Explainability-Driven Defense: Grad-CAM-Guided Model Refinement Against Adversarial Threats

Longwei Wang[1], Ifrat Ikhtear Uddin[1], Xiao Qin[2], Yang Zhou[2], KC Santosh[1]

[1] AI Research Lab, Department of Computer Science, University of South Dakota, Vermillion, SD 57069, USA
[2] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, USA
longwei.wang@usd.edu, ifratikhtear.uddin@coyotes.usd.edu, yangzhou@auburn.edu, xqin@auburn.edu, kc.santosh@usd.edu

Abstract

Deep learning models have excelled in tasks like image recognition and autonomous systems but remain vulnerable to adversarial attacks and spurious correlations, limiting their reliability in real-world and safety-critical settings. To address these challenges, we propose a novel framework that leverages explainable Artificial Intelligence (XAI) to enhance the robustness of Convolutional Neural Networks. Our approach integrates Grad-CAM insights into the model refinement process, guiding feature masking to reduce reliance on irrelevant or misleading features. We introduce three masking strategies: (1) binary masking to retain high-activation regions, (2) Gaussian-blurred masking to preserve contextual information while reducing noise, and (3) difference-based masking to remove unstable features unique to the baseline model. We evaluate these strategies against two common adversarial attack methods—Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Results show that all three strategies improve FGSM accuracy, with binary and difference-based masking providing consistent gains across perturbation levels. Gaussian-blurred masking delivers the highest improvement in PGD accuracy, particularly at higher perturbation strengths.

Introduction

Convolutional Neural Networks (CNNs) have achieved remarkable success in various computer vision tasks, including image classification, object detection, and segmentation (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Redmon et al. 2016; Wang and Liang 2019; Wang and Li 2019). Despite their widespread adoption, CNNs are highly susceptible to adversarial attacks, where small, often imperceptible perturbations to the input data can cause significant misclassifications (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015; Cui et al. 2024). This vulnerability raises critical concerns about the reliability and security of deep learning models, especially in high-stakes domains such as autonomous driving, healthcare, and security systems (Kurakin, Goodfellow, and Bengio 2018; Moosavi-Dezfooli et al. 2017; Shi et al. 2020).

To mitigate adversarial threats, several defense strategies have been proposed, including adversarial training (Madry et al. 2017; Waghela, Sen, and Rakshit 2024), input transformation (Guo et al. 2017; Wang et al. 2021b; Nayyem, Rakin, and Wang 2024), local linearization, symmetry enforcement, (Qin et al. 2019; Wang et al. 2024; Wang, Li, and Zhang 2024; Wang, Ghimire, and Santosh 2024) and model regularization (Ross and Doshi-Velez 2018). Among these methods, adversarial training—where the model is trained on adversarial examples—has been shown to be effective but computationally expensive and prone to overfitting to specific attack types (Tsipras et al. 2019). Input transformation techniques, such as image cropping, resizing, and blurring, offer an alternative approach by reducing the impact of perturbations, but they often degrade clean accuracy (Xie et al. 2019).

Recently, explainability methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017) have emerged as powerful tools for interpreting CNN decisions by identifying the most influential input regions (Chakraborty et al. 2022; Zhang et al. 2025; van Zyl, Ye, and Naidoo 2024). Grad-CAM provides a heatmap of high-activation regions, offering insights into which features the model relies on for prediction. While explainability techniques have primarily been used for model interpretation (Hassija et al. 2024; Wang et al. 2021a, 2020) and debugging (Lin, Lee, and Celik 2021; Baniecki and Biecek 2024), there is growing interest in leveraging them to enhance adversarial robustness. For example, guided masking of vulnerable features using Grad-CAM could enable the model to focus on more stable and less attack-prone features, improving overall resilience to adversarial perturbations.

In this work, we propose an explainability-driven defense strategy that employs Grad-CAM to guide feature masking as a mechanism for improving CNN robustness against adversarial attacks. Specifically, we explore three masking strategies: (1) **Binary feature masking** based on high-activation regions from Grad-CAM in a ResNet-50 model, (2) **Gaussian-blurred masking** to maintain contextual information while reducing sensitivity to high-frequency perturbations, and (3) **Difference-based masking** that removes features unique to the baseline model but absent in a stronger ResNet-50 model. We evaluate the effectiveness of these masking techniques against two well-established attack methods—Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015) and Projected Gradient

Descent (PGD) (Madry et al. 2017; Villegas-Ch, Jaramillo-Alcázar, and Luján-Mora 2024).

Our contributions can be summarized as follows:

- We propose a novel feature masking framework guided by Grad-CAM to enhance CNN robustness against adversarial attacks.
- We systematically compare three masking techniques—binary, Gaussian-blurred, and difference-based masking—highlighting their relative strengths under different attack scenarios.
- We demonstrate that feature masking improves adversarial robustness without requiring adversarial training, offering a lightweight and scalable defense mechanism.

Related Works

The vulnerability of deep neural networks, particularly Convolutional Neural Networks (CNNs), to adversarial attacks has been extensively studied in recent years. Adversarial attacks involve crafting small perturbations to input data that are imperceptible to humans but can significantly degrade the performance of CNNs. In this section, we review key approaches for defending CNNs against adversarial attacks, focusing on adversarial training, input transformation, model regularization, and explainability-guided defense mechanisms.

