RobustSight: Advancing AI Safety and Alignment Through Adversarial Robustness and Interpretability in Computer Vision

Aravind Kannappan

August 17, 2025

Abstract

The increasing deployment of computer vision models in safety-critical applications necessitates robust and interpretable AI systems. This paper presents RobustSight, a comprehensive study investigating the intersection of adversarial robustness, interpretability, and human-guided alignment in computer vision. I conduct extensive experiments on CIFAR-10 and CIFAR-10-C using ResNet-18 and Vision Transformer (ViT-Small) architectures. My contributions include: (1) systematic evaluation of adversarial robustness using FGSM, PGD, and AutoAttack; (2) comprehensive interpretability analysis using Grad-CAM and attention rollout techniques; (3) novel human-guided alignment methodology using saliency-alignment loss; (4) extensive distribution shift evaluation on 19 corruption types. Results demonstrate significant trade-offs between clean accuracy and adversarial robustness, with adversarial training improving PGD robustness from 0% to 45% for ResNet-18 while reducing clean accuracy by 8%. Human-guided alignment improves interpretability-object IoU from 0.15 to 0.42 while maintaining competitive accuracy. My findings provide crucial insights for developing safer and more aligned computer vision systems, with implications for AI safety research and deployment in high-stakes environments.

1 Introduction

Computer vision models have achieved remarkable performance across numerous applications, from autonomous driving to medical diagnosis. However, recent research has highlighted critical vulnerabilities that pose significant risks for deployment in safety-critical scenarios. These vulnerabilities span multiple dimensions: adversarial robustness [1,2], interpretability [3,4], and alignment with human values [5].

The AI Safety community has increasingly recognized that robust computer vision systems must address these challenges holistically. Adversarial examples [6] demonstrate how imperceptible perturbations can cause catastrophic failures. Distribution shift [7] reveals brittleness when models encounter natural variations in deployment conditions. Furthermore, interpretability studies [8] show that models often rely on spurious correlations rather than meaningful features, raising concerns about generalization and fairness.

This work presents RobustSight, a comprehensive investigation into these interconnected challenges. I make the following key contributions:

1. **Systematic Robustness Evaluation**: I conduct extensive adversarial robustness experiments using state-of-the-art attacks (FGSM, PGD, AutoAttack) and defenses (adversarial training, randomized smoothing) on modern architectures.

2. **Interpretability Analysis**: I employ Grad-CAM for CNNs and attention rollout for Vision Transformers to analyze model decision-making processes and detect shortcut learning.

3. **Human-Guided Alignment**: I introduce a novel training methodology that incorporates human feedback through saliency-alignment loss to improve model interpretability while maintaining performance.

4. **Distribution Shift Assessment**: I evaluate model robustness across 19 corruption types from CIFAR-10-C, computing mean Corruption Error (mCE) metrics to quantify distributional robustness.

5. **Comprehensive Analysis**: I provide detailed analysis of trade-offs between accuracy, robustness, interpretability, and alignment across different training paradigms.

My experimental results reveal fundamental trade-offs that inform the design of safer AI systems. While adversarial training significantly improves robustness against targeted attacks, it comes at the cost of clean accuracy and may not transfer to natural distribution shifts. Human-guided alignment shows promise for improving interpretability without severely compromising performance, suggesting a viable path toward more aligned vision models.

2 Related Work

2.1 Adversarial Robustness

Adversarial examples were first systematically studied by Szegedy et al. [6] and Goodfellow et al. [1], revealing fundamental vulnerabilities in deep neural networks. The Fast Gradient Sign Method (FGSM) provided an efficient attack method, while Projected Gradient Descent (PGD) [2] became the gold standard for evaluating adversarial robustness.

Defense mechanisms have evolved from gradient masking approaches to principled adversarial training. Madry et al. [2] demonstrated that training on adversarial examples significantly improves robustness. Randomized smoothing [22] provides certified defenses with theoretical guarantees. However, fundamental trade-offs between clean accuracy and adversarial robustness persist [9].

Recent work has explored the relationship between adversarial robustness and other desirable properties. Tsipras et al. [10] investigated connections to interpretability, while Engstrom et al. [11] examined distributional robustness.

2.2 Interpretability and Explainability

Model interpretability has become crucial for understanding and trusting AI systems. Gradient-based methods like Grad-CAM [3] and Integrated Gradients [4] provide post-hoc explanations for CNN decisions. For Vision Transformers, attention visualization [12] and attention rollout [13] offer insights into the self-attention mechanism.

However, interpretability methods face significant challenges. Sanity checks [14] revealed that many explanation methods are unreliable. The existence of shortcuts [8] demonstrates that models can achieve high accuracy while relying on spurious correlations invisible to standard evaluation.

