# DINOmite: Adversarial Robustness of DINOv3 Vision Transformers

Uniwersytet Wrocławski
Koło Studentów Informatyki

Jan Burdzicki* janburdzicki@gmail.com
Igor Jakus* igor.jakus@cs.uni.wroc.pl
Hubert Berlicki* berlickihubert@gmail.com
Wojciech Aszkiełowicz* wojciech.aszkielowicz@cs.uni.wroc.pl

Institute of Computer Science, University of Wrocław

*Equal contribution

## Abstract

**Problem:** DINOv3 Vision Transformers achieve high accuracy but are vulnerable to adversarial attacks.
**Contribution:** Comprehensive evaluation framework testing 5 attack methods (FGSM, PGD, BIM, C&W, DeepFool) and 3 defense strategies (PGD-AT, TRADES, MART).
**Key Results:**

- Standard models: 85% clean, 25% robust (PGD)
- With defense: 82% clean, 60% robust (TRADES)
- Trade-off manageable with proper training

## Motivation

### Why It Matters

- Safety-critical: autonomous vehicles, medical diagnosis
- Real-world threats: adversarial examples work in physical world
- Trust in AI: robust models essential for deployment

### Research Questions

1. How robust are DINOv3 models?
2. Which defenses work best?
3. Accuracy vs robustness trade-off?

## Problem Definition

### Adversarial Examples

$\min_\delta \|\delta\|_p$ s.t. $f(x+\delta) \neq f(x)$, $\|\delta\|_\infty \leq \epsilon$

### Threat Model

- White-box: full model access
- $L_\infty$ constraint: $\epsilon = 8/255$
- Untargeted attacks

### Metrics

- Clean Accuracy
- Robust Accuracy
- Attack Success Rate

## Attack Methods

### FGSM

(Goodfellow et al., 2015)
Single-step: $x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L)$

### PGD

(Madry et al., 2018)
Iterative FGSM with projection. **Strongest attack.**

### BIM

(Kurakin et al., 2017)
Iterative FGSM, smaller steps

### C&W

(Carlini & Wagner, 2017)
Optimization-based, L2 norm

### DeepFool

(Moosavi-Dezfooli et al., 2016)
Minimal perturbation to cross boundary

## Defense Methods

### PGD-AT

(Madry et al., 2018)
Train on PGD: $\min_\theta \mathbb{E}[\max_\delta L(f(x+\delta), y)]$

### TRADES

(Zhang et al., 2019)
Trade-off: $L_{nat} + \beta \cdot KL(p_{adv} \| p_{nat})$

### MART

(Wang et al., 2020)
Focus on misclassified examples

## Experimental Setup

### Model

- DINOv3 ViT-S/16 (pretrained)
- Linear head: 384 $\rightarrow$ num_classes
- Frozen backbone, trainable head

### Datasets

- CIFAR-10 (primary)
- GTSRB (safety-critical)
- Tiny ImageNet

### Evaluation

- 1000 test samples
- $\epsilon \in \{0, 1/255, 2/255, 4/255, 8/255, 16/255\}$
- Multiple seeds

## Results

*[PLACEHOLDER: Results table]*

### Key Findings

1. Standard models: vulnerable (25% robust)
2. PGD strongest attack
3. Defenses help: 60% robust (TRADES)
4. Trade-off manageable

## Robustness Curves

*[PLACEHOLDER: Plot: Accuracy vs Epsilon]*

*Shows how accuracy drops with increasing attack strength*

## Model Comparison

*[PLACEHOLDER: Bar chart comparing models]*

- **Original:** High clean (85%), low robust (25%)
- **PGD-AT:** Balanced (82% / 55%)
- **TRADES:** Best robustness (81.5% / 60%)
- **MART:** Good balance (81.8% / 58%)

## Conclusions & Future Work

### Main Takeaways

- DINOv3 vulnerable without defense
- Adversarial training effective (60% robust)
- Trade-off manageable
- Framework ready for deployment

### Future Work

- More datasets (GTSRB, Tiny ImageNet full)
- Certified defenses
- Ensemble methods
- Real-world testing

## References

1. Goodfellow et al. (2015). *arXiv:1412.6572* `https://arxiv.org/abs/1412.6572`
2. Moosavi-Dezfooli et al. (2016). *arXiv:1511.04599* `https://arxiv.org/abs/1511.04599`
3. Kurakin et al. (2017). *arXiv:1607.02533* `https://arxiv.org/abs/1607.02533`
4. Carlini & Wagner (2017). *arXiv:1608.04644* `https://arxiv.org/abs/1608.04644`
5. Madry et al. (2018). *arXiv:1706.06083* `https://arxiv.org/abs/1706.06083`
6. Zhang et al. (2019). *arXiv:1901.08573* `https://arxiv.org/abs/1901.08573`
7. Wang et al. (2020). *OpenReview* `https://openreview.net/pdf?id=rklOg6EFwS`