remote sensing

MDPI

Technical Note

Improve Adversarial Robustness of AI Models in Remote Sensing via Data-Augmentation and Explainable-AI Methods

Sumaiya Tasneem and Kazi Aminul Islam *

Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA; stasneem@students.kennesaw.edu

* Correspondence: kislam4@kennesaw.edu

Abstract: Artificial intelligence (AI) has made remarkable progress in recent years in remote sensing applications, including environmental monitoring, crisis management, city planning, and agriculture. However, the critical challenge in utilizing AI models in real-world remote sensing applications is maintaining their robustness and reliability, particularly against adversarial attacks. In adversarial attacks, attackers manipulate benign data to create a perturbation to mislead AI models into predicting incorrect decisions, posing a catastrophic threat to the security of their applications, particularly in crucial decision-making contexts. These attacks pose a significant threat to the integrity and comprehensiveness of AI models in remote sensing applications, as they can lead to inaccurate decisions with substantial consequences. In this paper, we propose to develop an adversarial robustness technique that will ensure the AI model's accurate prediction in the presence of adversarial perturbation. In this work, we address these challenges by developing a better adversarial training approach using explainable AI method-guided features and data augmentation techniques to strengthen the AI model prediction in remote sensing data against adversarial attacks. The proposed approach achieved the best adversarial robustness against Project Gradient Descent (PGD) attacks in EuroSAT and AID datasets and showed transferability of robustness against unseen attacks.

Keywords: deep learning; adversarial attack; adversarial robustness; explainable AI; model interpretability; remote sensing; data augmentation

Citation: Tasneem, S.; Islam, K.A. Improve Adversarial Robustness of AI Models in Remote Sensing via Data-Augmentation and Explainable-AI Methods. *Remote Sens.* **2024**, *16*, 3210. https://doi.org/10.3390/rs16173210

Academic Editors: Saeid Homayouni and Jiaojiao Li

Received: 16 June 2024
Revised: 11 August 2024
Accepted: 26 August 2024
Published: 30 August 2024

1. Introduction

In recent years, the integration of Artificial Intelligence (AI) and remote sensing has been successfully applied across various fields, including environmental monitoring, disaster management, building urban settlements, and farming [1,2]. Deep learning algorithms, particularly deep convolutional neural networks, He et al. [3] have demonstrated significant performance improvements over traditional methods [4]. Deep learning algorithm's advantages of these algorithms include direct usage of the feature vectors, rapid training and testing times, and superior generalization capabilities compared to traditional classification methods [5].

Despite these advancements, AI systems in remote sensing are vulnerable to adversarial attacks, which means intentionally adding perturbed input to the benign data to mislead machine learning models into making incorrect predictions. Adversarial attacks in remote sensing can significantly threaten the integrity of machine learning models used to inspect satellite imagery, aerial photographs, and other geospatial data [6–8]. Adversarial attacks deliberately manipulate benign data with malicious input, eventually creating erroneous AI model predictions. For instance, adversarial algorithms can deceive remote sensing models into misclassifying aircraft as birds, which has severe implications in military applications [7]. Researchers have employed various adversarial attack methods, e.g., Fast Gradient Sign Method (FGSM) [9], Basic Iterative Method (BIM) [10], Carlini & Wagner (C&W) [11], and Projected Gradient Descent (PGD) [12] to assess the vulnerability of remote sensing image scene classification systems [6,7]. Chan et al. [6] and Cheng et al. [13]

evaluated adversarial attack's impact in deep convolutional neural networks (DCNN) for scene classification, land cover mapping, and object detection in remote sensing. These models are susceptible to adversarial examples, leading to potential misclassifications and compromised model performance [6].

To counter these adversarial threats, several defense strategies have been introduced. Adversarial training [14] involves incorporating adversarial examples into the training process to improve model robustness. Adversarial regularization aims to enhance the model's resilience by adding regularization terms that penalize vulnerability to adversarial perturbations. Additionally, techniques like Progressive Generative Adversarial Networks (PSGAN) have been proposed to further bolster defenses against these sophisticated attacks [13], specifically tailored for remote sensing applications. PSGANs introduced reconstructed examples generated during image reconstruction, alongside clean and adversarial examples, to bolster classifier resilience against known and unknown adversarial attacks. These findings underscore the critical need for ongoing research to develop more resilient models capable of withstanding adversarial threats in the remote sensing domain. Most defense approaches lose clean data accuracy by achieving adversarial robustness.

Furthermore, the high level of complexity within the remote sensing data, which can change with the characteristics of lighting conditions, meteorology, and sensor systems, forms an extra barrier that makes it challenging to develop AI models that are robust enough [15]. Due to these challenges, it is crucial to develop highly-performing AI models for the field of remote sensing. Explainable AI (XAI) refers to methods and processes that provide insights into how machine learning models make decisions [16,17]. But those XAI methods can be manipulated by adversarial attacks [18] that create a need to train machine learning models to produce robust interpretations for their predictions [19]. Boopathy et al. [20] highlighted the integration of XAI into robust training frameworks to improve adversarial robustness by losing clean data accuracy.

Most of these defense approaches are developed for natural images that might not have a similar effectiveness in remote sensing. The high complexity and variability of remote sensing data pose additional challenges that are not fully addressed by current defense approaches. Existing defense methods often have a trade-off between adversarial robustness and clean data accuracy. It is crucial to develop new techniques that can enhance robustness without sacrificing performance on clean data. This paper addresses the research gap in achieving adversarial robustness in remote sensing.

We propose a novel adversarial robustness technique that combines robust interpretable features with data augmentation techniques. Our approach aims to enhance the robustness of AI models against adversarial attacks while maintaining high accuracy on clean data. We validate our method using the EuroSAT [5] and AID (Aerial Image Dataset) [21] datasets, demonstrating its effectiveness across diverse and complex remote sensing scenarios. Additionally, we apply the CAM [22] method to visualize its results on both clean and perturbed data after PGD attacks. Our experiments show satisfactory adversarial test accuracy (ATA) against PGD attacks, underscoring the potential of our approach to fill the existing research gaps. Additionally, we evaluate the transferability of the robustness against other attacks. Transferability refers to a model's ability to maintain its performance and robustness against newer or unseen types of adversarial attack. Our work aims to contribute to the development of more reliable, interpretable, and transferable AI models in remote sensing applications.

Our overall contributions can be summarized as follows:

- We proposed an adversarial robustness technique that uses robust interpretable features with data augmentation to enhance the robustness of AI models against adversarial attacks in remote sensing applications.
- We validated our approach using EuroSAT and AID datasets, demonstrating its effectiveness across diverse and complex remote sensing scenarios.
- We applied SaliencyMix [23] augmentation to improve adversarial robustness and clean data, which performed better than the traditional data-augmentation technique.

- We evaluated transferability of the robustness against FGSM and BIM attacks and achieved similar consistency as PGD attack.

