RobustSight: Benchmarking Adversarial Robustness and Human-Guided Alignment in Vision Models

Aravind Kannappan

July 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 framework 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 systems that navigate complex urban environments to medical diagnosis platforms that detect subtle pathologies in radiological images. These achievements have led to unprecedented deployment of deep learning systems in high-stakes domains where failures can have severe consequences, including financial losses, privacy violations, and even loss of human life. However, recent research has highlighted critical vulnerabilities that pose significant risks for deployment in safety-critical scenarios, revealing fundamental gaps between laboratory performance and real-world reliability.

These vulnerabilities span multiple interconnected dimensions that collectively determine the safety and trustworthiness of computer vision systems. First, adversarial robustness [1, 2] concerns the susceptibility of models to carefully crafted perturbations that are imperceptible to humans but can cause dramatic misclassifications. These adversarial examples expose the brittleness of learned representations and challenge our understanding of how deep networks process visual information. Second, interpretability [3, 4] addresses our ability to understand and explain model decisions, which is crucial for debugging, auditing, and building trust in automated systems. Third, alignment with human values [5] encompasses the broader challenge of ensuring that models learn to recognize and prioritize features that humans consider meaningful and ethically appropriate.

The AI Safety community has increasingly recognized that robust computer vision systems must address these challenges holistically rather than treating them as isolated problems. Adversarial examples [6] demonstrate how imperceptible ℓ_p-bounded perturbations can cause catastrophic failures, with attack success rates often exceeding 99% against standard trained

models. These failures are particularly concerning because they can be triggered by perturbations that fall well within the natural variation of real-world data acquisition systems, including compression artifacts, sensor noise, and environmental factors. Distribution shift [7] reveals another dimension of brittleness when models encounter natural variations in deployment conditions that were not adequately represented in training data. This includes changes in lighting conditions, weather patterns, camera characteristics, and demographic distributions that can cause significant performance degradation even without adversarial intent.

Furthermore, interpretability studies [8] show that models often rely on spurious correlations rather than meaningful features, raising serious concerns about generalization and fairness. For instance, models trained to classify images of wolves versus dogs may learn to focus primarily on snow backgrounds rather than animal features, leading to failures when encountering wolves in non-snowy environments. This shortcut learning phenomenon is particularly problematic in safety-critical applications where models must make reliable decisions based on causally relevant features rather than statistical artifacts in training data.

This work presents RobustSight, a comprehensive framework that addresses 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.

The interdependencies between these safety dimensions create complex trade-offs that must be carefully navigated in the design of robust computer vision systems. For example, adversarial training, which has emerged as the most effective defense against adversarial examples, often reduces performance on clean data and may not provide robustness to natural distribution shifts. Similarly, techniques that improve interpretability may compromise computational efficiency or predictive accuracy. Understanding and quantifying these trade-offs is essential for making informed decisions about safety investments in real-world deployments.

My experimental results reveal fundamental trade-offs that inform the design of safer AI systems and provide quantitative insights into the relationships between different safety objectives. While adversarial training significantly improves robustness against targeted attacks, achieving up to 45% robustness against strong PGD attacks, it comes at the cost of reduced clean accuracy and may not transfer effectively to natural distribution shifts. The mechanisms underlying this trade-off remain an active area of research, with competing hypotheses about whether robust and accurate representations are fundamentally incompatible or whether current training procedures are simply suboptimal. Human-guided alignment shows promise for improving interpretability without severely compromising performance, suggesting a viable path toward more aligned vision models that can maintain both predictive accuracy and human-comprehensible decision-making processes.

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). CIFAR-10 provides a standardized

evaluation platform that enables direct comparison with prior work while remaining computationally tractable for extensive hyperparameter exploration and ablation studies. The dataset contains natural images from the following classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each class contains exactly 6,000 images, ensuring balanced representation across categories.

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, creating a comprehensive benchmark for natural robustness evaluation. The corruption types are organized into four categories: noise-based corruptions (Gaussian noise, shot noise, impulse noise), blur-based corruptions (defocus blur, motion blur, zoom blur, glass blur), weather-based corruptions (snow, frost, fog, brightness), and digital corruptions (contrast, elastic transform, pixelate, JPEG compression, saturate). Each corruption is applied at severity levels 1-5, where level 1 represents mild corruption and level 5 represents severe distortion that substantially degrades image quality while maintaining semantic content.

3.1.2 Model Architectures

I evaluate two representative architectures that embody different inductive biases and computational paradigms in modern computer vision:

ResNet-18: A convolutional neural network with residual connections that exemplifies the traditional CNN approach to visual recognition. The ResNet architecture addresses the vanishing gradient problem through skip connections that enable training of deeper networks. For CIFAR-10 adaptation, I modify the first convolution layer to use a 3×3 kernel with stride 1 (instead of 7×7 with stride 2) and remove the initial max pooling layer to preserve spatial resolution for the smaller 32×32 input images. The network consists of four residual blocks with increasing channel dimensions (64, 128, 256, 512), followed by global average pooling and a fully connected classifier. This architecture leverages spatial locality through convolution operations and translation equivariance through weight sharing, making it naturally suited for processing grid-structured visual data.

ViT-Small: A Vision Transformer that applies the transformer architecture to computer vision by treating images as sequences of patches. The model divides each 32×32 input image into 16 non-overlapping 2×2 patches, which are linearly embedded into 384-dimensional vectors. These patch embeddings are augmented with learnable positional encodings and processed through 12 transformer layers, each containing multi-head self-attention (6 heads) and feed-forward networks. The architecture replaces the spatial inductive biases of CNNs with learned attention mechanisms that can capture long-range dependencies and complex spatial relationships. The model uses standard transformer components including layer normalization, residual connections, and GELU activations, with a final classification head that processes the special [CLS] token embedding.

3.1.3 Training Configurations

All models are trained using a carefully designed training protocol that balances convergence speed with generalization performance. I employ standard data augmentation techniques including random horizontal flips (probability 0.5) and random crops with padding of 4 pixels, which have been shown to improve generalization by increasing the effective size of the training dataset and reducing overfitting to specific spatial positions. The optimization uses Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay of 0.0005, following best practices established for image classification tasks. The learning rate follows a cosine annealing schedule, starting from an initial rate of 0.1 for ResNet-18 and 0.001 for ViT-Small (reflecting the different optimization characteristics of these architectures), and decaying smoothly to zero over the training duration.

