INTERPRETABLE ADVERSARIAL DEFENSE VIA CLASS-AWARE NEURON SENSITIVITY ATTRIBUTION AND RECALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Research on analyzing and enhancing model robustness at the neuron level has attracted significant attention due to its interpretability. However, existing methods are either based on neuron sensitivity evaluated by changes in activation values, or based on neuron importance measured by contributions to clean samples. Neither approach considers the intrinsic connection between the changes in neuron activations and the neural network's outputs. To bridge this gap, we propose a novel Neuron Sensitivity Attribution (NSA) method. By jointly calculating the dual contributions of intermediate neurons in suppressing the true label and promoting the adversarial label, this method provides a fine-grained, category-specific evaluation of neuron sensitivity under adversarial perturbations. Based on this metric, we propose a Sensitivity-Aware Neuron Recalibration mechanism to reform adversarial training strategies. Specifically, during adversarial training, stronger penalties are assigned to highly sensitive neurons, thereby enabling the model to rely more heavily on robust neurons. Extensive experiments across various models and adversarial training frameworks demonstrate that the proposed method can serve as an effective plug-and-play module to further enhance the robustness of deep learning models.

1 INTRODUCTION

Although deep neural networks have achieved massive success and been widely applied in numerous fields (Zhao et al., 2024; Mohamed et al., 2012; Chen & Manning, 2014; Gupta et al., 2019), their inherent lack of interpretability has raised severe security concerns. Extensive research has demonstrated that neural networks exhibit high vulnerability to adversarial attacks (Goodfellow et al., 2015; Carlini & Wagner, 2017): introducing merely minute and imperceptible perturbations to the input can cause the model to produce incorrect predictions.

To enhance model robustness against adversarial attacks, the academic community has proposed diverse defense mechanisms, primarily including adversarial training (Carlini & Wagner, 2017; Zhang et al., 2019; Cui et al., 2024), data augmentation (Zhang et al., 2017; Guo et al., 2018; Xie et al., 2018), and certified robustness (Zhang et al., 2018; Li et al., 2023). Meanwhile, some studies are dedicated to analyzing and enhancing model robustness at the neuron level in pursuit of better mechanistic interpretability. These existing methods can be mainly categorized into two classes: one is based on the sensitivity of neurons under adversarial perturbations (Zhang et al., 2021; Huang et al., 2024; Lian & Tian, 2023; Kim et al., 2023), another is based on the importance of neurons in predicting the true label (Bai et al., 2021; Yan et al., 2021; Won et al., 2025).

However, current sensitivity-based methods mostly evaluate neurons by simply calculating absolute activation changes. This magnitude-only approach is fundamentally flawed because intermediate activation shifts are highly complex; they encapsulate not only factors that suppress the ground-truth label or promote incorrect ones, but also mixed components exerting both effects simultaneously (Xu et al., 2019). Meanwhile, importance-based methods rely heavily on decision contributions to clean samples, ignoring neurons' susceptibility to adversarial manipulation. Crucially, both approaches fail to capture the intrinsic causal connection between intermediate activation variations and ultimate predictive deviations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address the aforementioned limitations, this paper introduces a novel Neuron Sensitivity Attribution (NSA) method. Unlike previous methods that provide a coarse, class-agnostic evaluation, NSA quantifies neuron sensitivity at a fine-grained, category-specific level. By jointly calculating the dual contributions of intermediate neurons in suppressing the true label and promoting the adversarial label, this method precisely characterizes their sensitivity under adversarial perturbations with respect to specific semantic classes. Consequently, it establishes a direct causal link between the changes in neuron activation values and the variations in model outputs, deeply revealing the internal adversarial sensitivity mechanism of neural networks.

Specifically, rather than utilizing attribution methods to merely evaluate neuron importance on clean samples, we extend this paradigm to quantify neuron sensitivity. Drawing inspiration from the core philosophy of DeepLIFT (Shrikumar et al., 2017), we define the clean sample as the reference point and the adversarial example as the current state. Under this formulation, NSA is capable of backpropagating the dual contributions—suppressing the ground-truth label and promoting the adversarial label—as attribution values to the intermediate-layer neurons. Based on this methodology, we can precisely calculate the sensitivity score of individual neurons with respect to specific class labels.

By conducting masking experiments on highly sensitive neurons, we observe a significant rebound in the ground-truth logits for adversarial examples, alongside a clear realignment of spatial attention back to semantically meaningful regions. This phenomenon demonstrates that these neurons serve as the primary targets of adversarial attacks. However, these experiments also reveal a critical limitation: rigid hard masking can lead to an accuracy collapse due to the severe entanglement between sensitive and core semantic neurons. This indicates that simply discarding these vulnerable neurons causes fatal semantic information loss, necessitating a more nuanced intervention.

Furthermore, we propose a Sensitivity-Aware Neuron Recalibration (SANR) training strategy. Based on the neuron sensitivity evaluated by NSA, this strategy applies differential penalty weights to varying neurons during adversarial training. By actively suppressing the highly sensitive neurons in the model, it accomplishes a deep recalibration of the feature space, thereby enabling the model to rely more heavily on robust neurons.

Seamlessly integrating SANR into existing adversarial training frameworks significantly enhances their adversarial robustness. This further validates the crucial role of our sensitivity attribution method in analyzing and understanding model robustness.

The main contributions of our work are summarized as follows:

- We introduce the Neuron Sensitivity Attribution (NSA) method. Unlike existing coarse and class-agnostic metrics, NSA provides a fine-grained, category-specific evaluation of neuron sensitivity under adversarial perturbations. By simultaneously quantifying the dual contributions of intermediate neurons in suppressing the ground-truth label and promoting the adversarial label, this formulation successfully establishes a direct causal link between intermediate activation shifts and ultimate predictive deviations.

- We demonstrate that highly sensitive neurons constitute the core vulnerabilities exploited by adversarial perturbations. Through ablation studies, we show that masking these neurons successfully drives a rebound in ground-truth logits and realigns spatial attention. Furthermore, this masking process exposes a critical phenomenon: the severe entanglement between sensitive and core semantic neurons. This insight reveals the inherent limitations of rigid hard masking and directly motivates our soft recalibration strategy.

- We propose a Sensitivity-Aware Neuron Recalibration (SANR) training strategy. Leveraging the sensitivity metrics derived from NSA, this strategy applies differential penalties to individual neurons during adversarial training, thereby actively guiding the model to rely more heavily on inherently robust neurons. Extensive experiments demonstrate that this plug-and-play strategy can be seamlessly integrated into existing mainstream adversarial training frameworks, significantly enhancing their overall robustness.

2 Related Work

Methods for analyzing and enhancing model robustness at the neuron level can be broadly categorized into two classes. The first category is based on the sensitivity of neurons to input perturbations. Zhang et al. (2021) pioneered the interpretation of adversarial examples from this perspective, evaluating neuron sensitivity by comparing activation differences between clean and adversarial inputs. They further enhanced robustness by constraining the activation values of the top-$k\%$ most sensitive neurons. Following a comparable rationale, Huang et al. (2024) evaluated neuron sensitivity based on activation variations caused by input mutations (e.g., rotation, shearing, blurring) to guide the generation of test cases. Furthermore, Lian & Tian (2023) measured sensitivity through the magnitude of change in a neuron's output range relative to the input perturbation radius, combining this with importance scores derived via the LRP algorithm (Bach et al., 2015) to identify vulnerable neurons. Kim et al. (2023) introduced the FSR module, which effectively bolsters adversarial robustness by disentangling and restoring the activations of non-robust (i.e., highly sensitive) neurons.

The second category of methods is based on the importance of neurons to the model's final prediction. For instance, Bai et al. (2021) introduced the CAS method, which dynamically evaluates the contribution of individual neurons during training and subsequently suppresses the activations of low-contribution ones. Building upon this neuron-level perspective, Yan et al. (2021) proposed the CIFS method to calculate each neuron's contribution to class prediction during training, amplifying high-contribution neurons while suppressing low-contribution ones. Additionally, Won et al. (2025) utilized Shapley values to quantify the precise contribution of individual neurons to the final prediction, ultimately enhancing the model's overall robustness by reducing its reliance on these critical components.

3 Methodology

In this section, we elaborate on the proposed methodology. First, we introduce Neuron Sensitivity Attribution (NSA) to precisely quantify internal neuron sensitivity under adversarial perturbations. Based on this metric, we further propose Sensitivity-Aware Neuron Recalibration (SANR) training, which effectively enhances adversarial robustness by applying weighted inhibition to highly sensitive neurons.

3.1 Neuron Sensitivity Attribution

As the first attribution analysis tailored specifically for neuron sensitivity, NSA comprehensively characterizes the dual effects of neurons under attack: suppressing the ground-truth label and promoting the adversarial label. Specifically, by separately calculating intermediate neurons' contributions to the logit variations of both classes and jointly evaluating them, we precisely quantify the sensitivity score for each individual neuron. Figure 1 illustrates the overall computation flow of NSA.

3.1.1 Revisiting the Definition of DeepLIFT

DeepLIFT (Deep Learning Important FeaTures) is a reference-based attribution method designed to explain the outputs of deep neural networks by comparing the activation states of neurons against a specific "reference state."

Let t and x_i ($i \in [1, n]$) denote the target output neuron and intermediate neurons, with t^0 and x_i^0 as their respective reference activations. DeepLIFT defines the difference-from-reference as $\Delta t = t - t^0$ and $\Delta x_i = x_i - x_i^0$. Its core summation-to-delta property assigns a contribution score $C_{\Delta x_i \Delta t}$ to each neuron, ensuring:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t. \tag{1}$$

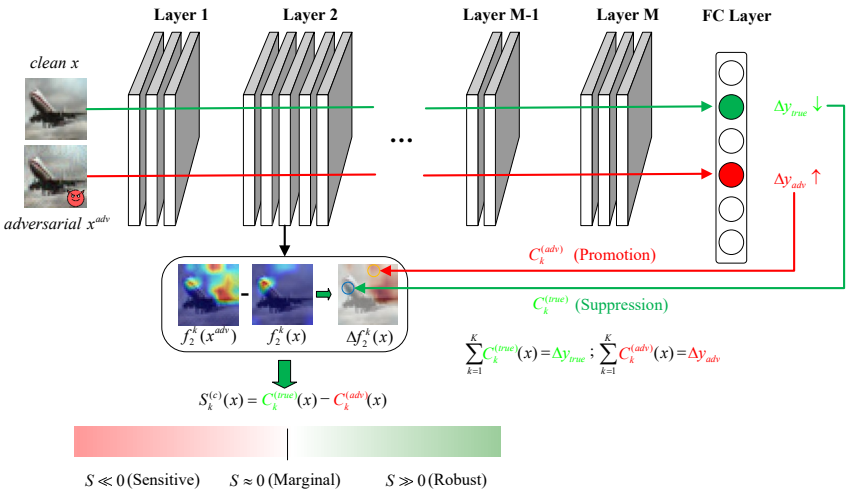


Figure 1: The specific computation flow of the proposed Neuron Sensitivity Attribution (NSA).

For efficient layer-wise backpropagation, DeepLIFT introduces a multiplier $m_{\Delta x,\Delta t} = \frac{C_{\Delta x,\Delta t}}{\Delta x}$, which obeys the chain rule:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta t}, \tag{2}$$

where y_j are direct downstream neurons of x_i.

Crucially, for non-linear computations, DeepLIFT decouples signal variations into positive (Δ^+) and negative (Δ^-) components and traces them independently. This prevents the implicit cancellation of these opposing components, forming the mathematical foundation for our independent quantification of adversarial dual effects.

