remote sensing

MDPI

Article

Defending against Poisoning Attacks in Aerial Image Semantic Segmentation with Robust Invariant Feature Enhancement

Zhen Wang [1], Buhong Wang [1], Chuanlei Zhang [2], Yaohui Liu [3,*] and Jianxin Guo [4]

[1] School of Information and Navigation, Air Force Engineering University, FengHao East Road, Xi'an 710082, China
[2] School of Artificial Intelligence, Tianjin University of Science and Technology, Dagu South Road, Hexi District, Tianjin 300457, China
[3] School of Surveying and Geo-Informatics, Shandong Jianzhu University, FengMing Road, LiCheng District, Jinan 250101, China
[4] School of Electronic Information, Xijing University, XiJing Road, Chang'an District, Xi'an 710123, China
* Correspondence: liuyaohui20@sdjzu.edu.cn; Tel.: +86-133-8531-9533

Abstract: The outstanding performance of deep neural networks (DNNs) in multiple computer vision in recent years has promoted its widespread use in aerial image semantic segmentation. Nonetheless, prior research has demonstrated the high susceptibility of DNNs to adversarial attacks. This poses significant security risks when applying DNNs to safety-critical earth observation missions. As an essential means of attacking DNNs, data poisoning attacks destroy model performance by contaminating model training data, allowing attackers to control prediction results by carefully crafting poisoning samples. Toward building a more robust DNNs-based aerial image semantic segmentation model, in this study, we proposed a robust invariant feature enhancement network (RIFENet) that can resist data poisoning attacks and has superior semantic segmentation performance. The constructed RIFENet improves the resistance to poisoning attacks by extracting and enhancing robust invariant features. Specifically, RIFENet uses a texture feature enhancement module (T-FEM), structural feature enhancement module (S-FEM), global feature enhancement module (G-FEM), and multi-resolution feature fusion module (MR-FFM) to enhance the representation of different robust features in the feature extraction process to suppress the interference of poisoning samples. Experiments on several benchmark aerial image datasets demonstrate that the proposed method is more robust and exhibits better generalization than other state-of-the-art methods.

Keywords: aerial images; semantic segmentation; deep neural networks (DNNs); robust invariant features; poisoning attack; adversarial defense

Citation: Wang, Z.; Wang, B.; Zhang, C.; Liu, Y.; Guo, J. Defending against Poisoning Attacks in Aerial Image Semantic Segmentation with Robust Invariant Feature Enhancement. *Remote Sens.* **2023**, *15*, 3157. https://doi.org/10.3390/rs15123157

Academic Editor: Giuseppe Scarpa

Received: 21 April 2023
Revised: 30 May 2023
Accepted: 7 June 2023
Published: 17 June 2023

1. Introduction

With the rapid development of airborne sensors, unmanned aerial vehicle (UAV) aerial imagery has become an important data source for many fields, such as remote sensing [1], disaster management [2], and urban planning [3]. However, the rich information in aerial imagery poses significant challenges for extracting valuable data. The application of semantic segmentation has become an effective technique for addressing this problem, as it enables fine-grained pixel-level classification of ground objects [4]. In this context, aerial image semantic segmentation has received extensive attention [5]. In recent years, owing to the powerful fitting ability of deep neural networks (DNNs), it has been widely used in various aerial image processing tasks, such as scene classification [6], object detection [7], and semantic segmentation [8]. DNNs can automatically learn complex features and abstract concepts [9] from data to improve the accuracy of semantic segmentation.

Because many aerial image processing tasks involve safe-critical applications, such as military [10] and defense [11], they require high precision, reliability, and security. Unfortunately, the vulnerability of DNNs leads to serious security risks when applied

to aerial image processing. For example, a noteworthy concern is that DNNs are highly vulnerable to adversarial example attacks [12]. Attackers can alter the prediction results of DNNs by adding intentionally designed but imperceptible adversarial perturbations to the image. For the aerial remote sensing community, adversarial attacks and defenses against DNNs have received attention. Czaja et al. [13] revealed the problem of adversarial examples in satellite remote sensing image (RSI) classification tasks, where a classifier can be fooled into making incorrect predictions by embedding adversarial noise into remote sensing images. Chen et al. [14] analyzed the impact of adversarial noise on multiple RSI recognition models and demonstrated the transferability of adversarial attacks. Xu et al. [15] demonstrated the threats posed by both targeted and untargeted attacks on RSI scene classification models and proposed an adversarial training strategy to train a more robust classifier. Chen et al. [16] employed the fast gradient sign method (FGSM) and basic iterative method (BIM) to attack RSI classification models. Ai et al. [17] explored the feasibility of black-box attacks on RSI scene classification models. Bai et al. [18] constructed a universal adversarial example generation method based on domain adaptation theory to attack an RSI classifier. Chen et al. [19] proposed a soft-threshold defense framework to enhance the robustness of RSI classification models. For RSI object detection, Wei et al. [20] proposed an adversarial pan-sharpening attack to destroy the performance of an object detector. Lian et al. [21] proposed a physically realizable adversarial patch generation method to attack the RSI object detector. Wang et al. [22] first systematically evaluated the adversarial example threat faced by RSI semantic segmentation models and proposed a global feature attention network to defend against various types of adversarial attacks.

The aforementioned studies mainly focus on adversarial attacks against DNNs in the inference stage. However, recent research has explored the possibility of conducting attacks in the model training process, with the most typical form being data poisoning attacks [23]. Different from adversarial example attacks, data poisoning attacks influence the training of DNNs by contaminating training data [24], so that the model outputs incorrect prediction results. In Figure 1, we provide an example of performing data poisoning attacks on the aerial image semantic segmentation network. Here, we use HFGNet [25] as the target model for the attack. As shown in Figure 1, although the difference between the original aerial image and the crafted poisoning sample is invisible to the human visual system, the poisoning sample severely misleads the model prediction results. HFGNet achieved a PA of 92.86% trained on the original clean samples, while its PA decreases to only 12.58% on the poisoning samples. This phenomenon undoubtedly increases the security risk level of DNN-based aerial image semantic segmentation models.

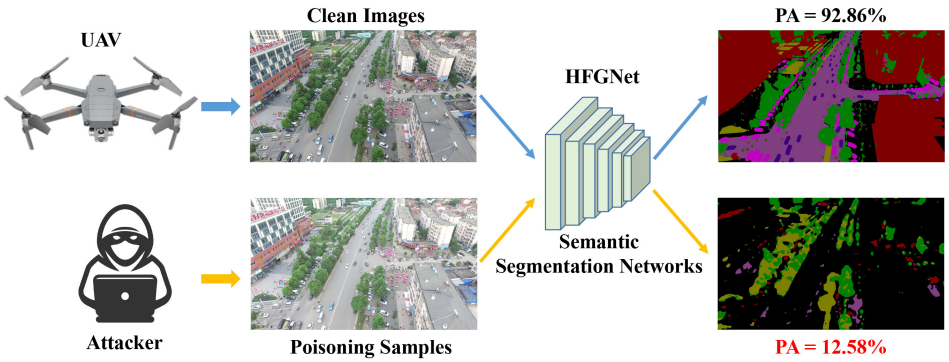


Figure 1. An illustration of poisoning attacks on aerial image semantic segmentation networks. Although the difference between the poisoning sample and the original clean image is imperceptible to the human visual system, the semantic segmentation model HFGNet [25] can be fooled by the poisoning sample to make wrong predictions.

The current widely used defense method against poisoning attacks is adversarial training [26], which generates poisoning samples under known attack methods and trains the model by mixing them with original clean samples, thereby improving the adversarial

robustness. However, there are significant drawbacks to using adversarial training for defense: (1) adversarial training requires additional training samples, which increases the model computational complexity, and (2) the trained model may be unable to resist unknown poisoning attacks, leading to poor generalization performance. Robust invariant features have a strong defense performance against adversarial attacks [27–29], such as solving adversarial examples and backdoor attacks by extracting robust invariant features. Inspired by the robust representation learning theory, we propose a robust invariant feature enhancement network (RIFENet) for defending against poisoning attacks in aerial image semantic segmentation. The network resists poisoning attacks by extracting and enhancing the robust invariant features in aerial images. RIFENet consists of a texture feature enhancement module (T-FEM), structural feature enhancement module (S-FEM), global feature enhancement module (G-FEM), and multi-resolution feature fusion module (MR-FFM). Specifically, T-FEM obtains robust invariant texture features by combining convolutional neural networks (CNNs) and the Transformer model. S-FEM obtains robust invariant structural features of ground objects by using the constructed coordinate attention mechanism. Inspired by the feature pyramid network [30], G-FEM extracts robust invariant global feature information in a bottom-up manner. MR-FFM selectively fuses and filters the obtained robust invariant features to further enhance the representation of robust features. In addition, we construct a hierarchical loss function to improve the training efficiency and generalization. The main contributions of this study can be summarized as follows.