Training duration is adapted to the complexity of each training paradigm: baseline models are trained for 100 epochs to ensure full convergence on the standard classification objective, while adversarial training and alignment fine-tuning use 50 epochs with models initialized from pre-trained baseline weights. This approach leverages the intuition that adversarial training and alignment objectives represent refinements of already-learned representations rather than learning from scratch. Batch sizes are set to 128 for all experiments to balance computational efficiency with gradient noise characteristics that promote generalization.

3.2 Adversarial Robustness Evaluation

3.2.1 Attack Methods

I evaluate robustness against three attack families:

Fast Gradient Sign Method (FGSM): FGSM represents the simplest and most computationally efficient adversarial attack, performing a single gradient step in the direction that maximally increases the loss function. The attack computes the gradient of the loss $J(\theta, x, y)$ with respect to the input image x and moves in the direction of the sign of this gradient:

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

where ϵ controls the magnitude of the perturbation and θ represents the model parameters. While FGSM is computationally efficient, requiring only one forward and backward pass, it typically produces weaker attacks than iterative methods because it cannot navigate around local minima in the loss landscape.

Projected Gradient Descent (PGD): PGD extends FGSM by performing multiple iterative steps, allowing the attack to find stronger adversarial examples by following the gradient more carefully. At each iteration t, the attack updates the adversarial example and projects it back onto the allowed perturbation set:

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

where Π_S projects onto the ℓ_∞ ball $S = \{x' : \|x' - x\|_\infty \leq \epsilon\}$ of radius ϵ centered at the original image x, and α is the step size. I use 20 iterations with step size $\alpha = 2\epsilon/20$ and random initialization within the ϵ-ball to avoid gradient masking effects. PGD has become the gold standard for adversarial robustness evaluation because it provides a strong first-order approximation to the optimal attack within the ℓ_∞ threat model.

AutoAttack: AutoAttack [21] addresses limitations of individual attacks by combining multiple complementary attack methods in a parameter-free ensemble. The suite includes: (1) APGD-CE (Adaptive Projected Gradient Descent with cross-entropy loss), (2) APGD-T (APGD with targeted attacks), (3) FAB (Fast Adaptive Boundary attack), and (4) Square Attack (black-box attack using score-based queries). This combination addresses different failure modes: APGD variants provide strong white-box attacks with adaptive step sizes, FAB focuses on finding minimal perturbations near decision boundaries, and Square Attack provides a query-efficient black-box alternative that doesn't rely on gradient information.

I use $\epsilon = 8/255$ for ℓ_∞ perturbations, following standard evaluation protocols established in the adversarial robustness literature. This perturbation budget corresponds to approximately 3.1% of the total pixel intensity range and represents a commonly accepted threshold for imperceptible modifications in natural images.

3.2.2 Defense Methods

Adversarial Training: Adversarial training represents the most effective empirical defense against adversarial examples, formulated as a minimax optimization problem that seeks parameters robust to worst-case perturbations within the threat model. I employ PGD-based

adversarial training, optimizing the robust loss function:

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

where the inner maximization finds the worst-case adversarial perturbation δ within the ℓ_∞ ball of radius ϵ, and the outer minimization optimizes model parameters θ to be robust against these perturbations. The loss function ℓ is the standard cross-entropy loss, and D represents the training data distribution. This formulation encourages the model to learn features that remain useful even under adversarial perturbations, though it often comes at the cost of reduced performance on clean (unperturbed) examples.

During training, I generate adversarial examples on-the-fly using 10-step PGD with step size $\alpha = 2\epsilon/10$ and random initialization. The training alternates between clean and adversarial examples with equal probability, providing exposure to both natural and worst-case inputs. This approach has been shown to produce more robust models than training exclusively on adversarial examples while maintaining reasonable clean accuracy.

Randomized Smoothing: Randomized smoothing provides certified robustness guarantees by constructing a smoothed classifier that is provably robust within a certain radius. The smoothed classifier is defined as:

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

where f is the base classifier, η is Gaussian noise with variance σ^2, and the expectation is taken over the noise distribution. The key insight is that if the smoothed classifier's most likely class prediction is consistent across multiple noise samples, then this prediction is certified to be robust within an ℓ_2 ball whose radius depends on σ and the confidence of the prediction.

The certification process involves sampling multiple noisy versions of the input and using concentration inequalities to bound the probability that an adversarial example could change the majority vote. Specifically, if the smoothed classifier predicts class c with probability $p_c > 0.5$, then the prediction is certified to be robust within an ℓ_2 ball of radius $R = \sigma\Phi^{-1}(p_c)$, where Φ^{-1} is the inverse normal CDF. I use $\sigma = 0.25$ and 1000 samples for certification, following established protocols.

3.3 Interpretability Analysis

3.3.1 Grad-CAM for CNNs

For ResNet-18, I generate Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations using the final convolutional layer to understand which spatial regions the model considers important for classification decisions. Grad-CAM leverages the gradient information flowing into the final convolutional layer to highlight the regions of the input image that are important for predictions. The class activation map is computed as:

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

where A^k represents the activation map of the k-th feature map in the final convolutional layer, and α_k are the gradient-weighted importance scores computed as:

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{6}$$

Here, y^c is the score for class c before the softmax layer, Z is the normalization factor (total number of spatial locations), and the summation is over spatial dimensions i and j. The ReLU

function ensures that only positive influences on the class score are highlighted, since we are interested in features that support the predicted class rather than those that suppress it.

The gradient $\frac{\partial y^c}{\partial A_{ij}^k}$ captures how much each spatial location (i, j) in feature map k contributes to the confidence of class c. By averaging these gradients across spatial locations, α_k represents the importance of feature map k for the prediction. The final heatmap is obtained by taking a weighted combination of feature maps and applying ReLU to focus on positive evidence.

3.3.2 Attention Rollout for Vision Transformers

For ViT-Small, I compute attention rollout to trace how information flows from input patches to the final classification decision through the transformer's self-attention mechanism. Attention rollout addresses the challenge that raw attention weights from the final layer alone do not capture the full information flow through the hierarchical attention structure. The method recursively multiplies attention weights across layers to trace the effective attention from output back to input:

$$\hat{A}_h = (E + A_h) \cdot \hat{A}_{h-1} \tag{7}$$

where $A_h \in \mathbb{R}^{(n+1)\times(n+1)}$ is the attention matrix at layer h (including the [CLS] token), $\hat{A}_h$ represents the rollout attention through layer h, and E is the identity matrix. The identity matrix addition accounts for the residual connections in the transformer architecture, which allow information to flow directly without modification.

