Compositional Corruption Robustness in Vision Models

Medha Srivastava, Siddarth Bhupathiraju

CS523: Deep Learning, Fall 2025

Introduction

Computer vision systems are deployed in settings where inputs are rarely clean: traffic cameras in fog, robots in dim lighting, or sensors on moving platforms. Robustness benchmarks such as ImageNet-C and CIFAR-10-C expose models to individual corruptions (e.g. noise or blur), but real images often contain multiple distortions at once.

We study compositional corruption robustness: how the accuracy of image classifiers degrades when *two* corruptions are applied simultaneously at varying severities. We focus on CIFAR-10 and compare CNNs (ResNet-20,VGG-16) and Transformers (ViT).

Prior work shows that even high-performing CNNs lose a large fraction of their accuracy under simple corruptions, and that architectural design (e.g. texture vs. shape bias) affects robustness. However, robustness is usually measured along a single axis at a time. We aim to understand how models behave when distortions combine.

Research Questions

- Do compositional corruptions interact non-linearly, or do they affect accuracy independently?
- How does model performance under compositional corruptions compare to individual corruption robustness?
- Do different architectures (CNNs vs. Vision Transformers) show different compositional robustness patterns?

We systematically measure accuracy as a function of corruption severity pairs to understand whether combined distortions amplify degradation or behave additively.

Methodology

Dataset: CIFAR-10 test set (10,000 images)

Models: ResNet-20 (92.1%), VGG-16 (93.6%), ViT-Base (95.1% clean accuracy)

Evaluation:

- 81 severity combinations per pair (9×9 grid: 0.0 to 1.0)
- 4 corruption pairs × 3 models
- Continuous severity (vs discrete 1-5 in prior work)

Corruptions: Blur, noise, brightness, saturation applied sequentially.

Key Metric - CRS:

$$\text{CRS} = A_{\text{pair}}(s_1, s_2) - \min(A_1(s_1), A_2(s_2))$$

Measures whether compositional accuracy is worse than expected from individual corruptions.

Results

Vision Transformers Show Superior Compositional Robustness

At moderate compositional severity (blur=0.5, brightness=0.5), Vision Transformer maintains **38% higher accuracy** than the best CNN.

Model	Accuracy	vs. ViT
ResNet-20	15.9%	-38%
VGG-16	14.1%	-44%
ViT-Base	**22.0%**	**–**

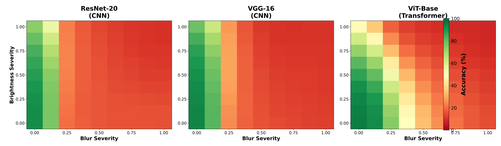

Figure 1: Accuracy heatmaps for blur+brightness. ViT maintains consistently higher accuracy across the severity grid compared to both CNNs.

This substantial gap reveals that architectural choice critically impacts robustness under compositional corruptions, despite similar clean accuracy (92-95%).
Under increasing compositional severity, CNNs and ViT show markedly different degradation patterns:

- **CNNs:** Steep degradation, plateau near 15% by severity 0.5
- **ViT:** Gentler slope, maintains 22% at severity 0.5

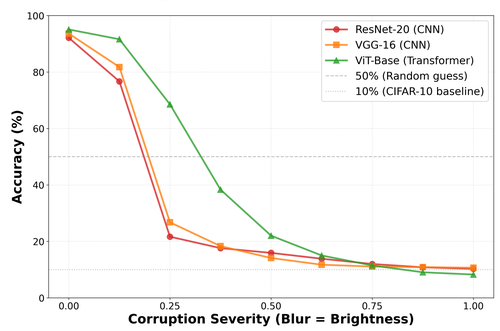

Figure 2: Degradation curves from clean to severity 0.5 for blur+brightness. ViT's transformer global attention model shows slower degradation than CNN local features.

ViT's global self-attention allows it to aggregate information across all patches simultaneously, emphasizing less-corrupted regions even when multiple areas are degraded. In contrast, CNN's hierarchical local features all degrade when multiple regions are corrupted.

Summary Comparison

Model	Clean	Comp.	Degrad.
ResNet-20	92.1%	15.9%	-76.2%
VGG-16	93.6%	14.1%	-79.5%
ViT-Base	**95.1%**	**22.0%**	**-73.1%**

Figure 3: Compositional robustness at blur=0.5, brightness=0.5. Despite similar clean accuracy, ViT shows 38% better compositional robustness than best CNN.

The architectural advantage is quantifiable and substantial:

- **38% relative improvement** at moderate severity (0.5, 0.5)
- **6 percentage points absolute** gain (22% vs 16%)
- **Consistent across severity levels:** ViT maintains advantage throughout degradation curve

This gap is significant for applications where worst-case performance under multiple multiple distortions matters. Results for the other corruption pairs were similar in trend, varying by scale.

Model Architecture

ViT performs better, because:
CNN local features:

- Hierarchical processing with limited receptive fields
- When multiple regions corrupted, local features all degraded
- No mechanism to emphasize cleaner patches

ViT global attention:

- Self-attention aggregates from all patches simultaneously
- Can identify and emphasize less-corrupted regions
- Robust integration across corrupted image

This architectural difference explains the 38% performance gap under compositional corruptions.

Key Points

1. CNN vs Transformer compositional robustness comparison

Quantified substantial ViT advantage (38%) under compositional corruptions—demonstrates that architectural choice critically impacts real-world robustness.

2. Continuous severity methodology

Our 0.0–1.0 framework (vs ImageNet-C's discrete 1–5) enables precise measurement of degradation trajectories, showing that ViT's advantage is consistent across severity levels, not just at specific thresholds.

Implications

Current robustness benchmarks test single corruptions—insufficient for evaluating real-world performance where multiple distortions co-occur. Compositional evaluation reveals architectural differences masked by single-corruption testing.

Global attention mechanisms provide substantial compositional robustness advantage over local CNN features. This 38% gap suggests fundamental limitations of localized processing under multi-corruption scenarios.

Models with similar clean accuracy can have vastly different compositional robustness. For autonomous vehicles and medical imaging where worst-case matters, Transformer architectures provide meaningfully better reliability.

Final Conclusion

Vision Transformers show 38% better compositional robustness than CNNs at moderate severity levels, revealing that compositional corruptions interact non-linearly and architectural choice critically impacts worst-case performance.

Global self-attention enables integration of information across corrupted image regions more effectively than hierarchical local features, explaining the substantial performance gap.

For safety-critical applications requiring worst-case guarantees, architectural choice critically impacts compositional robustness—a dimension current benchmarks fail to evaluate.

Future work:

- Analyze improved robustness methods under composed corruptions
- Develop compositional-aware training
- Scale to ImageNet