2. Methods

In this section, we have provided an overview of the methods and concepts used in our research. It includes a discussion on the threat model for adversarial example attacks, highlighting the techniques employed to generate adversarial perturbations. Key methods such as the FGSM and PGD are explained, demonstrating their application and impact on different datasets.

2.1. Threat Model: Adversarial Example Attack

Adversarial example attacks are specific input perturbations, or changes that make machine learning models predict incorrect information. The attacker cannot access the training process to poison the model f. However, the attacker can access or query the trained machine learning model's weight to generate the adversarial perturbations. The attacker can generate the adversarial example, x_{adv} by adding a perturbation δ to the clean image X. The attacker can choose to use any adversarial example generation method to achieve the best attack success rate by misclassifying the image into the wrong class, $f(x) \neq f(x_{adv})$.

FGSM, introduced by Goodfellow et al. in 2015 [9], is a white box and targeted attack. It adds subtle noise to the input data, aiming to maximize the loss function's value in a targeted way. To do that, it first calculates the loss function's gradient for the input image. This gradient represents the direction in which the loss function increases the fastest for small changes in the input data. FGSM then uses this gradient to generate a perturbation (adversarial noise) by multiplying it by a small constant value, ϵ. The positive or negative sign of the gradient indicates whether the perturbation is added to or subtracted from the input image. Conceptually, we can illustrate it as Equation (1), where x is the benign image, x_{adv} is the generated perturbed image, ϵ is the multiplication factor. ∇_x is the gradient with respect to the input x and $J(\Theta, x, y)$ is the loss function.

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

FGSM is a popular attack method because of its simplicity since it requires only one step to attack. It is suitable for scenarios requiring a quick generation of adversarial examples. However, it also has some limitations, including being less effective against models trained with robustness techniques or defenses specifically designed to mitigate gradient-based attacks.

PGD is the prolonged version of FGSM, which was developed by Madry et al. [12] to overcome the limitations of FGSM. This algorithm uses iterations of small perturbations to the input data. These perturbations are kept within a specified range. Similar to FGSM, PGD calculates the gradient of the loss function. However, instead of applying a single perturbation in one step, it applies the perturbations in multiple steps shown in Equation (2) where P denotes the projection operator.

$$\delta_{i+1} = P(\delta_i + \epsilon \cdot (\nabla_x J(\Theta, x + \delta_i, y)) \tag{2}$$

This iterative process allows PGD to explore the input space more comprehensively and find more effective adversarial examples compared to FGSM. Despite its effectiveness, PGD requires more computational resources due to its iterative nature. However, its ability to generate robust adversarial examples makes it a valuable technique for testing the resilience of machine learning models, particularly those used in critical applications such as medical image analysis. We demonstrate the impact of perturbation strengths ranging from $\epsilon = 2/255$ to $\epsilon = 10/255$ generated by the PGD attack on the clean samples of the EuroSAT and AID datasets in Figures 1 and 2 respectively. Although adversarial perturbations are added in these figures, the images still visually appear as benign.

(**a**) Original Sample (**b**) $\epsilon = 2/255$ (**c**) $\epsilon = 4/255$

(**d**) $\epsilon = 6/255$ (**e**) $\epsilon = 8/255$ (**f**) $\epsilon = 10/255$

Figure 1. Example of PGD attack for different perturbation strengths (ϵ) on sample from "Annual Crop" class in EuroSAT dataset.

(**a**) Original Sample (**b**) $\epsilon = 2/255$ (**c**) $\epsilon = 4/255$

(**d**) $\epsilon = 6/255$ (**e**) $\epsilon = 8/255$ (**f**) $\epsilon = 10/255$

Figure 2. Example of PGD attack for different perturbation strengths (ϵ) on sample from "Park" class in AID dataset.

2.2. Explainable AI Methods

Class Activation Map (CAM), proposed by Zhou et al. [22], works by generating a heatmap that highlights the regions of the input image that contributed the most to the final classification decision. This heatmap is created by examining the activations of the convolutional layers in a neural network. Formally, a CAM, M_c can be defined by

Equation (3). Here c refers to the class, w_k is the weight of the corresponding class and A_k represents the activation of the last convolutional layer at the spatial location (x, y).

$$M_c(x, y) = \sum_k w_k^c A_k(x, y) \tag{3}$$

2.3. Interpretation Discrepancy

Interpretation discrepancy refers to the difference in interpretation between a natural input example (x) and its corresponding adversarial example x' [20]. This can be quantified using a generic form of l_p norm-based interpretation discrepancy, denoted as $\mathcal{D}(x, x')$, where p can be either 1 or 2. We represent the interpretation discrepancy $\mathcal{D}(x, x')$ as follows:

$$\mathcal{D}(x, x') = \frac{1}{c} \sum_{i \in c} ||I(x, i) - I(x', i)||_p \tag{4}$$

where $I(x, i)$ represents the interpretation of the input x for the class label i, $I(x', i)$ represents the interpretation of the adversarial input x' for the class label i, and c denotes the class label.

Interpretation discrepancy can significantly impact the robustness and reliability of machine learning models, particularly in the context of adversarial attacks and model interpretability. Significant discrepancies between model interpretations for clean and perturbed inputs within the same class highlight potential vulnerabilities in the model's predictive capabilities. High interpretation discrepancy indicates inconsistent model behavior, undermining the reliability of its explanations and suggesting limited applicability across different inputs. Explanation methods, including CAM [22], GradCAM [16], and ScoreCAM [17] can be used to mitigate interpretation discrepancy, thereby enhancing the model's resilience against adversarial perturbations and improving the trustworthiness of model explanations.

3. Comparison Method for Adversarial Robustness

Adversarial robustness of a machine learning model refers to the fact that the machine learning model can maintain its performance in the event of adversarial attacks. Our proposed approach uses data augmentation and robust interpretable features to train the model, which ensures the model can correctly identify the object in the presence of adversarial perturbation. We compare several techniques, as follows:

3.1. Adversarial Training

Adversarial training is a common technique to improve adversarial robustness. It utilizes the training data with adversarially perturbed examples to expose the model to a diverse set of challenging inputs during training [12]. We repeatedly trained the model on clean samples and adversarial examples to achieve robust predictions against adversarial perturbations. We utilized adversarial training with the PGD (Projected Gradient Descent) based attack to create adversarial examples. The PGD attack iteratively perturbs the input data to maximize the loss within a specified perturbation budget mentioned in Section 2.1. The basic adversarial training [12] framework with PGD attack can be formulated as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \Delta} \mathcal{L}(\theta, x + \delta, y) \right] \tag{5}$$

Here θ represents the model parameters. The dataset $\mathcal{D}$ consists of input data x and corresponding labels y. Adversarial perturbations, denoted by δ, are added to the input x. Δ is the set of allowed perturbations, typically constrained by $\delta \leq \epsilon$. The loss function used for training is denoted as $\mathcal{L}$.