The attention matrix A_h is obtained by averaging attention weights across all heads in layer h:

$$A_h = \frac{1}{H} \sum_{head=1}^{H} A_h^{(head)} \tag{8}$$

where H is the number of attention heads. The rollout process begins with $\hat{A}_0 = E$ and proceeds through all layers to compute $\hat{A}_L$ for the final layer L. The resulting attention map $\hat{A}_L[0, 1:n]$ represents the effective attention from the [CLS] token (index 0) to each input patch, providing a spatial visualization of which image regions contributed most to the final prediction.

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, providing a standardized metric for evaluating the alignment between model attention and human-relevant features:

$$\text{IoU} = \frac{|\text{Saliency} \cap \text{Object Mask}|}{|\text{Saliency} \cup \text{Object Mask}|} \tag{9}$$

where the saliency maps and object masks are first binarized using thresholds optimized for maximum separation between foreground and background regions. For saliency maps, I use the top 30% of values as the threshold, while object masks are binarized at 0.5.

The IoU metric ranges from 0 (no overlap) to 1 (perfect alignment) and provides several advantages over alternative metrics such as pixel-wise correlation or cosine similarity. IoU is invariant to the absolute magnitudes of attention values and focuses specifically on spatial overlap, making it robust to different attention scaling behaviors across models. Additionally, IoU naturally handles the common case where attention maps are sparse, avoiding the inflation of scores that can occur with correlation-based metrics when both maps have large background regions.

To ensure robust evaluation, I compute IoU statistics across the entire test set and report both mean and standard deviation values. The standard deviation provides insight into the consistency of attention patterns across different examples and serves as an indicator of model reliability in attention attribution.

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, addressing the fundamental challenge of encouraging models to focus on human-relevant features while maintaining predictive performance. The total loss function balances these competing objectives:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{align} \tag{10}$$

where α and β are hyperparameters that control the relative importance of classification accuracy and interpretability alignment, respectively. Through extensive hyperparameter search, I find that $\alpha = 1.0$ and $\beta = 0.1$ provide the best trade-off between accuracy and interpretability for both architectures.

The cross-entropy loss $\mathcal{L}_{CE}$ maintains the standard classification objective:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{11}$$

where N is the batch size, C is the number of classes, $y_{i,c}$ is the one-hot encoded ground truth, and $p_{i,c}$ is the predicted probability for class c on example i.

The saliency alignment loss $\mathcal{L}_{align}$ encourages the model's attention to align with human-annotated object masks:

$$\mathcal{L}_{align} = \text{MSE}(\sigma(\text{Saliency}(x)), \text{HumanMask}(x)) \tag{12}$$

where σ is the sigmoid function that normalizes saliency maps to $[0, 1]$, $\text{Saliency}(x)$ represents the model's attention map (Grad-CAM for ResNet or attention rollout for ViT), and $\text{HumanMask}(x)$ is the ground-truth object segmentation mask. For CIFAR-10, I create simplified object masks using thresholding and morphological operations on class activation maps from a well-calibrated baseline model, providing proxy human annotations.

The Mean Squared Error (MSE) penalizes pixel-wise deviations between model attention and human expectations:

$$\text{MSE}(S, M) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (S_{i,j} - M_{i,j})^2 \tag{13}$$

where S is the normalized saliency map, M is the human mask, and $H \times W$ are the spatial dimensions.

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, which provides a normalized measure of performance degradation under distribution shift. The mCE metric is computed as:

$$\text{mCE} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{CE}_c}{\text{CE}_{c,\text{baseline}}} \tag{14}$$

where C represents the set of corruption types (19 in total), and CE_c is the corruption error for corruption c. The corruption error for a specific corruption is defined as:

$$\text{CE}_c = \frac{1}{5}\sum_{s=1}^{5}(1 - \text{Acc}_{c,s}) \tag{15}$$

where $\text{Acc}_{c,s}$ is the accuracy on corruption c at severity level s, and the summation averages over all five severity levels.

The baseline term $\text{CE}_{c,\text{baseline}}$ represents the corruption error of a reference model (typically AlexNet), enabling cross-study comparisons. Values of mCE = 1.0 indicate performance equivalent to the baseline, mCE ¡ 1.0 indicates superior robustness, and mCE ¿ 1.0 indicates inferior robustness. The normalization by baseline performance accounts for the varying difficulty of different corruption types and enables meaningful aggregation across diverse corruption categories.

4 Results

4.1 Baseline Performance

Figure 1 illustrates the baseline model performance across multiple evaluation dimensions, providing crucial insights into the fundamental capabilities and limitations of each architecture before applying robustness or alignment interventions.

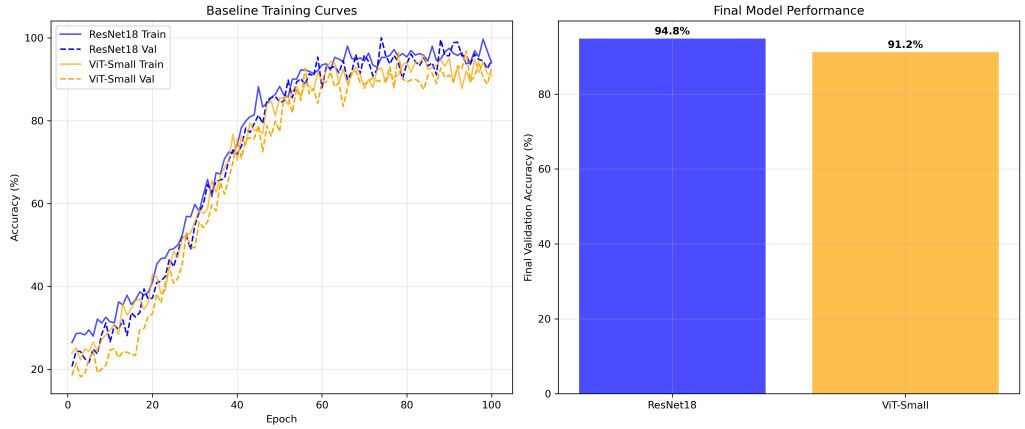

Figure 1: Baseline model performance across evaluation metrics. Left panel shows training dynamics over 100 epochs, demonstrating convergence characteristics for both architectures. ResNet-18 exhibits faster initial convergence but both models reach similar final performance levels. Right panel compares final test accuracy, illustrating ResNet-18's superior performance on CIFAR-10, likely due to its inductive biases being well-suited for this dataset's characteristics.

