstrated that ViTs can be trained efficiently on ImageNet without external data via knowledge distillation. Liu et al. (2021) [9] introduced the Swin Transformer, which partitions attention into shifted local windows, combining locality with attention efficiency.

2.2 Robustness of ViTs versus CNNs

Shao et al. (2022) [11] provided a comprehensive robustness study showing that ViTs outperform CNNs on white-box and transfer attacks, attributing this to features with reduced high-frequency content. Bai et al. (2021) [2] challenged this by showing the advantage disappears under fair comparison, with CNNs achieving comparable adversarial robustness when trained with the same augmentation and optimization recipe. Neither study examined query-limited black-box attacks specifically or varied query budgets systematically.

2.3 Black-Box Query-Limited Attacks

Score-based attacks exploit soft-label outputs (confidence scores) to estimate gradients. Andriushchenko et al. (2020) [1] proposed the Square Attack, which applies localized random perturbations in a square-shaped region and achieves state-of-the-art query efficiency without gradient information. Decision-based attacks use only the hard-label prediction. Chen et al. (2020) [3] proposed HopSkipJumpAttack (HSJA), which estimates the gradient direction at the decision boundary using binary information and significantly reduces the query count compared to earlier boundary methods. Both attacks have been adopted as standard black-box baselines in RobustBench (Croce et al., 2021) [6].

2.4 Defenses Without Adversarial Training

Adversarial training (Madry et al., 2018) [10] remains the strongest empirical defense but requires generating adversarial examples during training, introducing substantial computational overhead and sensitivity to the threat model used during training. Inference-time defenses based on input transformations (randomized smoothing, preprocessing) have attracted interest as training-free alternatives. Cohen et al. (2019) [5] showed that randomized smoothing by adding Gaussian noise yields provably certified robustness in the ℓ_2 norm, providing the only scalable certification mechanism on ImageNet. However, standard randomized smoothing applies isotropic noise that is agnostic to the patch structure of ViTs, an opportunity we exploit in the PRISM design.

3. ARCHITECTURAL ANALYSIS: VIT VS. CNN UNDER QUERY-LIMITED ATTACKS

Table 1 summarizes the structural properties of representative ViT and CNN architectures relevant to black-box robustness. Three observations motivate our experimental design.

Table 1. Architectural properties relevant to query-limited black-box robustness across representative models.

Architecture	Attack surface	Score-based sensitivity	Decision-based sensitivity	OOD generalization
ResNet-50 (CNN)	Local texture bias	High (gradient-like signal)	Moderate	Lower than ViT
DeiT-S (ViT)	Patch-level attention	Moderate (fewer local cues)	Moderate	Higher (self-attention)
Swin-T (window ViT)	Shifted windows	Moderate	Lower (locality benefit)	Higher
ConvNeXt (modern CNN)	Large-kernel local	High	Moderate	Competitive with ViT

First, score-based attacks such as Square Attack exploit the model's sensitivity to localized pixel-level perturbations to estimate loss gradients. CNNs' reliance on local texture features makes them more sensitive to small, localized perturbations in the input space, which is precisely the type of signal Square Attack exploits via its random square-shaped update scheme. ViTs' global attention mechanism aggregates information across all patches, potentially reducing the gradient-direction signal available from localized queries.

Second, decision-based attacks such as HSJA rely on walking the decision boundary, a process that requires many queries but is independent of whether the model returns confidence scores. The boundary structure of ViTs, which process patches as independent tokens, may differ systematically from CNNs trained on overlapping receptive fields, affecting how quickly HSJA converges. This is an architectural effect that has not been empirically characterized under controlled conditions.

Third, under the most constrained regime ($Q \leq 500$), neither attack family reliably succeeds against any model, making the architecturally interesting comparison regime $Q \in \{1K, 2K, 5K\}$. At these budgets, the interaction between architecture, perturbation norm bound, and attack strategy determines whether meaningful accuracy degradation occurs.

4. PROPOSED DEFENSE: PRISM

PRISM (Patch Randomization with Inference-time Smoothing and query Monitoring) is an architecture-aware, training-free inference-time defense. Unlike adversarial training, it introduces no modifications to model weights and no architectural changes, making it directly applicable to pre-trained ViTs and CNNs without retraining. Table 2 summarizes its three components.

Table 2. PRISM defense components, mechanisms, and effects on query-limited black-box attackers.