Adversarial Training

Adversarial training is one of the most widely studied and effective methods for improving adversarial robustness. In this approach, a model is trained on adversarially perturbed data to improve its ability to resist attacks. Goodfellow et al. (Goodfellow, Shlens, and Szegedy 2015) proposed the Fast Gradient Sign Method (FGSM) for generating adversarial examples and demonstrated that adversarial training could mitigate such attacks. Madry et al. (Madry et al. 2017) introduced Projected Gradient Descent (PGD) and showed that adversarial training with PGD-generated examples leads to state-of-the-art robustness. However, adversarial training is computationally expensive and can lead to overfitting on specific types of attacks, reducing generalization to unseen perturbations (Tsipras et al. 2019; Huang et al. 2020). Furthermore, adversarial training often degrades the clean accuracy of the model, creating a trade-off between robustness and generalization (Zhang et al. 2019).

Input Transformation

Input transformation techniques aim to preprocess input data in a way that disrupts adversarial perturbations while preserving the underlying signal. Guo et al. (Guo et al. 2017) proposed input transformations such as image cropping, resizing, and JPEG compression to counter adversarial attacks. Xie et al. (Xie et al. 2019) introduced feature denoising as a preprocessing step to enhance model robustness. While input transformation is lightweight and easy to implement, it often reduces clean accuracy and may not generalize well across different types of attacks (Wang et al. 2021c).

Model Regularization

Model regularization methods seek to enhance robustness by modifying the loss function or training process to improve the model's sensitivity to perturbations. Ross and Doshi-Velez (Ross and Doshi-Velez 2018)(Wang and Liang 2019) proposed gradient-based input regularization, where the model's sensitivity to input perturbations is reduced by penalizing large gradients. Jakubovitz and Giryes (Jakubovitz and Giryes 2018) introduced regularization terms based on the Lipschitz constant to improve adversarial robustness. While regularization-based approaches can enhance robustness, they require careful tuning and often increase the complexity of the training process.

Explainability-Guided Defenses

Explainability methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017)(Nayyem, Rakin, and Wang 2024) have been widely used to interpret CNN decisions by highlighting the most influential input regions. In our work, we extend the idea of explainability-guided defense by exploring three distinct feature masking strategies based on Grad-CAM activations. Unlike previous approaches, our method directly leverages Grad-CAM insights to refine the model's attention, improving robustness without requiring adversarial training. Furthermore, we conduct a detailed comparison of the effectiveness of each masking strategy against FGSM and PGD attacks.

Methodology

To evaluate the impact of feature masking on adversarial robustness, we propose a novel Grad-CAM-guided defense framework that leverages explainability to refine the model's feature sensitivity. We investigate three distinct feature masking strategies: (1) **binary feature masking**, which retains only high-activation regions; (2) **Gaussian-blurred masking**, which preserves contextual information while suppressing irrelevant regions; and (3) **difference-based masking**, which removes unstable features by comparing activations of a baseline model and a stronger reference model (ResNet-50).

Our approach is designed to selectively mask input features, retrain the model on the masked datasets as illustrated in Figure 1, and then evaluate the resulting robustness against adversarial attacks using FGSM and PGD. This section details the Grad-CAM-based feature extraction and masking strategies

Grad-CAM-Based Feature Extraction

Explainability methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017) provide a powerful way to visualize the contribution of input features to a model's predictions. Grad-CAM generates a heatmap that highlights the most influential input regions for a given class prediction. The core idea is to compute gradients of the class score y_c with respect to the feature maps from the last convolutional layer to determine which regions contribute most to the class decision.