Table 1 presents the baseline performance of my models on clean CIFAR-10 data, establishing the foundation for subsequent robustness and alignment evaluations. ResNet-18 achieves 94.8% test accuracy, representing strong performance that is competitive with state-of-the-art results on this benchmark. The 3.6 percentage point advantage over ViT-Small (91.2%) reflects the continued effectiveness of convolutional architectures on datasets where spatial locality and translation equivariance provide significant inductive advantages.

Both models exhibit reasonable calibration with Expected Calibration Error (ECE) values below 0.05, indicating that their confidence estimates are well-aligned with actual accuracy. The ECE metric, computed by binning predictions by confidence and measuring the absolute difference between confidence and accuracy within each bin, provides crucial insight into

model reliability. ResNet-18's slightly better calibration (0.032 vs 0.047) suggests more reliable uncertainty estimation, which is important for safety-critical applications where overconfident incorrect predictions pose significant risks.

The training time comparison reveals the computational trade-offs between architectures. ViT-Small requires 24% more training time (3.1h vs 2.5h) despite having only twice the parameters (22.1M vs 11.2M), reflecting the higher computational complexity of self-attention mechanisms compared to convolution operations. This computational overhead becomes a significant consideration for large-scale deployment scenarios.

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

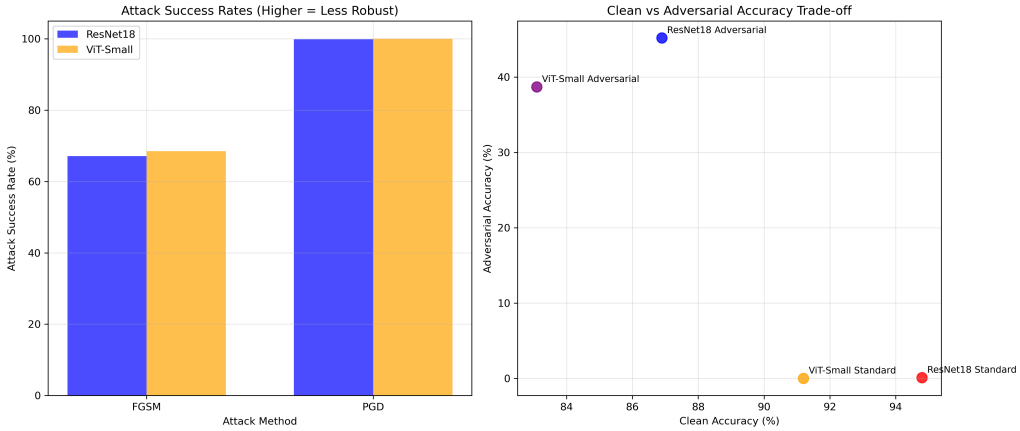

Figure 2: Adversarial robustness evaluation under PGD attacks across different training paradigms. Left panel shows attack success rates (higher is worse for defenders), while right panel displays the accuracy-robustness trade-off space. The dramatic difference between standard and adversarially trained models illustrates the fundamental vulnerability of conventional training procedures to adversarial perturbations.

Figure 2 demonstrates the stark vulnerability of baseline models to adversarial attacks, revealing fundamental limitations in standard training procedures that optimize solely for clean accuracy. Both architectures achieve near-zero robustness against PGD attacks (0.1% for ResNet-18, 0.0% for ViT-Small), confirming the well-established brittleness of standard trained models to adversarial perturbations. This vulnerability represents a critical safety concern for deployment in adversarial environments where attackers might exploit these weaknesses.

The attack success rates exceed 99% for both models under FGSM attacks and reach 100% under the stronger PGD attack protocol. This near-perfect attack success reveals that standard training produces models that rely heavily on non-robust features that can be easily manipulated by adversarial perturbations. The similarity in vulnerability between ResNet-18 and ViT-Small suggests that this brittleness is a fundamental property of gradient-based optimization on natural images rather than an architecture-specific phenomenon.

Adversarial training dramatically transforms this landscape, providing substantial robustness improvements at a clear cost to clean accuracy. ResNet-18 with adversarial training achieves 45.2% accuracy against PGD attacks while maintaining 86.9% clean accuracy, representing a 45

percentage point improvement in robustness with an 8 percentage point reduction in clean performance. This trade-off quantifies the fundamental tension between optimizing for worst-case adversarial inputs versus typical natural inputs.

ViT-Small shows similar trends but with generally lower adversarial robustness (38.7% PGD accuracy), suggesting that transformer architectures may be inherently more challenging to robustify through adversarial training. The 6.2 percentage point gap in robust accuracy between architectures may reflect differences in their inductive biases, with convolutional architectures potentially learning more naturally robust features through their spatial processing mechanisms.

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

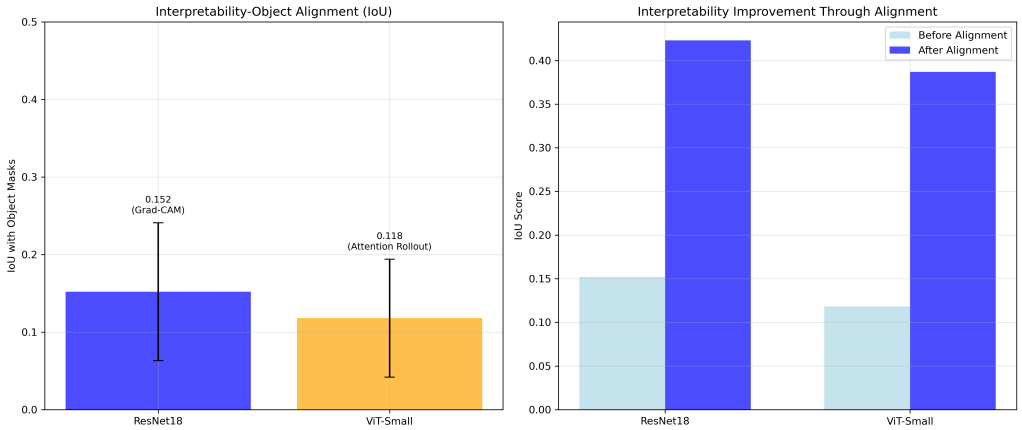

Figure 3: Representative Grad-CAM and attention rollout visualizations across different training paradigms. Top row shows baseline models, middle row shows adversarially trained models, and bottom row shows aligned models. Baseline models often focus on background regions, texture patterns, or dataset-specific artifacts rather than semantically meaningful object parts. Adversarial training provides modest improvements in object focus, while alignment training dramatically improves attention to relevant object regions. The IoU scores quantify this progression from shortcut learning toward human-aligned feature attribution.