3.2. Robustness Using Interpretability

We compared interpretability-aware robustness training methods proposed by Akhilan et al. [20] to improve adversarial robustness against adversarial attacks. Recalling from Section 2.3, the interpretability-aware defense method can reduce interpretation discrepancies to increase robustness. The target label-free interpretation discrepancy measure, denoted by Equation (6), quantifies the difference in interpretation between a natural example x and its adversarial example x'.

$$\begin{array}{ll} \tilde{\mathcal{D}}(\mathbf{x}, \mathbf{x}') & = (1/2)\|I(\mathbf{x}, y) - I(\mathbf{x}', y)\|_1 \\ + & (1/2)\sum_{i \neq t} \frac{e^{f(\mathbf{x}')_i}}{\sum_{i'} e^{f(\mathbf{x}')_{i'}}} \|I(\mathbf{x}, i) - I(\mathbf{x}', i)\|_1 \end{array} \tag{6}$$

The first term calculates the disparity in interpretation for the true label y, while the second term considers discrepancies in interpretations for other non-true labels, weighted by their importance in prediction. Based on this loss, interpretability-aware training methods were developed to train the classifier against the worst-case interpretation discrepancy. The following min-max optimization problem is used in interpretability-aware robustness training:

$$\min_{\theta} \mathbb{E}(x, t) \sim \mathcal{D}_{\text{train}}[f_{\text{train}}(\theta; x, y) + \gamma \tilde{\mathcal{D}}_{\text{worst}}(x, x')] \tag{7}$$

Here θ represents the model parameters, $\mathcal{D}_{\text{train}}$ indicates training data, f_{train} signifies the cross-entropy loss, $\tilde{\mathcal{D}}_{\text{worst}}$ measures the worst-case interpretation discrepancy between benign input x and perturbed input x', and γ regulates the balance between accuracy and interpretability robustness. Depending on the variation in measuring the worst-case interpretation discrepancy, two methods were proposed: *Int* and *Int*2.

3.2.1. *Int* and *Int* − *Adv*

This method aims to improve the robustness of the classifier by incorporating interpretability into the training process. It penalizes the interpretation discrepancy between natural and perturbed examples. The training process involves a min-max optimization problem (7) where the outer minimization aims to learn model parameters that reduce classification loss, and the inner maximization identifies the worst-case interpretation discrepancy within a defined perturbation bound. It uses the worst-case interpretation discrepancy measure defined in Equation (8) to maximize the interpretation discrepancy under l_∞ perturbations, where $\tilde{\mathcal{D}}(x, x')$ represents the discrepancy in interpretations between x and its perturbed version $x + \delta$.

$$\tilde{\mathcal{D}}_{\text{worst}}(\mathbf{x}, \mathbf{x}') := \max_{\|\delta\|_\infty \leq \epsilon} \tilde{\mathcal{D}}(\mathbf{x}, \mathbf{x} + \boldsymbol{\delta}) \tag{8}$$

This method focuses solely on interpretability discrepancy without directly incorporating adversarial examples designed to misclassify. In contrast, another variation of this method called *int* − *adv* enhances robustness by incorporating interpretability and integrating adversarial training to improve robustness against adversarial attacks. The training process involves minimizing classification loss and penalizing interpretation discrepancy while also including an adversarial loss (Equation (5)) component:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{\text{train}}} \left[f_{\text{train}}(\theta; x, y) + \gamma \tilde{\mathcal{D}}_{\text{worst}} + \text{adversarial Loss} \right] \tag{9}$$

The adversarial loss ensures the model is robust against inputs intentionally perturbed to cause misclassification. The main difference between *int* and *int* − *adv* methods is that *int* − *adv* method directly adds an adversarial loss in the training process.

3.2.2. $Int2$ and $Int2-Adv$

The $Int2$ method focuses on robustness by penalizing interpretation discrepancy while considering perturbations specifically aimed at causing misclassification. It uses a different interpretation discrepancy measure:

$$\tilde{\mathcal{D}}_{\text{worst}}\left(x, x'\right) := \tilde{\mathcal{D}}\left(x, x + \arg\max_{\|\delta\|_\infty \leq \epsilon} f_{\text{train}}\left(\theta; x+\delta, y\right)\right) \tag{10}$$

Here $f_{\text{train}}(\theta; x+\delta, y)$ is the adversarial loss that aims to maximize the difference between the predicted label and the true label, thereby causing misclassification. $\tilde{\mathcal{D}}(x, x+\delta)$ quantifies the difference between the interpretation maps of the natural example x and the perturbed example $x+\delta$.

$Int2-Adv$ combines robustness against interpretation discrepancy with a focus on misclassification perturbations and integrates adversarial training (Equation (5)) to enhance overall robustness. It utilizes a min-max optimization with an additional adversarial loss component.

Int and $Int2$ methods diverge in the focus and selection of perturbations during training. The Int method targets perturbations that maximize the interpretation discrepancy, aiming to generate adversarial examples that disrupt the model's interpretability. In contrast, the $Int2$ method targets perturbations that cause misclassification and maximizes the interpretation discrepancy for those misclassified examples. This dual focus ensures robustness against adversarial attacks that lead to incorrect predictions while maintaining consistent interpretations. Thus, Int primarily addresses interpretability robustness, while $Int2$ balances classification robustness and interpretability by targeting misclassification-induced interpretation discrepancies.

3.3. Traditional Data Augmentation

To mitigate the trade-off between adversarial robustness and clean data accuracy, we integrated data augmentation techniques with interpretability-aware robustness training to enhance performance against adversarial attacks. Initially, we employed the traditional data augmentation methods mentioned to verify their effectiveness in improving clean data accuracy and adversarial robustness.

Data augmentation is a technique widely utilized in machine learning and computer vision to artificially expand the size of the training dataset by applying various transformation techniques to the existing dataset [24]. The primary goal of data augmentation is to introduce diversity and variability into the training data, which can improve the model's ability to generalize and make accurate predictions on unseen data. Traditional data augmentation methods include rotation, translation, shearing, zooming, and flipping applied to images or data samples. These transformations, such as rotating images to simulate different viewpoints, shifting them horizontally or vertically to represent changes in perspective, or even mirroring them to introduce variation, serve to diversify the dataset.

4. Proposed Adversarial Robustness Method

We trained the model using both original examples and adversarial examples generated through the PGD attack. During training, we utilized cross-entropy loss for the classification task to measure the dissimilarity between the predicted probability distribution and the true label distribution. This loss optimizes classification performance, ensuring accurate predictions on natural examples.

Additionally, we generated explanation maps for both original and adversarial inputs using the CAM method. Interpretation discrepancy, as outlined in Boopathy et al.'s work [20], was calculated from these maps. While calculating the interpretation discrepancy, we incorporated a regularization term, which ensures that the model's explanations or interpretations remain consistent and reliable across different input variations, including adversarial perturbations. This explicit constraint minimizes interpretation differences

between natural and adversarial examples, enabling the model to provide consistent and reliable interpretations, ultimately leading to more trustworthy and robust machine learning systems. However, despite achieving the expected robustness against PGD attacks, we observed low accuracy in clean testing data. To mitigate this challenge, we propose a data augmentation-based adversarial robustness training approach that leverages both clean and augmented samples, as illustrated in Figure 3.