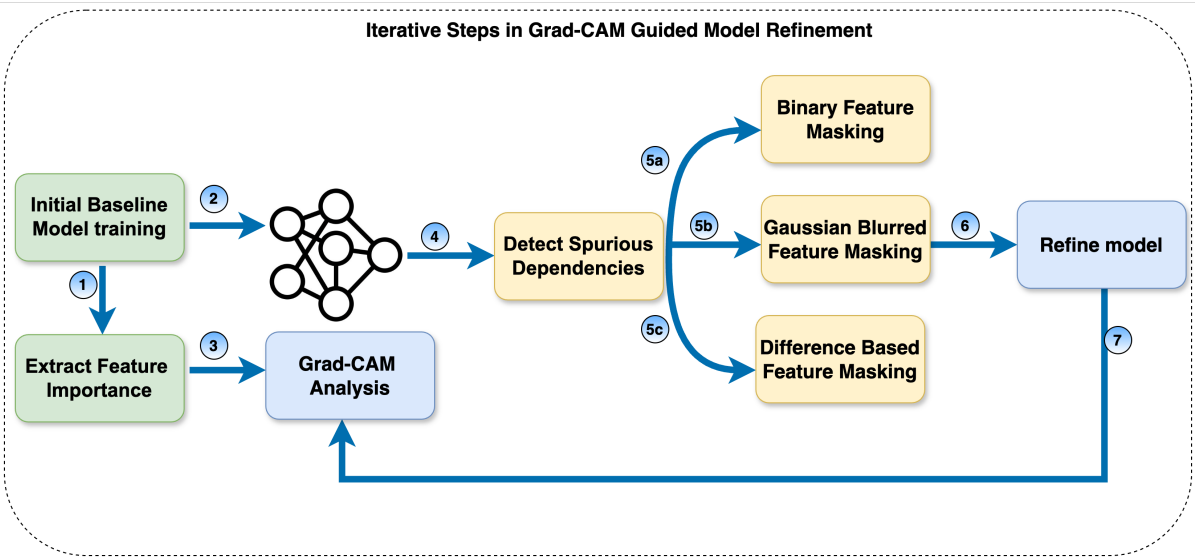

Figure 1: Iterative process for Grad-CAM guided model refinement. The workflow begins with training (1) an initial baseline model, followed by extracting feature importance through Grad-CAM analysis (2,3,4). Spurious dependencies in feature activations are identified, leading to the application of three feature masking strategies: Binary Masking (5a), Gaussian Blurred Feature Masking (5b), and Difference-Based Feature Masking (5c). The model is then refined using masked datasets, and the process iterates (6,7) to improve robustness.

Grad-CAM Computation Given an input image $I \in \mathrm{R}^{H \times W \times C}$, where H is the height, W is the width, and C is the number of channels, we pass the image through a CNN model to extract activation maps from the last convolutional layer. Let $A_k \in \mathrm{R}^{H' \times W'}$ denote the activation map for the k-th feature map, where H' and W' are the spatial dimensions of the feature map.

1. **Importance weight computation**

   The importance weight α_k^c for the target class c is computed by averaging the gradients of the class score with respect to the activation map:

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \frac{\partial y_c}{\partial A_k^{ij}} \tag{1}$$

   where $Z = H' \times W'$ is the total number of spatial locations in the feature map.

2. **Weighted sum of feature maps**

   The Grad-CAM heatmap is then computed as a weighted sum of the activation maps:

$$L^c_{\mathrm{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k \alpha_k^c A_k \right) \tag{2}$$

   The ReLU function ensures that only positive contributions are retained, based on the assumption that positive gradients correspond to features that positively contribute to the target class prediction.

3. **Normalization** To improve numerical stability and ensure consistent scaling across different samples, we normalize the heatmap as follows:

$$H_{\mathrm{norm}} = \frac{L^c_{\mathrm{Grad-CAM}} - \min(L^c_{\mathrm{Grad-CAM}})}{\max(L^c_{\mathrm{Grad-CAM}}) - \min(L^c_{\mathrm{Grad-CAM}}) + \delta} \tag{3}$$

   where δ is a small constant added to prevent division by zero.

Feature Masking Strategies

We explore three feature masking strategies that utilize the Grad-CAM heatmap to selectively suppress or retain features based on their importance for prediction.

Binary Feature Masking Binary feature masking involves creating a binary mask that retains only the most activated regions from the Grad-CAM heatmap while setting less relevant regions to zero. This approach directs the model's attention toward the most influential features and minimizes distractions from irrelevant patterns.

1. **Threshold-based masking**

   A binary mask is created by applying a threshold τ to the normalized heatmap:

$$\tau = 0.8 \times \mathrm{mean}(H_{\mathrm{norm}}) \tag{4}$$

   The mask is then computed as:

$$M_{\mathrm{binary}}(i,j) = \begin{cases} 1, & H_{\mathrm{norm}}(i,j) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

2. **Masked image computation**
   The masked image is computed as:

$$I_{\text{masked}}(i,j) = I(i,j) \cdot M_{\text{binary}}(i,j) \tag{6}$$

   Figure 2 shows binary masking sample from our three datasets

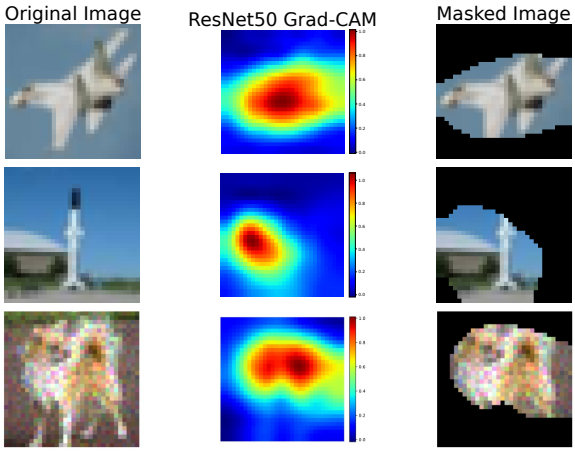

Figure 2: Visualization of resetnet50 based binary feature masking. (Top) CIFAR-10, (Middle) CIFAR-100, (Bottom) CIFAR-10C. Each sample shows Original Image (Left), Restnet50 Grad-CAM Heatmap (Middle) and Binary Masked Image (Right) based on the heatmap