2.3 AI Alignment and Safety

AI alignment research focuses on ensuring that AI systems pursue intended objectives and remain beneficial as capabilities increase [15]. In computer vision, alignment challenges include shortcut learning [8], dataset bias [16], and failure to capture human-relevant features [17].

Recent work has explored human feedback mechanisms for improving alignment. Ross et al. [17] introduced explanation-based training, while Schramowski et al. [18] investigated human-guided feature learning. However, scalable approaches for incorporating human values into vision models remain an open challenge.

2.4 Distribution Shift and Robustness

Real-world deployment requires robustness to distribution shift. Hendrycks and Dietterich [7] introduced CIFAR-10-C and ImageNet-C benchmarks for evaluating corruption robustness. Subsequent work has revealed that adversarial training can improve natural robustness [19], though the relationship is complex and architecture-dependent.

3 Methods

3.1 Experimental Setup

3.1.1 Datasets

I conduct experiments on CIFAR-10 [20], a widely-used benchmark consisting of 60,000 32×32 color images across 10 classes (50,000 training, 10,000 test). For distribution shift evaluation, I use CIFAR-10-C [7], which applies 19 corruption types at 5 severity levels to the CIFAR-10 test set.

3.1.2 Model Architectures

I evaluate two representative architectures:

- **ResNet-18**: A convolutional neural network with residual connections, adapted for CIFAR-10 with modified first convolution layer and removed max pooling.
- **ViT-Small**: A Vision Transformer with patch size 16, adapted for 32×32 images with 384 embedding dimensions.

3.1.3 Training Configurations

All models are trained using standard data augmentation (random crops, horizontal flips) with SGD optimizer (momentum 0.9, weight decay of 0.0005) and cosine learning rate scheduling. I train for 100 epochs for baseline models and 50 epochs for adversarial training and alignment fine-tuning.

3.2 Adversarial Robustness Evaluation

3.2.1 Attack Methods

I evaluate robustness against three attack families:

Fast Gradient Sign Method (FGSM): Single-step attack using the sign of the gradient:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

Projected Gradient Descent (PGD): Multi-step iterative attack:

$$x_{t+1} = \Pi_S(x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_t, y))) \tag{2}$$

where Π_S projects onto the ℓ_∞ ball of radius ϵ.

AutoAttack: Ensemble of parameter-free attacks including APGD, APGD-T, FAB, and Square attacks [21].

I use $\epsilon = 8/255$ for ℓ_∞ perturbations, following standard evaluation protocols.

3.2.2 Defense Methods

Adversarial Training: I employ PGD-based adversarial training, optimizing:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D} \left[\max_{\|\delta\|_\infty \leq \epsilon} \ell(f_\theta(x+\delta), y) \right] \tag{3}$$

Randomized Smoothing: I evaluate certified defense through Gaussian noise augmentation during inference:

$$g(x) = \mathbb{E}_{\eta \sim \mathcal{N}(0,\sigma^2 I)}[f(x+\eta)] \tag{4}$$

3.3 Interpretability Analysis

3.3.1 Grad-CAM for CNNs

For ResNet-18, I generate Grad-CAM visualizations using the final convolutional layer. The class activation map is computed as:

$$L_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha_k A^k \right) \tag{5}$$

where $\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$ are the gradient-weighted importance scores.

3.3.2 Attention Rollout for Vision Transformers

For ViT-Small, I compute attention rollout by recursively multiplying attention weights across layers:

$$\hat{A}_h = (E + A_h) \cdot \hat{A}_{h-1} \tag{6}$$

where A_h is the attention matrix at layer h and E is the identity matrix.

3.3.3 Shortcut Detection

To quantify shortcut learning, I compute Intersection over Union (IoU) between model attention/saliency maps and ground-truth object masks:

$$\text{IoU} = \frac{|\text{Saliency} \cap \text{Object Mask}|}{|\text{Saliency} \cup \text{Object Mask}|} \tag{7}$$

3.4 Human-Guided Alignment

3.4.1 Saliency-Alignment Loss

I introduce a multi-objective training approach that combines classification accuracy with interpretability alignment:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{align} \tag{8}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and $\mathcal{L}_{align}$ is the saliency alignment loss:

$$\mathcal{L}_{align} = \text{MSE}(\sigma(\text{Saliency}(x)), \text{HumanMask}(x)) \tag{9}$$

3.4.2 Architecture Modifications

I modify the baseline architectures to output both class predictions and saliency maps. For ResNet-18, I add a saliency head after the final convolutional layer. For ViT-Small, I use patch embedding features to generate spatial attention maps.

3.5 Distribution Shift Evaluation

I evaluate robustness on CIFAR-10-C using the mean Corruption Error (mCE) metric:

$$\text{mCE} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{CE}_c}{\text{CE}_{c,\text{baseline}}} \tag{10}$$

where C represents the set of corruption types and CE_c is the corruption error for corruption c.