- To the best of our knowledge, we introduce the concept of a poisoning attack into aerial image semantic segmentation for the first time and propose an effective defense framework against both targeted and untargeted poisoning attacks. Our research highlights the importance of enhancing the robustness of deep learning models in handling safety-critical aerial image processing tasks.
- To effectively defend against poisoning attacks, we propose a novel robust invariant feature enhancement framework based on the theory of robust feature representation. By obtaining robust invariant texture features, structural features, and global features, the proposed defense framework can effectively suppress the influence of poisoning samples on feature extraction and representation.
- To demonstrate the effectiveness of the proposed defense framework, we conducted extensive experiments to evaluate the adversarial defense performance against poisoning attacks. The experiments on the aerial image benchmark dataset in urban scenes show that the proposed framework can effectively defend against poisoning attacks and maintain better semantic segmentation performance.

The remainder of this article is organized as follows. Section 2 briefly reviews some related work. Section 3 describes the proposed defense framework. Section 4 presents the information on the datasets used in this study and the experimental results. The discussion and conclusion are summarized in Sections 5 and 6.

2. Related Works

In this section, we review the existing poisoning attacks, poisoning defense, and robust feature representation methods.

2.1. Poisoning Attacks

Poisoning attack refers to malicious tampering of training data or model parameters to mislead the model to produce incorrect prediction results [31]. The existing poisoning attack methods can be divided into white-box attacks [32] and black-box attacks [33]. The white-box attack defines that the attacker can access and modify the parameters and structure of the model arbitrarily, while the black-box attack sets that the attacker cannot access the internal structure information and parameter settings of the model. Pang et al. [34] used the influence function to select the training samples that significantly impact the model for label flipping and realized the data poisoning attack against the DNNs for the first time. Shafahi et al. [35] constructed the clean-label poisoning data by feature collision and care-

fully designed poisoning samples with high similarity to benign samples in the feature space. Zhao et al. [36] proposed a poisoning attack method with high stealthiness against image classification models based on generative adversarial networks. Kurita et al. [37] proposed a poisoning sample generation method for pre-trained models, which can cause destroy models dealing with different computer vision tasks. Muñoz-González et al. [38] used the back-propagation algorithm to generate poisoning samples, significantly degrading the performance of multiple DNNs models. As a novel data poisoning attack method, a backdoor attack [39] damages the model performance by embedding hidden triggers in test samples. Based on meta-learning theory, Huang et al. [40] proposed an approximate solution to the second-order optimization problem of data poisoning attacks to improve the attack efficiency. Aghakhani et al. [41] proposed a transferable clean-label poisoning attack, which achieved high attack success rates in multiple image classification tasks. In general, the poisoning attacks currently widely studied include adversarial noise, label flipping, and image tampering. These attacks pose significant security threats to DNNs models.

2.2. Poisoning Defense

The existence of poisoning attacks in DNNs has shown that ensuring the security and robustness of models is a significant challenge. To defend against poisoning attacks, many methods have been proposed and achieved better results. Currently, the most commonly used defense methods include adversarial training, data augmentation, detection-based methods, and ensemble defense strategies. Adversarial training [42] uses adversarial samples to train the model, which can improve the model robustness against poisoning attacks. Geiping et al. [43] constructed an adversarial training framework based on batch normalization, which can resist both targeted and untargeted poisoning attacks. Gao et al. [44] proposed a hybrid adversarial training strategy that can effectively enhance the model robustness against poisoning attacks with high stealthiness. Hallaji et al. [45] proposed a cascaded defense framework combining adversarial training and label noise analysis to defend against poisoning attacks. For the data augmentation-based defense methods, it expands the training set samples by image rotation, shearing, translation, and scaling to improve model robustness against poisoning attacks. Chen et al. [46] proposed a boundary feature augmentation method to suppress the impact of poisoning attacks on the model feature extraction process. Liu et al. [47] proposed a benign noise injection method to augment the original dataset to enhance the model robustness. Yang et al. [48] obtained more robust feature information by randomly erasing the training dataset images and using adversarial training to improve generalization ability. The detection-based method [49] is another effective strategy to defend against poisoning attacks, which reduces the success rate of poisoning attacks by detecting anomalous samples in the training and testing data. Additionally, ensemble defense methods [50] usually combine data augmentation, adversarial training, and detection-based methods to defend against poisoning attacks. Although these methods have provided possible solutions for poisoning attacks in aerial image semantic segmentation, they have not defended against attacks from the perspective of model architecture design, resulting in poor generalization performance.

2.3. Robust Feature Representation

For safety-critical aerial image semantic segmentation tasks, it is essential to achieve better semantic segmentation performance and ensure that the model can resist the negative impact of poisoning attacks. Different from adversarial training, the robust feature representation theory employs carefully designed robust feature extractors to obtain feature information that can improve model performance and enhance robustness. The current approach of defending against adversarial attacks by extracting robust features has received some attention. Zhang et al. [27] proposed an adversarial defense method based on feature scattering, which makes the model obtain more robust feature information by suppressing the representation of adversarial samples. Xu et al. [29] constructed the self-attention encoder for obtaining robust features to solve adversarial attacks in hyperspectral image

classification tasks. These existing studies have shown that carefully designed feature extractors can obtain robust invariant features with defensive effects. Zhang et al. [51] used the domain adaptation method to improve the similarity of features between adversarial and original domains and constructed the adversarial loss to obtain robust invariant feature information. Li et al. [52] first demonstrated that adversarial training is ineffective in defending against black-box attacks and proposed a robust feature-guided adversarial training method to enhance the model generalization. Kim et al. [53] filtered different feature information obtained by the model using knowledge distillation and information bottleneck to suppress the impact of adversarial features and enhance robust feature representation. Xie et al. [54] used denoising auto-encoders to filter adversarial features in the hidden space of feature extraction to enhance robust feature representation. Song et al. [55] demonstrated that local features can effectively enhance model robustness and guided the model to obtain robust local features by adversarial training. However, these methods ignore the improvement in the model defense capability by robust feature enhancement, so transferring them to semantic segmentation tasks cannot achieve the desired effect.

3. Methodology

Inspired by the encoder–decoder architecture constructed by U-Net [56], i.e., the encoder extracts multi-scale feature information from the input image, the decoder restores feature map resolution, and skip connections are used for feature transfer between encoder and decoder. By using the encoder–decoder architecture, valuable feature information can be extracted and fused layer-by-layer to improve semantic segmentation accuracy. Similarly, the proposed robust invariant feature enhancement network (RIFENet) uses the encoder–decoder as basic framework and integrates texture feature enhancement module (T-FEM), structural feature enhancement module (S-FEM), global feature enhancement module (G-FEM), and multi-resolution feature fusion module (MR-FFM) for resisting poisoning attacks and improving semantic segmentation accuracy. As shown in Figure 2, in the encoder structure, we use VGG16 as the backbone network to extract multi-scale feature information of aerial images. Then, the constructed T-FEM and S-FEM are used to enhance the texture and structure features of each layer output feature of the backbone network to improve the robust feature information representation and suppress the hidden triggers in the poisoning samples. For the decoder structure, we introduce the MR-FFM for fine-grained fusion of different scale feature maps, restore the original size of feature map resolution, and retain the detailed feature information contained in the aerial image. Between the encoder and decoder, we use the G-FEM to perform global correlation modeling and interaction of different features to enhance the perception of pixel position information and improve the semantic segmentation model robustness to poisoning attacks. In addition, to improve the model training process robustness and accelerate convergence, we use the deep supervision strategy for the decoder structure, i.e., add 1×1 convolution and sigmoid function to calculate the loss of each layer in the decoder.

3.1. Texture Feature Enhancement Module

Texture features reflect the spatial distribution properties of pixels, which have the characteristics of local irregularity but global regularity [57]. In addition, texture features have rotational invariant properties that can produce strong resistance to adversarial noise. Therefore, enhancing the representation of texture features can effectively resist the destroying of poisoning attacks. To obtain the texture features contained in the poisoning aerial image, we construct the texture feature enhancement module (T-FEM), which consists of CNNs-based hybrid attention block, Swin Transformer block [58], and feature fusion unit.