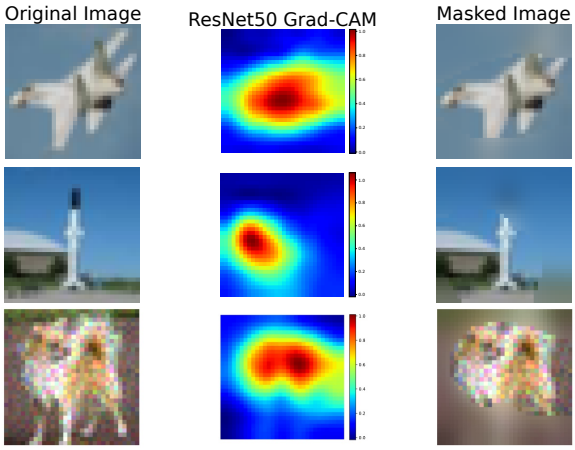

Figure 3: Visualization of Gaussian-Blurred feature masking.(Top) CIFAR-10, (Middle) CIFAR-100, (Bottom) CIFAR-10C. Each sample shows Original Image (Left), Restnet50 Grad-CAM Heatmap (Middle) and Gaussian-Blurred Masked Image (Right) based on the heatmap

Gaussian-Blurred Masking Binary masking may lead to the loss of spatial coherence, which can reduce generalization. To mitigate this, Gaussian-blurred masking replaces the masked-out regions with a smoothed version of the input image, shown in figure 3.

1. **Gaussian blur application**
   We generate a blurred version of the original image using a Gaussian kernel:

$$I_{\text{blurred}} = G(I,k) \tag{7}$$

   where k is the Gaussian kernel size controlling the blur intensity.
2. **Soft-masking strategy**
   The masked image is computed as a weighted combination of the original and blurred image:

$$I_{\text{masked}}(i,j) = I(i,j) \cdot M(i,j) + I_{\text{blurred}}(i,j) \cdot (1 - M(i,j)) \tag{8}$$

Difference-Based Masking Difference-based masking aims to eliminate unstable features by comparing Grad-CAM activations from the baseline model with those from a stronger ResNet-50 model.

1. **Difference heatmap**
   The difference heatmap is computed as:

$$H_{\text{diff}} = H_{\text{baseline}} - H_{\text{resnet}} \tag{9}$$

2. **ReLU-like filtering**
   Negative values (indicating shared features) are removed using ReLU-like filtering:

$$H_{\text{diff}}(i,j) = \max(H_{\text{diff}}(i,j), 0) \tag{10}$$

3. **Mask generation**
   The mask is computed as:

$$M(i,j) = 1 - H_{\text{diff}}^{\text{norm}} \tag{11}$$

4. **Masked image**
   The final masked image is computed as:

$$I_{\text{diffmasked}}(i,j) = I(i,j) \cdot M(i,j) \tag{12}$$

   Figure 4 shows samples of difference based masking on each three of our datasets.

Experimental Results

In this section, we present the details of our experiments, including the configuration of adversarial attacks and the evaluation process for each masking strategy. We analyze the impact of each approach on model robustness and discuss the results in detail.

Baseline Model Architecture

The baseline model is a custom CNN architecture designed for image classification tasks. It consists of six convolutional layers with filter sizes of 64, 128, 128, 256, 256, and 512. Each convolutional layer is followed by batch normalization to stabilize training and improve convergence. MaxPooling is applied after the second and fourth convolutional layers to reduce the spatial dimensions and prevent overfitting. Dropout is incorporated at three different stages with rates

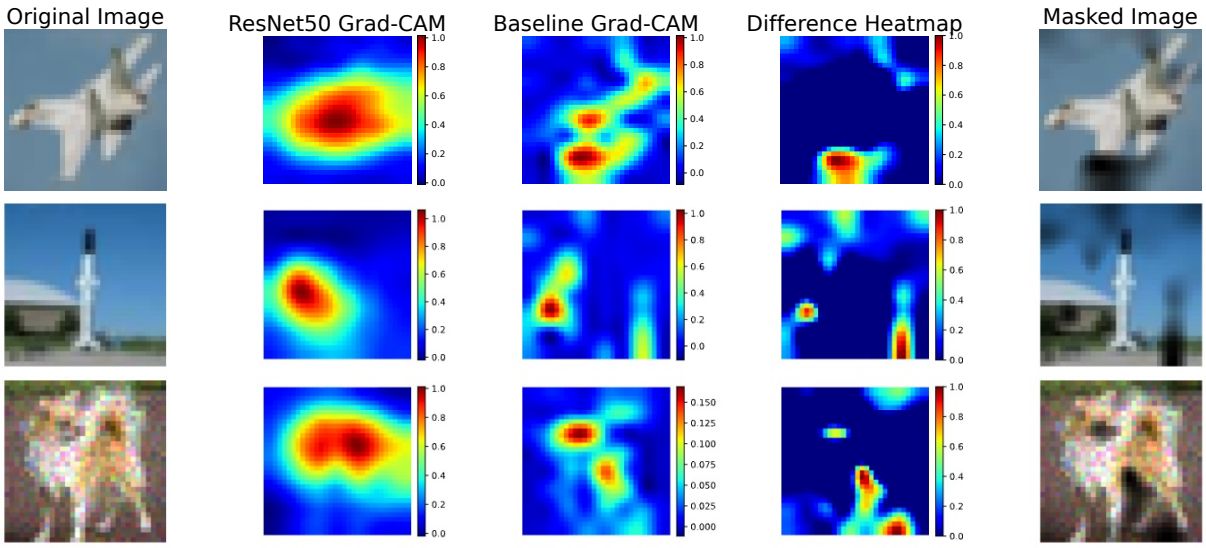

Figure 4: Visualization of difference based feature masking. (Top) CIFAR-10, (Middle) CIFAR-100, (Bottom) CIFAR-10C. Each sample shows Original Image (Left), followed by three Grad-CAM heatmap of Restnet50, Baseline Model and generated difference heatmap (Middle) and Difference Masked Image (Right) based on the heatmap