3.1.2 Dual-Target Attribution

We propose using the clean sample x as the "reference state" and its adversarial counterpart x_{adv} as the input. This allows us to measure the state variations induced by the adversarial example—namely, the suppression of the ground-truth label and the promotion of the adversarial label (Xu et al., 2019)—and backpropagate them to intermediate neurons. Unlike traditional single-label saliency maps, we construct a "Dual-Target Attribution" mechanism. For an intermediate layer L with K neurons, DeepLIFT satisfies the summation-to-delta property for a target class c:

$$\sum_{k=1}^{K} C_k^{(c)} = f_c(x_{\text{adv}}) - f_c(x), \tag{3}$$

where $C_k^{(c)}$ denotes the contribution of the k-th neuron to the logit variation of class c.

To independently quantify the disruption of the ground-truth class and the promotion of the adversarial class, we perform two separate attribution backpropagations per sample. Specifically: (1) **Ground-Truth Contribution ($C_k^{(\textbf{true})}$):** backward attribution starts from the ground-truth logit y_{true} to extract each neuron's contribution to its confidence decline; (2) **Adversarial Contribution ($C_k^{(\textbf{adv})}$):** backward attribution starts from the logit of the adversarial class $y_{\text{adv}} = \arg\max f(x_{\text{adv}})$ to quantify each neuron's contribution to the adversarial confidence increase. Finally, these attribution scores are spatially summed over dimensions (H, W) to obtain the macroscopic neuron-level contribution.

3.1.3 Neuron Sensitivity Quantification

Neuron sensitivity is a composite result of ground-truth disruption and adversarial promotion. Since C_{true} and C_{adv} implicitly encompass logit variations via the summation-to-delta property, we first define the joint sensitivity score for the k-th neuron given a single sample x as:

$$S_k^{(c)}(x) = C_k^{(\text{true})}(x) - C_k^{(\text{adv})}(x). \tag{4}$$

This metric intuitively categorizes neurons based on their distinct response patterns: (1) **Sensitive Neuron** ($S_k^{(c)}(x) \ll 0$): $C_k^{(\text{adv})}$ is significantly positive while $C_k^{(\text{true})}$ is negative, indicating the neuron's response to ground-truth features is severely suppressed, while it strongly promotes the adversarial class; (2) **Robust Neuron** ($S_k^{(c)}(x) \gg 0$): $C_k^{(\text{true})}$ remains positive and unaffected by adversarial noise ($C_k^{(\text{adv})} \leq 0$), demonstrating strong anti-interference capability; (3) **Marginal Neuron** ($S_k^{(c)}(x) \approx 0$): The neuron is insensitive to both ground-truth and adversarial signals, or their opposing effects cancel each other out.

To eliminate sample-specific bias, we calculate the expected sensitivity $S_k^{(c)}$ for the k-th neuron across all training samples of class c ($\mathcal{D}_c$):

$$S_k^{(c)} = \mathbb{E}_{x \sim \mathcal{D}_c}\left[S_k^{(c)}(x)\right]. \tag{5}$$

3.2 Sensitivity-Aware Neuron Recalibration (SANR)

We propose Sensitivity-Aware Neuron Recalibration (SANR) to enhance adversarial robustness. By actively suppressing neurons based on prior sensitivity scores, SANR deeply recalibrates the feature space, forcing the model to shift its reliance from vulnerable dependencies to robust core neurons during optimization.

Sensitivity Weight Generation. To smoothly integrate the quantified neuron sensitivity into the subsequent optimization phase, it is necessary to apply non-linear scaling. To this end, we introduce a temperature coefficient γ to compute the final weight mask $\alpha_k^{(c)} \in [-1, 1]$ for the k-th neuron:

$$\alpha_k^{(c)} = \tanh(\gamma \cdot S_k^{(c)}). \tag{6}$$

These class-wise coefficients form a mask vector $\boldsymbol{\alpha}^{(c)}$ that serves directly to guide the robustness optimization of the network.

Neuron-Level Recalibration. Traditional feature alignment usually forces adversarial features to perfectly fit clean features (i.e., $f_{\text{adv}} \to f_{\text{clean}}$), but this ignores the vulnerability discrepancy exhibited by different neurons when facing adversarial attacks. To address this, we innovatively introduce a neuron-level modulation factor based on $\boldsymbol{\alpha}^{(y)}$, applying differentiated penalty weights to individual neurons to construct safe target features $\tilde{f}_{\text{target}}^{(l)}$ at layer l:

$$\tilde{f}_{\text{target}}^{(l)} = f_{\text{clean}}^{(l)} \odot \left(1 + \lambda \boldsymbol{\alpha}^{(y)}\right), \tag{7}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\alpha}^{(y)}$ is the sensitivity mask vector based on the ground-truth label y, and λ is a hyperparameter controlling the intervention intensity. Essentially, this term acts as a neuron-specific modulator that selectively suppresses sensitive neurons ($\alpha_k^{(y)} < 0$) while amplifying the activation of robust ones ($\alpha_k^{(y)} > 0$).

Recalibration Loss. To guide internal neuron activations towards these robust targets, we compute the L_1 distance across M target mounted layers as the recalibration loss L_{recal}:

$$L_{\text{recal}} = \frac{1}{M} \sum_{l=1}^{M} \left\| f_{\text{adv}}^{(l)} - \tilde{f}_{\text{target}}^{(l)} \right\|_1. \tag{8}$$

Joint Optimization Objective. Finally, we integrate this intrinsic recalibration with a base adversarial training loss (L_{base}) to construct the complete robust optimization objective:

$$L_{\text{total}} = L_{\text{base}} + \beta \cdot L_{\text{recal}}, \tag{9}$$

where β balances the recalibration intensity. This joint optimization enables the network to resist output-level deception while acquiring intrinsic structural security via neuron-level self-suppression and reinforcement. The complete training procedure is detailed in Algorithm 1 of Appendix A.1.

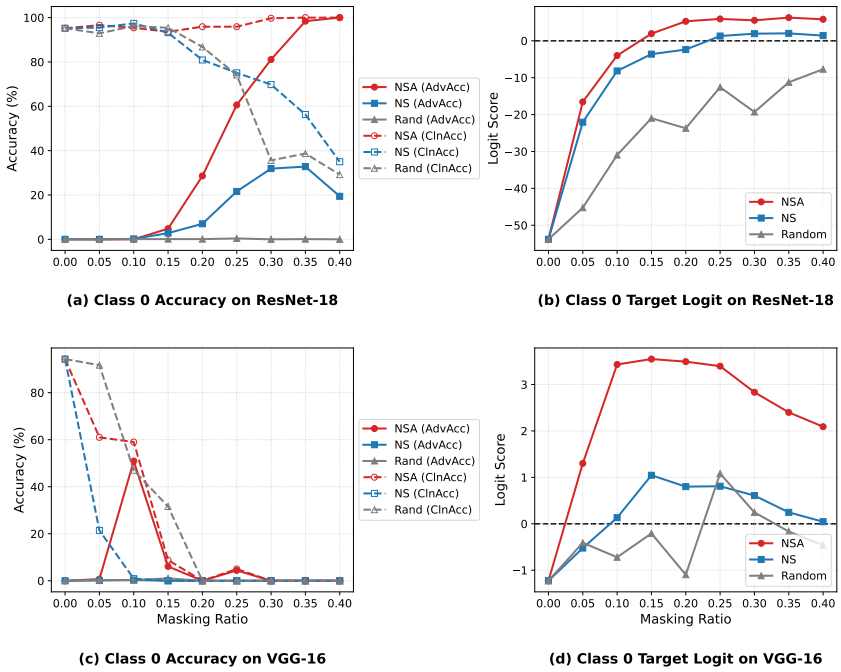


Figure 2: Defense performance evaluation of different neuron masking strategies. (a) and (b) compare post-masking adversarial accuracy and ground-truth logit evolution on ResNet-18. (c) and (d) show the corresponding results on VGG-16. While NSA-guided masking optimally balances robustness and generalization on ResNet-18, the accuracy collapse at higher masking ratios on VGG-16 (shown here for Class 0) exposes the severe entanglement between sensitive and core semantic neurons.

4 Validation of Sensitivity Attribution Effectiveness

To comprehensively validate the effectiveness of the proposed NSA method, we evaluate the quantified neuron sensitivity across two complementary dimensions: (1) a quantitative neuron masking evaluation to assess how suppressing highly sensitive neurons restores model accuracy, and (2) a qualitative Grad-CAM (Selvaraju et al., 2017) spatial attribution analysis to reveal how this masking guides the model's attention to refocus on the target class.

4.1 Experimental Setup

We evaluate the VGG-16 (Simonyan & Zisserman, 2015) and ResNet-18 (He et al., 2016) architectures on the CIFAR-10 (Krizhevsky et al., 2009) dataset, which serves as a standard benchmark in adversarial robustness to ensure fair baseline comparisons.

To ensure clarity, we standardize the layer naming conventions across both architectures. Specifically, VGG-16 comprises five convolutional blocks (B1–B5) with internal layers denoted as *conv1*, *conv2*, etc., followed by fully connected layers (*fc1–fc3*). ResNet-18 consists of an initial stem (initial convolution and activation), four residual blocks (*Layer1–Layer4*), and a final fully connected layer (*fc*). Based on these structures, our neuron masking evaluations in this section are applied to all 13 convolutional layers of VGG-16, and the four primary residual blocks (*Layer1–Layer4*) of ResNet-18, strictly excluding the fully connected layers and the final classification head.

4.2 Quantitative Evaluation via Neuron Masking

In this ablation study, we zero out the top-$k\%$ class-specific sensitive neurons (evaluated by NSA) for each adversarial example to observe adversarial accuracy recovery. We compare against random masking and NS-based masking (Zhang et al., 2021) (which relies on absolute activation variations

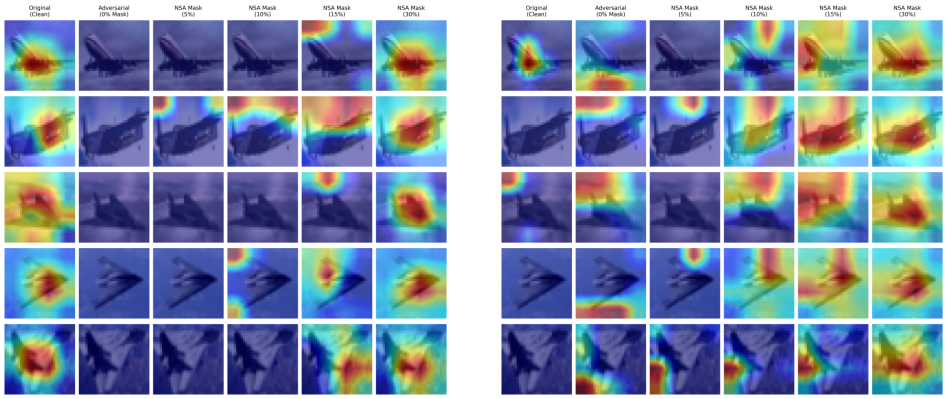


Figure 3: Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 Class 0 (Airplane). (a) On ResNet-18 (Layer4), the suppressed attention successfully re-emerges from a globally dormant state and refocuses on the target's core semantics, restoring accuracy. (b) On VGG-16 (B4_conv3), a paradoxical decoupling phenomenon occurs: while progressive masking visually realigns attention with the core semantics and corrects the misaligned or vanished initial heatmaps caused by shortcut bias, the actual classification accuracy collapses.

and is similarly evaluated class-wise for fairness). Figure 2 illustrates results for CIFAR-10 Class 0 on ResNet-18 and VGG-16; comprehensive results for the remaining classes, which exhibit notable class-wise variances, are detailed in Appendix A.2.

