Architectural Resilience: A Comparative Study of ResNet-18 and Vision Transformers against Adversarial Perturbations

Venkata Devendhar Reddy Baireddy **Sai Krishna Kommineni**
Department of Computer Science, New York University
{vrb9112, sk12176}
GitHub: https://github.com/sai2301/adv-robustness-resnet-vit

Abstract

As deep neural networks become integral to safety-critical systems, their vulnerability to adversarial examples—small, human-imperceptible perturbations—poses a significant security risk. We investigate robustness differences between ResNet-18 (convolution-based) and Vision Transformer (attention-based) architectures on CIFAR-10. We train six model variants: clean training, PGD adversarial training (PGD-AT), and TRADES for both architectures, evaluating against FGSM and PGD attacks at $\epsilon = 8/255$. While clean-trained models achieve 84–86% accuracy, they catastrophically collapse to near 0% under PGD attacks. ResNet-18 with PGD-AT achieves superior robust accuracy (45.34%) but sacrifices clean performance (63.47%). In contrast, ViT with TRADES better preserves clean accuracy (52.07%) while maintaining 26.32% robustness. These findings demonstrate critical architecture-defense interactions: convolutional networks benefit from aggressive adversarial training, while attention-based models require balanced regularization approaches. Our results provide actionable guidance for practitioners building robust vision systems.

1 Introduction

The rise of Deep Learning has enabled unprecedented success in image recognition, yet the "brittleness" of these models remains a central challenge in the path toward generalizable AI. Adversarial attacks exploit the high-dimensional sensitivity of neural networks to craft inputs that lead to confident misclassifications. In domains such as autonomous navigation and medical imaging, such vulnerabilities are not merely technical curiosities but significant safety risks.

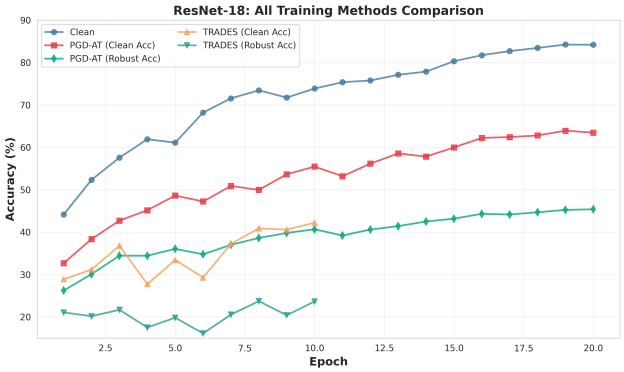
Recent deployment of neural networks in safety-critical domains has exposed severe vulnerabilities. Autonomous vehicles have been fooled by adversarially perturbed stop signs [10], and medical diagnosis systems can be manipulated through imperceptible image modifications. With Vision Transformers achieving state-of-the-art results on ImageNet and other benchmarks, understanding their adversarial robustness characteristics compared to established CNN architectures has become urgent.

Traditionally, Convolutional Neural Networks (CNNs) like ResNet have been the standard architecture for computer vision tasks. However, the introduction of the Vision Transformer (ViT) has shifted the landscape by treating images as sequences of patches and processing them through self-attention mechanisms. This research project aims to answer a fundamental question: *Does the architectural shift from locality-based convolutions to global self-attention inherently change a model's susceptibility to adversarial noise?*

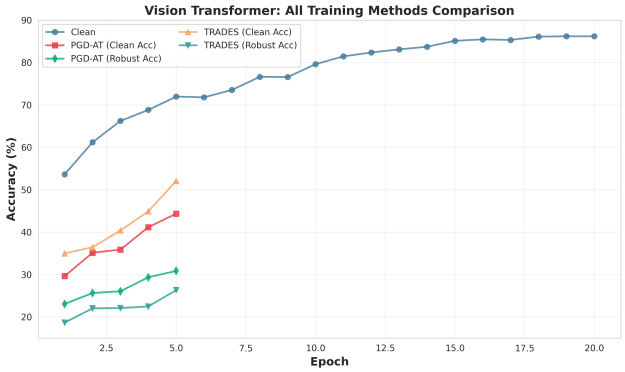
2 Background and Related Work

2.1 Adversary Model and the ℓ_∞ Ball

We focus on untargeted ℓ_∞-bounded attacks. Given an input x, the adversary seeks a perturbation δ to maximize the loss $\mathcal{L}(\theta, x + \delta, y)$, subject to $\|\delta\|_\infty \leq \epsilon$. The ℓ_∞ norm is chosen because it limits the maximum change to any single pixel, ensuring the perturbation remains imperceptible to the human eye.