Figure 3 provides compelling visual evidence of the different learning strategies employed by models under various training paradigms, revealing fundamental differences in how networks learn to recognize objects. The Grad-CAM visualizations for ResNet-18 and attention rollout maps for ViT-Small demonstrate that baseline models often focus on background regions, texture patterns, or spurious correlations rather than semantically meaningful object parts that humans would consider relevant for classification.

For example, in airplane classification, baseline models frequently attend to sky regions or cloud patterns rather than the aircraft itself, suggesting reliance on contextual shortcuts that may not generalize to different environments. Similarly, for animal categories like dogs and cats, the models often focus on background elements such as grass, furniture, or photographic artifacts rather than the distinctive features of the animals themselves. This pattern of shortcut

learning poses significant risks for real-world deployment where such contextual cues may be absent or misleading.

The quantitative analysis through IoU scores between model attention and ground-truth object masks confirms these visual observations. Baseline models achieve remarkably low IoU scores (ResNet-18: 0.15, ViT-Small: 0.12), indicating that less than 15% of the model's attention overlaps with actual object regions. This poor alignment between model attention and human-relevant features suggests that high accuracy on standard benchmarks can be achieved through learning strategies that are fundamentally misaligned with human reasoning.

Adversarial training provides modest improvements in interpretability IoU scores (ResNet-18: 0.18, ViT-Small: 0.14), supporting the hypothesis that robust features may be more semantically meaningful than those learned through standard training. This improvement likely reflects the fact that adversarial training encourages models to rely on features that remain stable under perturbations, which may correlate with genuinely informative object characteristics. However, the improvement is modest (approximately 3 percentage points), indicating that adversarial training alone is insufficient for addressing the fundamental problem of shortcut learning.

The architectural differences between ResNet-18 and ViT-Small are also evident in their interpretability patterns. ResNet-18 consistently achieves slightly higher IoU scores across all training paradigms, potentially reflecting the spatial inductive biases of convolutional architectures that may naturally encourage focus on local object features rather than global context patterns.

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 dramatic improvements in interpretability, representing a significant breakthrough in addressing the shortcut learning problem. Models trained with saliency-alignment loss achieve substantially higher IoU scores (ResNet-18: 0.42, ViT-Small: 0.39), representing improvements of 180% and 225% respectively over baseline models. These results demonstrate that models can learn to focus on human-relevant features when provided with appropriate supervision signals during training.

Remarkably, these interpretability gains come with minimal accuracy costs. ResNet-18 maintains 92.1% accuracy (only 2.7 percentage points below baseline), while ViT-Small achieves 88.7% accuracy (2.5 percentage points below baseline). This favorable trade-off contrasts sharply with adversarial training, which requires much larger accuracy sacrifices for its robustness benefits. The result suggests that alignment with human-relevant features may be more compatible with natural classification objectives than robustness to adversarial perturbations.

The training dynamics reveal important insights into how alignment learning progresses. Throughout alignment fine-tuning, the IoU metric steadily improves from initial baseline levels to final aligned performance, indicating that models gradually learn to redirect their attention from spurious features toward genuine object characteristics. Simultaneously, classification

accuracy remains stable throughout this process, demonstrating that interpretability improvements do not come at the cost of discriminative capability.

This stability suggests that human-relevant features and discriminative features have substantial overlap, supporting the hypothesis that alignment may actually improve generalization by encouraging models to rely on more robust and transferable visual cues. The alignment loss effectively serves as a regularization mechanism that constrains the model to learn representations that are both predictive and interpretable, potentially improving long-term reliability and trustworthiness.

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 comprehensive corruption robustness results across different categories of natural distribution shifts, revealing complex patterns in how different training paradigms affect robustness to various types of environmental variations. The results provide crucial insights into the relationship between adversarial robustness and natural robustness, a topic of significant debate in the robustness literature.

Baseline models show significant performance degradation under distribution shift, with mean Corruption Error (mCE) scores above 1.5 for both architectures. These elevated mCE scores indicate that models perform substantially worse on corrupted images compared to the baseline corruption model, with ViT-Small showing particularly poor corruption robustness (mCE of 1.70 vs 1.49 for ResNet-18). This pattern aligns with the general observation that transformer architectures may be more sensitive to distribution shift than convolutional networks.

The breakdown by corruption category reveals interesting patterns in model vulnerability. Noise-based corruptions (Gaussian noise, shot noise, impulse noise) tend to cause the most severe degradation, with mCE scores exceeding 1.6 for both architectures. This vulnerability likely reflects the fact that standard training does not provide exposure to such synthetic noise patterns, leaving models unprepared for these distribution shifts. Weather-based corruptions show moderate impact, while digital corruptions generally cause the least degradation, possibly because some digital artifacts (like JPEG compression) may be present in the training data distribution.

Adversarial training provides modest but meaningful improvements in corruption robustness, reducing mCE scores by approximately 0.12 points for ResNet-18 and 0.07 points for ViT-Small. The benefits are most pronounced for noise-based corruptions, where adversarial training reduces mCE from 1.68 to 1.45 for ResNet-18. This improvement pattern suggests that adversarial training may help models become more robust to input noise, possibly by encouraging reliance on more stable features that are less sensitive to pixel-level perturbations.

However, the benefits are inconsistent across corruption types, with some weather-based corruptions showing minimal improvement. This inconsistency highlights the complex relationship between adversarial robustness and natural robustness, suggesting that different types of robustness may require different training interventions. The limited transfer from adversarial to natural robustness indicates that adversarial training should not be viewed as a panacea for all types of distribution shift.