As shown in Figure 3, for the input feature $X^i \in \mathbb{R}^{H\times W\times C}$, T-FEM first splits it into $X^i_{S_1} \in \mathbb{R}^{H\times W\times C_1}$ and $X^i_{S_2} \in \mathbb{R}^{H\times W\times C_2}$, where C_1 and C_2 represent the number of channels of feature $X^i_{S_1}$ and $X^i_{S_2}$. The splitting rule is based on spatial information and aims to capture different aspects of the input feature. Specifically, $X^i_{S_1}$ is obtained by applying a spatial convolutional operation with a small receptive field, while $X^i_{S_2}$ is obtained by

applying a global pooling operation to aggregate the feature information across the entire spatial domain. Then, feature $X_{S_1}^i \in \mathbb{R}^{H\times W\times C_1}$ is input into the hybrid attention block composed of spatial attention and channel attention for feature extraction and interaction.

$$\tilde{X}_{S_1}^i = \mathcal{F}_{\text{sigmoid}}\left(\mathcal{F}_{\text{BN}}\left(\mathcal{K}_{1\times 1}\left(X_{S_1}^i\right)\right)\right) \otimes X_{S_1}^i \tag{1}$$

$$\hat{X}_{S_1}^i = \left(\mathcal{F}_{\text{mean}}\left(\tilde{X}_{S_1}^i\right) \oplus \mathcal{F}_{\max}\left(\tilde{X}_{S_1}^i\right)\right) \otimes \tilde{X}_{S_1}^i \tag{2}$$

where $\mathcal{K}_{1\times 1}(\cdot)$ denotes 1×1 convolution, $\mathcal{F}_{\text{BN}}(\cdot)$ indicates the batch normalization function, $\mathcal{F}_{\text{sigmoid}}(\cdot)$ is the nonlinear activation function. $\mathcal{F}_{\text{mean}}(\cdot)$ and $\mathcal{F}_{\max}(\cdot)$ denote average pooling and max pooling operations on the feature map channel dimension. $\tilde{X}_{S_1}^i$ and $\hat{X}_{S_1}^i$ represent the output feature maps of spatial attention and channel attention, respectively. For feature map $X_{S_2}^i \in \mathbb{R}^{H\times W\times C_2}$, Swin Transformer is used to establish the global correlation of texture features, which can be defined as follows.

$$\hat{X}_{S_2}^i = \mathcal{F}_{\text{dim\_recover}}\left(\mathcal{F}_{\text{swin\_attention}}\left(\mathcal{F}_{\text{dim\_slice}}\left(X_{S_2}^i\right)\right)\right) \tag{3}$$

where $\mathcal{F}_{\text{dim\_recover}}(\cdot)$ and $\mathcal{F}_{\text{dim\_slice}}(\cdot)$ denote the slice and recover on feature channel dimension, and $\mathcal{F}_{\text{swin\_attention}}(\cdot)$ indicates the use of Swin Transformer block to obtain global attention feature information. The constructed T-FEM uses the global texture feature correlation modeling unit that consists of Swin Transformer blocks, each of which includes regular window multi-head self-attention (RW-MSA), shifted window multi-head self-attention (SW-MSA), residual connection, multi-layer perception (MLP), and layer normalization. The specific calculation process of Swin Transformer block is as follows.

$$\hat{z}^l = \mathcal{F}_{\text{RW\_MSA}}\left(\mathcal{F}_{\text{LN}}\left(z^{l-1}\right)\right) + z^{l-1}; z^l = \mathcal{F}_{\text{MLP}}\left(\mathcal{F}_{\text{LN}}\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{4}$$

$$\hat{z}^{l+1} = \mathcal{F}_{\text{SW\_MSA}}\left(\mathcal{F}_{\text{LN}}\left(z^l\right)\right) + z^l; z^{l+1} = \mathcal{F}_{\text{MLP}}\left(\mathcal{F}_{\text{LN}}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{5}$$

where $\hat{z}^l$ and z^l are the output of RW-MSA and MLP for the lth Swin Transformer block, and $\hat{z}^{l+1}$ and z^{l+1} indicate the output of SW-MSA and MLP for the $l+1$th Swin Transformer block. $\mathcal{F}_{\text{RW\_MSA}}(\cdot)$ denotes the function of RW-MSA, $\mathcal{F}_{\text{SW\_MSA}}(\cdot)$ indicates the function of SW-MSA, $\mathcal{F}_{\text{LN}}(\cdot)$ is the layer normalization operation. The multi-head attention mechanism of Swin Transformer is defined as follows.

$$\mathcal{F}_{\text{Atten}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{soft}\max\left(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B}\right)\boldsymbol{V} \tag{6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V} \in \mathbb{R}^{M^2\times d}$ denote query, key, and value matrices. M and d indicate the number of patches in the window and the dimension of $\boldsymbol{Q}$ and $\boldsymbol{K}$, respectively. The values of matrix $\boldsymbol{B}$ are obtained by calculating the bias matrix $\hat{\mathbf{B}} \in \mathbb{R}^{(2M-1)\times(2M+1)}$. To fuse the texture features obtained by hybrid attention and Swin Transformer blocks, we construct the feature fusion unit. As shown in Figure 3, the feature fusion unit first uses the feature concatenation function to splice features $\hat{X}_{S_1}^i$ and $\hat{X}_{S_2}^i$ on the channel dimension to ensure that the spliced features have the same dimension as the original features, and then uses the residual unit to fuse spliced features to obtain the fusion texture features. The specific calculation is as follows.

$$Y_C^i = \mathcal{F}_{\text{ResConv}}\left(\mathcal{F}_{\text{cat}}\left(\hat{X}_{S_1}^i, \hat{X}_{S_2}^i\right)\right) \tag{7}$$

where $\mathcal{F}_{\text{cat}}(\cdot)$ denotes the feature concatenation function for splice features $\hat{X}_{S_1}^i$ and $\hat{X}_{S_2}^i$, and $\mathcal{F}_{\text{ResConv}}(\cdot)$ indicates the residual unit for feature fusion.

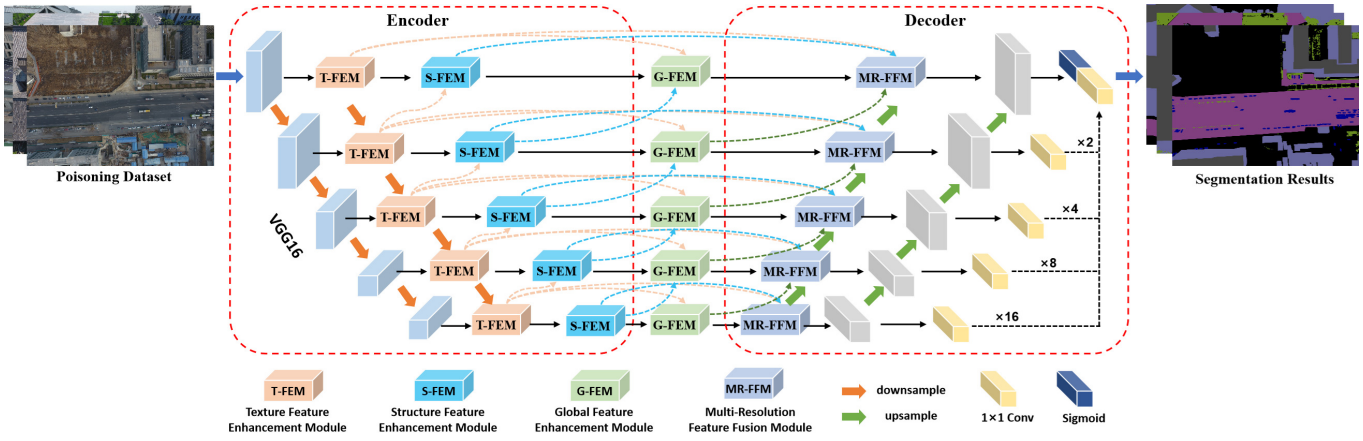


Figure 2. Overall framework of the proposed robust invariant feature enhancement network (RIFENet). The VGG16 is used as the backbone network to extract multi-scale features. Then, we use texture feature enhancement module (T-FEM), structural feature enhancement module (S-FEM), and global feature enhancement module (G-FEM) to enhance the representation of robust invariant features. Finally, the multi-resolution feature fusion module is adopted to perform fine-grained fusion of different scale feature maps and output the aerial image semantic segmentation results.

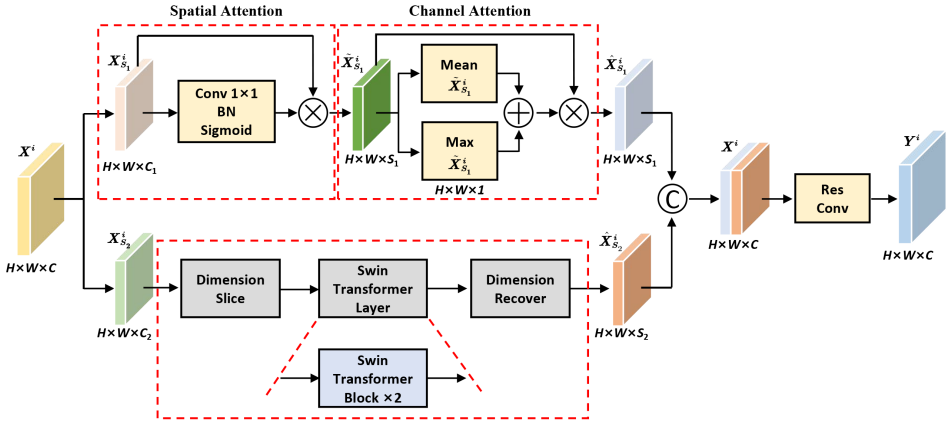


Figure 3. The detailed structure of texture feature enhancement module (T-FEM).

3.2. Structural Feature Enhancement Module

The structural features contained in the image consist of contour and region features. The contour features can describe the boundary information of the ground object [59], while the region features can represent the complete object information [60].