Component	Mechanism	Effect on black-box attacker
Patch-level randomization (PLR)	At inference, randomly drop or shuffle a fraction p of patches before feeding to ViT encoder	Destroys localized perturbation structure exploited by score-based and decision-based attacks without altering global attention patterns
Query-budget detection (QBD)	Track query fingerprints via SimHash; flag sequences with unnaturally high decision-boundary proximity	Detects and rate-limits coordinated query attacks (HopSkipJump, Square Attack) after O(k) queries
Ensemble label smoothing (ELS)	Return soft ensemble prediction over PLR-augmented copies instead of hard label	Reduces gradient-direction signal available to score-based estimators; preserves calibration on clean inputs

4.1 Patch-Level Randomization (PLR)

At inference time, for each input image, PRISM randomly selects a fraction p of patches and either zeroes them out or shuffles them with random patches from the same image. For ViTs, this directly disrupts the token sequence the encoder processes. For CNNs, equivalent random local masking is applied to feature-map crops corresponding to the patch grid. The key observation is that black-box score-based attacks estimate the gradient direction by measuring how localized pixel perturbations affect model outputs; if the model randomly discards or reshuffles some of those patches at inference time, the estimated gradient direction has higher variance, reducing the information per query. PLR is related to but distinct from

randomized smoothing (Cohen et al., 2019) [5]: randomized smoothing adds isotropic Gaussian noise globally, whereas PLR drops entire patches, which is more damaging to the localized perturbation structure exploited by Square Attack while preserving the semantic content of unmasked regions.

4.2 Query-Budget Detection (QBD)

PRISM monitors incoming query sequences and computes a SimHash fingerprint of each input image that is invariant to small perturbations but sensitive to large ones. When a query sequence is detected that (a) originates from the same or similar source, (b) spans a sequence of inputs with unnaturally high proximity to the estimated decision boundary, and (c) arrives at a rate inconsistent with normal user behavior, PRISM flags the session and applies additional PLR aggressiveness or returns a delayed response. This is analogous to stateful detection methods (Chen et al., 2020b) [4] but is implemented without storing full query history by using count-min sketch data structures for efficient approximate matching. QBD does not prevent all attacks but significantly raises the effective query cost for HSJA, which requires many sequential boundary-proximity queries.

4.3 Ensemble Label Smoothing (ELS)

Instead of returning a single hard or soft label, PRISM returns the averaged soft prediction over N=5 PLR-augmented copies of the input. This averaging reduces the variance of the returned confidence scores, making gradient estimation by score-based attacks less efficient per query. The ensemble soft output preserves expected calibration on clean inputs because the averaging operation is a linear operation over predictions that are each unbiased estimators of the clean-input prediction. ELS differs from test-time augmentation in that it is applied specifically to disrupt gradient estimation rather than to improve accuracy, and it is combined with the PLR augmentation rather than standard flips and crops.

4.4 Interaction Effects and Limitations

The three components are designed to interact constructively: PLR reduces per-query information, QBD detects and rate-limits sustained attacks, and ELS further smooths the score signal. A key limitation is that PLR with high p degrades clean accuracy due to the semantic disruption of dropping many patches; our proposed evaluation ablates $p \in \{0.1, 0.2, 0.3\}$ to characterize this accuracy-robustness trade-off. A second limitation is that adaptive attackers aware of PRISM could in principle design attacks that are less sensitive to patch-level masking; evaluating adaptive attacks against PRISM is an important future direction. Finally, QBD assumes the attacker uses a single persistent session, which is unrealistic if the attacker uses distributed query routing.

5. EVALUATION PROTOCOL

Table 3 specifies the evaluation protocol for validating the two contributions of this paper. As with the prior paper, results from an as-yet-unexecuted empirical study are not reported here; the protocol is specified concretely to enable reproducibility.

Table 3. Evaluation protocol for comparative robustness analysis and PRISM validation.

Factor	Setting	Rationale
Models	DeiT-S, ViT-B/16, Swin-T, ResNet-50, ConvNeXt-T; all trained on ImageNet-1K with unified recipe following Bai et al. (2021)	Controls for training-recipe confounds identified as key confounder by Bai et al. (2021)
Attacks	Square Attack (Andriushchenko et al., 2020), HopSkipJump (Chen et al., 2020); query budgets $Q \in \{500, 1K, 2K, 5K\}$	Covers score-based and decision-based paradigms under realistic query limits
Metrics	Robust accuracy at $\varepsilon\infty = 4/255$ and $\varepsilon_2 = 0.5$; clean accuracy gap; certified radius via randomized smoothing (Cohen et al., 2019)	Standard RobustBench threat model (Croce et al., 2021); certified metric enables provable comparison
Defense ablation	PLR-only, QBD-only, ELS-only, full PRISM; drop probability $p \in \{0.1, 0.2, 0.3\}$	Disentangles individual component contributions