NSA-guided masking on ResNet-18 steadily increases adversarial accuracy, reaching 100% at a 40% ratio (Figure 2a). Regarding logit dynamics (Figure 2b), the ground-truth logit rebounds from -53.81 to -3.99 during the initial 0%–10% phase despite near-zero accuracy, and turns positive (+1.94) at the 15% tipping point, initiating accuracy recovery (4.80%). Conversely, NS-based masking yields limited robustness (32.80% peak) and degrades clean accuracy to 56.30%, as absolute variations fail to distinguish ground-truth suppression from adversarial promotion. Random masking destroys normal feature extraction (clean accuracy 29.20%) with persistently negative logits. Notably, NSA restores robustness while maintaining $> 95\%$ clean accuracy, acting as highly efficient pruning of "non-robust features" (Ilyas et al., 2019) without disrupting core semantics.

Evaluating across topologies, VGG-16 (Figure 2c) shows NSA masking surging adversarial accuracy from 0% to 50.90% at a 10% ratio. The ground-truth logit (Figure 2d) climbs from -1.22 to +1.30 at 5% masking (crossing the decision boundary) and peaks at +3.54 at 15%. However, clean accuracy simultaneously drops to 61.00% at the 5% ratio. Beyond 10%, aggressive masking causes a near-complete network collapse. We attribute this to severe entanglement between highly sensitive and core semantic neurons; hard masking thus inevitably causes fatal semantic information loss while removing adversarial noise. The NS baseline incorrectly masks core semantics, yielding merely 0.3% peak adversarial accuracy and devastating clean accuracy (21.4%) at 5% masking.

In summary, this ablation study demonstrates that adversarial impacts are highly polarized across neurons, validating NSA's capability in precisely quantifying neuron sensitivity. Furthermore, the VGG-16 results reveal the inherent limitations of hard masking. As exemplified by the accuracy collapse in Class 0, rigid masking exposes a severe entanglement between sensitive and core semantic neurons—a vulnerability that exhibits notable class-wise variance (as extensively detailed in Appendix A.2). This insight serves as the primary motivation for our Sensitivity-Aware Neuron Recalibration (SANR) training mechanism, which seeks to softly suppress sensitive neurons without discarding critical semantic information.

4.3 Qualitative Analysis via Spatial Attribution

To further investigate the internal mechanisms of the NSA strategy within the model, we utilize Grad-CAM to extract class activation heatmaps from the deep network layers (Layer4 of ResNet-18 and B4_conv3 of VGG-16), conducting a qualitative analysis of the multi-layer neuron masking

experiments from the perspective of spatial attribution. Taking CIFAR-10 Class 0 as a representative example, this section uses five randomly selected images to illustrate the dynamic evolution of the model's attention distribution when sensitive neurons are forcibly masked at various ratios (5%, 10%, 15%, 30%) under PGD attacks (Madry et al., 2018) (10 steps, step size $\alpha = 2/255$), as depicted in Figure 3(a) and (b) (comprehensive visualization results for the remaining classes, which further reveal notable class-wise variances, are detailed in Appendix A.3)

For ResNet-18 (Figure 3(a)), the clean sample's heatmap clearly focuses on the target's core semantics. Conversely, when the adversarial sample is input without any neuron masking (0% Mask), target-class activations collapse, and the heatmap exhibits a globally dormant state. This indicates that adversarial perturbations thoroughly suppress ground-truth neuron activations. As the highly sensitive neurons evaluated by NSA are progressively masked, the misled attention steadily regroups toward the target region. Notably, at a 30% masking ratio, the focus almost perfectly reverts to the core semantics.

Interestingly, for VGG-16 (Figure 3(b)), the clean sample heatmaps for certain images exhibit incorrect focal shifts or vanish entirely, revealing that the vanilla model relies on non-robust shortcuts (Ilyas et al., 2019) rather than capturing the primary object. While progressive masking seemingly corrects this bias by visually realigning attention with the true semantic regions at a 30% ratio, this visual correction is accompanied by a near-complete collapse in classification accuracy (as demonstrated quantitatively in Section 4.2). We attribute this paradox to Grad-CAM's normalization: while relative spatial gradients localize on core semantics, aggressive hard masking fatally destroys the absolute magnitude of feature vectors required by the fully connected layers. This profound "spatial-decision decoupling", perfectly mirroring the Class 2 anomaly (Appendix A.3), proves that rigid masking inevitably causes fatal semantic loss, thereby necessitating our flexible SANR mechanism.

5 Robustness Evaluation of Sensitivity-Aware Neuron Recalibration

In this section, we evaluate the robustness improvements achieved by integrating the proposed SANR strategy into existing adversarial training frameworks, utilizing both ResNet-18 and VGG-16 architectures on the CIFAR-10 dataset.

5.1 Experimental Setup

Adversarial Attacks. We evaluate robustness against FGSM (Goodfellow et al., 2015), PGD (20 and 100 steps, step size $\alpha = 2/255$) (Madry et al., 2018), C&W (optimized via PGD for 10 steps) (Carlini & Wagner, 2017), and Auto-Attack (Croce & Hein, 2020) under the ℓ_∞ norm ($\epsilon = 8/255$).

Adversarial Training Frameworks. We integrate the proposed SANR strategy into three baselines: AT (Madry et al., 2018), TRADES (Zhang et al., 2019), and DKL (Cui et al., 2024), comparing the robustness before and after integration. For all models, we report the robustness evaluated at the final epoch of training in the main text. Additionally, further details on parameter settings, the evaluation results at the best-performing epoch, and supplementary ablation studies are provided in Appendix A.4, A.5, and A.6, respectively.

5.2 Robustness Enhancement on Adversarial Training

As shown in Table 1, integrating the proposed SANR strategy achieves significant robustness improvements across various network architectures and adversarial training frameworks under diverse attack scenarios. This thoroughly demonstrates the universality of the proposed method, indicating that it can serve as an effective plug-and-play module to enhance the defense capabilities of existing frameworks. More importantly, even when applied to a highly optimized framework like DKL, our method yields further improvements (e.g., boosting PGD-20 accuracy by 1.05% and Auto-Attack robustness by 0.44% on ResNet-18), pushing the boundaries of existing defense capabilities and elevating the model's overall robustness to new heights. Crucially, the consistent gains observed under the rigorous Auto-Attack evaluation compellingly demonstrate that the obtained defense capability is substantial, rather than a false sense of robustness caused by overfitting effects or gradient obfuscation.

Table 1: Performance comparison of different adversarial training frameworks before and after integrating SANR on CIFAR-10. Best results are highlighted in bold.

Model	Method	Accuracy (%)					
		Clean	FGSM	PGD-20	PGD-100	C&W	AA
ResNet-18	AT	85.12	57.73	49.90	49.45	55.41	46.64
	AT + SANR	**85.40**	**58.49**	**50.42**	**49.96**	**56.15**	**47.26**
	TRADES	83.26	59.19	52.38	52.11	57.65	49.32
	TRADES + SANR	**83.71**	**59.28**	**53.28**	**52.82**	**58.83**	**49.78**
	DKL	**87.23**	59.67	52.87	52.55	57.76	49.75
	DKL + SANR	86.97	**60.58**	**53.92**	**53.49**	**58.34**	**50.19**
VGG-16	AT	**81.63**	54.59	46.55	46.01	55.44	42.82
	AT + SANR	81.51	**54.70**	**47.08**	**46.64**	**55.94**	**43.20**
	TRADES	81.12	55.49	48.38	47.90	56.36	44.94
	TRADES + SANR	**81.28**	**55.76**	**48.79**	**48.41**	**56.75**	**45.36**
	DKL	85.18	56.54	49.07	48.63	56.61	44.85
	DKL + SANR	**85.41**	**56.85**	**49.79**	**49.26**	**57.14**	**45.27**

Table 2: Robustness comparison between SANR and existing neuron-level defense methods (ResNet-18 trained under AT on CIFAR-10). Best results are in bold.

Method	Accuracy (%)					
	Clean	FGSM	PGD-20	PGD-100	C&W	AA
AT	85.12	57.73	49.90	49.45	55.41	46.64
+ SNS	84.00	57.91	50.15	49.74	56.14	46.63
+ CAS	**86.41**	53.98	44.89	43.14	50.26	44.39
+ CIFS	83.02	58.02	48.68	48.11	51.05	44.67
+ FSR	83.06	57.30	**50.73**	**50.25**	51.79	46.33
+ SANR	85.40	**58.49**	50.42	49.96	**56.15**	**47.26**

5.3 Comparison with Existing Methods

To further validate the effectiveness of SANR, we compare our strategy against existing neuron-level defenses (SNS (Zhang et al., 2021), CAS (Bai et al., 2021), CIFS (Yan et al., 2021) and FSR (Kim et al., 2023)) in Table 2. Notably, while FSR achieves the highest accuracy under PGD attacks, it suffers a severe performance drop under C&W and the rigorous Auto-Attack. Moreover, this localized PGD gain comes at a significant cost to clean accuracy, which drops to 83.06%. This discrepancy exemplifies overfitting or gradient obfuscation against specific attack patterns. In contrast, SANR maintains a highly competitive clean accuracy (85.40%) and achieves the best defense performance under the most rigorous Auto-Attack evaluation. This confirms that by softly recalibrating sensitive neurons, SANR successfully avoids localized overfitting, providing genuine and well-rounded robustness.

6 Conclusion

In this paper, we investigate adversarial robustness from an internal neuron perspective and propose the Neuron Sensitivity Attribution (NSA) method. By jointly calculating the dual contributions of neurons in suppressing the true label and promoting the adversarial label, NSA provides a fine-grained, category-specific evaluation of sensitivity, successfully establishing a direct causal link between intermediate activations and model outputs. Based on this, we propose the Sensitivity-Aware Neuron Recalibration (SANR) strategy, which applies differential penalties to highly sensitive neurons, compelling the model to rely on intrinsically robust features. Extensive experiments demonstrate that this plug-and-play module significantly enhances existing adversarial training frameworks.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, grammar correction, and improving the overall fluency of the manuscript. We have not used generative AI tools for generating core research ideas, formulating scientific claims, deriving theoretical analyses, or fabricating experimental results, and the generation of primary datasets is not applicable to this work.

Additionally, we used generative AI tools for assisting with utility script generation and debugging within our PyTorch experimental framework. We have reviewed all AI-assisted work. Specifically, we carefully proofread all AI-edited text to ensure the original scientific meaning and technical nuances of our methodology were accurately preserved. Furthermore, all AI-assisted code snippets were thoroughly reviewed, executed, and empirically validated by the authors to ensure correctness during the neuron sensitivity attribution and robustness evaluation processes. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research focuses on analyzing neuron sensitivity to enhance the adversarial robustness of deep neural networks from an interpretability perspective. While the analysis of neuron sensitivity inherently involves dual-use considerations, our primary objective is strictly defensive: to develop neuron sensitivity attribution and recalibration techniques that mitigate these risks and protect AI models against malicious exploitation. All empirical evaluations in this study are conducted on standard, publicly available benchmark datasets. This research does not involve human subjects, nor does it collect, process, or release any personally identifiable information (PII). We believe our contributions adhere to standard research ethics guidelines and will promote the development of safer, more transparent, and trustworthy artificial intelligence systems. Furthermore, to ensure research integrity and reproducibility, the source code and pre-trained model weights associated with our proposed defense methodology will be made publicly available upon publication.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our empirical results, we have provided the complete PyTorch source code in the supplementary materials, including the implementations of our neuron sensitivity attribution and recalibration methods. Detailed experimental settings—such as exact hyperparameter configurations, specific target layers for calibration, and the parameters for the adversarial attack benchmarks (e.g., PGD, CW, and AutoAttack)—are comprehensively described in Sections 4.1 and 5.1, as well as in Appendix A.4. Furthermore, the formal mathematical definitions and algorithmic details of our proposed neuron sensitivity attribution and recalibration methods are thoroughly documented in the main text.

REFERENCES

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=zQTezqCCtNx.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.49. URL https://doi.org/10.1109/SP.2017.49.

