# DINOmite: Adversarial Robustness of DINOv3 Vision Transformers

Wojciech Aszkiełowicz* Hubert Berlicki* Jan Burdzicki* Igor Jakus*

w.aszkielowicz@proton.me berlickihubert@gmail.com janburdzicki@gmail.com igorjakus@protonmail.com

Institute of Computer Science, University of Wrocław

*Equal contribution

## Abstract

**Problem:** DINOv3 Vision Transformers achieve state-of-the-art accuracy but remain vulnerable to adversarial attacks.
**Contribution:** We present a comprehensive evaluation framework for DINOv3, testing robustness against gradient-based and optimization-based attacks. We compare standard fine-tuning against advanced adversarial defense strategies.
**Key Results:**

- Standard models: **85% clean** vs **25% robust** (PGD).
- TRADES defense: **60% robust accuracy**.

## Motivation

### Why It Matters

- **Safety-Critical Systems:** Autonomous driving, medical imaging.
- **Physical Threats:** Adversarial patches work in reality.

### Research Questions

1. Does self-supervised DINOv3 offer inherent robustness?
2. Which defense strategy minimizes the accuracy trade-off?

## Problem Definition

We formulate the attack as a constrained optimization problem:

$$\max_{\delta} L(f(x+\delta), y) \quad \text{s.t.} \quad \|\delta\|_\infty \le \epsilon$$

### Threat Model

- **White-box:** Full access to gradients.
- **Constraint:** $\epsilon = 8/255$.
- **Goal:** Untargeted misclassification.

## Experimental Setup

### Model Architecture

- **Backbone:** DINOv3 ViT-S/16 (Frozen).
- **Head:** Linear Classification Head (Trainable).

### Datasets

- **CIFAR-10:** Primary benchmark.
- **GTSRB:** Traffic signs (Safety-critical).

### Evaluation

- **Metric:** Clean vs. Robust Accuracy (PGD-10).
- **Epsilon:** $\epsilon \in \{0, \ldots, 8/255\}$.

## Conclusions

### Main Takeaways

- **Vulnerability:** DINOv3 has no inherent robustness against gradient attacks.
- **Defense:** TRADES successfully recovers 60% accuracy under strong PGD attacks.
- **Trade-off:** The robustness gain justifies the small drop in clean accuracy.

### Future Work

- Evaluating Certified Robustness.
- Scaling to ImageNet-1k.

## Attack Methods

We focus on three primary attack vectors representing different threat levels:

### FGSM (Fast Gradient Sign Method)

A single-step attack assuming linear loss surface.

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$$

*Note: Fast, but weak against iterative training.*

### PGD (Projected Gradient Descent)

The "Universal First-Order Adversary". Iterative FGSM with projection.

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \alpha \cdot \text{sign}(\nabla_x L(\theta, x^t, y)))$$

*Note: The strongest defense benchmark. We use 10 steps.*

### C&W (Carlini & Wagner)

Optimization-based attack minimizing distance and loss term.