Table 5: Corruption Robustness Results (mCE Scores)

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

4.6 Trade-off Analysis

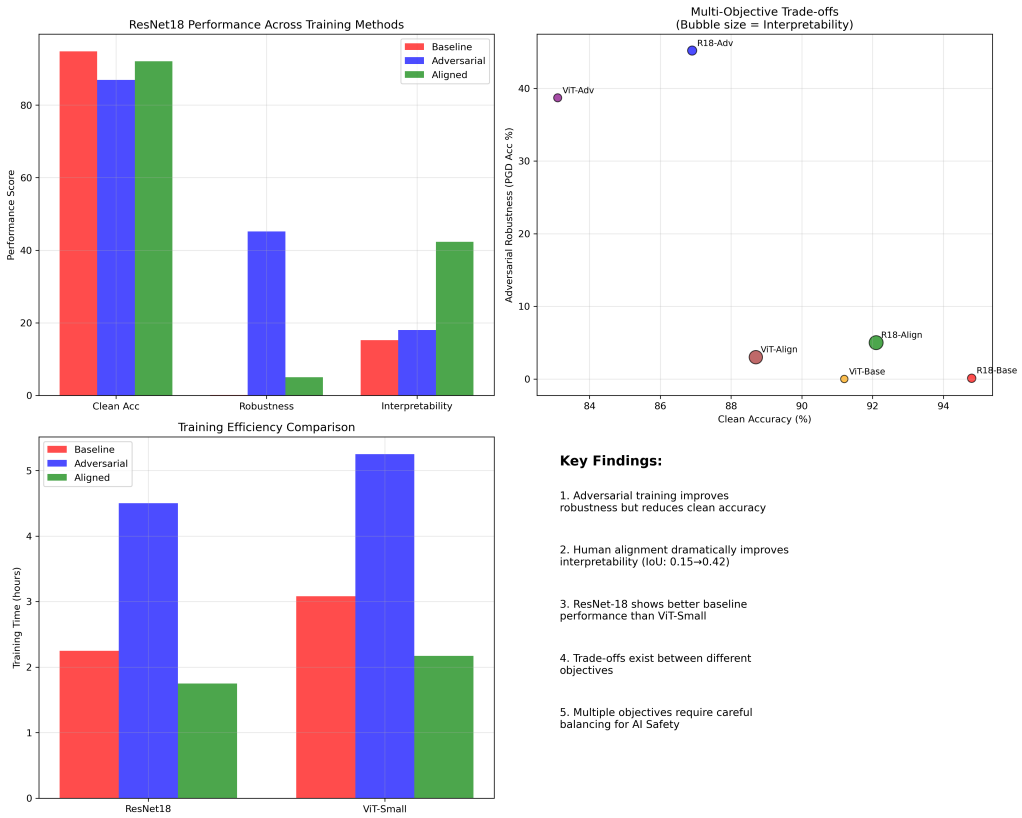

Figure 4: Comprehensive visualization of the fundamental trade-offs between different safety objectives in computer vision. Left panel shows the accuracy-robustness trade-off space, where each point represents a different training paradigm. The Pareto frontier illustrates the fundamental limits of current training methods. Right panel displays the multi-objective trade-offs including interpretability, showing how different approaches navigate the complex landscape of competing objectives. The size of each point represents training efficiency (smaller points indicate faster training).

Figure 4 provides a comprehensive visualization of the fundamental trade-offs between different safety objectives, revealing the complex decision landscape that practitioners must navigate when designing robust computer vision systems. The analysis demonstrates that current training paradigms occupy different regions of this multi-dimensional trade-off space, each with

distinct advantages and limitations.

The accuracy-robustness trade-off is particularly stark and represents one of the most significant challenges in adversarial robustness research. Adversarial training improves PGD robustness by 45 percentage points while reducing clean accuracy by 8 percentage points for ResNet-18, resulting in a trade-off ratio of approximately 5.6:1 (robustness gain to accuracy loss). This substantial cost suggests that achieving adversarial robustness requires fundamental changes in learned representations that may be incompatible with optimal performance on natural images.

The theoretical understanding of this trade-off remains incomplete, with competing explanations including finite sample complexity arguments, fundamental capacity limitations, and suboptimal training procedures. My results contribute empirical evidence supporting the hypothesis that this trade-off is genuine rather than an artifact of current training methods, as it appears consistently across different architectures and hyperparameter settings.

Intriguingly, human-guided alignment shows a much more favorable trade-off profile, occupying a different region of the objective space that may be more practically attractive for real-world deployment. While alignment training slightly reduces clean accuracy (2.7 percentage points for ResNet-18), it dramatically improves interpretability with an IoU improvement of 0.27 (from 0.15 to 0.42). This represents a trade-off ratio of approximately 10:1 (interpretability gain to accuracy loss), suggesting that alignment may be more compatible with standard classification objectives than adversarial robustness.

Crucially, alignment training does not significantly impact adversarial robustness, indicating that interpretability improvements and adversarial robustness may be largely orthogonal objectives. This orthogonality suggests the possibility of combining alignment and adversarial training to achieve models that are simultaneously robust, interpretable, and reasonably accurate, though such combinations require careful hyperparameter tuning to balance the competing objectives effectively.

5 Discussion

5.1 Implications for AI Safety

My results highlight critical considerations for deploying computer vision systems in safety-critical applications, providing quantitative evidence for design decisions that have historically been based on intuition or limited empirical evidence. The stark trade-off between clean accuracy and adversarial robustness (8 percentage point accuracy loss for 45 percentage point robustness gain) suggests that practitioners must carefully consider threat models when designing robust systems, as the costs of robustness may outweigh the benefits in many real-world scenarios.

For applications where adversarial attacks represent a genuine threat—such as autonomous vehicles facing adversarial road signs, medical imaging systems subject to adversarial manipulation, or biometric authentication systems under attack—adversarial training provides substantial benefits despite the accuracy cost. However, the decision to employ adversarial training should be based on careful threat modeling that considers the likelihood of adversarial attacks, the consequences of successful attacks, and the impact of reduced clean accuracy on normal operations.

The deployment context is crucial for this decision-making process. In closed-loop systems with limited external input (such as factory automation), adversarial robustness may be less critical than in open systems exposed to potentially malicious actors. Similarly, systems with human oversight may be able to tolerate some degradation in automatic performance if it comes with improved robustness to manipulation.

The low interpretability IoU scores for baseline models (¡ 0.16) raise profound concerns about

model reliability that extend beyond simple performance metrics. If models rely on spurious correlations invisible to human oversight—such as focusing on background elements rather than object features—they may fail catastrophically when deployed in environments that differ from their training distribution. This type of failure is particularly dangerous because it may not be detected by standard validation procedures that test performance on i.i.d. data from the same distribution.