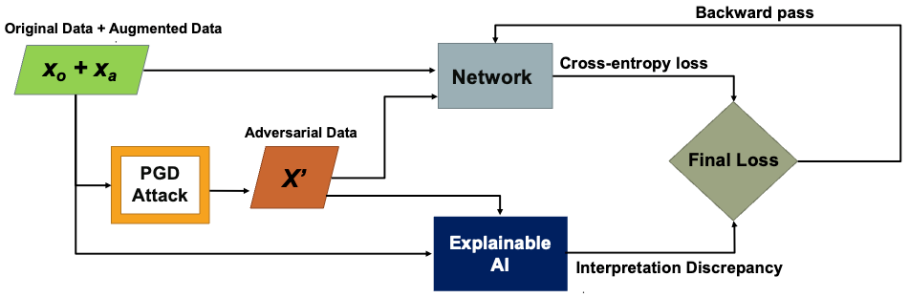

Figure 3. The proposed adversarial Robustness Approach via ExplainableAI and Data Augmentation.

The SaliencyMix [23] data augmentation method focuses on selecting image patches based on the saliency (explanation) information to enhance model training. It begins with a saliency detection algorithm that generates a saliency map for a given source image. The most salient region within this map is identified, allowing the selection of a patch that contains significant object information. Then this selected source patch is combined with a target image using a binary mask to create a mixed image sample, represented as:

$$X_{\text{mix}} = M \odot X_{\text{source}} + (1 - M) \odot X_{\text{target}} \tag{11}$$

Here, X_{mix} is the augmented image, M is the binary mask, X_{source} is the source image patch, and X_{target} is the target image. The multiplication $\odot$ represents element-wise multiplication between the binary mask and the images. In addition to mixing images, SaliencyMix also mixes the labels of the source and target images based on the sizes of the patches. The mixed label is defined as Equation (12), where Y_{mix} is the mixed label and Y_{target} are the labels of the source and target images, respectively, and α is the mixing ratio based on patch sizes.

$$Y_{\text{mix}} = \alpha Y_{\text{source}} + (1 - \alpha) Y_{\text{target}} \tag{12}$$

The objective function for training models with SaliencyMix includes the standard cross-entropy loss L_{CE} and a regularization term L_{reg}, combined as Equation (13) where λ controls the strength of the regularization.

$$L = L_{CE} + \lambda L_{reg} \tag{13}$$

Thus, SaliencyMix enhances model performance and robustness by integrating saliency-guided patch selection, image and label mixing, and a well-structured objective function.

5. Experimental Setup

5.1. Datasets

In this experiment, we have used two remote sensing datasets EuroSAT [5] and AID (Aerial Image Dataset) [21]. EuroSAT dataset is a collection of satellite images of European land cover. The images, acquired from the Sentinel-2 satellite, consist of 27,000 labeled 64 × 64 image patches. These patches represent ten types of land cover, including urban areas, farms, forests, and water bodies, where each class contains images ranging from 2000 to 3000. The AID dataset is a collection of aerial images of diverse land cover types in China. The images, acquired from Google Earth, consist of 10,000 labeled 600 × 600 image patches. These patches represent thirty types of land cover, including

residential areas, farmlands, forests, and water bodies, with each class containing between 220 and 420 images. The EuroSAT dataset is a collection of satellite images of European land cover. The AID dataset offers high-resolution RGB images, while the EuroSAT dataset offers both RGB and multispectral images containing 13 bands. To ensure fair comparisons, we used image patches with a dimension of $600 \times 600 \times 3$ for AID and $64 \times 64 \times 3$ for EuroSAT. However, during implementation, we resized the AID dataset images to 200×200 to reduce computational complexity and ensure efficient processing without significantly compromising the spatial resolution. To ensure equitable representation, we have balanced both datasets through a $70 - 30$ split, dedicating 70% of the data for training purposes and 30% for testing.

5.2. Convolutional Neural Network (CNN) Architecture

For training and evaluating our experiments with the EuroSAT and AID datasets, we utilized a small CNN architecture consisting of three convolutional layers with padding, which was used to maintain the spatial dimensions of the input feature map throughout the network. The first convolution layer has a 4×4 kernel size and a stride of 2 with a 16 number of filters; the second convolution layer has a 4×4 kernel size and a stride of 2 with a 32 number of filters; and the third convolution layer has a 7×7 kernel size and a stride of 1 with a 100 number of filters. Then, we use global average max pooling, followed by a 1×1 convolutional. Next, we use a flatten layer to convert the features into a vector representation, and finally, we apply a softmax cross-entropy for classification.

5.3. Hyper-Parameters

In our work, we have experimented and fine-tuned the overall performance with different hyper-parameters. We set the learning rate according to the number of epochs, starting at 0.001 for the initial 50 epochs, reducing to 0.0001 until epoch 100. We trained the model for 100 epochs with a batch size of 64. In the case of adversarial training, we specified, step size of $2/255$ and 10 adversarial steps. We applied a regularization parameter λ of 0.001 to prevent overfitting. ReLU [25] has been used as an activation function through the entire network, while Adam [26] was used as an optimizer. This hyperparameter configuration was selected to ensure optimal performance and generalization.

5.4. Evaluation Matrix: Adversarial Test Accuracy (ATA)

We used Adversarial Test Accuracy (ATA) for evaluating adversarial robustness. It measures the model's ability to correctly classify adversarial examples. To calculate ATA, adversarial examples are first generated using an attack method, such as FGSM, PGD, BIM, etc., from a set of clean (non-adversarial) inputs. The clean inputs and the adversarial examples are then passed through the model to obtain predictions. The number of correct predictions on the adversarial examples by the robust model compared to total adversarial examples used as input. The ATA is calculated using the following formula:

$$\text{ATA} = \left(\frac{\text{Number of Correct Predictions on Adversarial Examples}}{\text{Total Number of Adversarial Examples}} \right) \times 100\% \quad (14)$$

ATA is a quantitative metric assessing a model's resilience to adversarial attacks. A higher ATA indicates superior robustness, as the model can maintain accurate predictions even when presented with maliciously perturbed inputs.

6. Results

We trained the model with the proposed interpretability-aware training methods mentioned in Section 4. We also compared the models in normal and applied PGD-based adversarial training [18] settings to compare with the performance of the interpretability-aware training methods Int, $Int - adv$, $Int2$, and $Int2 - adv$. We also used only cross-entropy loss to train the model and labeled it as a normal training method. In our experiment, we generated adversarial examples using the PGD attack shown in Figures 4 and 5.