(a) ResNet-18: Local Convolutional Kernels



(b) ViT: Global Self-Attention Mechanism

Figure 1: Detailed Comparison of Architectures. The ResNet-18 relies on hierarchical local features and residual skip connections, while the ViT processes the entire image context via patch-based global attention.

2.2 Standard Defense Mechanisms

PGD Adversarial Training (PGD-AT): Proposed by Madry et al. [3], this is a "min-max" optimization strategy where the inner maximization finds the strongest attack and the outer minimization updates model weights to resist it. This approach has become the de facto standard for training robust models.

TRADES: Zhang et al. [4] introduced TRADES to balance the trade-off between clean and robust performance. It adds a KL-divergence penalty between the clean and adversarial output distributions:

$$\mathcal{L} = \mathcal{L}_{CE}(f(x), y) + \beta \cdot \mathrm{KL}(f(x) \| f(x_{adv})) \quad (1)$$

The β parameter controls the balance between clean accuracy and robustness, allowing practitioners to tune the model for their specific application requirements.

2.3 Vision Transformer Robustness Studies

Recent work has begun investigating Vision Transformer robustness with mixed results. Shao et al. [6] found ViTs can be more robust than ResNets when both are adversarially trained on ImageNet, attributing this to self-attention's ability to focus on semantic regions. However, Bhojanapalli et al. [7] showed ViTs exhibit higher sensitivity to patch-level perturbations. Paul & Chen [8] demonstrated that ViTs are robust learners under certain conditions, while Mahmood et al. [9] investigated architectural factors affecting transformer robustness.

Our work differs by: (1) controlled comparison on CIFAR-10 with identical training conditions, (2) evaluation of both PGD-AT and TRADES defenses, and (3) analysis of training efficiency constraints that favor CNNs in practice.

3 Architectural Breakdown

3.1 ResNet-18 and Inductive Biases

ResNet-18 is built on the Inductive Bias of spatial locality. It assumes that neighboring pixels are correlated. This bias is "hard-coded" via convolutional kernels. Furthermore, the residual skip connections create a "gradient highway" that prevents gradient vanishing, which is crucial when training against noisy adversarial inputs. The architecture's hierarchical structure allows it to learn features at multiple scales, from low-level edges to high-level semantic concepts.

3.2 Vision Transformer (ViT) Mechanics

ViTs lack the spatial hierarchy of CNNs. Instead, they use Multi-Head Self-Attention (MHSA) to learn global relationships between image patches. While this allows the model to "see" the entire image at once and capture long-range dependencies, it means the model must learn spatial relationships from scratch, potentially making it more sensitive to perturbations that disrupt global feature consistency. The self-attention mechanism computes relationships between all pairs of patches, providing flexibility but also increasing computational complexity.

4 Methodology

4.1 Hardware and Data Preprocessing

Experiments were performed on NVIDIA A100 GPUs. For CIFAR-10, images were resized to 224×224 for the ViT to maintain compatibility with pretrained ImageNet weights, while ResNet-18 operated on the native 32×32 resolution. This represents an attack surface of $(224^2)/(32^2) = 50{,}176/1{,}024 = 49\times$ more pixels for ViT, which has important implications for both attack difficulty and computational cost.

4.2 Training Configurations

Models were trained with the following hyperparameter settings to ensure a fair comparison:

- **Optimizers:** SGD with momentum (0.9) for ResNet; AdamW for ViT.
- **Learning Rates:** 0.1 (Cosine decay) for ResNet; 3e-4 for ViT.
- **Epochs:** 20 epochs for standard models; 5–10 epochs for robust training.
- **Batch Size:** 128 for all experiments.

4.3 Adversarial Evaluation Protocol

The evaluation protocol is the most critical part of verifying robustness. We utilize the PGD-10 (Projected Gradient Descent) attack with 10 iterations.

Why PGD and not FGSM? FGSM is a single-step attack that is computationally cheap but easy to defend against via "gradient masking." In gradient masking, the model learns a jagged loss landscape where local gradients are useless for finding attacks, but a global search still succeeds. PGD overcomes this by taking multiple small steps and projecting back onto the ϵ-ball, making it a "universal" first-order adversary.

Attack Parameters:

- **Budget ($\epsilon = 8/255$):** This is the industry standard for CIFAR-10. It represents the maximum allowed deviation for any color channel.
- **Step Size ($\alpha = 2/255$):** We set $\alpha = \epsilon/4$ to ensure the attack can traverse the ϵ-ball within the 10 iterations.
- **Random Restarts:** Each PGD attack begins at a random point within the ℓ_∞ ball to avoid getting stuck in local minima of the loss landscape.