The tampering of structural features requires reversing the gradient information of the original image, so the poisoning samples constructed by the poisoning attackers have difficulty destroying the extraction and representation of structural features by the semantic segmentation model. In addition, the enhanced representation of structural features can be regarded as adversarial noise suppression, which strengthens valuable feature information and invalidates hidden backdoor triggers. Inspired by the coordinate attention mechanism, we construct the structural feature enhancement module (S-FEM) to extract and enhance structural features. Different from the coordinate attention mechanism that calculates spatial information in the X and Y directions of the feature map, S-FEM introduces channel information in the Z direction of the feature map; the structure of S-FEM is shown in Figure 4. The constructed S-FEM uses different convolution blocks to learn the structural feature information in X, Y, and Z directions, and then uses the weighted fusion to obtain the feature weights and achieve the structural feature interaction. Formally, for the input feature map $I \in \mathbb{R}^{C\times H\times W}$, the structural feature information in the X, Y, and X directions is calculated as follows.

$$S_C^X(I) = \mathcal{F}_{\text{Avg}}^{1\times 1\times W}(\mathcal{F}_{\text{ResConv}}(I)) \tag{8}$$

$$S_C^Y(I) = \mathcal{F}_{\text{Avg}}^{1\times H\times 1}(\mathcal{F}_{\text{ResConv}}(I)) \tag{9}$$

$$S_C^Z(I) = \mathcal{F}_{\text{sigmoid}}\left(\mathcal{K}_{1\times 1}\left(\mathcal{F}_{\text{GAP}}^{1\times H\times W}(\mathcal{F}_{\text{ResConv}}(I))\right)\right) \tag{10}$$

where $S_C^X(I) \in \mathbb{R}^{C\times H\times 1}$, $S_C^Y(I) \in \mathbb{R}^{C\times 1\times W}$, and $S_C^Z(I) \in \mathbb{R}^{C\times 1\times 1}$ correspond to the extracted structural feature information in X, Y, and Z directions, respectively. $\mathcal{F}_{\text{ResConv}}(\cdot)$ denotes residual unit function, $\mathcal{F}_{\text{Avg}}^{1\times 1\times W}(\cdot)$ denotes average pooling with size of $1\times 1\times W$, $\mathcal{F}_{\text{Avg}}^{1\times H\times 1}(\cdot)$ denotes average pooling with size of $1\times H\times 1$, and $\mathcal{F}_{\text{GAP}}^{1\times H\times W}(\cdot)$ denotes global average pooling with size of $1\times H\times W$. $\mathcal{K}_{1\times 1}(\cdot)$ indicates 1×1 convolution with batch normalization and ReLU layer. After the extraction of structural features in different directions, the Z direction features are fused with the X and Y direction features. The specific calculation is as follows.

$$Z_X = S_C^X(I) \times S_C^Z(I) \tag{11}$$

$$Z_Y = S_C^Y(I) \times S_C^Z(I) \tag{12}$$

$$Z = \mathcal{K}_{3\times 3}\left(\mathcal{F}_{\text{cat}}\left(Z_X^T, Z_Y\right)\right) \tag{13}$$

where $Z_X \in \mathbb{R}^{C\times H\times 1}$ and $Z_Y \in \mathbb{R}^{C\times H\times 1}$ denote the channel weight feature maps in X and Y directions. $Z \in \mathbb{R}^{C/r\times 1\times (W+H)}$ indicates the fusion feature map of Z_X and Z_Y, where r is the reduction coefficient used to reduce the number of channels. $\mathcal{K}_{3\times 3}(\cdot)$ denotes 3×3 convolution with batch normalization and Silu activation function. Then, the fusion feature Z is split into $Z_X^{'}$ and $Z_Y^{'}$, and the convolution and sigmoid function are used to further enhance the structural feature representation. The specific calculation is as follows.

$$Z_X^{''} = \mathcal{F}_{\text{sigmoid}}\left(\mathcal{K}_{1\times 1}\left(Z_X^{'}\right)\right) \tag{14}$$

$$Z_Y^{''} = \mathcal{F}_{\text{sigmoid}}\left(\mathcal{K}_{1\times 1}\left(Z_Y^{'}\right)\right) \tag{15}$$

where $Z_X^{''}$ and $Z_Y^{''}$ denote the activation features in X and Y directions, and $\mathcal{K}_{1\times 1}(\cdot)$ indicates the 1×1 convolution used to restore the number of channels in the feature map. Then, the dot multiplication is used to calibrate and fuse the feature I and the direction features $Z_X^{''}$ and $Z_Y^{''}$. The specific calculation is as follows.

$$S_{\text{FEM}} = I \cdot Z_X^{''} \cdot Z_Y^{''} \tag{16}$$

where S_{FEM} denotes the structural enhancement feature. The constructed S-FEM can capture the structural feature information of the ground object in X, Y, and Z directions and achieves feature interaction to enhance the representation of structural feature information.

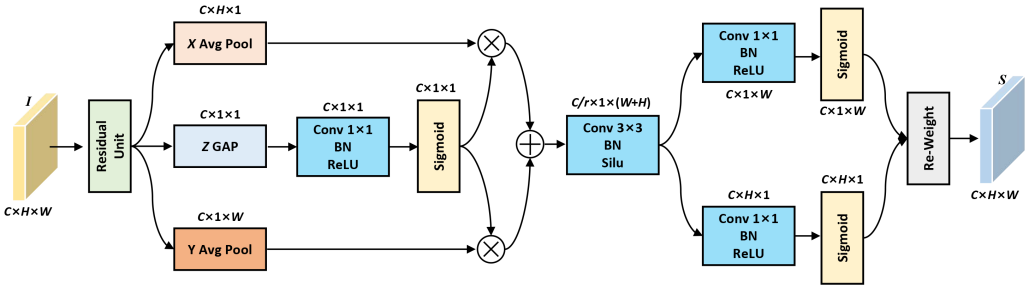


Figure 4. The detailed structure of structural feature enhancement module (S-FEM).

3.3. Global Feature Enhancement Module

Because global features can establish a fixed relationship between the given pixel and all related pixels in the image, it is difficult for malicious attackers to construct poisoning samples that affect global feature extraction and representation. In addition, if an incorrect label is assigned to a certain class of pixels by the attacker, the incorrect loss at this pixel can be passed to all other related pixels through back propagation. In this case, anomalies can be easily found in the model training process, so that the attacked can detect the poisoning samples. Therefore, enhancing the representation of global feature information can effectively suppress the influence of poisoning samples on the semantic segmentation model training process. In the use of CNNs for feature extraction, shallow features contain rich spatial detail information, while deep features contain semantic information. The fusion of spatial and semantic information can effectively enhance the representation of global features. Inspired by the feature pyramid network (FPN) structure [30], we construct the global feature enhancement module (G-FEM) for fusing spatial and semantic feature information, and the structure is shown in Figure 5.

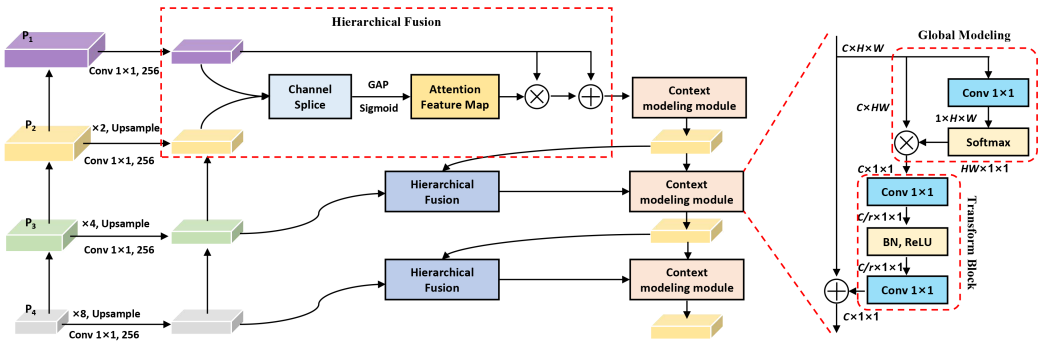


Figure 5. The detailed structure of global feature enhancement module (G-FEM).

G-FEM uses a bottom-up fusion strategy, which uses the shallow feature $\mathbf{P_1}$ as the initial fusion layer. First, feature $\mathbf{P_1}$ is fused with $\mathbf{P_2}$, and then global average pooling and sigmoid functions are used to obtain attention weight maps of the fusion feature, which contains semantic information that can guide shallow features to obtain global correlation. Similar to the fusion process of $\mathbf{P_1}$ and $\mathbf{P_2}$, features $\mathbf{P_2}$, $\mathbf{P_3}$, and $\mathbf{P_4}$ are fused, and the obtained attention weight map is used to guide the shallow feature reconstruction and maintain the same size as the original feature map resolution. In addition, to prevent the problem of feature loss, G-FEM maintains a consistent number of channels in the feature fusion process. The hierarchical fusion strategy in G-FEM is to calculate the pixel (x, y) of the feature map, which is defined as follows.

$$H_i^{(x,y)} = \mathcal{F}_{\text{conv}}\left(G^{(x,y)}, \lambda^{(x,y)}\right) + G^{(x,y)} \tag{17}$$

where $H_i^{(x,y)}$ denotes the ith-layer fusion feature, $G^{(x,y)}$ denotes the initial shallow feature, $\mathcal{F}_{\text{conv}}(\cdot)$ indicates the convolution operation, and $\lambda^{(x,y)}$ indicates the attention feature map. For each pixel in the feature map, it has the corresponding mapping position with the pixel in the input image. Given s denotes the stride value of the feature map, the pixel mapping method is defined as follows.