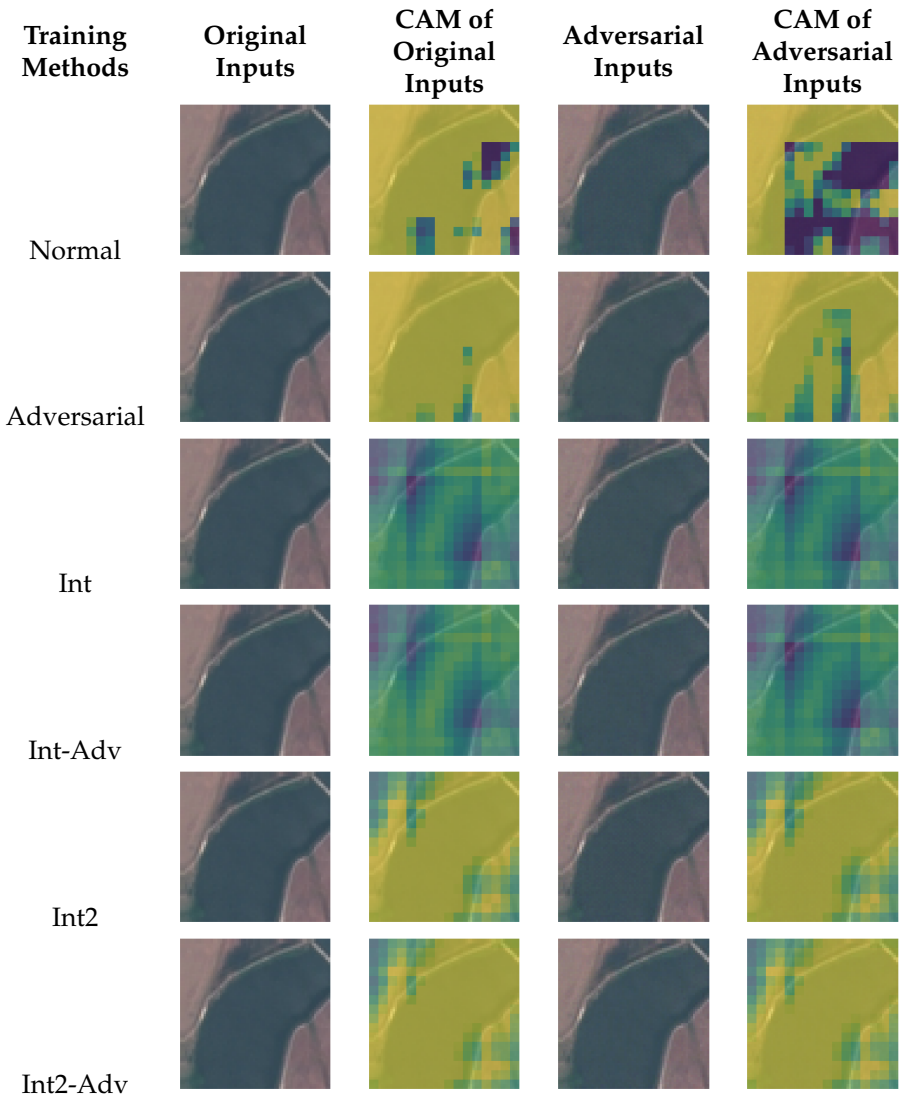

Figure 4. Class activation maps (CAMs) of the original input of the river class from the EuroSAT dataset and their corresponding adversarial inputs for different training methods in the proposed SaliencyMix-based data augmentation method.

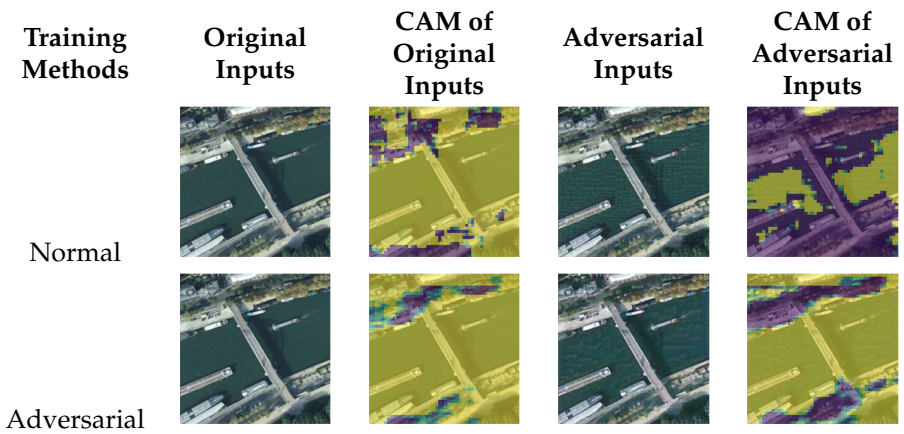

Figure 5. *Cont.*

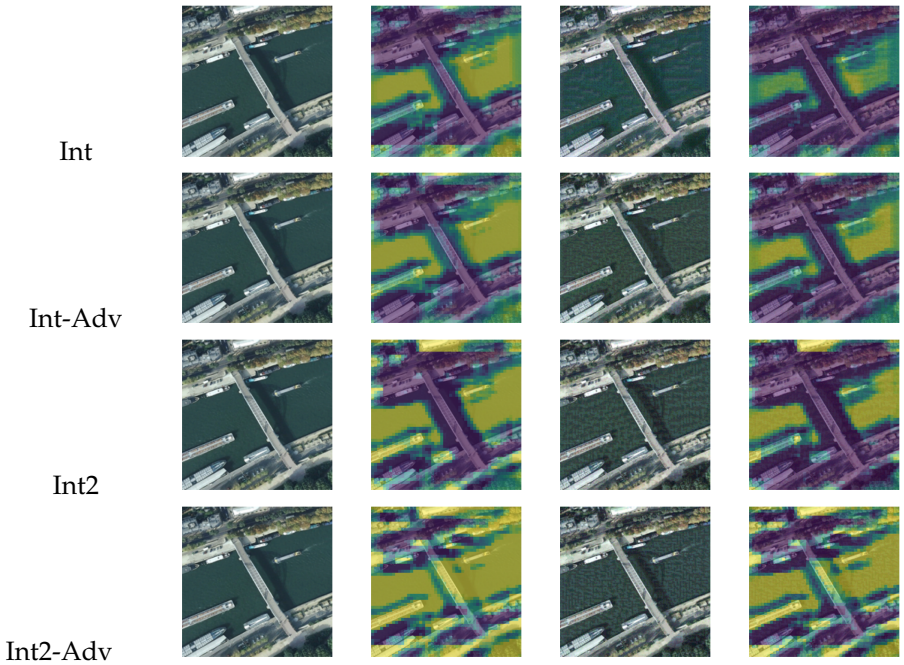

Figure 5. Class activation maps (CAMs) of the original input of the bridge class from the AID dataset and their corresponding adversarial inputs for different training methods in the proposed SaliencyMix-based data augmentation method.