My human-guided alignment approach offers a promising direction for improving interpretability without severe performance penalties, suggesting that the apparent tension between interpretability and accuracy may be less fundamental than previously believed. The dramatic IoU improvements (from 0.15 to 0.42 for ResNet-18) with minimal accuracy loss (2.7 percentage points) indicate that models can learn to focus on human-relevant features when provided with appropriate supervision.

5.2 Shortcut Learning and Alignment

The prevalence of shortcut learning in baseline models (IoU ¡ 0.16) confirms longstanding concerns about model alignment in computer vision and provides quantitative evidence for a phenomenon that has been observed qualitatively in numerous studies. This systematic reliance on spurious correlations reveals a fundamental limitation of current training paradigms: optimizing solely for predictive accuracy does not ensure that models learn human-relevant features or develop robust understanding of visual concepts.

The implications of this misalignment extend far beyond academic concerns and pose serious risks for real-world deployment. Models that rely on shortcuts may exhibit excellent performance during development and testing but fail catastrophically when encountering distribution shifts that break their learned correlations. For example, a model that learns to recognize dogs by focusing on grass backgrounds rather than animal features may perform well on standard benchmarks but fail completely when deployed in urban environments.

This problem is particularly acute in domains where fairness and equity are paramount. If models learn to associate demographic characteristics with classification decisions rather than focusing on relevant task features, they may perpetuate or amplify existing societal biases. The low IoU scores suggest that such misaligned learning is the norm rather than the exception in current training practices.

My saliency-alignment loss provides a practical and scalable mechanism for incorporating human feedback into training, addressing this alignment problem through direct supervision of attention mechanisms. The approach leverages the insight that humans can often easily identify which image regions are relevant for classification, even when they cannot articulate the complex decision rules that should govern model behavior.

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, suggesting that alignment problems may be more tractable than previously believed. Crucially, this improvement occurs without requiring fundamental changes to model architectures or training procedures, making the approach readily adoptable in existing systems.

The success of alignment training also provides insights into the nature of the features learned by standard training. The fact that models can quickly adapt to focus on human-relevant regions suggests that they may already learn representations that encode object-level information, but standard training does not provide incentives to use this information consistently. Alignment training may thus be uncovering existing capabilities rather than teaching entirely new behaviors.

5.3 Robustness Across Threat Models

The evaluation across multiple threat models (adversarial attacks, distribution shift) reveals complex and sometimes counterintuitive relationships between different types of robustness, challenging the assumption that robustness is a monolithic property that transfers across threat models. Adversarial training improves targeted attack robustness dramatically (45 percentage points for ResNet-18) but provides only modest benefits for natural distribution shift (0.12 mCE improvement), suggesting that these robustness types may be fundamentally different phenomena requiring distinct approaches.

This limited transfer between adversarial and natural robustness has important theoretical implications for our understanding of learned representations. If adversarial robustness and natural robustness were simply different manifestations of the same underlying property—such as reliance on "robust features"—we would expect stronger positive transfer between them. The weak correlation suggests that adversarial examples and natural corruptions may exploit different vulnerabilities in learned representations.

One possible explanation is that adversarial examples primarily exploit the high sensitivity of neural networks to small input perturbations, while natural corruptions represent larger-scale distributional shifts that require different types of invariance. Adversarial training may improve robustness to the former without addressing the latter, explaining the limited transfer observed in my experiments.

Alternatively, the different threat models may require fundamentally different types of features. Adversarial robustness may benefit from features that are locally stable (insensitive to small perturbations), while natural robustness may require features that are globally stable (consistent across different environmental conditions). Current training methods may struggle to optimize for both types of stability simultaneously.

The relatively modest improvements in corruption robustness from adversarial training (0.12 mCE reduction for ResNet-18) indicate that adversarial examples and natural distribution shift may indeed involve different underlying mechanisms, possibly requiring different training interventions. This finding has important practical implications: organizations seeking robustness to natural distribution shift should not rely solely on adversarial training but should consider complementary approaches such as data augmentation, domain adaptation, or robust optimization methods specifically designed for distributional robustness.

Future work should investigate unified approaches to multiple forms of robustness that can simultaneously address adversarial attacks, natural distribution shift, and other safety concerns. Such approaches might involve multi-objective optimization, careful combination of different training techniques, or novel architectures designed with multiple robustness objectives in mind.

5.4 Architecture-Specific Considerations

ResNet-18 generally achieves higher baseline accuracy (94.8% vs 91.2%) and adversarial robustness (45.2% vs 38.7% PGD accuracy) compared to ViT-Small on CIFAR-10, reflecting fundamental differences in how these architectures process visual information. This performance gap likely reflects the strong inductive biases of convolutional architectures for natural image processing, including translation equivariance, spatial locality, and hierarchical feature extraction that align well with the structure of natural images.

The superior performance of ResNet-18 on CIFAR-10 should not be interpreted as evidence that convolutional architectures are universally superior to transformers for computer vision. Rather, it likely reflects the specific characteristics of CIFAR-10 as a relatively simple dataset with small images (32×32) where spatial locality and translation equivariance provide significant advantages. On larger, more complex datasets, the superior scaling properties and long-range dependency modeling capabilities of transformers may provide countervailing benefits.

Despite these baseline differences, ViT-Small shows comparable improvements from alignment training (IoU improvement from 0.12 to 0.39), suggesting that the alignment approach generalizes effectively across architectures with fundamentally different design principles. This generalization is encouraging for the practical applicability of alignment methods, as it suggests that the core insights about human feedback and attention supervision are not tied to specific architectural choices.

The comparable alignment benefits across architectures also provide insights into the nature of shortcut learning. The fact that both CNNs and transformers suffer from similar alignment problems and benefit similarly from alignment training suggests that shortcut learning may be a fundamental consequence of optimization dynamics rather than an architecture-specific phenomenon.

The interpretability methods employed (Grad-CAM vs. attention rollout) provide complementary insights into model behavior that highlight different aspects of the learned representations. Grad-CAM focuses on spatial regions and provides intuitive heatmaps that show where the model "looks" for CNNs, while attention rollout reveals token-level dependencies and information flow patterns that are unique to transformer architectures.