$$f(x) = \mathcal{F}\lfloor\frac{s}{2}\rfloor + xs; f(y) = \mathcal{F}\lfloor\frac{s}{2}\rfloor + ys \tag{18}$$

where $\mathcal{F}\lfloor\cdot\rfloor$ denotes the floor function. After hierarchical feature fusion, G-FEM uses the context modeling module to further extract global information and fuse with hierarchical features. The context modeling module consists of global modeling and transform block, as shown in Figure 5. For the global modeling part, 1×1 convolution is used to convert

the hierarchical feature H_i into the size of $HW \times 1 \times 1$, and softmax function is used to obtain the global weight attention feature map that can represent the importance of each pixel position. Then, the global attention feature is multiplied by the input feature reconstructed to $C \times HW$ size to obtain the context feature information. The calculation of global modeling is defined as follows.

$$\boldsymbol{A}(x) = \sum_{j=1}^{HW} \frac{\exp(w_k x_{i,j})}{\sum_{m=1}^{HW} \exp(w_k x_{i,m})} \times x_j \tag{19}$$

where x_i denotes the current layer feature, w_k denotes the linear transformation matrix obtained by 1×1 convolution, and $\boldsymbol{A}(x)$ indicates the context feature obtained by global modeling. For the transform block in G-FEM, it is defined as follows.

$$C(x_i) = x_i + w_v \otimes \boldsymbol{A}(x_i) \tag{20}$$

where $C(x_i)$ denotes the output feature of the context modeling module. The 1×1 convolution and ReLU activation function in the transform block can increase the number of network layers and obtain the linear transformation matrix w_v. The inner product operation of matrix w_v and feature $\boldsymbol{A}(x)$ can further enhance the global feature representation.

3.4. Multi-Resolution Feature Fusion Module

To enhance the representation of robust features (texture feature, structural feature, and global feature) and achieve fine-grained feature information fusion, we construct the multi-resolution feature fusion module (MR-FFM). The use of MR-FFM can effectively enhance the robust feature representation to suppress the interference of adversarial noise in the poisoning samples.

As shown in Figure 6, MR-FFM interacts and fuses low-resolution and high-resolution features of different scales in parallel. In the feature fusion process, MR-FFM maintains the original size of the low-resolution features, reduces the high-resolution features to 1/2 of the original size to expand the receptive field range, and adaptively aggregates the feature information of different receptive fields by using multi-resolution selection fusion (MRSF) strategy. Formally, given the input feature map as $X \in \mathbb{R}^{H \times W \times C}$, the feature map X is preprocessed by upsampling and downsampling operations with residual structure to obtain the reconstructed feature maps $\mathcal{F}_{\mathrm{Up}}(X) \in \mathbb{R}^{2H \times 2W \times 1/2C}$ and $\mathcal{F}_{\mathrm{Down}}(X) \in \mathbb{R}^{1/2H \times 1/2W \times 2C}$. In the sampling process, we use bilinear interpolation upsampling and anti-aliasing downsampling operations, and we use Gaussian error linear units (GELU) as activation functions to prevent feature information loss caused by sampling operations. The specific calculation process is as follows.

$$\mathcal{F}_{\mathrm{Up}}(X) = \mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=3}}\left(\mathcal{F}_{\mathrm{CT}_{\mathrm{RF}=3}}\left(\mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=3}}\left(\mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=1}}(X)\right)\right)\right) + \mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=1}}\left(\mathcal{F}_{\mathrm{Bilinear}}(X)\right) \tag{21}$$

$$\mathcal{F}_{\mathrm{Down}}(X) = \mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=1}}\left(\mathcal{F}_{\mathrm{AD}}\left(\mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=3}}\left(\mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=1}}(X)\right)\right)\right) + \mathcal{F}_{\mathrm{CG}_{\mathrm{RF}=1}}\left(\mathcal{F}_{\mathrm{AD}}(X)\right) \tag{22}$$

where $\mathcal{F}_{\mathrm{CG}}(\cdot)$ is composed of convolution and GELU activation function. $\mathcal{F}_{\mathrm{CT}}(\cdot)$ denotes deconvolution function, RF denotes the convolution kernel size, $\mathcal{F}_{\mathrm{Bilinear}}(\cdot)$ denotes bilinear interpolation upsampling function, and $\mathcal{F}_{\mathrm{AD}}(\cdot)$ indicates anti-aliasing downsampling function. For the multi-resolution selection fusion strategy in MR-FFM, it first fuses parallel convolution feature information of different resolution, and then uses global average pooling to obtain global feature information. The specific calculation is as follows.

$$X_f = X_h + X_l \tag{23}$$

$$X_G = \mathcal{F}_{\mathrm{GAP}}\left(X_f\right) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_f(i,j) \tag{24}$$

where X_h and X_l represent high-resolution and low-resolution features. Then, the obtained global feature X_G is used as the input of fully connected layer to fuse different feature information. The specific calculation is as follows.

$$X_z = \mathrm{FC}(X_G) = \mathcal{F}_{\mathrm{Sigmoid}}(\mathrm{Conv}_{\mathrm{RF}=1}(X_G)) \tag{25}$$

where $X_z \in \mathbb{R}^{1\times1\times C}$ denotes the inter-layer fusion feature. The parallel 1×1 convolution is used to restore the number of inter-layer fusion feature X_z and generate feature vectors $v_1 \in \mathbb{R}^{1\times1\times1/2C}$ and $v_2 \in \mathbb{R}^{1\times1\times1/2C}$. Then, the weight matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with different receptive fields are calculated by the softmax function, and the feature maps with different resolutions are calibrated by the weight matrix. The calibrated feature map is weighted fusion to obtain the fusion feature map $\bar{X}$. The specific calculation is defined as follows.

$$\bar{X} = \boldsymbol{A}_c \cdot X_l + \boldsymbol{B}_c \cdot X_h \tag{26}$$

where $\boldsymbol{A}_c$ and $\boldsymbol{B}_c$ denote the weight matrix for channel calibration of different resolution feature maps. Through selective fusion of different resolution features, MR-FFM can enhance the representation of robust features, suppress the interference of backdoor triggers in poisoning samples, and improve the accuracy of aerial image semantic segmentation.

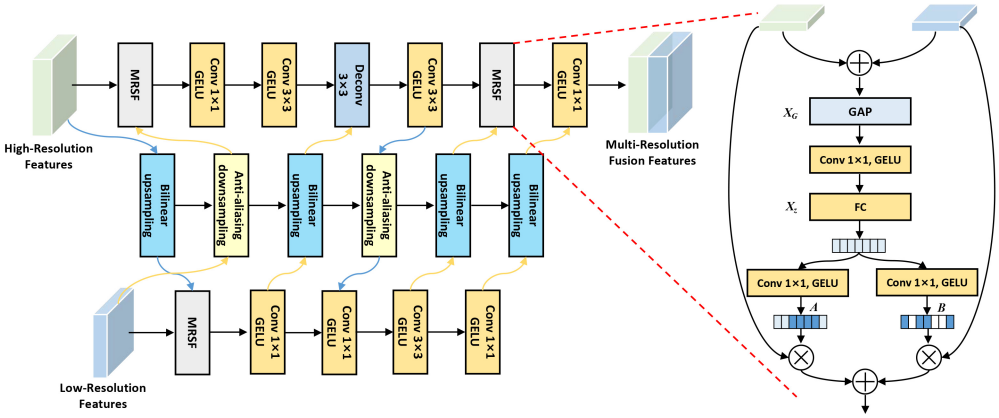


Figure 6. The detailed structure of multi-resolution feature fusion module (MR-FFM), where the blue arrow represents the information flow of high-resolution features, and the yellow arrow represents the information flow of low-resolution features.

3.5. Hierarchical Loss Function

In the proposed RIFENet, it consists of the symmetric encoder–decoder architecture. The first set of encoder–decoder structure is defined as shallow unit Z^s, the last set of encoder–decoder structure is defined as deep unit Z^d, and the rest is the middle unit Z^m. To better train and optimize the model parameters, we set different weight information for the encoder–decoder units of different layers. Formally, W is defined as the model weight, and W^s, W^d, and W^d indicate the weight information of encoder–decoder units of different layers. The cross entropy loss [61] is used to calculate the encoder–decoder units of different layers, and the specific definition is as follows.

$$\mathcal{L}(X;W) = \sum\nolimits_{x_i \in X} -\log p(t_i|x_i;W,w^c) \tag{27}$$

where X denotes the number of train samples, and $p(y_i = t(x_i)|x_i;W,w^c)$ indicates the probability that category x_i is correctly classified as the corresponding label $t(x_i)$; $c \in \{s,m,d\}$ denotes the index of different encoder–decoder units, and the loss function for introducing hierarchical structures is defined as follows.

$$\mathcal{L}\left(X;W,w^s,w^m,w^d\right) = \sum\nolimits_{c\in\{s,m,d\}} \alpha_c \mathcal{L}(X;W,w^c) \tag{28}$$

where α_c denotes the weight coefficient used to adjust the optimization process of different layer encoder–decoder units. In addition, in the feature extraction process of RIFENet, we first fuse the shallow and middle-layer units, and then splice them with the deep unit. The splice loss is calculated using the cross entropy function, which is defined as follows.

$$\mathcal{L}_s = -\frac{1}{N}\sum_{n=1}^{N}\left[g_{n,i}\log p_{n,i} + (1-g_{n,i})\log(1-p_{n,i})\right] \tag{29}$$

where N denotes the number of ground object categories, $p_{n,i}$ denotes the prediction probability that pixel i belongs to the nth class object, and $g_{n,i}$ indicates the annotation information corresponding to pixel i. The total loss function for RIFENet is defined as

$$\mathcal{L}_{total} = \lambda\mathcal{L}_s + (1-\lambda)\sum_{c\in\{s,m,d\}}\alpha_c\mathcal{L}(X;W,w^c) \tag{30}$$

where λ denotes the weight coefficient. In addition, inspired by the deep supervision strategy [62], we use the sigmoid function to calculate the loss for each layer of encoder–decoder structure to improve the training efficiency and generalization performance of the semantic segmentation network.