6.1. Base Method

The base method applies Int, $Int - Adv$, $Int2$, and $Int2 - Adv$ using the original, unaugmented dataset and labeled as Base. The standard base method achieved a clean data accuracy of 88% ($\epsilon = 0$) for the EuroSAT dataset shown in Table 1. However, the model was highly susceptible to adversarial attacks, with ATA dropping to 0% for an adversarial perturbation of $\epsilon \geq 4/255$ for the normal training method. Adversarial training improved robustness significantly, achieving 30.5% ATA at an adversarial perturbation of $\epsilon = 4/255$, but resulting in a lower clean data accuracy of 80.5%. The interpretability-aware methods (Int, $Int - Adv$, $Int2$, and $Int2 - Adv$) also showed improved robustness, with $Int2 - Adv$ achieving 21.9% ATA at an adversarial perturbation of $\epsilon = 10/255$, even though with a clean data accuracy of 48.4%.

Table 1. Adversarial test accuracy (ATA) after evaluation of 200 step PGD attack under different perturbation sizes ϵ in the EuroSAT dataset with Convolutional Neural Network (CNN) architecture. $\epsilon = 0$ indicates without any adversarial perturbation.

	Training Methods	$\epsilon = 0$	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$	$\epsilon = 8/255$	$\epsilon = 9/255$	$\epsilon = 10/255$
Base	Normal	88%	3%	0%	0%	0%	0%	0%
	Adv	80.5%	62.5%	30.5%	9.9%	3.5%	2%	1.5%
	Int	52.5%	47%	41%	28%	19%	18.5%	14%
	Int-Adv	45.5%	41%	36%	29%	24.5%	22.5%	21.5%
	Int2	51.5%	44.5%	37.5%	30.5%	23.5%	19.9%	17.5%
	Int2-Adv	48.4%	41.5%	35%	30%	26%	24.5%	21.9%
Trad Aug	Normal	91%	3%	0.2%	0.05%	0%	0%	0%
	Adv	75%	59.8%	36%	15.1	6.7%	04.8%	3.7%
	Int	57%	51.6%	43.6%	34.8%	27%	24%	22%
	Int-Adv	52%	47.7%	42.4%	36.9%	31.2%	0.28.7%	26.3%
	Int2	60%	49.7%	43%	0.36%	29%	26%	23%
	Int2-Adv	53%	48%	42.9%	37%	32%	29%	26%
SaliencyMix	Normal	91%	0.67%	0.05%	0.04%	0.02%	0.02%	0.02%
	Adv	76%	59%	35%	14%	7%	6%	4 %
	Int	80%	55%	26%	9.6%	4%	3%	2%
	Int-Adv	53%	47.9%	41.2%	34.1%	28.7%	26.5%	24.5%
	Int2	80%	56%	42.2%	34.3%	27.6%	24.5%	22.1%
	Int2-Adv	52%	47%	41%	35%	29%	26%	24%

For the AID dataset, Table 2 shows that the normal training method achieved a clean data accuracy of 71.5%, with a significant drop in robustness against adversarial attacks. Adversarial training improved robustness but resulted in a lower clean data accuracy of 61% and ATA of 1.9% accuracy at $\epsilon = 10/255$. The interpretability-aware methods (Int, $Int - Adv$, $Int2$, and $Int2 - Adv$) demonstrated varied results, with $Int2 - Adv$ achieving ATA of 6.7% accuracy at an adversarial perturbation of $\epsilon = 10/255$ and 40.7% accuracy on clean data.

Table 2. Adversarial test accuracy (ATA) after evaluating 200 step PGD attack under different perturbation sizes ϵ in the AID dataset with Convolutional Neural Network (CNN) architecture. $\epsilon = 0$ indicates without any adversarial perturbation.

	Training Methods	$\epsilon = 0$	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$	$\epsilon = 8/255$	$\epsilon = 9/255$	$\epsilon = 10/255$
Base	Normal	71.5%	3.8%	0.7%	0.2%	0%	0%	0%
	Adv	61%	28.4%	9.9%	3.2%	1.7%	1.4%	1.9%
	Int	46.2%	0.343	0.219	0.138	0.079	0.059	0.045
	Int-Adv	43.1%	32.7%	23.6%	16.2%	8.9%	6.6%	5.8%
	Int2	44.5%	34.2%	22.9%	15.4%	7.8%	6.4%	4.7%
	Int2-Adv	40.7%	31.4%	24.2%	17.2%	10.6%	8.3%	6.7%
Trad Aug	Normal	73.8%	4.1%	1.3%	0.2%	0%	0%	0%
	Adv	59.9%	28.8%	10.1%	0.03%	1.9%	1.5%	2%
	Int	46.9%	33.6%	22%	13.9%	6.9%	6.2%	4.5%
	Int-Adv	43.7%	1.8%	22.9%	14.9%	9.2%	6.9%	6%
	Int2	45.7%	34.1%	22.6%	13.4%	7.9%	6.2%	5%
	Int2-Adv	42.7%	32.7%	24.9%	17%	10.7%	9.2%	6.9%
SaliencyMix	Normal	75.3%	7.6%	2.3%	0.7%	0.1%	0.1%	0.1%
	Adv	60%	29.1%	18.3%	2.9%	2.4%	1.1%	2.8%
	Int	47.6%	35.1%	22.4%	4.2%	6.2%	4.7%	4.8%
	Int-Adv	44%	31.9%	23.9%	15.6%	9.3%	7.7%	6.4%
	Int2	46.8%	34.7%	24.1%	14.8%	8.3%	6.8%	5.7%
	Int2-Adv	42.9%	34.2%	26.5%	17.6%	11%	9.8%	7%

6.2. Traditional Data-Augmentation

When we introduced traditional data augmentation, labeled as Trad Aug in Table 1, the clean data accuracy improved for all methods. The normal training method's accuracy increased to 91%, though its robustness against adversarial attacks remained poor. The interpretability-aware methods demonstrated significant improvements in clean data accuracy and maintained robustness. For example, the Int2-Adv method's clean data accuracy rose to 53%, and ATA of 26% at an adversarial perturbation of $\epsilon = 10/255$.

For the AID dataset, applying traditional data augmentation led to improved clean data accuracy, as indicated in Table 2 and labeled as Trad Aug. The normal training method's accuracy increased to 73.8%, but its robustness remained low. The interpretability-aware methods showed enhancements in both clean data accuracy and robustness. For instance, the $Int2 - Adv$ method's clean data accuracy increased to 42.7%, with ATA of 6.9% at an adversarial perturbation of $\epsilon = 10/255$.

6.3. SaliencyMix Based Data-Augmentation

The most notable improvements were observed with the application of SaliencyMix data augmentation for the EuroSAT dataset, labeled as SaliencyMix in Table 1. The normal training method's clean data accuracy remained at 91%, with slight improvements in adversarial robustness. The interpretability-aware methods, particularly $Int2$ and $Int2 - Adv$, showed substantial enhancements in both clean data accuracy and robustness. For instance, the $Int2$ method achieved 80% accuracy on clean data and 22.1% ATA at an adversarial perturbation of $\epsilon = 10/255$.