$$\min_{\delta} \|\delta\|_2 + c \cdot f(x+\delta)$$

*Note: Finds adversarial examples with minimal visible perturbation.*

## Adversarial Examples

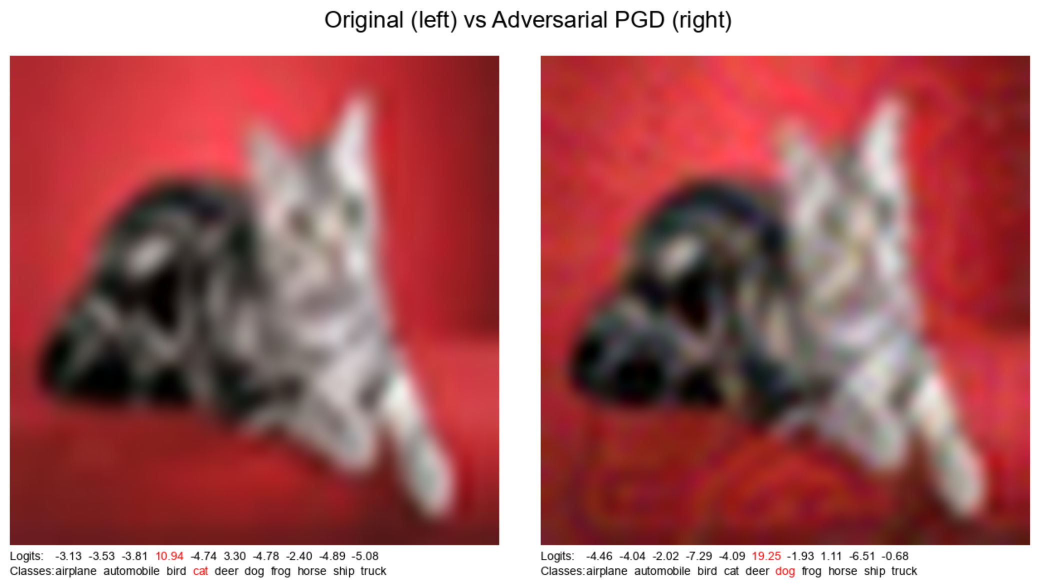
**C&W** ($L_2$) - Minimal Perturbation

Original (left) vs Adversarial (right)

**FGSM** ($\epsilon = 8/255$) - Patterned Noise

**PGD** ($\epsilon = 8/255$) - High Noise

Original (left) vs Adversarial PGD (right)



## Defense Strategies

Standard training yields 0% robustness. We compare:

### PGD-AT (Adversarial Training)

Min-Max game training on PGD examples.

$$\min_{\theta} \mathbb{E}[\max_{\delta \in S} L(f(x+\delta), y)]$$

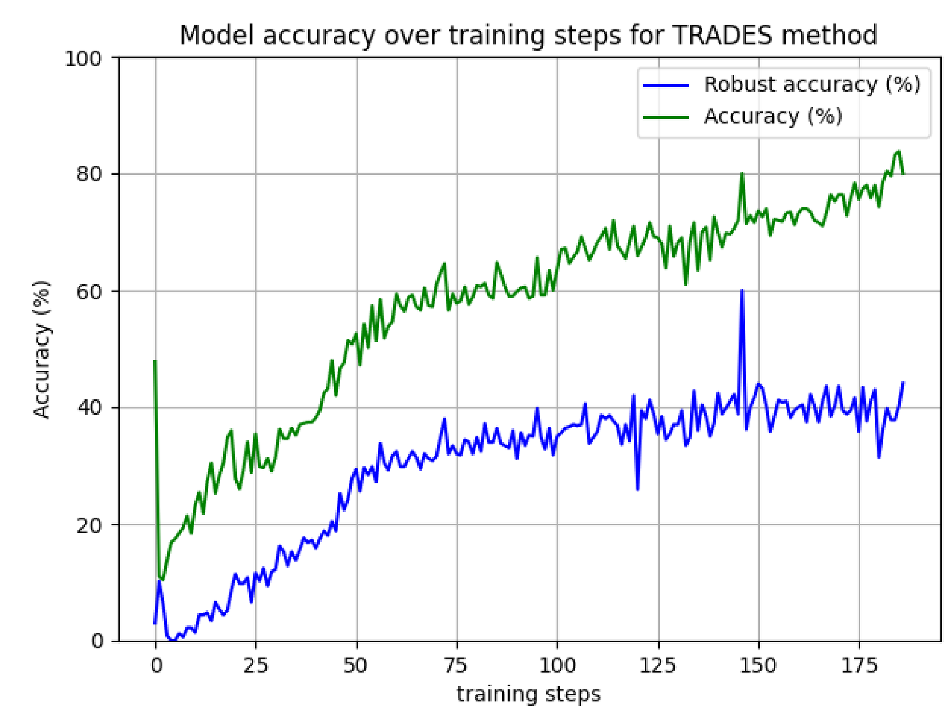
### TRADES

Separates clean accuracy and stability (KL-divergence).