4. Experiments and Analysis

In this section, we first present the dataset and specific parameter settings used in the experiments, and then demonstrate the effectiveness of the proposed defense framework by conducting various poisoning attacks on the aerial image dataset against different aerial image semantic segmentation networks. Finally, we perform ablation studies to demonstrate the effectiveness of each component in the proposed method.

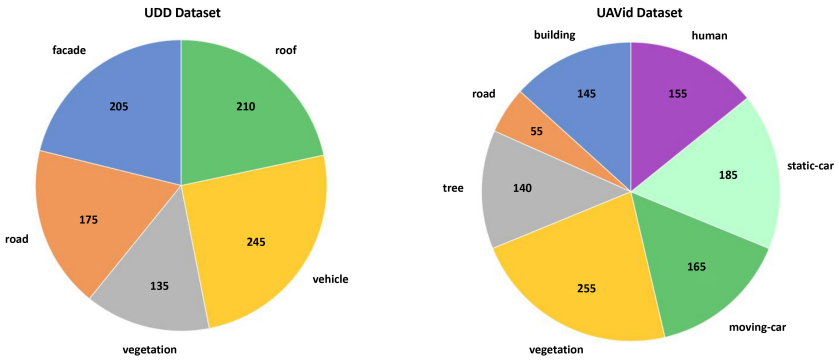
4.1. Data Descriptions

To verify the effectiveness and feasibility of the proposed method, we conduct experimental verification on the aerial image semantic segmentation benchmark datasets UDD [63] (https://github.com/MarcWong/UDD, accessed on 11 November 2020) and UAVid [64] (https://uavid.nl/, accessed on 15 July 2020) collected in urban scenes. The details of the dataset used are elaborated as follow.

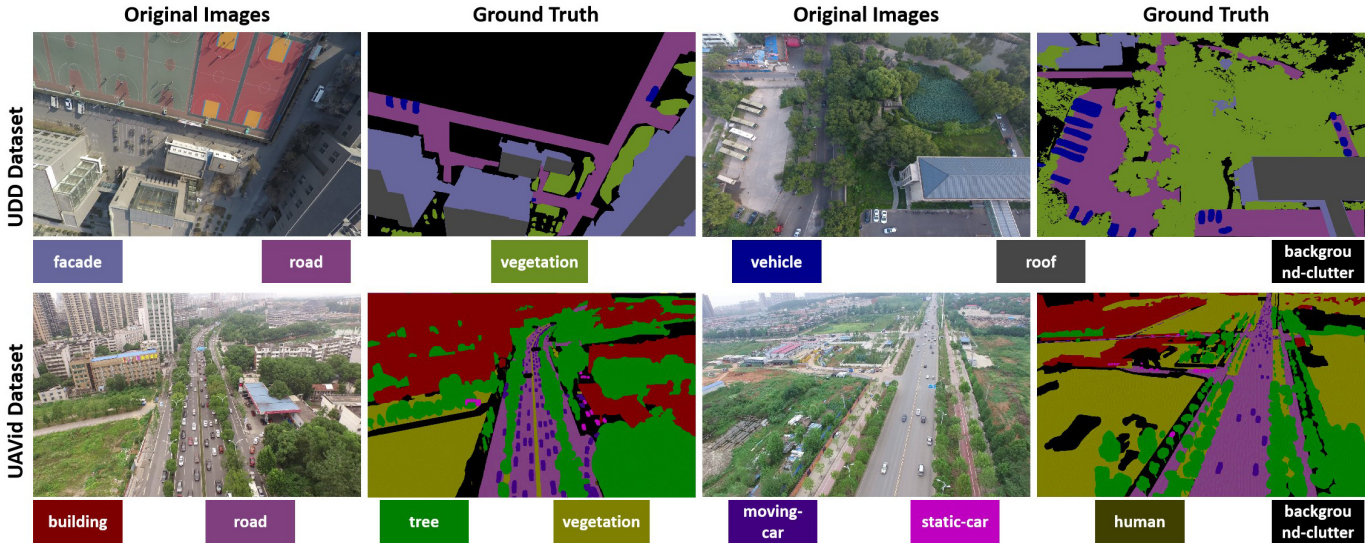
UDD Dataset: The dataset is collected by the professional-grade drone DJI-Phantom4 equipped with a 4K high-resolution camera at an altitude between 60 and 100 m. In the process of data collection, the camera shoot mode is set as interval shoot, and the panoramic image is obtained at the interval of 120 s. The original image resolution in this dataset is 4096 × 2160 or 4000 × 3000 and provides manual annotation information for semantic segmentation. Because the image is mainly collected in the urban region, the dataset contains common ground objects such as facade, road, vegetation, vehicle, and roof in the urban scene. In Figure 7, we provide the number of instances for each category and some sample examples. The dataset provides 205 high-resolution aerial images. We use 145 images as training set, 20 images as validation set, and the remaining 40 images as testing set. Limited by the computing resources of the hardware device, we scale the training set image to 1024 × 512, maintaining the original image size for the validation and testing sets.

UAVid Dataset: The dataset uses the 10 m/s stable flying drone DJI-Phantom3 as the data collection device, flying at the altitude of around 50 m and using a camera with 4k resolution for continuous shoot. The original resolution of the collected images is 4096 × 2160 or 3840 × 2160, and each image contains urban objects in different scenes. The dataset provides fine-grained manual annotation information, and the labeled object categories include building, road, tree, vegetation, moving-car, static-car, and human. Because the dataset is mainly collected in the urban center region, the image scene is more complex. The number of instances for each category and some sample images are shown in Figure 7. For the semantic segmentation task, the dataset provides 300 high-resolution aerial images. We use 210 images as training set, 30 images as validation set, and the remaining

60 images as testing set. In addition, we scale the original image size to 2048 × 1024 to reduce the computational burden of hardware devices and accelerate model training.



(**a**) The number of instances for each category



(**b**) Sample images and corresponding ground truth

Figure 7. Detailed analysis of the UDD [63] dataset and Semantic Drone [64] dataset.

4.2. Experimental Setup and Implementation Details

Poisoning Attack Settings: To demonstrate the effectiveness of the proposed defense framework against poisoning attacks, we use different poisoning sample generation strategies, including clean-label attack [35], back-gradient attack [38], generative attack [65], feature selection attack [66], transferable clean-label attack [67], and concealed poisoning attack [68]. For different attack methods, we only consider the untargeted attack scenario, which destroys the prediction results of the target model for all categories of pixels. We assume that the attacker has sufficient knowledge of the target model, including model structure and training samples.

In the process of conducting attacks, we set the attack ratio of all poisoning attacks algorithms to 30% of the number of training set samples to maximize the attack efficiency. To systematically verify the performance of different semantic segmentation models in poisoning attack scenario, we conduct clean-label, back-gradient, and generative attacks on the UDD dataset and feature selection, transferable clean-label, and concealed poisoning attacks on the UAVid dataset. Algorithm 1 provides detailed steps to attack the proposed RIFENet by poisoning samples. The purpose of poisoning attack on aerial image semantic segmentation network is to use poisoning samples to maximize destroying the prediction results of semantic segmentation model.

Algorithm 1: Poisoning Attack on Aerial Image Semantic Segmentation

Input:
Aerial image x and corresponding ground truth y.
Semantic segmentation model f with parameters θ.
Poisoning sample x_p, training epochs τ, and learning rate η.

Output:
The prediction on the poisoning sample x_p.

Initial model parameters θ with uniform distribution.
for t *in range*$(0, \tau)$ **do**
  Compute the texture enhancement features $\boldsymbol{X}_T$ via Equations (1)–(7).
  Compute the structural enhancement features $\boldsymbol{X}_S$ via Equations (8)–(16).
  Compute the global enhancement features $\boldsymbol{X}_G$ via Equations (17)–(20).
  Compute the multi-resolution fusion features $\boldsymbol{X}_M$ via Equations (21)–(26).
  Compute the hierarchical loss function $\mathcal{L}_S$ via Equations (27)–(30).
  Update θ by descending its stochastic gradients.
end
Generate the poisoning sample dataset $\mathcal{D}_{poison}$ via Ref. [35,38,65–68].
Feed the poisoning sample x_p to the model f to achieve the segmentation.