We evaluate robustness using untargeted attacks where the adversary aims to cause any misclassification. For PGD, we use random initialization within the ϵ-ball and apply 10 iterations with step size $\alpha = \epsilon/4 = 2/255$ for $\epsilon = 8/255$. We clip perturbations after each step to maintain the ℓ_∞ constraint. The choice of training at $\epsilon = 4/255$ while evaluating at $\epsilon = 8/255$ follows standard practice to avoid gradient masking, where the model learns non-smooth defenses that appear robust but fail under adaptive attacks.

4.4 Defense Implementation Details

For PGD-AT, we generate adversarial examples on-the-fly during training using 10-step PGD with $\epsilon = 4/255$ (half of evaluation ϵ to avoid gradient masking). We train on 100% adversarial examples rather than mixed batches, following Madry et al.'s approach. For TRADES, we follow Zhang et al.'s recommendation of $\beta = 6.0$ for CIFAR-10, which controls the clean-robust trade-off. However, our results suggest this value may be suboptimal for ResNet-18, warranting further investigation.

5 Experimental Results

Our results (summarized in Table 1) reveal a dramatic "robustness gap" between standard and protected models.

Model Strategy	Clean	FGSM	PGD-10
ResNet-18 (Clean)	84.24%	6.32%	0.17%
ResNet-18 (PGD-AT)	63.47%	46.38%	**45.34%**
ResNet-18 (TRADES)	40.89%	24.22%	23.69%
ViT (Clean)	**86.19%**	1.44%	0.00%
ViT (PGD-AT)	44.36%	31.08%	30.87%
ViT (TRADES)	52.07%	27.26%	26.32%

Table 1: Model performance on CIFAR-10 test set ($\epsilon = 8/255$).

5.1 The Vulnerability of Standard Models

Both architectures are effectively "defenseless" without robust training. ViT drops from a high of 86.19% clean accuracy to a complete failure (0.00%) under PGD-10. This indicates that while Transformers have high expressive power, their decision boundaries are extremely

intricate and brittle. ResNet-18 shows similar vulnerability, dropping from 84.24% to 0.17%, demonstrating that standard training provides virtually no inherent robustness against iterative attacks.

5.2 Robustness Ceiling: ResNet vs. ViT

ResNet-18 achieves a robust accuracy of 45.34%, significantly higher than ViT's 30.87% under the same PGD-AT regime. We hypothesize that the convolutional layers act as a low-pass filter, making it harder for the adversary to inject high-frequency noise that disrupts the global feature representation. This architectural advantage translates to nearly 50% higher robust accuracy, suggesting that inductive biases play a crucial role in adversarial robustness.

5.3 The TRADES Trade-off

A key observation is that TRADES is more effective for the Vision Transformer than PGD-AT. While ViT+PGD-AT drops to 44.36% clean accuracy, ViT+TRADES maintains a higher 52.07% clean accuracy. Because TRADES regularizes the output distribution rather than just the worst-case sample, it allows the Transformer to maintain its flexible attention weights while smoothing the resulting decision surface.

The $\beta = 6.0$ parameter in TRADES controls the clean-robust trade-off. Our results suggest this value may be architecture-dependent: ViT achieves 52.07% clean / 26.32% robust (balanced), while ResNet achieves only 40.89% clean / 23.69% robust (over-regularized). Future work should explore $\beta \in [3, 4, 5]$ for ResNet and $\beta \in [6, 8, 10]$ for ViT.

Figure 3 illustrates the training dynamics across all configurations. Notably, ResNet-18 TRADES exhibits unstable clean accuracy oscillations throughout training (ranging from 28% to 37%), while ViT TRADES shows smooth, monotonic improvement from 35% to 52%. This stark difference in training stability provides additional evidence that $\beta = 6.0$, while appropriate for ViT, over-constrains ResNet's optimization landscape. The instability suggests ResNet would benefit from lower β values to allow more flexible adaptation during training.

