Diagnose and Defend: Lightweight Behavior-Aware Attention Gating for Robust Vision Transformers

Yanny Gao [1,3], Sara Kothari [1,2], Natalie Kuo [1,4]

[1]Department of Computer Science, [2]Department of Electrical Engineering, [3]Department of Design, Stanford, [4]Department of Art and Art History, Stanford

Introduction & Problem Statement

Vision Transformers (ViTs) achieve strong performance on clean images but are highly vulnerable to adversarial attacks—even small, imperceptible perturbations can cause drastic accuracy drops. This limits their use in security-critical settings.

Most defenses target CNNs; ViT-specific work is limited and often costly (e.g., full-model adversarial training), intrusive (e.g., architecture changes), or non-explanatory (e.g., binary detection). We aim to develop lightweight, interpretable defenses effective across diverse attacks without requiring retraining.

We propose a lightweight, modular defense framework that detects, diagnoses, and mitigates adversarial attacks without modifying the base ViT or requiring retraining. Our approach combines:

Key Contributions

- **Universal adversarial patch** optimized to suppress object detection across diverse images using DETR
- **Multi-attack vulnerability assessment** across twelve attack types (including our novel universal patch)
- **CLS-token based detector** that generalizes to unseen attacks using only attention block 6 output.
- **Unsupervised clustering** to categorize attention disruptions into spatial shifting, attention attenuation, and clean-like patterns
- **Cluster-guided adaptive gating** that applies corrective transformations to attention matrices in blocks 7–12 while preserving clean performance

Dataset & Experimental Setup

Primary Dataset: ImageNet-1k (ILSVRC 2012) with 12,000 training images, 375 validation images, 16,000 test images across 1,000+ object categories. Images resized to 224×224 for ViT input.

Universal Patch Dataset: Open Images V7 subset with 400 images (300 train, 75 val, 25 test), focusing on "Person" class and resized to 640×480 for DETR requirements.

Adversarial Attacks: Training set uses 7 attacks (FGSM, PGD, BIM, C&W, EOTPGD, APGD, APGDT). Testing evaluates on 12 attacks total, including 5 unseen attacks (FAB, VANILA, GN, DeepFool + others) for generalization assessment.

Methods

Attention-Based Defense Architecture

- **Multi-Attack Vulnerability Assessment:** We benchmark ViT robustness against 12 adversarial attacks from the `torchattacks` library (FGSM, PGD, CW, etc.), as well as our own universal patch attack, quantifying accuracy drops and identifying differential sensitivity across attack types.
- **CLS-Token Based Adversarial Detector:** Using only the CLS token embedding from attention block 6, we train a lightweight MLP classifier to distinguish clean from adversarial inputs: $p_{\text{adv}} = \text{MLPdetector}(\mathbf{h}\text{cls}^{(6)})$ Block 6 balances noise sensitivity with semantic awareness. Trained on a subset of attacks, it generalizes well to previously unseen ones.
- **Unsupervised Attention Diagnoser:** For detected adversarial inputs, we apply PCA followed by K-means clustering on internal attention statistics to categorize disruptions into three types: (1) *spatial shifting*, (2) *attention attenuation*, and (3) *clean-like miscellaneous* patterns. A separate MLP predicts cluster membership for real-time diagnosis.
- **Cluster-Guided Adaptive Gating:** Each disruption type is mapped to a small MLP gate network that applies a corrective transformation to attention matrices in blocks 7–12: $\mathbf{A}\text{gated} = \mathbf{G}c_i(\mathbf{A})$ where c_i is the predicted cluster ID. Clean inputs bypass this mechanism entirely, ensuring no performance degradation on benign examples.

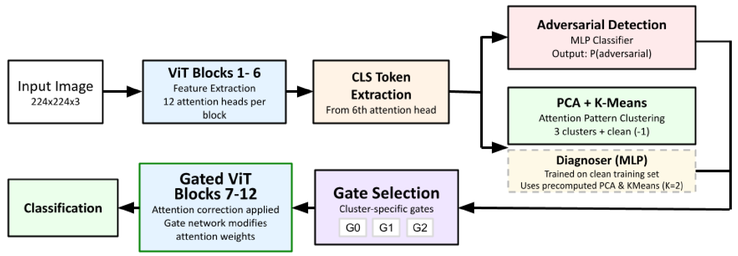

Figure 1. Adversarial Defense Architecture for Vision Transformers

Universal Adversarial Patch We construct a universal patch $\mathcal{P}$ optimized to suppress object detection across diverse images using DETR ResNet-50:

- **Training:** Patch undergoes random transformations (translation, rotation $\pm 30°$, scaling $0.5\times$–$1.5\times$)
- **Loss Function:**

$$\mathcal{L} = \alpha \cdot \left(-\text{KL}\left(\text{softmax}(\mathbf{y}_{\text{clean}}) \parallel \log \text{softmax}(\mathbf{y}_{\text{patched}})\right)\right) + \beta \cdot \left(-\text{ShiftDistance}(\mathbf{b}_{\text{clean}}, \mathbf{b}_{\text{patched}})\right)$$

  - Maximize shifts in bounding boxes & differences in predictions.
- **Combined Objective:** KL divergence + spatial displacement to disrupt both classification and localization

Results and Discussion

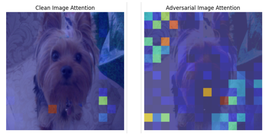

Figure 2. Attention map before and after adversarial perturbation.