Application Details: In the experiment, we use Pytorch 1.11.0 and Python 3.8.0 to construct the proposed defense framework. All experiments are run on Dell workstations with Intel i9-12900T CPU, 64GB RAM, NVIDIA GTX Geforce 3090 GPU, Ubuntu 18.04 operation system. For model parameter optimization, we set the initial learning rate as 0.001, use the stochastic gradient descent (SGD) with momentum of 0.9 as the optimizer, and use the poly learning strategy to automatically adjust the model learning rate. The training epoch of the model is set as 2000, and the batch size is set as 16. To ensure the credibility of the experimental results, we randomly selected samples in the dataset used as training set, validation set, and testing set and repeated the experimental process 20 times. In addition, limited by the number of training set samples, we use data augmentation methods including size clipping, random inversion, brightness transformation, and random erasure to increase the number of samples to improve the model generalization capability. To assure the fairness of the comparison results, for all the compared aerial image semantic segmentation methods, we use the source code provided by the author to conduct experiments, consistent with the original hyper-parameters setting and optimization strategy.

Evaluation Metrics: To quantitatively evaluate the experimental results, we use **PA**, **mPA**, **F1_score**, and **mIoU** typically used in semantic segmentation as evaluation metrics. Specifically, we first define tp, fp, fn, and tn as true positives, false positives, false negatives, and true negatives, respectively. The definitions of different evaluation metrics are as follows.

- Pixel Accuracy (PA): This metric is defined as the proportion of correctly classified pixels to the total number of pixels, that is, $\mathbf{PA} = (tp + tn)/(tp + tn + fp + fn)$.
- Mean Pixel Accuracy (mPA): This metric is the weighted average of pixel accuracy, which calculates the pixel accuracy for each category, and then averages the pixel accuracy of all categories.
- F1 Measure (F1_score): This metric is the harmonic mean of precision (**P**) and recall (**R**) of each class. Formally, $\mathbf{F1\_score} = 2 \times (\mathrm{P} \times \mathrm{R})/(\mathrm{P} + \mathrm{R})$, where $\mathbf{P} = tp/(tp + fp)$ and $\mathbf{R} = tp/(tp + fn)$.
- Mean Intersection over Union (mIoU): This metric is the mean of IoU, and the IoU is calculated as $\mathbf{IoU} = |P_i \cap G_i|/|P_i \cup G_i|$, where P_i and G_i denote the set of prediction pixels and ground truth for the ith class.

These evaluation metrics can effectively analyze the performance of different aerial image semantic segmentation networks in poisoning attacks.

4.3. Comparison with State-of-the-Art Methods

To demonstrate the advantages of the proposed method in defending against poisoning attacks and completing accurate semantic segmentation, we compare the proposed method with several state-of-the-art methods, including the CNNs-based methods and the Transformer-based methods. For the CNNs-based methods, the proposed RIFENet is compared with AFNet [69], SBANet [70], MANet [71], SSAtNet [72], and HFGNet [25]. For the Transformer-based methods, RIFENet is compared with STUFormer [73], EMRFormer [74], CONFormer [75], ATTFormer [76], and DSegFormer [77]. For the generation of poisoning samples, as described in Section 4.2, we conduct clean-label attack [35], back-gradient attack [38], and generative attack [65] on the UDD dataset and perform feature selection attack [66], transferable clean-label attack [67], and concealed poisoning attack [68] on the UAVid dataset. The details of different compared methods are as follows.

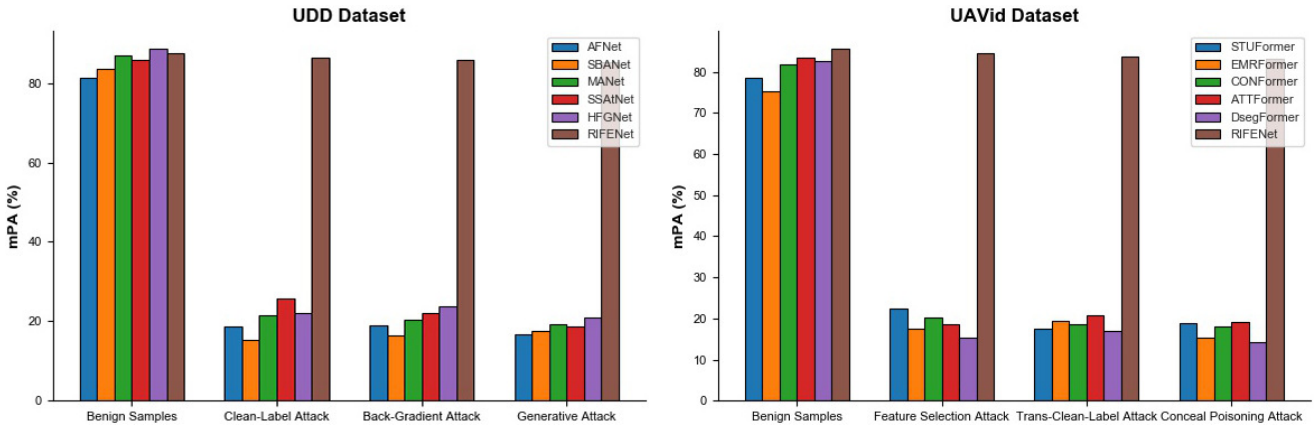
1. AFNet [69]: This network uses the hierarchical cascade structure to enhance different scale features and uses the scale-feature attention mechanism to establish the context correlation of multi-scale feature information.
2. SBANet [70]: This network uses the boundary attention module to enhance the feature representation of the boundary region and uses the multi-task learning strategy to guide the model to mine valuable feature information.
3. MANet [71]: This network uses discriminative feature learning to obtain fine-grained feature information and uses the multi-scale feature calibration module to filter redundant features to enhance feature representation.
4. SSAtNet [72]: This network uses the pyramid attention pooling module to adaptively enhance multi-scale feature representation and uses the pooling index correlation module to restore the loss of detailed feature information.
5. HFGNet [25]: This network enhances the representation of different feature information by mining hidden attention feature maps and uses the local channel attention mechanism to establish feature correlation.
6. STUFormer [73]: This model uses the spatial interaction module to establish the pixel-level correlation and uses the feature compression module to reduce the loss of detail feature and restore the feature map resolution.
7. EMRFormer [74]: This model uses multi-layer Transformer structure to extract local feature information, uses spatial attention mechanism to obtain global information, and uses feature alignment module to achieve feature fusion.
8. CONFormer [75]: This model uses context Transformer to adaptively fuse local feature information and uses the two-branch semantic correlation module to establish the correlation between local features and global features.
9. ATTFormer [76]: This model uses atrous Transformer to enhance multi-scale feature representation and uses channel and spatial attention mechanism to enhance the fine-grained representation of global feature information.
10. DSegFormer [77]: This model uses position-encoder attention mechanism to extract valuable feature information from different categories of pixel regions and uses skip connections for feature interaction and fine-grained fusion.

These methods include multi-scale feature extraction, fine-grained feature fusion, feature enhancement, and feature correlation modeling techniques commonly used in aerial image semantic segmentation. Therefore, comparing the proposed methods with existing methods can demonstrate the effectiveness and advantages of the proposed methods.

4.3.1. Experimental Results on UDD Dataset

The quantitative and qualitative results of the proposed method and all the CNNs-based methods on the UDD dataset are shown in Figure 8a, Figure 9, and Table 1. From Figure 8a, it can be observed that all compared methods, including the proposed RIFENet, achieved satisfactory performance on the benign sample dataset that are not interfered with by poisoning attacks. Nevertheless, on different poisoning sample test sets, the performance of all the compared CNNs-based methods is significantly reduced, while only our method

can maintain relatively better segmentation accuracy. The quantitative comparison results in Table 1 further show that the poisoning samples significantly reduce the accuracy of all the CNNs-based methods that perform well on the benign sample test sets. For instance, HFGNet [25] (with the best performance on the benign sample test set of the UDD dataset) decreased the mPA to 31.35% and the mIoU to 24.81% on the clean-label [10] poisoning sample test set. In addition, for some ground object categories such as facade, road, and vehicle, the PA values obtained by some CNNs-based methods are only close to 20%, indicating that many of the state-of-the-art aerial image semantic segmentation networks are highly vulnerable to poisoning attacks. SSAtNet [72] and HFGNet [25] achieved relatively better results on the poisoning sample test sets, indicating that enhancing the multi-scale features extracted by CNNs can suppress the negative impact of poisoning attacks. Compared with all the CNNs-based methods, the proposed RIFENet achieves the best performance on different poisoning sample test sets. For instance, on the poisoning sample test set generated by generative attack [65], the mPA value is 57.18% higher than that of the second-best method SSAtNet [72], demonstrating that extracting and enhancing the robust invariant features contained in aerial images can effectively improve the robustness against poisoning attacks. The semantic segmentation visualization results shown in Figure 9 reveal that all the CNNs-based methods fail to accurately predict the pixel-wise segmentation of different ground object regions under the influence of poisoning attacks. For instance, under the back-gradient attack [38], AFNet [69] suffers from severe pixel classification mistakes. The performance of the proposed method on different poisoning sample test sets is consistent with the ground-truth information provided by the UDD dataset. This further indicates that the proposed RIFENet can effectively defend against poisoning attacks and maintain better semantic segmentation performance.