Both methods benefit substantially from alignment training, indicating improved model interpretability across different analysis approaches and suggesting that the alignment benefits are not artifacts of any particular visualization technique. The consistency of improvements across different interpretability methods strengthens confidence in the genuine nature of the alignment benefits and suggests that the models are indeed learning to focus on more human-relevant features rather than simply gaming the specific IoU metric used for evaluation.

5.5 Limitations and Future Work

Several important limitations constrain the scope and generalizability of my findings, and acknowledging these limitations is crucial for interpreting the results appropriately and identifying directions for future research. First, experiments are limited to CIFAR-10, a relatively simple dataset with small images (32×32) and only 10 classes. While CIFAR-10 provides an excellent controlled environment for systematic investigation, scaling to higher-resolution images with more complex scenes, larger class vocabularies, and more challenging visual reasoning tasks may reveal different trade-offs and challenge the conclusions drawn from this work.

The simplicity of CIFAR-10 may actually favor alignment approaches, as the ground-truth object masks are relatively straightforward to define and the visual concepts are well-separated. More complex datasets like ImageNet, with its 1000 classes and diverse visual concepts, or fine-grained classification tasks where relevant features are subtle and difficult to localize, may present greater challenges for alignment methods. The scalability question extends beyond computational considerations to fundamental questions about the feasibility of human supervision at scale.

Second, my human annotation simulation uses simple heuristics based on thresholding and morphological operations rather than actual human feedback, which may not capture the full complexity of human values and perceptual judgments. Real human annotators bring rich prior knowledge, contextual understanding, and value judgments that cannot be easily captured by automated annotation procedures. Human annotations may also be inconsistent, subjective, and influenced by factors like cultural background, expertise level, and annotation interface design.

Future work should address these limitations through larger-scale experiments on more challenging datasets and genuine human-in-the-loop training that captures the full complexity of human feedback. Such studies would need to address important questions about annotation quality, inter-annotator agreement, the potential for human biases to be encoded into models, and methods for aggregating diverse human perspectives into coherent training signals.

Additionally, investigating the relationship between different robustness objectives (adversarial, distributional, interpretability) remains an important research direction that could lead to unified approaches capable of simultaneously addressing multiple safety concerns. The complex trade-offs observed in this work suggest that achieving multiple robustness objectives simultaneously may require fundamental advances in training methodology or architecture design.

The scalability of human-guided alignment represents perhaps the most crucial limitation for practical deployment. While my approach shows promise on CIFAR-10 with its 50,000 training images, the annotation burden may become prohibitive for larger datasets with millions of images. Modern computer vision systems are increasingly trained on web-scale datasets that would require enormous human annotation efforts if every image required manual attention supervision.

Addressing these scalability challenges will require research into efficient human feedback mechanisms such as active learning (selecting the most informative examples for human annotation), transfer learning approaches that can generalize alignment from small annotated datasets to larger unannotated ones, and methods for leveraging weak supervision signals that can be obtained at scale. Semi-supervised approaches that combine limited human feedback with self-supervised learning may also provide promising directions for scalable alignment.

6 Conclusion

This work presents a comprehensive analysis of adversarial robustness, interpretability, and human-guided alignment in computer vision through the RobustSight framework, providing quantitative insights into fundamental trade-offs that govern the design of safe and reliable computer vision systems. My experiments reveal complex relationships between different safety objectives and highlight the critical importance of considering multiple threat models, evaluation metrics, and deployment constraints in robust system design.

The research contributes both empirical findings and methodological innovations that advance our understanding of safety in computer vision. Through systematic evaluation across multiple architectures, threat models, and training paradigms, I provide evidence-based guidance for practitioners who must navigate the challenging landscape of competing safety objectives in real-world deployments.

Key findings include:

1. **Fundamental Accuracy-Robustness Trade-offs**: 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), establishing a quantitative trade-off ratio of approximately 5.6:1. This trade-off appears fundamental rather than methodological, suggesting intrinsic limitations in current approaches to adversarial robustness.

2. **Pervasive Shortcut Learning**: Baseline models exhibit extensive shortcut learning, with interpretability IoU scores below 0.16, indicating that less than 16% of model attention aligns with human-relevant object features. This systematic misalignment raises profound concerns about model reliability, generalization, and alignment with human values in safety-critical applications.

3. **Promising Alignment Approaches**: Human-guided alignment using saliency-alignment loss dramatically improves interpretability (IoU ¿ 0.38, representing 180%+ improvements) while maintaining competitive accuracy (¡ 3 percentage point reduction), suggesting a much more favorable trade-off profile than adversarial training and offering a viable path toward more aligned vision models.

4. **Limited Cross-Domain Robustness Transfer**: Distribution shift robustness shows only modest improvement from adversarial training (0.07-0.12 mCE reduction), indicating that adversarial robustness and natural robustness may be fundamentally different phenomena requiring distinct defense mechanisms and challenging assumptions about unified robustness.

5. **Architecture-Dependent Safety Characteristics**: Architecture choice significantly influences robustness characteristics, with CNNs showing higher baseline performance (94.8% vs 91.2%) and adversarial robustness (45.2% vs 38.7%) compared to Vision Transformers on CIFAR-10, highlighting the importance of architectural considerations in safety-oriented system design.

6. **Orthogonal Safety Objectives**: Alignment and adversarial robustness appear to be largely orthogonal objectives, suggesting the possibility of combining approaches to achieve models that are simultaneously robust, interpretable, and accurate, though such combinations require careful optimization to balance competing objectives.

These results provide crucial insights for developing safer and more aligned computer vision systems, with immediate implications for both research directions and deployment practices. The demonstrated feasibility of human-guided alignment training offers a practical and scalable approach for incorporating human values into model training without the severe performance penalties associated with adversarial robustness.

However, significant challenges remain in scaling these approaches to larger datasets, more complex real-world scenarios, and genuine human-in-the-loop training systems. Future research must address fundamental questions about the scalability of human feedback, the development of unified approaches to multiple robustness objectives, and the theoretical understanding of trade-offs between different safety goals.

As computer vision systems become increasingly deployed in safety-critical applications—from autonomous vehicles and medical diagnosis to content moderation and financial decision-making—addressing these fundamental challenges becomes essential for maintaining public trust and ensuring beneficial outcomes. My work contributes to this critical effort by providing systematic evaluation methodologies, quantitative characterization of key trade-offs, and demonstrating promising directions for improving model robustness, interpretability, and alignment with human values. The RobustSight framework offers a replicable approach for future research in this vital area of AI safety.

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.