Danqi Chen and Christopher D. Manning. A fast and accurate dependency parser using neural networks. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October*

25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 740–750. ACL, 2014. doi: 10.3115/V1/D14-1082. URL https://doi.org/10.3115/v1/d14-1082.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/87ee1bbac4635e7c948f3eea83c1f262-Abstract-Conference.html.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

Rahul Gupta, Aditya Kanade, and Shirish K. Shevade. Neural attribution for semantic bug-localization in student programs. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11861–11871, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f29a179746902e331572c483c45e5086-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Dong Huang, Qingwen Bu, Yichao Fu, Yuhao Qing, Xiaofei Xie, Junjie Chen, and Heming Cui. Neuron sensitivity-guided test case selection. *ACM Trans. Softw. Eng. Methodol.*, 33(7):188:1–188:32, 2024. doi: 10.1145/3672454. URL https://doi.org/10.1145/3672454.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 125–136, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration for adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 8183–8192. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00791. URL https://doi.org/10.1109/CVPR52729.2023.00791.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pp. 1289–1310. IEEE, 2023. doi: 10.1109/SP46215.2023.10179303. URL https://doi.org/10.1109/SP46215.2023.10179303.

Zhichao Lian and Fengjun Tian. Deepsi: a sensitive-driven testing samples generation method of whitebox cnn model for edge computing. *Tsinghua Science and Technology*, 29(3):784–794, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Abdel-rahman Mohamed, George E. Dahl, and Geoffrey E. Hinton. Acoustic modeling using deep belief networks. *IEEE Trans. Speech Audio Process.*, 20(1):14–22, 2012. doi: 10.1109/TASL.2011.2109382. URL https://doi.org/10.1109/TASL.2011.2109382.

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 *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Soyoun Won, Hyeon Bae Kim, Yong Hyun Ahn, Hong Joo Lee, and Seong Tae Kim. Understanding adversarial robustness of deep neural networks via decision reliance. *Image Vis. Comput.*, 163: 105743, 2025. doi: 10.1016/J.IMAVIS.2025.105743. URL https://doi.org/10.1016/j.imavis.2025.105743.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Sk9yuql0Z.

Kaidi Xu, Sijia Liu, Gaoyuan Zhang, Mengshu Sun, Pu Zhao, Quanfu Fan, Chuang Gan, and Xue Lin. Interpreting adversarial examples by activation promotion and suppression, 2019. URL https://arxiv.org/abs/1904.02057.

Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Y. F. Tan, and Masashi Sugiyama. CIFS: improving adversarial robustness of cnns via channel-wise importance-based feature selection. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11693–11703. PMLR, 2021. URL http://proceedings.mlr.press/v139/yan21e.html.

Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Trans. Image Process.*, 30:1291–1304, 2021. doi: 10.1109/TIP.2020.3042083. URL https://doi.org/10.1109/TIP.2020.3042083.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 7472–7482. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 4944–4953, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d04863f100d59b3eb688a11f95b0ae60-Abstract.html.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017. doi: 10.1109/TIP.2017.2662206. URL https://doi.org/10.1109/TIP.2017.2662206.

Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan D. Parmar. A review of convolutional neural networks in computer vision. *Artif. Intell. Rev.*, 57 (4):99, 2024. doi: 10.1007/S10462-024-10721-6. URL https://doi.org/10.1007/s10462-024-10721-6.

A Appendix

A.1 Complete Training Algorithm

Herein, we present the complete training procedure of the proposed Sensitivity-Aware Neuron Recalibration (SANR) in Algorithm 1.

Algorithm 1 Sensitivity-Aware Neuron Recalibration (SANR) robust training

Require: Training dataset $\mathcal{D} = \{(x_i, y_i)\}$, initial model F_θ, perturbation budget ϵ, balance coefficient β, intervention intensity λ, pre-computed sensitivity masks $\boldsymbol{\alpha}$, learning rate η, number of target layers M

Ensure: Robust model F_θ^*

for epoch $= 1$ to N_{epoch} **do**
  for each batch (x, y) in $\mathcal{D}$ **do**
    Generate adversarial sample x_{adv} via standard PGD bounded by ϵ
    Initialize losses: $L_{\text{base}} \leftarrow L_{\text{CE}}(F_\theta(x_{\text{adv}}), y), \quad L_{\text{recal}} \leftarrow 0$
    for $l = 1$ to M **do**
      Extract GAP features: $f_{\text{clean}}^{(l)} \leftarrow \text{GAP}(F_\theta^{(l)}(x)), \quad f_{\text{adv}}^{(l)} \leftarrow \text{GAP}(F_\theta^{(l)}(x_{\text{adv}}))$
      Construct target feature: $\tilde{f}_{\text{target}}^{(l)} \leftarrow \text{Detach}(f_{\text{clean}}^{(l)}) \odot \left(1 + \lambda \boldsymbol{\alpha}^{(y)}\right)$
      Accumulate layer loss: $L_{\text{recal}} \leftarrow L_{\text{recal}} + \left\| f_{\text{adv}}^{(l)} - \tilde{f}_{\text{target}}^{(l)} \right\|_1$
    end for
    Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \left(L_{\text{base}} + \beta \cdot \frac{L_{\text{recal}}}{M}\right)$
  end for
end for

A.2 Additional Quantitative Evaluation via Neuron Masking

As an extension to the corresponding quantitative evaluation presented in the main text, which demonstrated the effectiveness of our NSA-guided neuron masking using CIFAR-10 Class 0 as a representative example, this section provides comprehensive masking ablation results for the remaining 9 classes (Class 1 to Class 9).

We adopted the exact same experimental settings as in the main text to compare the recovery of adversarial classification accuracy and the evolution of the ground-truth logits after masking the top $k\%$ most sensitive neurons guided by the NSA method, the NS-based method, and the random masking strategy.

As shown in Figures 4 to 12, the overall trends of the different masking strategies largely align with those observed in the main text, though presenting notable class-wise variances. Specifically, for ResNet-18, the accuracy on adversarial examples generally exhibits a steady increase as the proportion of NSA-guided masked neurons grows. Meanwhile, for VGG-16, the defense performance demonstrates highly class-dependent behaviors.

However, several exceptional cases warrant particular attention. First, for ResNet-18 on Class 1 (see Figure 4(a) and (b)), the recovery of adversarial accuracy following NSA-guided masking of sensitive neurons is less than ideal under conventional masking ratios: the adversarial accuracy peaks at only 33.3% when 35% of the neurons are masked, while the clean accuracy drops sharply to 25.2%. Similar but slightly milder phenomena are observed in Class 7 (see Figure 10(a) and (b)) and Class 9 (see Figure 12(a) and (b)). To investigate the underlying causes of this phenomenon, we further extended the masking ratio for these specific classes. As illustrated in Figure 13, the experimental results are striking: for ResNet-18 on Class 1, when the NSA-guided masking ratio is increased to 70%, the adversarial accuracy surges to 100%, and the clean accuracy remarkably recovers to 99.6%. Classes 7 and 9 exhibit highly similar trends. We hypothesize that for these three classes, the sensitive neurons exploited by adversarial attacks account for a substantially larger proportion of the entire model. Consequently, a much higher masking ratio is strictly required to thoroughly cut off the forward propagation paths of the adversarial perturbations. This also demonstrates that even in extreme cases with a remarkably high proportion of sensitive neurons, NSA can still accurately evaluate their sensitivity, thereby firmly proving its effectiveness.

Secondly, for the VGG-16 model, a significant class-wise variance is observed among the different categories. Specifically, Classes 2, 3, 4, and 5 exhibit a recovery trajectory highly consistent with that of ResNet-18: when the masking ratio of sensitive neurons reaches 40%, the adversarial robustness is successfully restored to 100%. Conversely, the remaining classes align with the findings for the representative class (Class 0) detailed in the main text, where a higher masking ratio leads to a near-complete network collapse. As discussed in the main text, we attribute this degradation to the severe entanglement between highly sensitive neurons and core semantic neurons within these specific classes. This phenomenon further highlights that the tolerance for hard masking varies drastically across different classes.

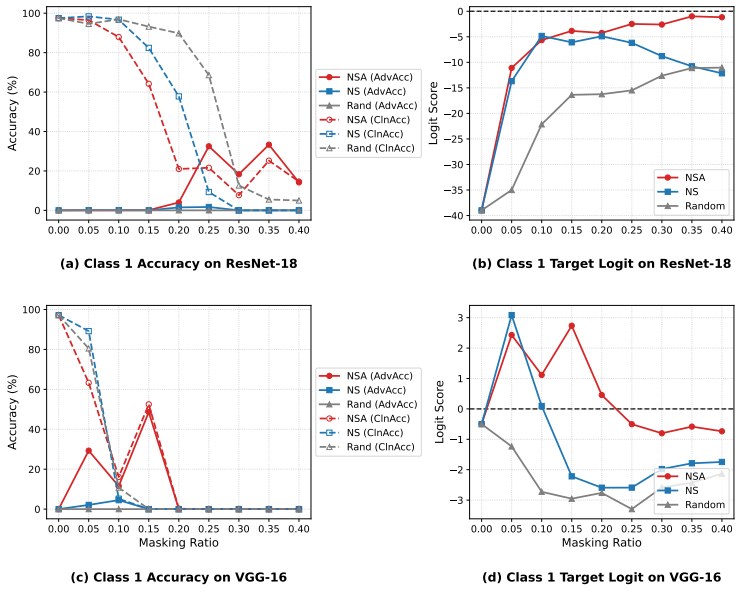


Figure 4: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 1**.

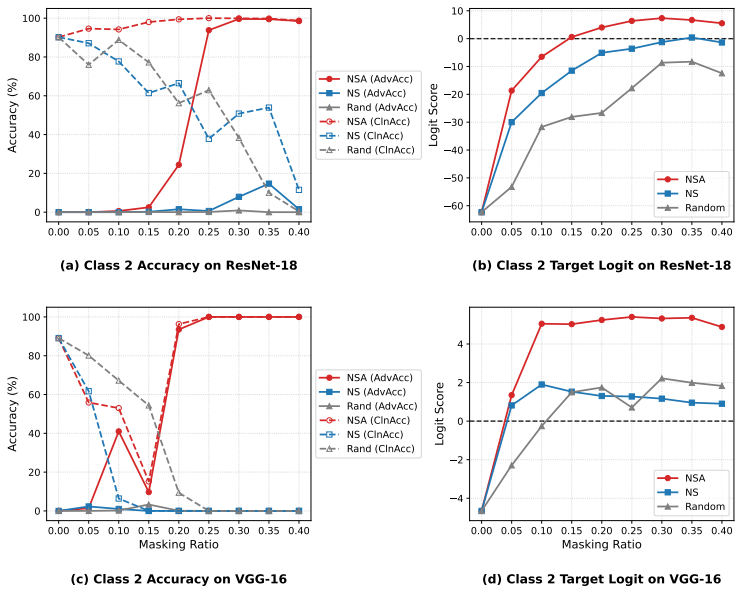


Figure 5: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 2**.

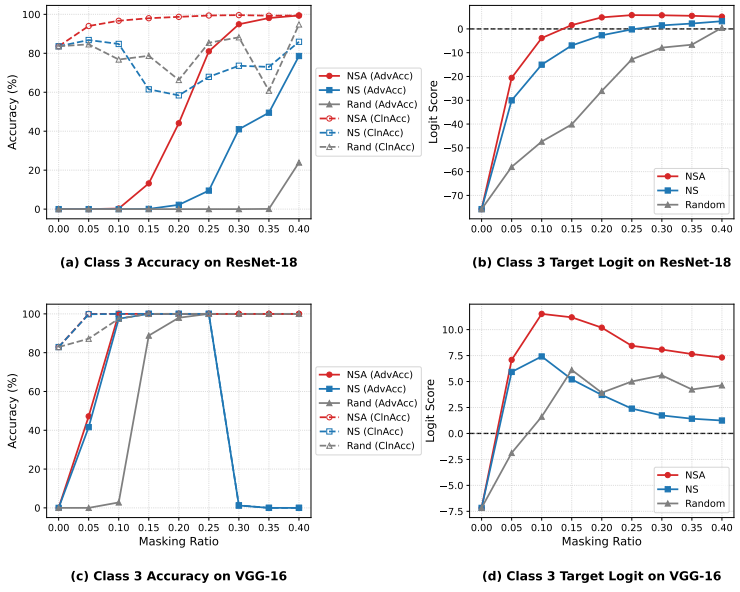


Figure 6: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 3**.