of 0.25, 0.3, and 0.4 to improve generalization by reducing co-adaptation of units. Global Average Pooling (GAP) is applied before the fully connected layers to reduce the number of parameters and aggregate spatial information. The final dense layer consists of 512 neurons followed by a softmax activation function for multi-class classification. The architecture is adjusted for two classification tasks: a 10-class output for CIFAR-10 and CIFAR-10-C datasets and a 100-class output for CIFAR-100 dataset. This model architecture provides a balance between expressiveness and computational efficiency, making it suitable for evaluating adversarial robustness under different masking strategies.

Datasets

We evaluate the proposed masking strategies using three widely used image classification datasets: CIFAR-10, CIFAR-100, and CIFAR-10-C. CIFAR-10 and CIFAR-100 each consist of 50,000 training images and 10,000 test images, each with a resolution of 32×32 pixels and three color channels (RGB). CIFAR-10 contains 10 object categories, while CIFAR-100 includes 100 object categories. The images are evenly distributed across all categories, ensuring a balanced classification task.

CIFAR-10-C is a corrupted version of CIFAR-10 that includes 950,000 labeled images with 19 different types of corruption (e.g., Gaussian noise, motion blur, brightness shifts). Each corruption type is presented at five different severity levels. For our experiments, we selected 90,000 images for training and 20,000 images for testing. To maintain a balanced dataset, we ensured that each corruption type contributed an equal number of samples to both the training and test sets.

Masked datasets were generated by applying each of the three masking techniques (binary, Gaussian-blurred, and difference-based) to the training and test sets of CIFAR-10, CIFAR-100, and CIFAR-10-C. This resulted in a total of nine masked datasets (three masking strategies applied to each of the three datasets). These masked datasets were used to retrain the baseline model and evaluate its robustness under adversarial attacks.

Model Training

We first trained the baseline CNN models on the original training images from CIFAR-10, CIFAR-100, and CIFAR-10-C for 200 epochs using the categorical cross-entropy loss function and a batch size of 32. The Adam optimizer was used for optimization with an initial learning rate of 0.001, which was decayed by 5% every 10 epochs. Batch normalization was applied after each convolutional layer to prevent internal covariate shift and improve training stability. Dropout was used at different levels to reduce overfitting and improve model generalization.

Adversarial Evaluation

To evaluate the impact of feature masking on adversarial robustness, we followed a two-stage process. First, we generated masked datasets by applying the three masking strategies—binary masking, Gaussian-blurred masking, and difference-based masking—to the training sets of CIFAR-10, CIFAR-100, and CIFAR-10-C. This resulted in a total of nine masked datasets. After generating the masked datasets, we retrained the baseline CNN model separately on each masked dataset for 100 epochs using the same optimizer, loss function, and learning rate settings as in the baseline training. This produced nine distinct models (three masked datasets for each of the three original datasets).

After retraining, we evaluated the adversarial robustness of the masked models using two common attack methods:

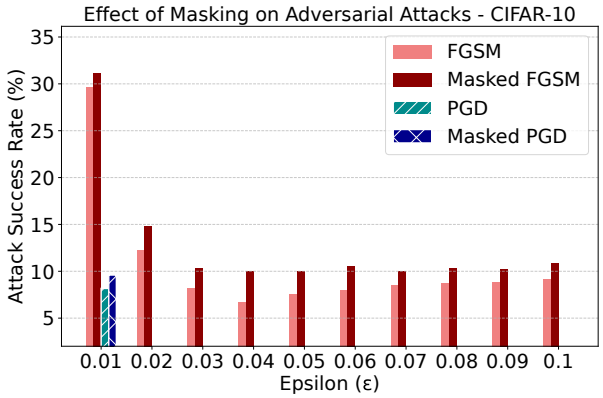

Figure 5: Adversarial attack results on CIFAR10 datasets. This bar chart compares adversarial accuracy of FGSM and PGD attacks with and without binary feature masking across different ϵ values.

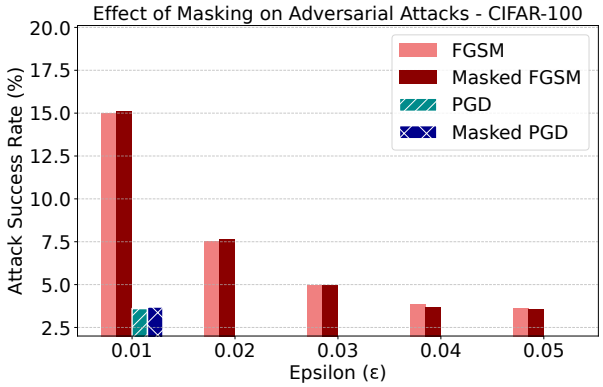

Figure 6: Adversarial attack results on CIFAR100 datasets. This bar chart compares adversarial accuracy of FGSM and PGD attacks with and without binary feature masking across different ϵ values.

Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM generates adversarial examples by perturbing the input in the direction of the gradient of the loss function. Specifically, adversarial examples are generated using the equation:

$$X_{\text{adv}} = X + \epsilon \cdot \text{sign}(\nabla_X \mathcal{L}) \tag{13}$$

where X is the original input, ϵ is the perturbation magnitude, and $\nabla_X \mathcal{L}$ is the gradient of the loss function with respect to the input.

PGD generates adversarial examples iteratively using a step size α according to the equation:

$$X_{\text{adv}}^{t+1} = \text{Proj}_\epsilon(X_{\text{adv}}^t + \alpha \cdot \text{sign}(\nabla_X \mathcal{L})) \tag{14}$$

where α is the step size and Proj_ϵ is the projection operator that ensures the perturbed sample remains within the L_∞ ball of radius ϵ.

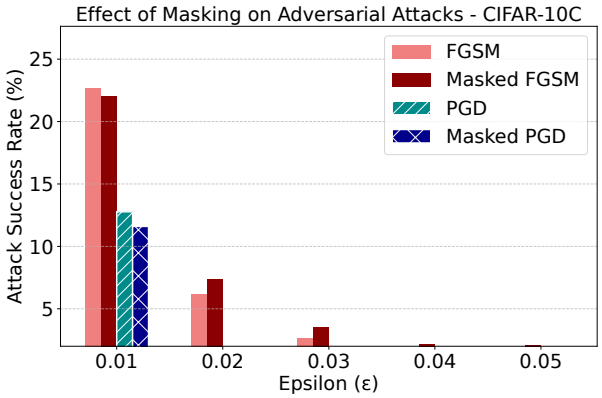

Figure 7: Adversarial attack results CIFAR10C datasets. This bar chart compares adversarial accuracy of FGSM and PGD attacks with and without binary feature masking across different ϵ values.

Results and Discussion

To evaluate the effectiveness of different feature masking methods, we conducted adversarial attack experiments using FGSM and PGD across a range of perturbation levels (ϵ). The results for binary masking, Gaussian-blurred masking, and difference-based masking are presented in Figure 5, 8 and in Tables 1, and 2. The analysis focuses on how each masking technique influences adversarial robustness under varying attack strengths and different datasets.

Binary Masking Results Binary masking is the most aggressive masking technique among the three approaches, as it completely removes low-activation regions identified by Grad-CAM, retaining only the highest-importance features. As shown in Figure 5 , 6 and 7, binary feature masking consistently improves adversarial robustness against FGSM attacks across all tested ϵ values for CIFAR-10, CIFAR-100, and CIFAR-10-C datasets. This indicates that binary masking helps the model focus on the most critical features, reducing sensitivity to noise and improving resilience against single-step gradient-based attacks.

However, the effectiveness of binary masking is less consistent against PGD attacks. While the results for FGSM are positive across all perturbation levels, binary masking improves PGD robustness only at lower perturbation strengths (e.g., $\epsilon = 0.01$ and $\epsilon = 0.02$). At higher perturbation levels, binary masking shows diminishing returns and eventually fails to defend against stronger iterative attacks like PGD. This suggests that while binary masking helps mitigate weaker adversarial attacks by sharpening the model's focus on high-saliency regions, it struggles to defend against more adaptive and iterative attacks, where masked regions might be exploited to craft stronger perturbations.

The overall pattern indicates that binary masking enhances adversarial robustness by reinforcing the model's reliance on the most influential features, which explains the consistent improvement in FGSM accuracy. However, the lack of contextual information due to aggressive feature removal appears to limit its ability to handle complex pertur-

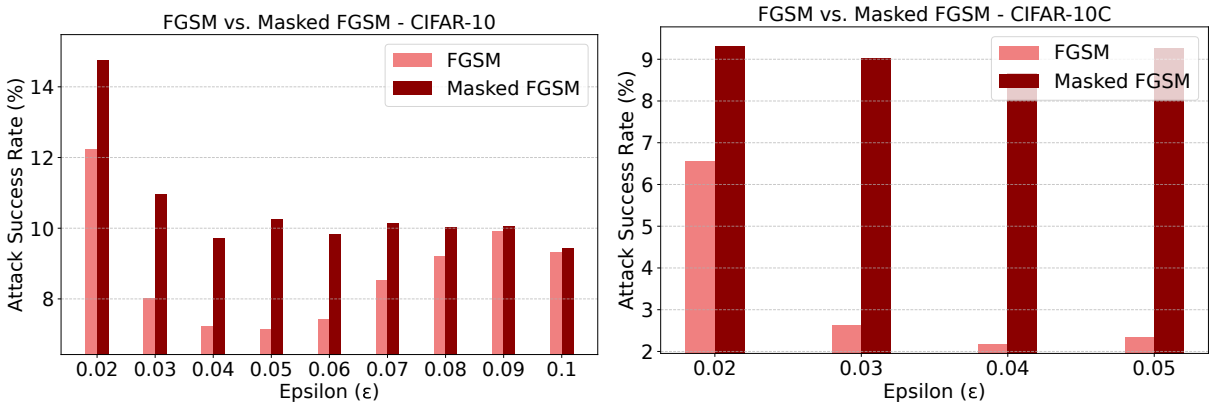

Figure 8: Adversarial attack results across different datasets. This bar chart compares adversarial accuracy of FGSM attacks with and without Gaussian-Blurred feature masking across different ϵ values.

bations generated by iterative methods like PGD. This highlights a fundamental trade-off between model specificity and generalization when employing binary masking as a defense mechanism.