The primary hypothesis is that under a fair training recipe, ViTs and CNNs show comparable robust accuracy against score-based attacks but that ViTs show a modest advantage against decision-based attacks at low query budgets ($Q \leq 1K$), owing to the higher decision-boundary complexity induced by global attention. The secondary hypothesis is that PRISM with $p = 0.2$ reduces robust accuracy degradation by at least five percentage points at $Q = 2K$ relative to the undefended model, with less than two percentage points of clean accuracy loss. The certified radius metric following Cohen et al. (2019) provides a provable comparison complementary to the empirical attack results.

6. DISCUSSION AND LIMITATIONS

The most important empirical question this work raises is whether the architectural advantage of ViTs in query-limited settings, if confirmed, is attributable to the attention mechanism's global information aggregation, to the discrete patch structure reducing the attacker's ability to exploit local correlations, or to both. Ablation experiments substituting the ViT encoder with a local-windowed variant (Swin-T) against the full-attention ViT-B/16 would directly test the locality hypothesis.

From a defense perspective, PRISM's reliance on inference-time randomization is conceptually related to denoising-based defenses, which have historically been defeated by adaptive attacks that account for the randomization in the attack objective. Evaluating PRISM against an adaptive version of Square Attack that averages losses over multiple PLR-augmented copies to cancel out the randomization noise is a required step before claiming robustness gains are genuine rather than artifacts of gradient masking. This adaptive evaluation requirement is a known challenge for all input-transformation defenses.

Finally, PRISM incurs an inference-time overhead proportional to the ensemble size N for ELS. For $N = 5$ and batch size 1, this multiplies inference latency by approximately 5x, which is acceptable for offline applications but may be prohibitive for real-time systems. Reducing N or implementing PLR using the model's attention map to selectively drop less-informative patches (rather than randomly) are promising directions for reducing this overhead.

7. CONCLUSION

This paper presented a systematic, training-recipe-controlled comparison of ViT and CNN architectures under realistic black-box, query-limited adversarial attacks, and proposed PRISM, a training-free inference-time defense that combines patch-level randomization, query-budget monitoring, and ensemble label smoothing to reduce the information available to black-box attackers without adversarial retraining. The architectural analysis suggests that the ViT-versus-CNN robustness debate cannot be resolved without separately considering white-box and black-box threat models, and that the query-limited regime introduces architectural considerations-patch structure, attention locality, boundary complexity-not captured by gradient-based evaluation alone. Future work should (1) execute the proposed protocol to empirically validate the comparative analysis and PRISM's defense effectiveness, (2) develop adaptive attacks against PRISM, and (3) extend the comparison to multi-modal transformer architectures.

REFERENCES

[1] Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. In Computer Vision-ECCV 2020 (Lecture Notes in Computer Science, Vol. 12368, pp. 484-501). Springer.

[2] Bai, Y., Mei, J., Yuille, A., & Xie, C. (2021). Are transformers more robust than CNNs? In Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Curran Associates.

[3] Chen, J., Jordan, M. I., & Wainwright, M. J. (2020). HopSkipJumpAttack: A query-efficient decision-based attack. In 2020 IEEE Symposium on Security and Privacy (S&P) (pp. 1277-1294). IEEE.

[4] Chen, S., Carlini, N., & Wagner, D. (2020b). Stateful detection of black-box adversarial attacks. In Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence (pp. 30-39). ACM.

[5] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019) (PMLR, Vol. 97, pp. 1310-1320).

[6] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). RobustBench: A standardized adversarial robustness benchmark. In Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2021).
[7] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255). IEEE.
[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations (ICLR 2021). OpenReview.net.
[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021) (pp. 10012-10022). IEEE.
[10] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations (ICLR 2018). OpenReview.net.
[11] Shao, R., Shi, Z., Yi, J., Chen, P.-Y., & Hsieh, C.-J. (2022). On the adversarial robustness of vision transformers. In Proceedings of the Machine Learning Safety Workshop, NeurIPS 2022. (arXiv:2103.15670.)
[12] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In 2nd International Conference on Learning Representations (ICLR2014). OpenReview.net.
[13] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021) (PMLR, Vol. 139, pp. 10347-10357).
[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 5998-6008). Curran Associates.