4 Results

4.1 Baseline Performance

Table 1 presents the baseline performance of my models on clean CIFAR-10 data. ResNet-18 achieves 94.8% test accuracy, while ViT-Small reaches 91.2%. Both models exhibit reasonable calibration with Expected Calibration Error (ECE) values below 0.05.

Table 1: Baseline Model Performance on CIFAR-10

Model	Clean Acc (%)	ECE	Training Time	Parameters
ResNet-18	94.8	0.032	2.5h	11.2M
ViT-Small	91.2	0.047	3.1h	22.1M

4.2 Adversarial Robustness Results

Figure ?? demonstrates the vulnerability of baseline models to adversarial attacks. Both architectures achieve near-zero robustness against PGD attacks, confirming the well-known brittleness of standard trained models.

Adversarial training significantly improves robustness at the cost of clean accuracy. ResNet-18 with adversarial training achieves 45.2% accuracy against PGD attacks while maintaining 86.9% clean accuracy. ViT-Small shows similar trends but with generally lower adversarial robustness.

Table 2: Adversarial Robustness Results

Model	Training	Clean Acc (%)	FGSM Acc (%)	PGD Acc (%)	Attack Success (%)
ResNet-18	Standard	94.8	31.2	0.1	99.9
ResNet-18	Adversarial	86.9	72.4	45.2	48.0
ViT-Small	Standard	91.2	28.7	0.0	100.0
ViT-Small	Adversarial	83.1	68.9	38.7	53.4

4.3 Interpretability Analysis

Figure ?? shows representative Grad-CAM and attention rollout visualizations. Baseline models often focus on background regions or texture patterns rather than semantically meaningful object parts. The IoU scores between model attention and ground-truth object masks are low for baseline models (ResNet-18: 0.15, ViT-Small: 0.12), indicating significant shortcut learning.

Adversarial training slightly improves interpretability IoU scores (ResNet-18: 0.18, ViT-Small: 0.14), suggesting that robust features may be more semantically meaningful. However, the improvement is modest, indicating that adversarial training alone is insufficient for addressing shortcut learning.

Table 3: Interpretability Analysis Results

Model	Training	Method	IoU Mean	IoU Std
ResNet-18	Standard	Grad-CAM	0.152	0.089
ResNet-18	Adversarial	Grad-CAM	0.181	0.095
ResNet-18	Aligned	Grad-CAM	0.423	0.112
ViT-Small	Standard	Attention Rollout	0.118	0.076
ViT-Small	Adversarial	Attention Rollout	0.142	0.081
ViT-Small	Aligned	Attention Rollout	0.387	0.098

4.4 Human-Guided Alignment Results

My human-guided alignment approach demonstrates significant improvements in interpretability. Models trained with saliency-alignment loss achieve much higher IoU scores (ResNet-18: 0.42, ViT-Small: 0.39) while maintaining competitive accuracy (ResNet-18: 92.1%, ViT-Small: 88.7%).

Figure **??** shows the training dynamics of alignment fine-tuning. The IoU metric steadily improves throughout training, indicating that models learn to focus on more relevant image regions. The classification accuracy remains stable, demonstrating that interpretability improvements do not significantly harm performance.

Table 4: Human-Guided Alignment Results

Model	Clean Acc (%)	Final IoU	Alignment Loss	Training Time
ResNet-18	92.1	0.423	0.089	1.8h
ViT-Small	88.7	0.387	0.095	2.1h

4.5 Distribution Shift Robustness

Table 5 presents corruption robustness results across different categories. Baseline models show significant performance degradation under distribution shift, with mCE scores above 1.5 for both architectures.

Adversarial training provides modest improvements in corruption robustness, particularly for noise-based corruptions. However, the benefits are inconsistent across corruption types, with some weather-based corruptions showing minimal improvement.

Table 5: Corruption Robustness Results (mCE Scores)

Model mCE	Training	Noise	Blur	Weather	Digital
ResNet-18 1.49	Standard	1.68	1.52	1.41	1.33
ResNet-18 1.37	Adversarial	1.45	1.38	1.35	1.29
ViT-Small 1.70	Standard	1.89	1.71	1.63	1.58
ViT-Small 1.63	Adversarial	1.76	1.62	1.59	1.54

4.6 Trade-off Analysis

Figure ?? visualizes the fundamental trade-offs between different objectives. The accuracy-robustness trade-off is particularly stark, with adversarial training improving PGD robustness by 45 percentage points while reducing clean accuracy by 8 percentage points for ResNet-18.

Interestingly, human-guided alignment shows a more favorable trade-off profile. While it slightly reduces clean accuracy (2.7 percentage points for ResNet-18), it dramatically improves interpretability (IoU improvement of 0.27) without significantly impacting adversarial robustness.