Dataset	ϵ	PGD	Masked PGD
CIFAR-10	0.01	8.97%	10.06%
	0.02	0.00%	10.17%
	0.03	0.00%	9.91%
	0.04	0.00%	9.31%
	0.05	0.00%	10.31%
	0.06	0.00%	10.00%
	0.07	0.00%	10.46%
	0.08	0.00%	9.97%
	0.09	0.00%	10.29%
	0.10	0.00%	9.89%
CIFAR100	0.02	0.11%	1.09%
	0.03	0.00%	0.74%
	0.04	0.00%	1.20%
	0.05	0.00%	1.09%
CIFAR-10C	0.01	12.66%	8.83%
	0.02	0.83%	9.03%
	0.03	0.14%	8.86%
	0.04	0.00%	8.89%
	0.05	0.60%	9.20%

Table 1: Adversarial attack results across different datasets. The table compares adversarial accuracy of PGD attacks with and without Gaussian Blurred feature masking across different ϵ values.

Gaussian-Blurred Masking Gaussian-blurred masking preserves contextual information by applying a blurred transformation to the less relevant regions identified by Grad-CAM. Unlike binary masking, which completely removes low-activation regions, Gaussian-blurred masking softens these areas, allowing the model to retain global structure and context while focusing on discriminative features.

Table 1 shows that Gaussian-blurred masking significantly improves robustness against PGD attacks across all tested ϵ values for CIFAR-10, CIFAR-100, and CIFAR-10-C datasets. Notably, at a higher perturbation level of $\epsilon = 0.1$, Gaussian-blurred masking achieves a PGD accuracy of **9.89%** on CIFAR-10, compared to **0.0%** for the baseline model. This demonstrates that introducing a controlled level of smoothing in non-salient regions enhances the model's ability to withstand more aggressive iterative attacks.

Gaussian-blurred masking also shows consistent improvements against FGSM attacks as illustrate in Figure 8, albeit at a smaller scale than PGD. This contrasts with binary masking, which primarily benefits FGSM attacks but struggles with stronger, iterative perturbations. The ability of Gaussian-blurred masking to achieve uniform gains for both FGSM and PGD attacks—particularly at low perturbation levels—highlights its advantage in promoting better generalization. Soft attenuation of non-salient features appears to help the model maintain stable feature representations, thereby improving resilience against both single-step and multi-step adversarial attacks.

Dataset	ϵ	FGSM	Masked FGSM	PGD	Masked PGD
CIFAR-10	0.01	28.26%	30.37%	8.91%	8.26%
	0.02	11.80%	11.91%	0.00%	0.00%
	0.03	7.34%	8.51%	0.00%	0.00%
	0.04	7.06%	7.77%	0.00%	0.00%
	0.05	7.94%	8.29%	0.00%	0.00%
	0.06	8.23%	8.91%	0.00%	0.00%
	0.07	8.60%	9.29%	0.00%	0.00%
	0.08	8.89%	10.20%	0.00%	0.00%
	0.09	9.03%	9.89%	0.00%	0.00%
	0.10	9.56%	10.66%	0.00%	0.00%
CIFAR-100	0.02	6.60%	7.31%	0.14%	0.14%
	0.03	4.69%	4.94%	0.00%	0.00%
CIFAR-10C	0.02	7.20%	7.43%	0.83%	1.00%
	0.03	2.17%	3.34%	0.14%	0.09%
	0.04	2.14%	2.63%	0.20%	0.00%
	0.05	1.97%	2.37%	0.09%	0.00%

Table 2: Adversarial Attack Results Across Different Datasets. The table compares adversarial accuracy of FGSM and PGD attacks with and without difference based feature masking across different ϵ values.

Difference-Based Masking Difference-based masking removes the least amount of pixel information among all

the masking strategies, masking only a small portion of the image. As shown in Figure 4, despite the minimal modification to the input, this approach consistently improves FGSM accuracy across all tested ϵ values. The steady increase in FGSM accuracy suggests that eliminating unstable, low-saliency features helps the model focus on more robust and generalizable patterns, enhancing its resistance to single-step gradient-based attacks. However, the effectiveness of difference-based masking against PGD attacks is more limited. Table 2 shows that this masking technique improves robustness against weaker PGD attacks (e.g., at lower perturbation levels), but its defensive capability diminishes under stronger iterative attacks. This outcome reflects the relatively conservative nature of difference-based masking—while it retains most of the original image information, it may fail to sufficiently suppress attack-prone regions under more aggressive perturbations. Nevertheless, the consistent improvement in FGSM accuracy demonstrates that even small-scale feature refinement can enhance the model's resilience to single-step adversarial attacks.