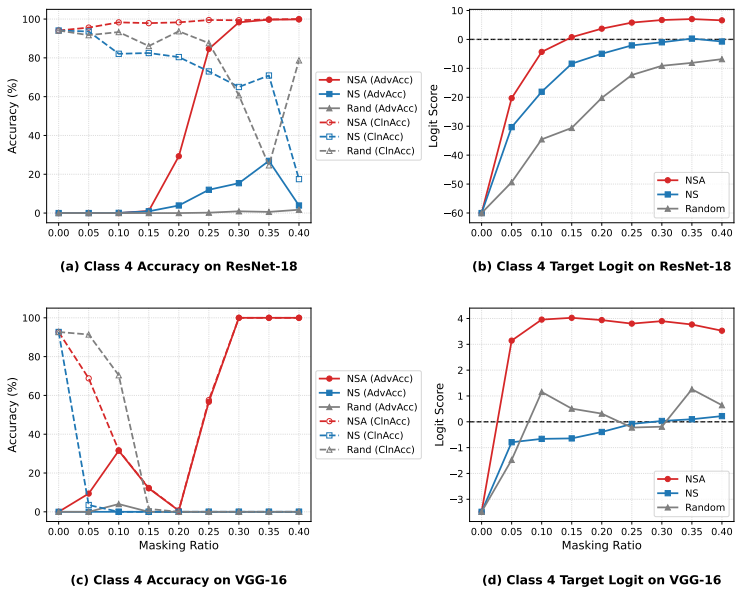


Figure 7: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 4**.

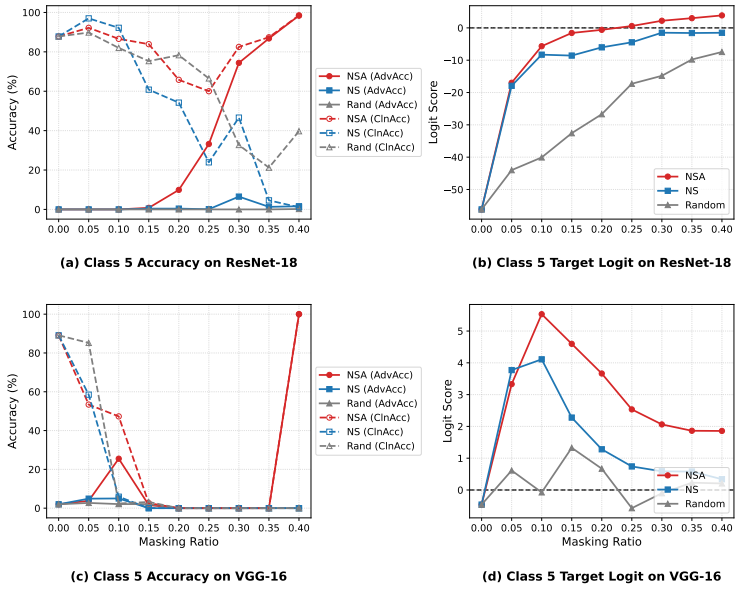


Figure 8: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 5**.

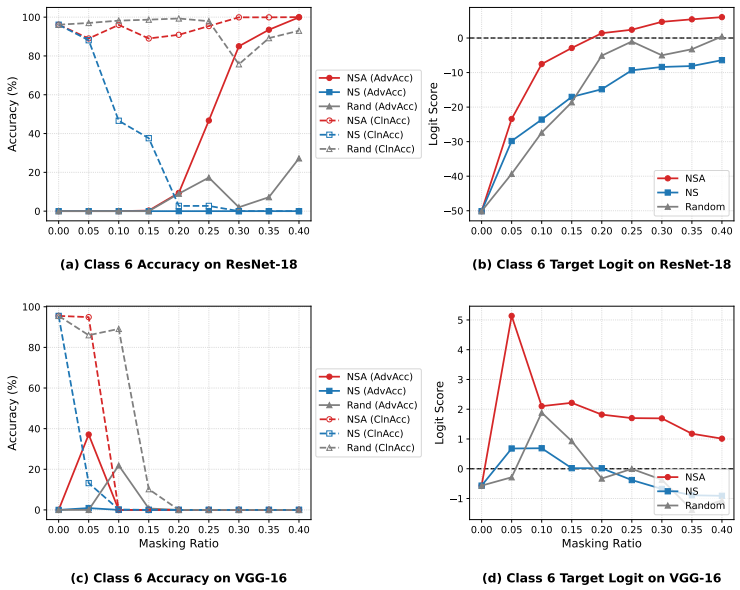


Figure 9: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 6**.

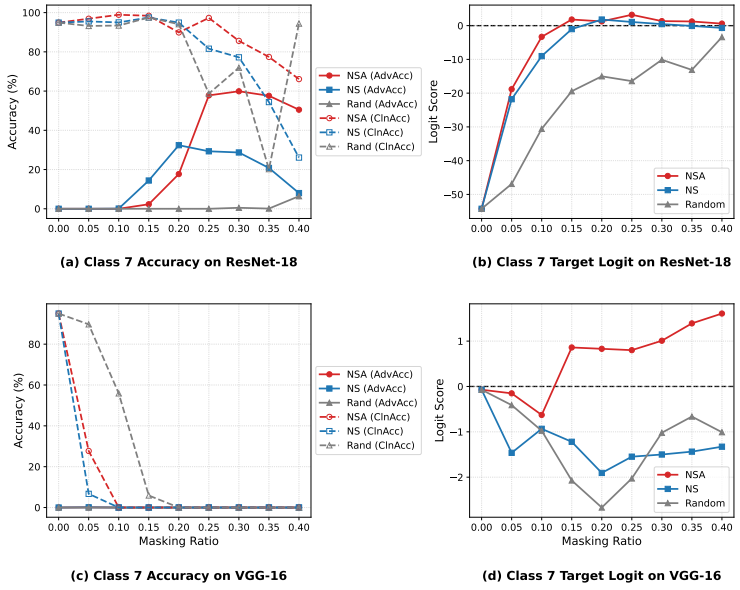


Figure 10: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 7**.

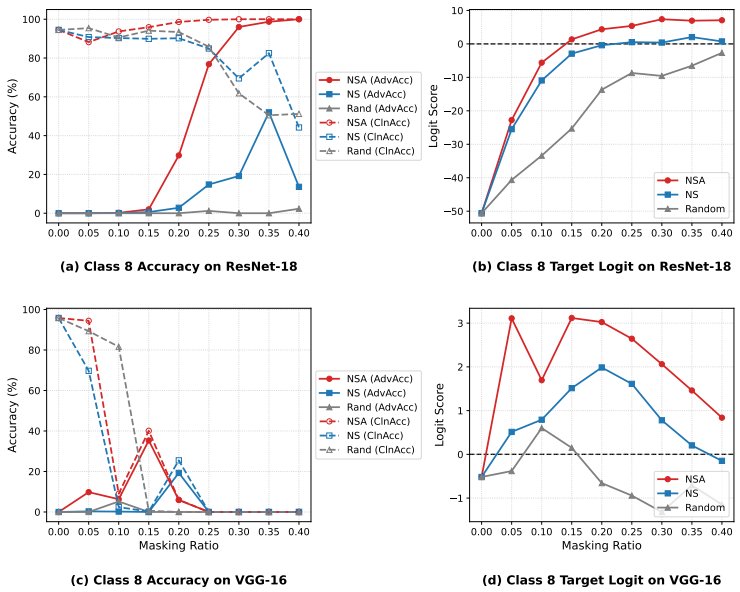


Figure 11: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 8**.

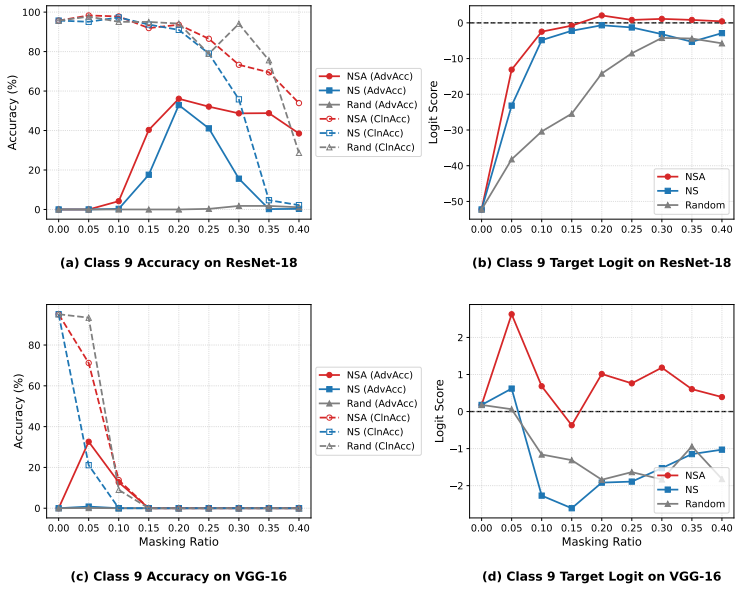


Figure 12: Quantitative evaluation of different neuron masking strategies on CIFAR-10 **Class 9**.

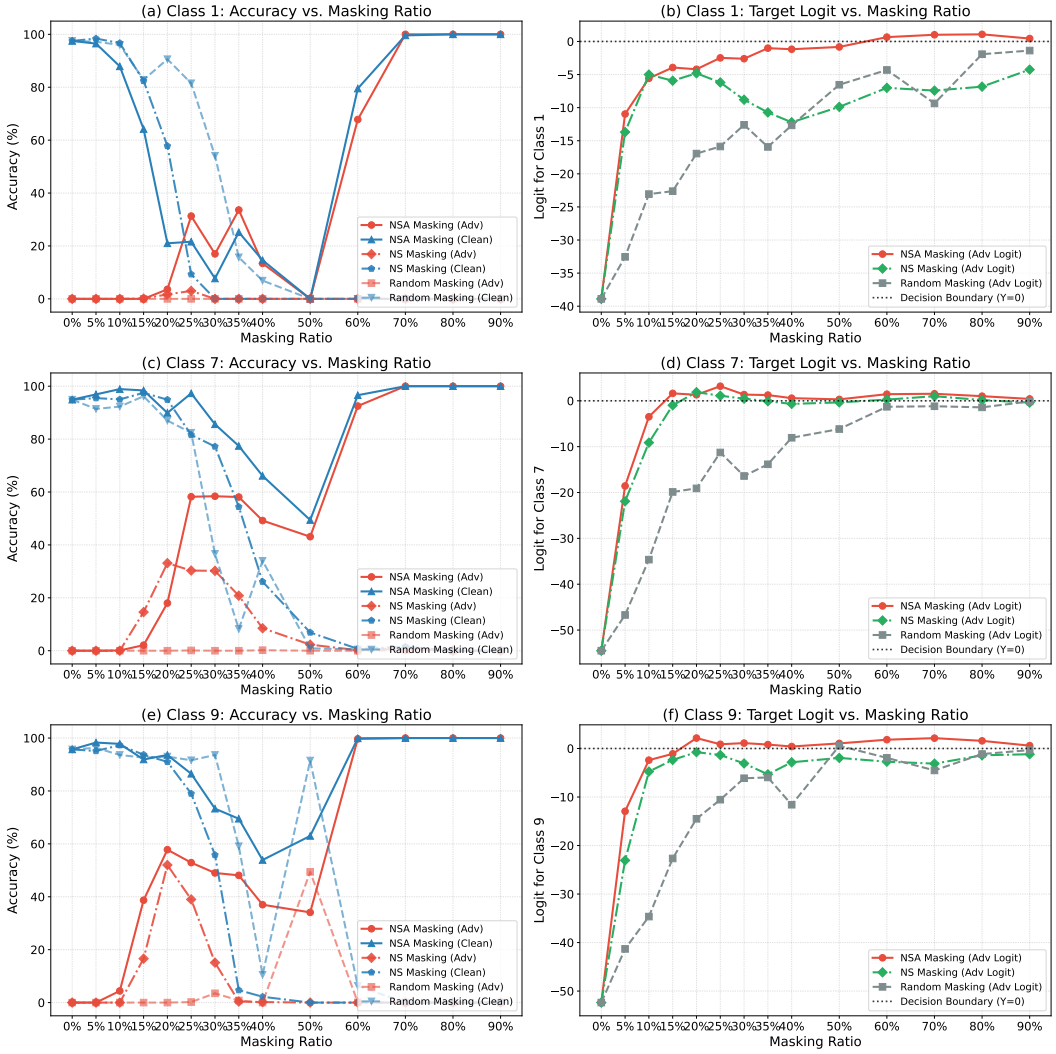


Figure 13: Quantitative evaluation with extended neuron masking ratios on CIFAR-10 **Classes 1, 7, and 9** for ResNet-18. Subfigures (a), (b), and (c) correspond to Class 1, Class 7, and Class 9, respectively. By increasing the NSA-guided masking ratio up to 70%, the adversarial and clean accuracy for all three exceptional classes recover remarkably to near 100%. This shared trend indicates that the sensitive neurons exploited by attacks in these classes account for a larger proportion of the network, requiring extensive masking to thoroughly block the adversarial forward propagation.