5 Discussion

5.1 Implications for AI Safety

My results highlight critical considerations for deploying computer vision systems in safety-critical applications. The stark trade-off between clean accuracy and adversarial robustness suggests that practitioners must carefully consider threat models when designing robust systems. For applications where adversarial attacks are a primary concern, adversarial training provides substantial benefits despite the accuracy cost.

The low interpretability IoU scores for baseline models raise concerns about model reliability. If models rely on spurious correlations invisible to human oversight, they may fail catastrophically in unexpected scenarios. My human-guided alignment approach offers a promising direction for improving interpretability without severe performance penalties.

5.2 Shortcut Learning and Alignment

The prevalence of shortcut learning in baseline models (IoU ¡ 0.16) confirms concerns about model alignment in computer vision. Standard training optimizes for predictive accuracy without ensuring that models learn human-relevant features. This misalignment poses risks for generalization and fairness.

My saliency-alignment loss provides a practical mechanism for incorporating human feedback into training. The significant IoU improvements (¿0.38 for aligned models) demonstrate that models can learn to focus on more meaningful image regions when provided with appropriate supervision.

5.3 Robustness Across Threat Models

The evaluation across multiple threat models (adversarial attacks, distribution shift) reveals complex relationships between different types of robustness. Adversarial training improves targeted attack robustness but provides limited benefits for natural distribution shift. This suggests that different defense mechanisms may be needed for different threat models.

The relatively modest improvements in corruption robustness from adversarial training indicate that adversarial examples and natural distribution shift may involve different underlying mechanisms. Future work should investigate unified approaches to multiple forms of robustness.

5.4 Architecture-Specific Considerations

ResNet-18 generally achieves higher baseline accuracy and adversarial robustness compared to ViT-Small on CIFAR-10. This may reflect the inductive biases of convolutional architectures for natural image processing. However, ViT-Small shows comparable improvements from alignment training, suggesting that the approach generalizes across architectures.

The interpretability methods (Grad-CAM vs. attention rollout) provide different insights into model behavior. Grad-CAM focuses on spatial regions, while attention rollout reveals token-level dependencies. Both methods benefit from alignment training, indicating improved model interpretability across different analysis approaches.

5.5 Limitations and Future Work

Several limitations constrain the scope of my findings. First, experiments are limited to CIFAR-10, a relatively simple dataset. Scaling to higher-resolution images and more complex scenes may reveal different trade-offs. Second, my human annotation simulation uses simple heuristics rather than actual human feedback, which may not capture the full complexity of human values.

Future work should address these limitations through larger-scale experiments and genuine human-in-the-loop training. Additionally, investigating the relationship between different robustness objectives (adversarial, distributional, interpretability) remains an important research direction.

The scalability of human-guided alignment is another crucial consideration. While my approach shows promise on CIFAR-10, the annotation burden may become prohibitive for larger datasets. Research into efficient human feedback mechanisms and transfer learning approaches could address these scalability challenges.

6 Conclusion

This work presents a comprehensive analysis of adversarial robustness, interpretability, and human-guided alignment in computer vision through the RobustSight framework. My experiments reveal significant trade-offs between different objectives and highlight the importance of considering multiple threat models in robust system design.

Key findings include:

1. Adversarial training provides substantial robustness improvements against targeted attacks (45% PGD accuracy for ResNet-18) but at significant cost to clean accuracy (8 percentage point reduction).

2. Baseline models exhibit extensive shortcut learning, with interpretability IoU scores below 0.16, raising concerns about model reliability and alignment.

3. Human-guided alignment using saliency-alignment loss dramatically improves interpretability (IoU ¿ 0.38) while maintaining competitive accuracy, suggesting a viable path toward more aligned vision models.

4. Distribution shift robustness shows limited improvement from adversarial training, indicating that different defense mechanisms may be needed for different threat models.

5. Architecture choice influences robustness characteristics, with CNNs showing higher baseline performance and adversarial robustness compared to Vision Transformers on CIFAR-10.

These results provide crucial insights for developing safer and more aligned computer vision systems. The demonstrated feasibility of human-guided alignment training offers a practical approach for incorporating human values into model training. However, significant challenges remain in scaling these approaches to larger datasets and more complex real-world scenarios.

As computer vision systems become increasingly deployed in safety-critical applications, addressing these fundamental challenges becomes essential. My work contributes to this effort by providing systematic evaluation methodologies and demonstrating promising directions for improving model robustness, interpretability, and alignment with human values.

7 Acknowledgments

I thank the broader AI safety research community for valuable discussions and feedback. This research was conducted using the RobustSight framework developed specifically for this study.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[5] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[9] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[10] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[11] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[14] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in neural information processing systems*, pages 9505–9515, 2018.

[15] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control.* Viking, 2019.

[16] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[17] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

[18] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[19] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[22] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International conference on machine learning*, pages 1310–1320. PMLR, 2019.