Conclusion

In this work, we proposed an explainability-driven defense framework that leverages Grad-CAM-guided feature masking to enhance the adversarial robustness of Convolutional Neural Networks. We introduced and systematically evaluated three distinct masking strategies: binary masking, Gaussian-blurred masking, and difference-based masking. Each strategy was designed to selectively retain or suppress input features based on their saliency, encouraging the model to focus on stable and discriminative patterns while mitigating the impact of adversarial perturbations. Our experimental results demonstrate that all three masking strategies improve adversarial robustness to varying degrees across different datasets and attack methods. While the study highlights the potential of feature masking as a lightweight adversarial defense, several areas warrant further exploration. A key direction is to investigate whether counterfactual feature masking can enhance robustness further, as well as how varying the strength of blurring and masking thresholds impacts performance. Integrating explainability techniques to analyze the influence of masking on model decision-making could provide deeper insights into adversarial robustness.

References

Baniecki, H.; and Biecek, P. 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107: 102303.

Chakraborty, T.; Trehan, U.; Mallat, K.; and Dugelay, J.-L. 2022. Generalizing adversarial explanations with Grad-CAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 187–193.

Cui, X.; Aparcedo, A.; Jang, Y. K.; and Lim, S.-N. 2024. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24625–24634.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Guo, C.; Rana, M.; Cisse, M.; and Van Der Maaten, L. 2017. Countering adversarial images using input transformations.

Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; and Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, T.; Menkovski, V.; Pei, Y.; and Pechenizkiy, M. 2020. Bridging the performance gap between fgsm and pgd adversarial training. *arXiv preprint arXiv:2011.05157*.

Jakubovitz, D.; and Giryes, R. 2018. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European conference on computer vision (ECCV)*, 514–529.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Lin, Y.-S.; Lee, W.-C.; and Celik, Z. B. 2021. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*, 1027–1035. ACM.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Nayyem, N.; Rakin, A.; and Wang, L. 2024. Bridging Interpretability and Robustness Using LIME-Guided Model Refinement. *arXiv preprint arXiv:2412.18952*.

Qin, C.; Martens, J.; Gowal, S.; Krishnan, D.; Dvijotham, K.; Fawzi, A.; De, S.; Stanforth, R.; and Kohli, P. 2019. Adversarial robustness through local linearization. *Advances in neural information processing systems*, 32.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ross, A. S.; and Doshi-Velez, F. 2018. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 1660–1669. AAAI Press.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. 618–626.

Shi, M.; Wang, R.; Liu, E.; Xu, Z.; and Wang, L. 2020. Deep reinforcement learning based computation offloading for mobility-aware edge computing. In *Communications and Networking: 14th EAI International Conference, ChinaCom 2019, Shanghai, China, November 29–December 1, 2019, Proceedings, Part I 14*, 53–65. Springer.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. arXiv:1805.12152.

van Zyl, C.; Ye, X.; and Naidoo, R. 2024. Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, 353: 122079.

Villegas-Ch, W.; Jaramillo-Alcázar, A.; and Luján-Mora, S. 2024. Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW. *Big Data and Cognitive Computing*, 8(8): 1–23.

Waghela, H.; Sen, J.; and Rakshit, S. 2024. Robust Image Classification: Defensive Strategies against FGSM and PGD Adversarial Attacks. arXiv:2408.13274.

Wang, L.; Chen, P.; Wang, C.; and Wang, R. 2020. Layer-wise entropy analysis and visualization of neurons activation. In *Communications and Networking: 14th EAI International Conference, ChinaCom 2019, Shanghai, China, November 29–December 1, 2019, Proceedings, Part II 14*, 29–36. Springer International Publishing.

Wang, L.; Ghimire, A.; and Santosh, K. 2024. Enhanced Model Robustness by Integrated Local and Global Processing. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, 234–239. IEEE.

Wang, L.; Ghimire, A.; Santosh, K.; Zhang, Z.; and Li, X. 2024. Enhanced robustness by symmetry enforcement. *IEEE CAI*.

Wang, L.; Li, X.; and Zhang, Z. 2024. Dense cross-connected ensemble convolutional neural networks for enhanced model robustness. *arXiv preprint arXiv:2412.07022*.

Wang, L.; and Li, Y. 2019. Information theory and representation learning inspired multimodal data fusion. *IEEE MMTC Frontier*.

Wang, L.; and Liang, Q. 2019. Representation learning and nature encoded fusion for heterogeneous sensor networks. *IEEE Access*, 7: 39227–39235.

Wang, L.; Wang, C.; Li, Y.; and Wang, R. 2021a. Explaining the behavior of neuron activations in deep neural networks. *Ad Hoc Networks*, 111: 102346.

Wang, L.; Wang, C.; Li, Y.; and Wang, R. 2021b. Improving robustness of deep neural networks via large-difference transformation. *Neurocomputing*, 450: 411–419.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021c. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7639–7648.

Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 501–509.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy.

Zhang, Y.; Zhu, Y.; Liu, J.; Yu, W.; and Jiang, C. 2025. An Interpretability Optimization Method for Deep Learning Networks Based on Grad-CAM. *IEEE Internet of Things Journal*, 12(4): 3961–3970.