A.3 Additional Grad-CAM Visualizations for Other Classes

In the main text, we used Class 0 of the CIFAR-10 dataset as an example to illustrate the dynamic evolution of the model's attention distribution when sensitive neurons are forcibly masked. To comprehensively verify the effectiveness and generalizability of the NSA strategy across different classes, this section supplements the Grad-CAM class activation heatmaps for the remaining 9 classes (Class 1 to Class 9) in the CIFAR-10 dataset.

Consistent with the experimental settings in the main text, we extract the feature maps from Layer4 of ResNet-18 and B4_conv3 of VGG-16. The adversarial examples are likewise generated from the original unmasked model using the PGD attack (10 iterations, step size $\alpha = 2/255$). Figures 14 through 22 detail the spatial attribution variations of the model's attention mechanism under different NSA-guided sensitive neuron masking ratios (5%, 10%, 15%, 30%).

It is observed that for Classes 1, 7, and 9 of the ResNet-18 model, the attention in some images fails to fully recover to the clean sample state under low masking ratios, unlike other classes. This visual phenomenon perfectly echoes the quantitative evaluation results in Section A.2. Consequently, we further extended the neuron masking ratio specifically for these classes. As shown in Figure 23, when the NSA-guided masking ratio of sensitive neurons exceeds 60%, the model's attention almost entirely refocuses on the core semantic regions. This phenomenon intuitively corroborates our earlier hypothesis: for these three classes, the sensitive neurons exploited by adversarial attacks account for a large proportion of the entire network, necessitating a strictly high masking ratio to thoroughly cut off the forward propagation paths of adversarial perturbations. In summary, for the ResNet-18 model, the attention distribution across all classes successfully refocuses on the core semantic features of the objects following NSA-guided masking.

For the VGG-16 model, the supplementary visualization results further illuminate the complex, class-dependent behaviors and the "spatial-decision decoupling" paradox introduced in the main text. For several classes, the visual evolution initially aligns with the core semantics at moderate masking ratios. However, as the masking ratio increases to 30% or beyond, the severe entanglement causes the attention focus to shift or vanish (e.g., Class 8), visually corroborating the fatal semantic information loss caused by aggressive hard masking.

A highly counter-intuitive phenomenon emerges in Class 2, which serves as the exact inverse of the Class 0 paradox discussed in the main text. When the masking ratio exceeds 10%, although the classification accuracy remains at an optimal 100% (as detailed in Appendix A.2), the attention heatmaps vanish completely. We analyze that this occurs due to extreme spatial-decision decoupling under aggressive masking. The hard masking disrupts the continuous spatial semantic features in the intermediate layer (B4_conv3), causing the positive gradient contributions in Grad-CAM to be zeroed out by the ReLU activation. Consequently, the downstream fully connected layers are forced to rely on extremely sparse, non-interpretable features, or even activated classifier biases, to make a mathematically correct but spatially uninterpretable decision. These phenomena profoundly reveal that hard masking inevitably destroys the spatial interpretability and structural integrity of the model's decision-making process—whether by causing accuracy collapse despite visual alignment (Class 0), or by erasing spatial attribution despite optimal accuracy (Class 2).

A.4 Additional Implementation Details

To mitigate robust overfitting, all models evaluated under the AT framework are trained for 100 epochs, whereas models under the TRADES and DKL frameworks are trained for 200 epochs. The initial learning rate is uniformly set to 0.1. During the training phase, adversarial examples are generated using the PGD method under the ℓ_∞ norm constraint, with a maximum perturbation of $\epsilon = 8/255$, 10 iterations, and a step size of $\alpha = 2/255$.

Regarding other neuron-level robustness enhancement methods, CAS, CIFS and FSR are implemented strictly following the settings in their original papers. For the SNS method, we set its hyperparameter $\lambda = 1$ to achieve optimal robustness for the ResNet-18 model on the CIFAR-10 dataset. Regarding the implementation of our proposed SANR, we specify the target layers for neuron calibration based on the model architecture. For the VGG-16 model, all 13 convolutional layers are selected as the target layers. Similarly, for the ResNet-18 model, all 20 convolutional layers (including those in the downsampling shortcut connections) are selected as calibration targets.

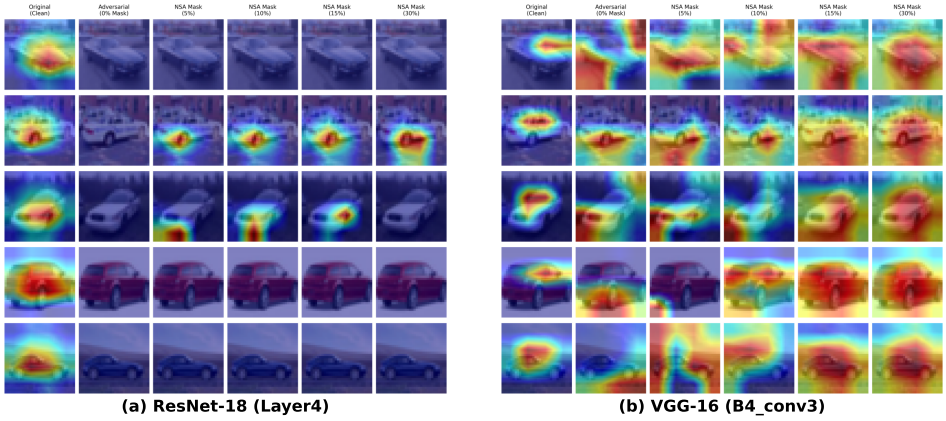


Figure 14: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 1 (Automobile)**. (a) On ResNet-18, unlike most other classes, the attention fails to fully recover to the clean state under conventional masking ratios (up to 30%), visually echoing the quantitative performance degradation and indicating a higher proportion of sensitive neurons. (b) On VGG-16, the focal shift triggered by aggressive masking visually demonstrates the severe feature entanglement discussed in the main text.

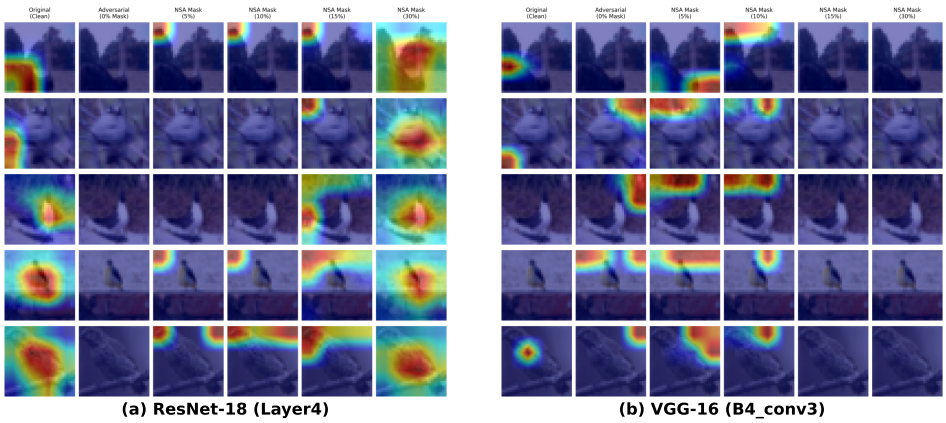


Figure 15: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 2 (Bird)**. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics. (b) On VGG-16, a counter-intuitive phenomenon occurs: attention aligns most closely with the core semantics at a 10% masking ratio, but aggressive masking (e.g., 30%) causes the heatmap to vanish completely despite maintaining 100% accuracy. This illustrates the decoupling of spatial attribution and classification decision under extreme masking.

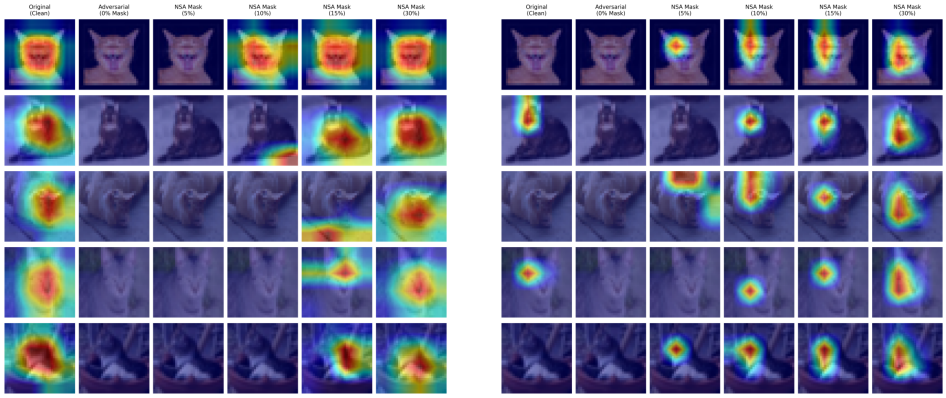


Figure 16: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 3 (Cat)**. The observed phenomena are highly consistent with the main text. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics of the target. (b) On VGG-16, attention partially recovers at low ratios, but aggressive masking (e.g., 30%) again triggers a focal shift, exposing the severe, class-dependent entanglement between sensitive and core semantic neurons.

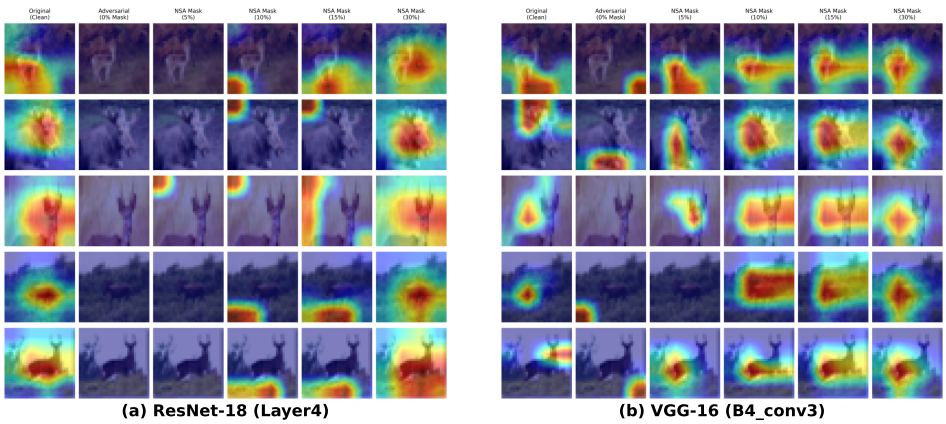


Figure 17: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 4 (Deer)**. The observed phenomena are highly consistent with the main text. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics of the target. (b) On VGG-16, attention partially recovers at low ratios, but aggressive masking (e.g., 30%) again triggers a focal shift, exposing the severe, class-dependent entanglement between sensitive and core semantic neurons.