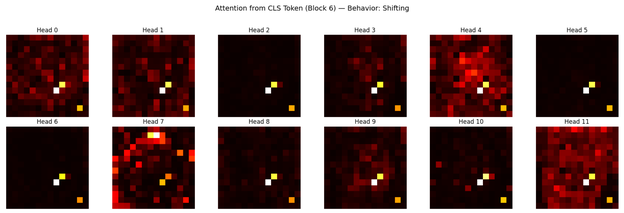

Figure 3. Block 6 CLS-to-patch attention maps

Figure 2 shows that clean image attention is tightly focused on the dog's face, while the adversarial version disperses attention across irrelevant background areas—revealing **semantic drift** in the final block's CLS-to-patch attention. Figure 3 displays head-wise attention from Block 6 for the same adversarial input. The scattered and inconsistent activation patterns, especially in Heads 4 and 7, support the diagnoser's classification of this case as a *shifting* behavior. Together, these visualizations demonstrate how adversarial noise disrupts both global focus and per-head alignment in ViT.

- As shown in Table 1 below, our **CLS-token-based MLP classifier achieves 93.30% training and 87.63% validation accuracy**. This supports the effectiveness of CLS embeddings for adversarial detection and confirms that adversarial inputs are linearly separable from clean samples in embedding space.
- The diagnoser MLP, trained on 9K adversarial examples with cluster-based pseudo-labels, achieves 86.95% training and **99.75%** validation accuracy. This suggests that attention disruption modes are linearly separable in CLS embedding space.
- Table 2 shows that our **defense improves classification accuracy across all attacks**, with post-defense accuracy exceeding 98% in every case. Gains range from +39.1% (FGSM) to +43.4% (GN), demonstrating strong generalization. However, clean accuracy drops by 98.2%, indicating the detector may be overly aggressive on benign inputs.

Task	Train Acc	Val Acc
ViT + Classifier	93.30%	87.63%

Table 1. Defense Results: Classifier Performance

Attack Type	Baseline Accuracy	Post-Defense Accuracy	% Change
CLEAN	0.8713	0.0156	-98.2%
FGSM	0.7089	0.9862	+39.1%
PGD	0.6942	0.9853	+41.9%
BIM	0.6887	0.9853	+43.1%
CW	0.7016	0.9835	+40.2%
EOTPGD	0.7043	0.9853	+39.9%
APGD	0.7016	0.9835	+40.2%
APGDT	0.7016	0.9835	+40.2%
FAB	0.7016	0.9835	+40.2%
VANILA	0.7016	0.9835	+40.2%
GN	0.6961	0.9982	+43.4%
DEEPFOOL	0.7016	0.9835	+40.2%

Table 2. Comparison of ImageClassification Accuracy Before and After Defense (ViT)

- Table 3 shows that, although overall detection confidence remained stable, certain adversarial patches—particularly the checkerboard and universal designs—induced noticeable spatial shifts in object localization. Among the basic patterns, the **checkerboard patch was most effective**, likely due to its high-frequency contrast disrupting convolutional filters and misleading feature extraction, especially in DETR's attention-based architecture. The **universal patch caused slightly larger average shifts** (0.415 vs. 0.39), suggesting marginally greater disruption, with a minimally higher variance (0.0427 vs. 0.038) that may reflect slightly reduced consistency. These shifts often appeared far from the patch itself, indicating that DETR's **global self-attention propagates local perturbations**. This localization instability, despite preserved detection confidence, highlights potential for stronger disruption through improved patch optimization.

Patch Type	 Noise	 Black	 Checker	 Universal
Averages	0.11	0.13	0.39	0.415

Table 3. Shift Distances for 25 Images. Patch examples shown above column headers.

Conclusion and Future

This work proposes a lightweight defense for Vision Transformers using cluster-aware adaptive gating, achieving 30-43% accuracy gains on adversarial examples without costly full-model retraining. The method generalizes to unseen attacks via unsupervised behavior discovery and uses selective attention gating for efficiency. Despite improved robustness, there's a trade-off with clean performance, motivating future work on more precise detection, alternative conditioning, and broader ViT integration.

sarako@stanford.edu, rgao1218@stanford.edu, nskuo@stanford.edu