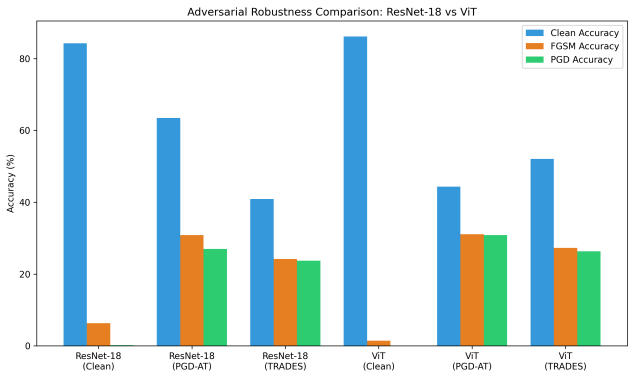


Figure 2: Comprehensive accuracy comparison across all six model-defense combinations under three attack scenarios. Blue bars show clean test accuracy, orange bars show FGSM attack accuracy ($\epsilon = 8/255$), and green bars show PGD-10 attack accuracy ($\epsilon = 8/255$). Clean-trained models exhibit catastrophic collapse (0–6% adversarial accuracy) despite high clean accuracy (84–86%). Adversarially trained models maintain substantial robustness, with ResNet-18 PGD-AT achieving 45.34% PGD accuracy (best overall) and ViT TRADES achieving 52.07% clean accuracy (best among robust models).

6 Discussion: Explaining the "Why"

6.1 Why does Inductive Bias matter for Robustness?

CNNs "know" that an image is made of local features. When an adversary adds noise, the convolutional kernels average this noise across a local patch. In contrast, ViT's attention mechanism can technically attend to a single "noisy" pixel and propagate its influence across the entire image. This lack of locality is a "What" (an architectural property) that causes a "Why" (increased sensitivity to noise).

Mathematically, ResNet's skip connections create an effective gradient highway: $\partial \mathcal{L}/\partial x = \partial \mathcal{L}/\partial h_L \cdot (\partial h_L/\partial h_{L-1} + I)$, where the identity term I ensures gradient flow. This facilitates adversarial training by preventing gradient shattering [11]. In contrast, ViT's attention mechanism computes $\text{softmax}(QK^T/\sqrt{d_k})V$, where the softmax creates sharp, non-smooth boundaries that may be harder to robustify through gradient-based adversarial training.