SaliencyMix data augmentation led to significant improvements for the AID dataset, labeled as SaliencyMix in Table 2. The normal training method's clean data accuracy increased to 75.3%. The interpretability-aware methods demonstrated the best performance

with SaliencyMix. For example, the $Int2$ method achieved 46.8% accuracy on clean data and 5.7% ATA at an adversarial perturbation of $\epsilon = 10/255$, while $Int2 - Adv$ achieved 42.9% on clean data and 7.0% ATA at an adversarial perturbation of $\epsilon = 10/255$.

The CAM explanation maps further support these findings. The differences in the explanation maps between original and adversarial examples for regular methods highlight the effectiveness of interpretability-aware methods in minimizing discrepancies between original and adversarial inputs. This alignment in explanations is crucial for maintaining model performance under adversarial conditions. Figure 4 illustrates the CAM of original and adversarial inputs for a "River" class image from the EuroSAT dataset. The CAM explanation maps in the 2nd and last columns of the figure show purple and green portions indicating the important region classification identified by the CAM method. Notably, there are differences in the explanation map between the original and adversarial examples in regular normal and adversarial methods, where the original "River" sample was misclassified as the "AnnualCrop" class. However, for the interpretability-aware methods (Int, $Int - Adv$, $Int2$, and $Int2 - Adv$), there are no noticeable differences in the CAM between the original input and adversarial inputs. During training in these methods, the model minimizes the discrepancies between the interpretations of original and adversarial inputs. For the AID dataset, similar trends were observed in the explanation maps shown in Figure 5.

6.4. Robustness Transferability

In the trained models using EuroSAT and AID datasets with SaliencyMix augmentation, we applied PGD attacks in all the training methods, including the interpretability-aware training such as Int, $Int - adv$, $Int2$, and $Int2 - adv$. Then, we evaluated these models using other attacks, including FGSM and BIM, to assess robustness transferability. The FGSM attack induces small, one-step perturbations to input features based on the gradient sign of the loss function. We used the ATA metric to evaluate model robustness. The ATA metric indicated varying levels of robustness against the FGSM attack across different training methods. The ATA metric for these training methods was lower than the PGD attack results for both FGSM and BIM attacks for the EuroSAT dataset. These results are shown in the last column of Table 3 with an adversarial perturbation of $\epsilon = 10/255$. However, for the AID dataset, Table 4 shows that FGSM attacks resulted in better ATA than PGD, and BIM yielded higher ATA than both PGD and BIM.

Table 3. Evaluation of robustness transferability using adversarial test accuracy (ATA) in unseen FGSM and BIM attacks under different perturbation sizes ϵ in the EuroSAT dataset with Convolutional Neural Network (CNN) architecture and SaliencyMix augmentation. $\epsilon = 0$ indicates without any adversarial perturbation.

	Training Methods	$\epsilon = 0$	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$	$\epsilon = 8/255$	$\epsilon = 9/255$	$\epsilon = 10/255$
FGSM	Normal	91%	98%	3.3%	2.7%	2.8%	3.4%	**3.8%**
	Adv	76%	60.2%	43.1%	28.4%	18.4%	15.1%	12.3%
	Int	80%	56.9%	34.8%	21.8%	13.3%	11.1%	9.5%
	Int-Adv	53%	48.1%	43.1%	38.7%	34.3%	32.5%	31%
	Int2	80%	57.2%	34.5%	20.9%	12%	9.7%	7.9%
	Int2-Adv	52%	58.8%	44.8%	33%	24.6%	21.2%	18.3%
BIM	Normal	91%	0.7%	0.02%	0%	0%	0%	0%
	Adv	76%	59.1%	34.5%	14.1%	7.2%	5.6%	4.5%
	Int	80%	54.7%	26.3%	9.6%	3.9%	2.9%	2.1%
	Int-Adv	53%	47.9%	41.2%	34.1%	28.7%	26.5%	24.5%
	Int2	80%	55.6%	25.6%	8.9%	3.8%	2.8%	2.1%
	Int2-Adv	52%	58%	40%	22%	9.2%	6.9%	5.5%

Table 4. Evaluation of robustness transferability using adversarial test accuracy (ATA) in unseen FGSM and BIM attacks under different perturbation sizes ϵ in the AID dataset with Convolutional Neural Network (CNN) architecture and SaliencyMix augmentation. $\epsilon = 0$ indicates without any adversarial perturbation.

	Training Methods	$\epsilon = 0$	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$	$\epsilon = 8/255$	$\epsilon = 9/255$	$\epsilon = 10/255$
FGSM	Normal	75.3%	05.6%	1.7%	1.4%	1.2%	0.9%	0.9%
	Adv	61%	31.3%	14.3%	6.5%	3.6%	2.8%	2.4%
	Int	47.6%	34.9%	23.9%	16.5%	12.6%	10.8%	8.9%
	Int-Adv	44%	33.1%	25.2%	19.3%	13.9%	11.9%	10.5%
	Int2	46.8%	34.3%	24.5%	17.3%	11.9%	10.7%	8.7%
	Int2-Adv	43%	31.7%	25%	19.2%	14.5%	12.4%	11.1%
BIM	Normal	75.3%	3.8%	0.7%	1.4%	0.2%	0%	0%
	Adv	61%	28.4%	9.9%	3.2%	1.7%	1.4%	1%
	Int	47.6%	34.3%	21.9%	13.8%	7.9%	5.9%	4.5%
	Int-Adv	44%	32.7%	23.6%	16.2%	9.8%	7.6%	5.8%
	Int2	46.8%	34.2%	22.3%	15.4%	8.7%	6.4%	4.7%
	Int2-Adv	43%	31.4%	24.2%	17.2%	10.6%	8.3%	6.7%

7. Discussion

Our experimental results on the EuroSAT and AID datasets provide a comprehensive analysis of various training methods aimed at enhancing adversarial robustness while maintaining or improving accuracy on clean data. The baseline results highlight a common trade-off in adversarial training, where increased robustness against attacks typically results in reduced performance on clean data. For instance, the standard adversarial training methods improved robustness but could not withstand a large adversarial noise in the EuroSAT dataset, as seen with an accuracy drop from 80.5% to 1.5% against an adversarial perturbation of $\epsilon = 10/255$ from PGD-based attack, as shown in Table 1. Whereas, interpretability-aware training methods (Int, $Int - Adv$, $Int2$, and $Int2 - Adv$) yielded better robustness, particularly at higher perturbation levels (ϵ). Notably, the $Int2 - Adv$ method demonstrated superior robustness, maintaining 21.9% ATA at an adversarial perturbation of $\epsilon = 10/255$ for EuroSAT, although this came with a lower clean data accuracy of 48.4%. This indicates that while these methods enhance robustness, there is still a trade-off with clean data accuracy.

The integration of traditional data augmentation techniques resulted in a notable improvement in clean data accuracy across all methods. This enhancement was evident in the increase of clean data accuracy to 53% from 48.4% in $int2 - adv$ training for EuroSAT dataset. However, robustness improvements were limited, which indicates the necessity for more sophisticated augmentation strategies.