$$\min_{\theta} \mathbb{E}[L(f(x), y) + \beta \cdot \text{KL}(f(x) \| f(x+\delta_{adv}))]$$

*Result: Smoother decision boundaries.*

## Defense Results



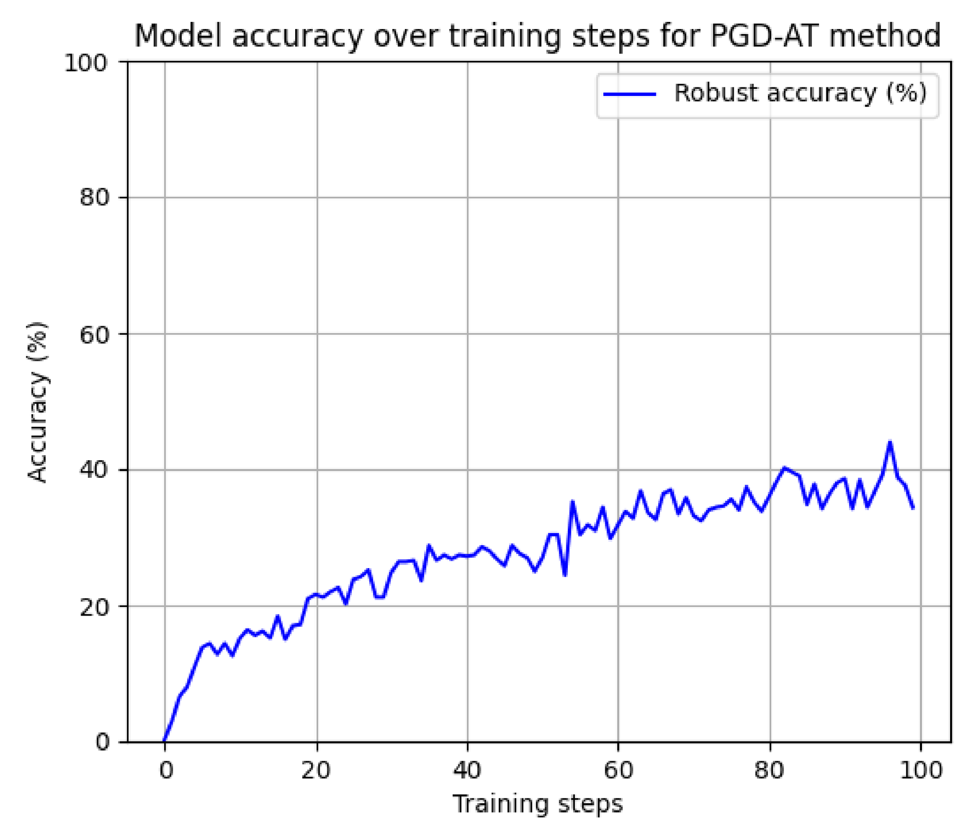




**Figure 1:** Robustness and accuracy across defense methods.

## Quantitative Results

Comparison on CIFAR-10 ($\epsilon = 8/255$).

| Model | Clean | FGSM | PGD |
|---|---|---|---|
| Standard | **85.0%** | 45.0% | 25.0% |
| PGD-AT | 82.0% | 65.0% | 55.0% |
| TRADES | 81.5% | **68.0%** | **60.0%** |

## References

1. Goodfellow et al. (2015). *Explaining and Harnessing Adversarial Examples.* arXiv:1412.6572
2. Moosavi-Dezfooli et al. (2016). *DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks.* arXiv:1511.04599
3. Madry et al. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks.* arXiv:1706.06083
4. Carlini & Wagner (2017). *Towards Evaluating the Robustness of Neural Networks.* arXiv:1608.04644
5. Zhang et al. (2019). *Theoretically Principled Trade-off between Robustness and Accuracy.* arXiv:1901.08573
6. Kurakin et al. (2017). *Adversarial Examples in the Physical World.* arXiv:1607.02533
7. Wang et al. (2020). *Improving Adversarial Robustness Requires Revisiting Data Augmentation and Training.* OpenReview: rklOg6EFwS