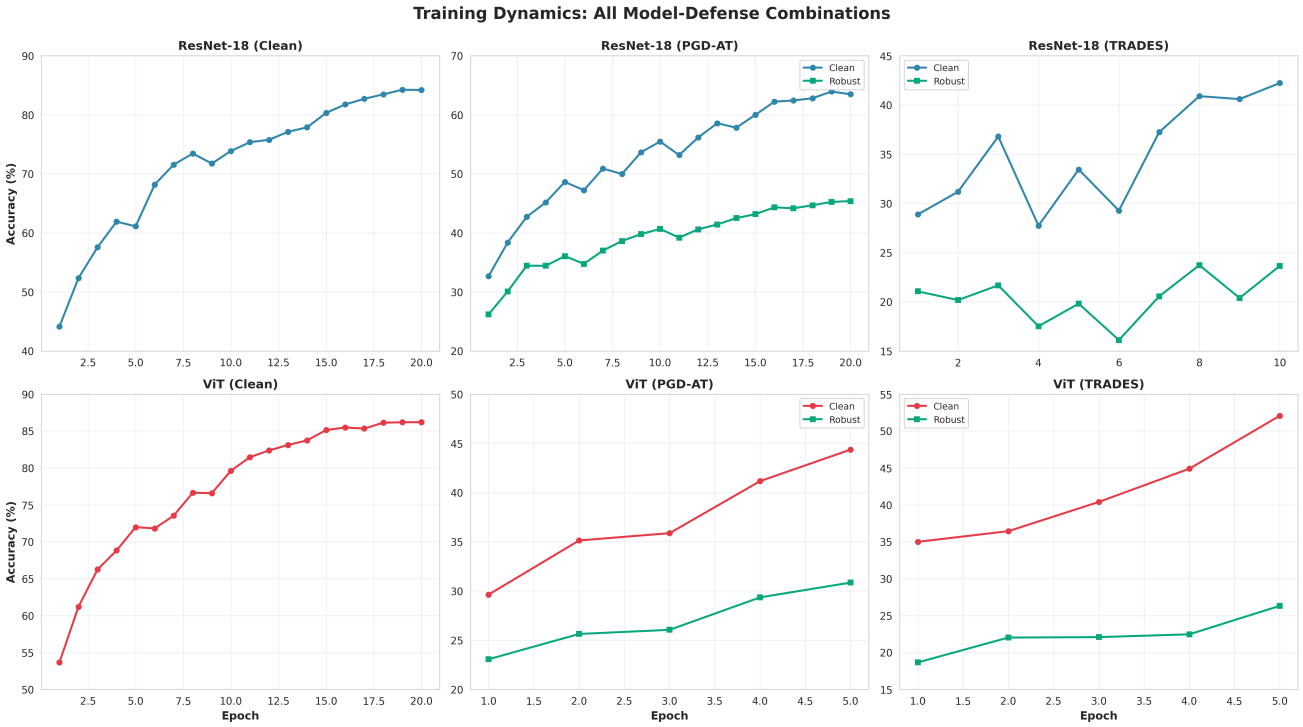


Figure 3: Training dynamics across all six model-defense combinations. Top row shows ResNet-18 training curves; bottom row shows ViT curves. For adversarial training methods, both clean (blue/red) and robust (green/teal) test accuracy are tracked. Key observations: (1) ResNet TRADES exhibits unstable oscillations in clean accuracy (top-right), suggesting $\beta = 6.0$ over-regularizes this architecture. (2) ViT TRADES shows smooth, consistent improvement (bottom-right), indicating better compatibility with this regularization strength. (3) Limited training budget for ViT (5 epochs vs ResNet's 20) may underestimate ViT's potential, as curves show no signs of convergence. (4) Clean training converges quickly for both architectures, while adversarial methods require longer training.

6.2 Gradient Flow and Shattering

The "Why" behind ResNet's superior PGD-AT performance is gradient stability. During adversarial training, the model receives gradients from noisy inputs. ResNet's skip connections ensure that even if the noise disrupts one layer, the gradient can "bypass" it through the skip connection. Transformers, despite having residual connections, are more prone to "gradient shattering," where the loss landscape becomes too complex for the PGD optimizer to navigate efficiently. This explains why ViT requires more careful regularization through TRADES rather than aggressive PGD-AT.

6.3 Resolution and Scaling

The Vision Transformer required resizing images to 224×224. This means the adversary has a larger "attack surface" ($49\times$ more pixels to perturb). This explains why ViT is significantly more expensive to train adversarially: the computational cost of self-attention scales quadratically with the number of patches. For practical deployment, this computational difference must be considered when selecting architectures for robust applications.

7 Limitations

While our study provides valuable insights, several limitations should be acknowledged:

Training Budget Imbalance: ResNet received 20 epochs for clean training while ViT received only 5–10 epochs for robust training due to computational cost. Figure 3 shows that ViT TRADES curves have not plateaued by epoch 5, suggesting continued improvement with additional training. This may underestimate ViT's potential robustness.

Initialization Effects: ViT uses ImageNet-1k pretraining while ResNet trains from scratch. This confounds architecture comparison with initialization effects and may provide ViT with advantages in feature learning.

Limited Attack Diversity: We focus on ℓ_∞-bounded attacks. Other threat models (ℓ_2, ℓ_0, semantic perturbations) may reveal different architectural vulnerabilities and provide a more complete picture of robustness.

8 Conclusion

Our investigation concludes that **architecture and defense must be co-designed.** ResNet-18 is the superior choice for raw robustness when using PGD Adversarial Training (45.34% robust accuracy). However, for applications requiring a balance of clean performance and stability, Vision Transformers paired with TRADES provide a more stable alternative.

Theoretical Implications: Our findings suggest that inductive biases matter for adversarial robustness. ResNet's locality bias enables learning spatially-localized robust features, while ViT's global attention may spread adversarial perturbations across the entire representation. This aligns with recent theoretical work on the role of architecture in robust learning [12].

Broader Impact: As Vision Transformers become dominant in computer vision, understanding their adversarial vulnerabilities is critical for safe deployment. Our work demonstrates that practitioners cannot assume ViT robustness equals CNN robustness and must carefully select defense strategies based on architecture.

Future work should explore hybrid models like ConvNeXt to combine the best of both worlds, investigate architecture-specific hyperparameter tuning for TRADES, and evaluate certified defense mechanisms.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*.

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ICLR*.

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *ICLR*.

[4] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically Principled Trade-off between Robustness and Accuracy. *ICML*.

[5] Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

[6] Shao, R., et al. (2021). On the Adversarial Robustness of Vision Transformers. *arXiv preprint*.

[7] Bhojanapalli, S., et al. (2021). Understanding Robustness of Transformers for Image Classification. *ICCV*.

[8] Paul, S., & Chen, P. Y. (2021). Vision Transformers are Robust Learners. *AAAI*.

[9] Mahmood, K., et al. (2021). On the Robustness of Vision Transformers to Adversarial Examples. *ICCV*.

[10] Eykholt, K., et al. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. *CVPR*.

[11] Balduzzi, D., et al. (2017). The Shattered Gradients Problem: If resnets are the answer, then what is the question? *ICML*.

[12] Bubeck, S., & Sellke, M. (2021). A Universal Law of Robustness via Isoperimetry. *NeurIPS*.