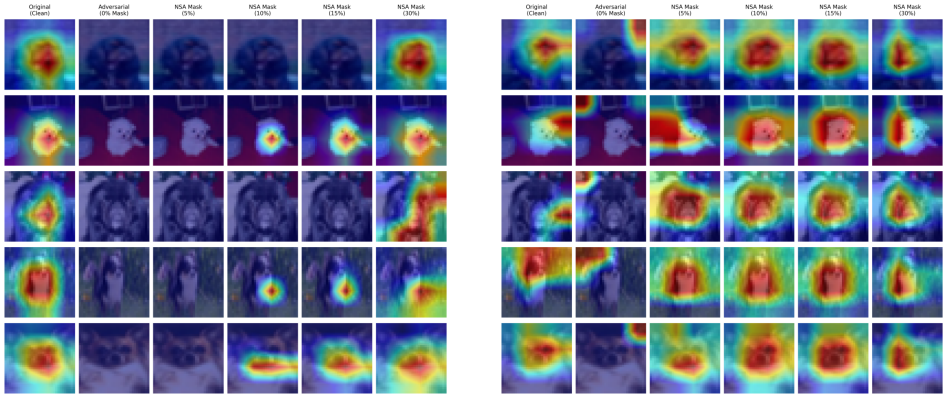


Figure 18: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 5 (Dog)**. The observed phenomena are highly consistent with the main text. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics of the target. (b) On VGG-16, attention partially recovers at low ratios, but aggressive masking (e.g., 30%) again triggers a focal shift, exposing the severe, class-dependent entanglement between sensitive and core semantic neurons.

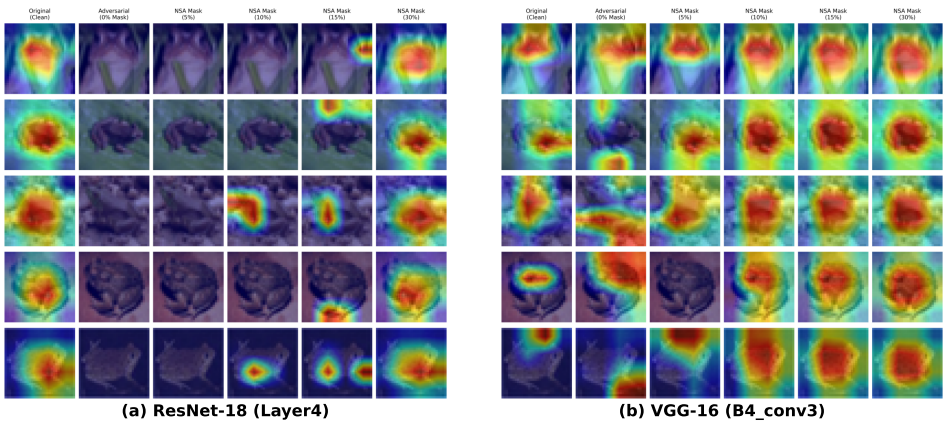


Figure 19: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 6 (Frog)**. The observed phenomena are highly consistent with the main text. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics of the target. (b) On VGG-16, attention partially recovers at low ratios, but aggressive masking (e.g., 30%) again triggers a focal shift, exposing the severe, class-dependent entanglement between sensitive and core semantic neurons.

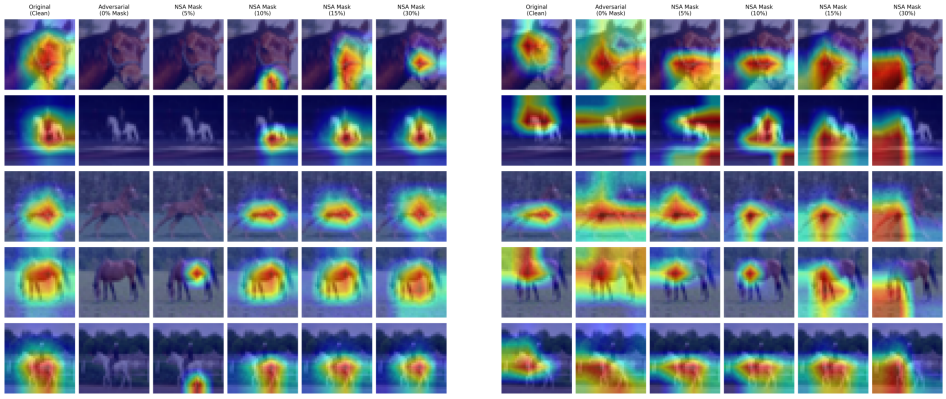


Figure 20: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 7 (Horse)**. (a) On ResNet-18, unlike most other classes, the attention fails to fully recover to the clean state under conventional masking ratios (up to 30%), visually echoing the quantitative performance degradation and indicating a higher proportion of sensitive neurons. (b) On VGG-16, the focal shift triggered by aggressive masking visually demonstrates the severe feature entanglement discussed in the main text.

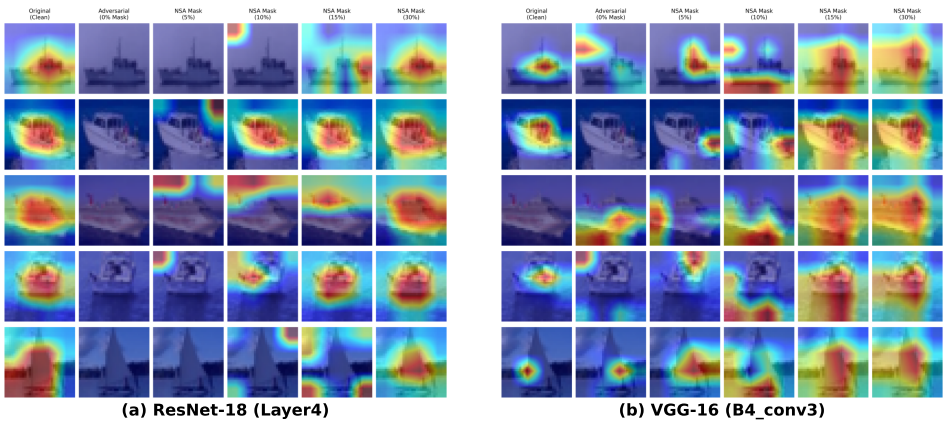


Figure 21: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 8 (Ship)**. (a) On ResNet-18, the suppressed attention successfully re-emerges and refocuses on the core semantics of the target. (b) On VGG-16, the attention aligns most closely with the core semantics at a 10% masking ratio. Aggressive masking (e.g., 30%) triggers a focal shift, explicitly demonstrating the exceptionally high degree of entanglement between core semantic and sensitive neurons for this specific class.

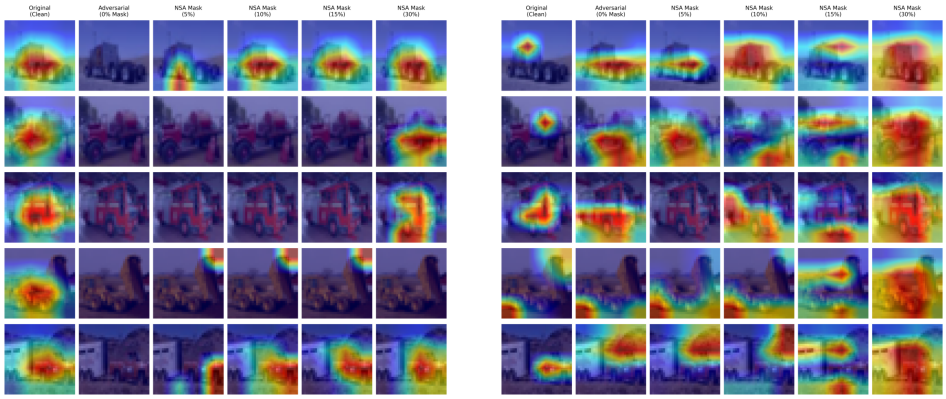


Figure 22: Additional Grad-CAM visualizations of spatial attribution under progressive NSA-based masking for CIFAR-10 **Class 9 (Truck)**. (a) On ResNet-18, unlike most other classes, the attention fails to fully recover to the clean state under conventional masking ratios (up to 30%), visually echoing the quantitative performance degradation and indicating a higher proportion of sensitive neurons. (b) On VGG-16, the focal shift triggered by aggressive masking visually demonstrates the severe feature entanglement discussed in the main text.

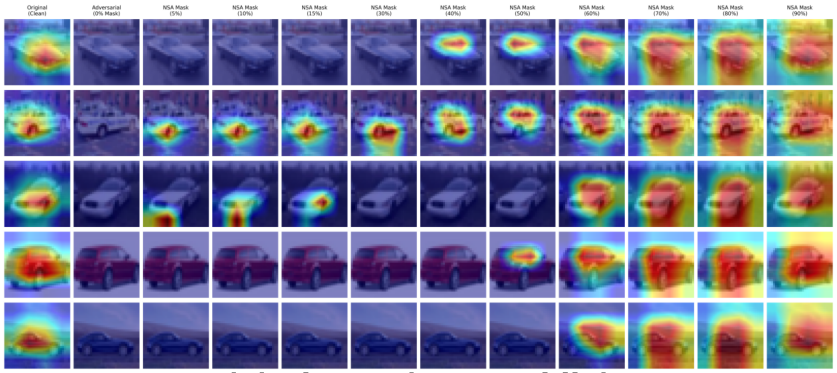
The specific hyperparameter settings for the different base frameworks are detailed as follows:

- **AT**: This framework directly optimizes the cross-entropy loss on adversarial examples. We set the weight coefficient of the standard cross-entropy loss to 1. For the ResNet-18 model, we set $\beta = 0.5$ and $\lambda = 0.1$. To balance clean accuracy and robustness, the weight decay is set to 5×10^{-4} both before and after integration. For the VGG-16 model, we set $\beta = 0.01$ and $\lambda = 1$, with the weight decay set to 2×10^{-4} before and after integration.
- **TRADES**: This method utilizes Kullback-Leibler (KL) divergence to constrain the discrepancy in prediction distributions between clean and adversarial examples. The weight coefficient of the KL divergence regularization term is fixed at 6 both before and after integrating SANR. For the ResNet-18 model, we set $\beta = 1$ and $\lambda = 1$, with the weight decay set to 5×10^{-4}. For the VGG-16 model, we set $\beta = 0.1$ and $\lambda = 1$, with the weight decay set to 2×10^{-4}.
- **DKL**: This framework comprises a Margin Mean Squared Error (MSE) term and a Distillation Cross-Entropy (CE) term. We maintain the original parameter settings unchanged before and after integration, setting the Margin MSE weight to 7 and the CE weight to 1. For the ResNet-18 model, we set $\beta = 0.1$ and $\lambda = 1$; for the VGG-16 model, we set $\beta = 1$ and $\lambda = 1$. The weight decay is uniformly set to 5×10^{-4} for both models before and after integration.