In summary, from Tables 1 and 2, we can see that the ATA score has increased when data augmentation, specifically SaliencyMix, is applied to clean data ($\epsilon = 0$) in both normal training and interpretability-aware robustness training methods, Int, $Int - Adv$, $Int2$, and $Int2 - Adv$. Therefore, we can conclude that the SaliencyMix method improved clean data accuracy while maintaining or enhancing robustness against adversarial attacks. The interpretability-aware methods, when combined with SaliencyMix, provided the best balance between clean data accuracy and adversarial robustness. Similarly, in the AID dataset, we found that the clean data accuracy improved from 40.7% to 42.9% and ATA accuracy improved from 6.7% to 7% against an adversarial perturbation of $\epsilon = 10/255$ from the PGD-based attack, as shown in Table 2. These results underscore the effectiveness of combining interpretability-aware training with advanced data augmentation techniques to achieve robust and accurate models for remote sensing image classification.

To evaluate the robustness transferability, when we applied FGSM and BIM attacks on our proposed adversarial robustness model (SaliencyMix), we observed consistent patterns of accuracy improvement across normal and interpretability-aware methods. For example, comparing the ATA of the PGD-based attack shown in Table 1 (labeled as SaliencyMix) and FGSM-based attack from Table 3, we see that for PGD, the normal training accuracy

drops from 91% to 0.02% at the highest perturbation level (adversarial perturbation of $\epsilon = 10/255$). Similarly, for FGSM, the accuracy drops from 91% to 3.8%, which has a similar drop in accuracy. We observe similar trends in the case of interpretability-aware training methods, such as $Int-2$ and $Int2-Adv$. $Int2-Adv$ has a ATA of 24% under PGD-based attack (Table 1) and ATA of 18.3% under FGSM-based attack (Table 3). We also found similar adversarial robustness performance in Adv, Int, $Int-Adv$, and $Int-2$ training methods. The robustness method performed worse against BIM-based attacks, as shown in Table 3 compared to the FGSM-based attack. For the AID dataset, we observed similar adversarial robustness performance in Adv, Int, $Int-Adv$, $Int-2$, and $Int-2-adv$ training methods against FGSM and BIM adversarial attacks shown in Table 4. This suggests that the robustness gained from these training methods is transferable against other unseen attacks.

Overall, these results underscore the effectiveness of combining interpretability-aware training with advanced data augmentation techniques like SaliencyMix to achieve robust and accurate models for remote sensing image classification.

8. Conclusions

Our study demonstrates that combining interpretability-aware training with advanced data augmentation techniques such as SaliencyMix can significantly enhance the robustness and clean data accuracy of models trained on remote sensing datasets. While adversarial training improves robustness, it often does so at the cost of clean data accuracy. However, integrating saliency-guided data augmentation methods provides the best approach, yielding models that are not only robust to adversarial perturbations but also highly accurate on unperturbed data. Interpretability-aware techniques, particularly when paired with SaliencyMix, stand out by ensuring reliable and consistent model explanations, which further contribute to their robustness and trustworthiness.

These findings highlight the importance of advanced data augmentation techniques in adversarial training paradigms. However, our approach is not without limitations. The computational efficiency of the proposed methods remains a challenge, as the training process can be time-consuming and resource-intensive. Additionally, our models have been tested against only three types of attacks (PGD, FGSM, and BIM), leaving uncertainty about their robustness against other sophisticated adversarial techniques, such as adversarial patches [27].

In the future, we can explore further refinements in augmentation strategies and interpretability constraints to push the boundaries of robust and accurate model training. Additionally, extending these techniques to other datasets and exploring their applicability in real-world scenarios will be crucial for broader adoption. Our work underscores a promising direction for developing resilient machine-learning models capable of maintaining high performance in the face of adversarial challenges.

Author Contributions: Conceptualization, K.A.I.; methodology, K.A.I. and S.T.; software, S.T.; validation, S.T. and K.A.I.; formal analysis, K.A.I. and S.T.; investigation, S.T.; resources, K.A.I.; data curation, S.T.; writing—original draft preparation, S.T. and K.A.I.; writing—review and editing, K.A.I. and S.T.; visualization, S.T.; supervision, K.A.I.; project administration, K.A.I.; funding acquisition, K.A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Full-Form	Abbreviation
Artificial Intelligence	AI
Class Activation Map	CAM
Fast Gradient Sign Method	FGSM
Basic Iterative Method	BIM
Projected Gradient Descent	PGD
Aerial Image Dataset	AID
Adversarial Test Accuracy	ATA

References

1. Navalgund, R.R.; Jayaraman, V.; Roy, P. Remote sensing applications: An overview. *Curr. Sci.* **2007**, *93*, 1747–1766.
2. Van Westen, C. Remote sensing for natural disaster management. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 1609–1617.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
4. Özyurt, F.; Avcı, E.; Sert, E. UC-Merced Image Classification with CNN Feature Reduction Using Wavelet Entropy Optimized with Genetic Algorithm. *Trait. Signal* **2020**, *37*, 347–353. [CrossRef]
5. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [CrossRef]
6. Chan-Hon-Tong, A.; Lenczner, G.; Plyer, A. Demotivate adversarial defense in remote sensing. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 3448–3451.
7. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial example in remote sensing image recognition. *arXiv* **2019**, arXiv:1910.13222.
8. Xu, Y.; Du, B.; Zhang, L. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1604–1617. [CrossRef]
9. Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
10. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2017**, arXiv:1607.02533.
11. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
12. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2017**, arXiv:1706.06083.
13. Cheng, G.; Sun, X.; Li, K.; Guo, L.; Han, J. Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
14. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
15. Zhang, Y.; Zhang, Y.; Qi, J.; Bin, K.; Wen, H.; Tong, X.; Zhong, P. Adversarial patch attack on multi-scale object detection for uav remote sensing images. *Remote Sens.* **2022**, *14*, 5298. [CrossRef]
16. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
17. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. *arXiv* **2020**, arXiv:1910.01279.
18. Dombrowski, A.K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.R.; Kessel, P. Explanations can be manipulated and geometry is to blame. *arXiv* **2019**, arXiv:1906.07983.
19. Chen, J.; Wu, X.; Rastogi, V.; Liang, Y.; Jha, S. Robust attribution regularization. *arXiv* **2019**, arXiv:1905.09957.
20. Boopathy, A.; Liu, S.; Zhang, G.; Liu, C.; Chen, P.Y.; Chang, S.; Daniel, L. Proper network interpretability helps adversarial robustness in classification. *arXiv* **2020**, arXiv:2006.14748.
21. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. *arXiv* **2016**, arXiv:1512.04150.
23. Uddin, A.; Monira, M.; Shin, W.; Chung, T.; Bae, S.H. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv* **2020**, arXiv:2006.01791
24. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
25. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.

26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial Patch. *arXiv* **2018**, arXiv:1712.09665.