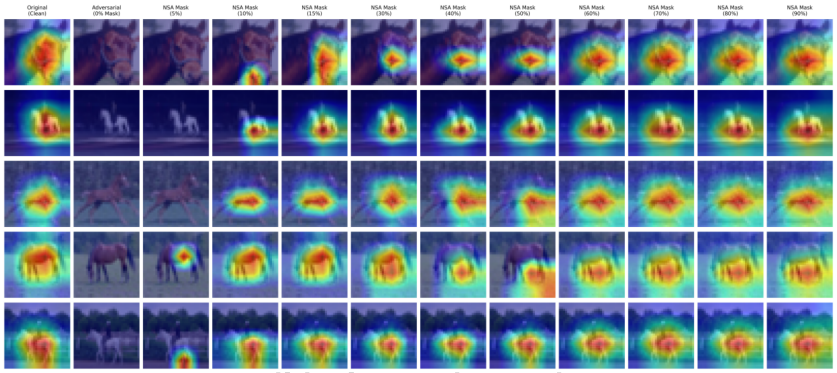
A.5 Robustness Results of the Best Checkpoint

As shown in Table 3, we additionally report the robustness evaluation results based on the best training epoch. The best model is selected based on the highest PGD-10 adversarial accuracy on the validation set, serving as a comparison and supplement to the "Last Epoch" results presented in the main text. It is worth noting that for the DKL framework, the best model and the last model are identical.

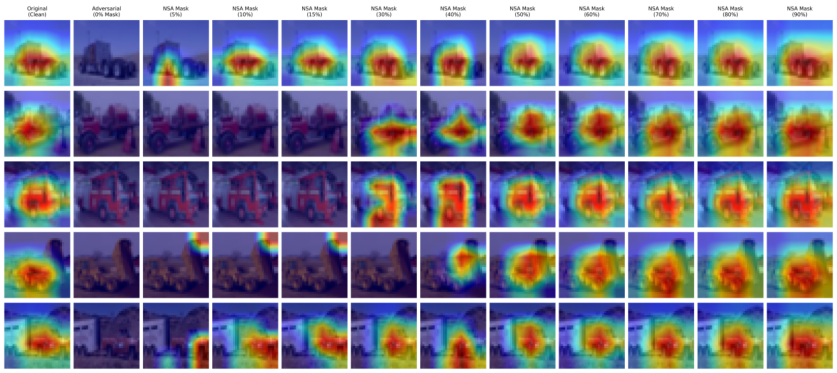
The experimental results demonstrate that under the vast majority of adversarial training frameworks and model architectures, SANR continues to significantly break through the robustness upper bounds of baseline methods. Particularly in the TRADES and DKL frameworks, whether evaluated on ResNet-18 or VGG-16, the introduction of SANR achieves comprehensive outperformance across all adversarial robustness metrics (FGSM, PGD, C&W, AA) while consistently preserving highly competitive clean accuracy. This further validates the consistent enhancement effect of SANR on model robustness throughout the entire training cycle.



(a) Class 1 (Automobile)



(b) Class 7 (Horse)



(c) Class 9 (Truck)

Figure 23: Grad-CAM visualizations of spatial attribution under **extended NSA-guided masking ratios (over 60%)** for CIFAR-10 **Classes 1, 7, and 9** on ResNet-18. Subfigures (a), (b), and (c) correspond to Class 1 (Automobile), Class 7 (Horse), and Class 9 (Truck), respectively. Consistent with the quantitative findings, when the masking ratio is sufficiently increased, the model's suppressed attention successfully re-emerges and almost entirely refocuses on the core semantic regions across all three exceptional classes. This visual evidence intuitively corroborates that effectively cutting off the adversarial forward propagation paths in these classes requires masking a substantially larger proportion of sensitive neurons.

Table 3: Performance comparison of different adversarial training frameworks before and after integrating SANR on CIFAR-10 based on the **best epoch** checkpoints. Best results within each framework pair are highlighted in bold.

Model	Method	Accuracy (%)					
		Clean	FGSM	PGD-20	PGD-100	C&W	AA
ResNet-18	AT	83.88	57.63	**51.75**	**51.42**	57.60	47.60
	AT + SANR	**84.10**	**58.15**	51.63	51.37	**57.79**	**48.27**
	TRADES	82.57	58.72	52.88	52.66	58.35	49.70
	TRADES + SANR	**83.46**	**59.33**	**53.60**	**53.27**	**59.13**	**50.14**
	DKL	**87.23**	59.67	52.87	52.55	57.76	49.75
	DKL + SANR	86.97	**60.58**	**53.92**	**53.49**	**58.34**	**50.19**
VGG-16	AT	79.81	**54.41**	48.40	48.05	58.13	**44.20**
	AT + SANR	**79.96**	54.25	**48.87**	**48.56**	**58.19**	44.14
	TRADES	80.92	55.84	49.03	48.59	57.46	45.30
	TRADES + SANR	**81.01**	**56.27**	**49.36**	**48.99**	**57.63**	**45.81**
	DKL	85.18	56.54	49.07	48.63	56.61	44.85
	DKL + SANR	**85.41**	**56.85**	**49.79**	**49.26**	**57.14**	**45.27**

Furthermore, the experimental data objectively captures a slight oscillation phenomenon in a very few metrics under the standard adversarial training (AT) framework. For the ResNet-18 model, the peak performance of AT+SANR against PGD-20 and PGD-100 slightly lags behind the original AT (by a marginal gap of approximately 0.05% to 0.12%); similarly, for the VGG-16 model, the peak values of AT+SANR against FGSM and AA also experience minor pullbacks. From the perspective of deep learning optimization, this phenomenon is well within expectations: lacking structural regularization constraints, the original AT method is highly susceptible to local overfitting towards a specific adversarial perturbation pattern (such as a single PGD step size) during specific training stages, thereby exhibiting an inflated single-point peak at a given epoch. In contrast, the core mechanism of SANR lies in globally reshaping the feature space through the suppression of sensitive neurons; thus, its optimization trajectory is smoother and prioritizes the balance of overall defense capabilities. Consequently, although SANR does not cater to specific attack patterns at an extreme truncation moment, it endows the model with stronger consistency and generalization ability (a conclusion fully corroborated by the Last Epoch results in the main text, where SANR consistently outperforms the baseline across all adversarial robustness metrics).

A.6 Ablation Studies

To investigate the impact of the hyperparameters in SANR, Figure 24 and Figure 25 present the ablation study results for parameters β and λ, respectively, evaluated on the ResNet-18 model under the DKL training framework. In our experiments, we first fix the intervention intensity $\lambda = 1.0$ to search for the optimal loss balancing factor β. To intuitively illustrate the sensitivity of the model's performance to this loss weight, Figure 24 plots the trends of several core evaluation metrics (Clean, C&W, PGD-20, AA) as β varies. As shown, as β gradually increases, the model's robustness against various attacks steadily improves, reaching an optimal defense balance around $\beta = 0.1$; however, assigning an excessively large weight to this loss (e.g., $\beta = 5.0$) leads to a severe performance collapse. For more comprehensive and detailed raw test data across all attacks, please refer to Table 4 in the Appendix.

In a similar vein, we conduct another set of ablation experiments to evaluate the impact of the intervention intensity λ, which controls the scaling magnitude of the sensitivity mask. In this phase, we fix the loss balancing factor β to its optimal setting (i.e., $\beta = 0.1$) and vary λ from 0 to 2.0. To intuitively illustrate the sensitivity of the model's performance to this intervention intensity, Figure 25 plots the trends of core evaluation metrics, while Table 5 provides the comprehensive testing data. Notably, when $\lambda = 0$, the SANR mechanism degenerates into traditional feature alignment, which treats all neurons equally and forces adversarial features to blindly match clean features. Un-

der this setting, the model achieves sub-optimal robustness (e.g., 49.95% under Auto-Attack). As λ increases, the model's defense capability steadily improves, strongly validating the necessity of applying differentiated, sensitivity-aware modulation at the neuron level. The model reaches its peak adversarial robustness across almost all advanced attacks when $\lambda = 1.0$, securing 50.19% against Auto-Attack and 58.34% against C&W. However, further increasing the intervention intensity (e.g., $\lambda = 1.5$ or 2.0) causes a slight decline or fluctuation in robust accuracy, suggesting that an excessively aggressive intervention may overly distort the intrinsic feature distribution and disrupt the robust decision boundaries.

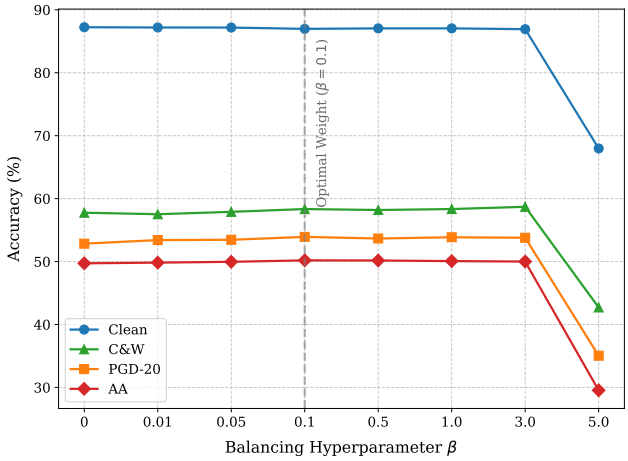


Figure 24: Ablation study of the balancing hyperparameter β. The optimal robustness trade-off is achieved at $\beta = 0.1$.

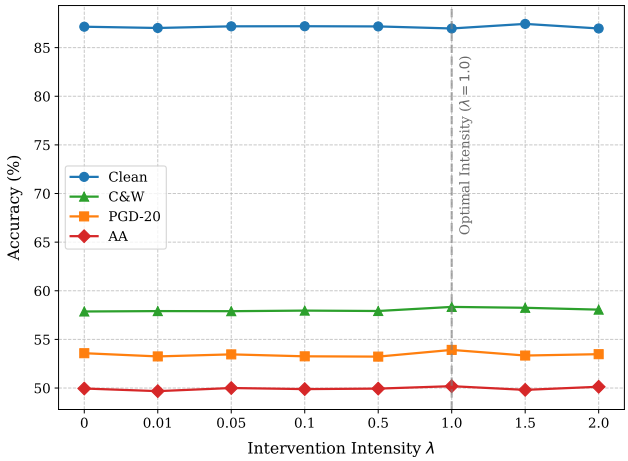


Figure 25: Ablation study of the intervention intensity λ. The model achieves the best adversarial robustness when $\lambda = 1.0$.

Table 4: Ablation study on the balancing hyperparameter β. Experiments are conducted with ResNet-18 trained using the DKL framework on CIFAR-10. Best results are in bold.

β	Accuracy (%)					
	Clean	FGSM	PGD-20	PGD-100	C&W	AA
0 (Baseline)	**87.23**	59.67	52.84	52.53	57.76	49.72
0.01	87.19	59.63	53.41	53.05	57.52	49.83
0.05	87.18	59.73	53.46	53.24	57.90	49.96
0.1	86.97	**60.58**	**53.92**	**53.49**	58.34	**50.19**
0.5	87.05	60.15	53.66	53.33	58.19	50.17
1.0	87.05	60.12	53.86	53.48	58.34	50.07
3.0	86.92	59.91	53.78	53.39	**58.70**	50.00
5.0	67.97	39.78	35.04	34.81	42.71	29.54

Table 5: Ablation study on the intervention intensity λ. Best results are in bold.

λ	Accuracy (%)					
	Clean	FGSM	PGD-20	PGD-100	C&W	AA
0 (Baseline)	87.15	59.97	53.58	53.25	57.87	49.95
0.01	87.02	59.91	53.25	53.01	57.91	49.68
0.05	87.19	59.60	53.46	53.17	57.90	50.00
0.1	87.20	60.03	53.26	52.97	57.96	49.89
0.5	87.18	59.86	53.23	53.07	57.92	49.94
1.0	86.97	**60.58**	**53.92**	**53.49**	**58.34**	**50.19**
1.5	**87.44**	60.00	53.34	52.99	58.25	49.81
2.0	86.97	59.94	53.48	53.23	58.06	50.13