(**a**) Quantitative results on UDD dataset (**b**) Quantitative results on UAVid dataset

Figure 8. Quantitative comparison results of benign samples and different poisoning attacks on UDD and UAVid datasets.

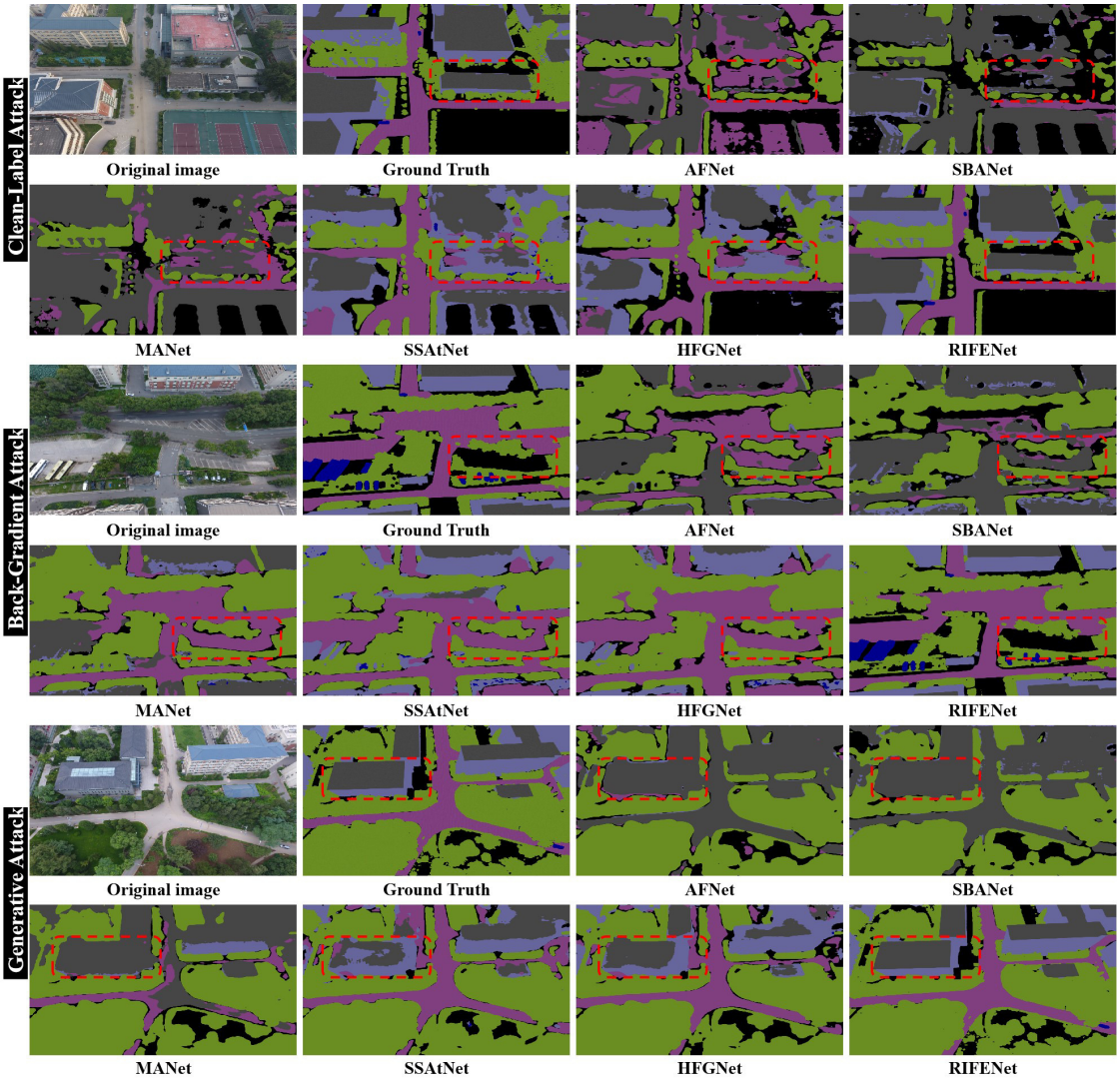


Figure 9. Semantic segmentation visualization results of different CNNs-based aerial image semantic segmentation methods encountering clean-label attack [35], back-gradient attack [38], and generative attack [65]. The color scheme used in the visualization is consistent with the color mapping provided in Figure 7b, where each color represents a specific category.

Table 1. Quantitative results of poisoning sample (Clean-Label [35]/Back-Gradient [38]/Generative [65] Attacks) test sets of the UDD dataset [63]. The best results are shown in **bold**, and the accuracy of each data sample category is reported by **PA** value.

Category	AFNet	SBANet	MANet	SSAtNet	HFGNet	RIFENet
facade	15.73/13.52/10.87	12.82/9.53/7.65	19.58/16.31/13.82	26.35/21.97/19.85	24.43/22.56/18.64	**82.75/79.62/78.43**
road	18.52/16.37/14.75	14.87/11.34/9.46	22.38/18.75/12.53	32.67/28.64/25.41	25.17/27.85/23.71	**85.97/84.31/82.45**
vegetation	22.58/20.86/18.42	19.63/16.35/12.02	25.86/22.13/15.64	35.25/31.46/28.52	36.87/28.14/25.42	**86.53/85.26/84.17**
vehicle	11.24/9.75/5.62	8.52/6.34/4.21	11.28/8.37/6.98	22.45/18.64/16.35	24.91/17.24/15.84	**83.62/82.46/82.05**
roof	23.71/18.34/16.57	21.62/17.58/14.35	24.56/20.62/18.95	39.41/32.52/29.78	36.54/28.92/24.83	**87.26/86.31/85.42**
background	32.64/28.73/26.45	26.38/20.35/18.73	33.84/30.98/28.34	42.56/38.23/35.64	40.15/35.06/33.27	**88.15/87.12/86.07**
mPA (%)	20.74/17.93/15.45	17.31/13.58/11.07	22.91/19.53/16.04	33.12/28.57/25.92	31.35/26.63/23.62	**85.71/84.18/83.10**
F1_score (%)	18.51/15.32/13.41	15.12/11.24/9.65	19.13/17.12/13.89	29.38/26.52/24.39	27.58/22.46/20.54	**82.46/81.52/79.38**
mIoU (%)	16.72/14.58/13.24	13.89/10.41/8.36	17.42/15.21/11.35	27.04/24.75/22.84	24.81/20.24/18.37	**81.63/80.75/78.24**
Runtime (s)	19.42/21.65/22.36	17.28/18.52/19.35	24.38/26.75/28.97	21.57/22.86/24.15	18.24/19.85/21.36	**16.52/17.58/18.46**

4.3.2. Experimental Results on UAVid Dataset

Different from the UDD dataset, the UAVid dataset contains more complex scenes and ground object categories, which increase the difficulty of semantic segmentation. On the UAVid dataset, we compared the proposed method with several existing Transformer-based aerial image semantic segmentation methods. From Figure 8b, it can be seen that the Transformer-based methods achieved better semantic segmentation performance on the benign sample test set, while the performance on the poisoning sample test set is significantly decreased. This indicates that the Transformer-based methods are also more vulnerable to poisoning attacks. The quantitative results of the different methods on the UAVid dataset are shown in Table 2. It can be seen that ATTFormer [76] (with excellent performance on the benign sample test set) has an mPA of 17.69% in the poisoning sample test set generated by feature selection attack [21], while the mPA only reaches 15.68% and 14.69% under the transferable clean-label attack [35] and concealed poisoning attack [68], respectively. For STUFormer [73], which obtained relatively better performance on the poisoning sample test set, its mPA reached 21.35% on the test set generated by the concealed poisoning attack [68], indicating that enhancing the representation of local and global features can suppress the impact of poisoning samples on the feature extraction process. Compared with all the Transformer-based methods, the proposed RIFENet has significant advantage on the poisoning sample test set. For instance, under the feature selection attack [66], EMRFormer [74] only achieved an mPA of 16.42%, while the proposed RIFENet achieved an mPA of 77.86%, further illustrating the superiority of the proposed method in defending against poisoning attacks. In addition, it can be seen from Table 2 that the transferable clean-label attack [67] has a greater negative impact on semantic segmentation networks. The reason is that the attack method can destroy the model feature extraction process, causing irreparable impact on the extraction of shallow and deep features, while the proposed RIFENet only achieved an mPA of 76.62% under this attack. Figure 10 presents the visualization results of different semantic segmentation models on the poisoning sample test set of the UAVid dataset. It can be seen that all the compared Transformer-based methods have significant discrepancies with the ground-truth information provided by the UAVid dataset. In contrast, our proposed method achieves better performance for different ground object categories, which further demonstrates that the proposed method can suppress the negative impact of poisoning samples by enhancing robust feature representation.

Table 2. Quantitative results of poisoning sample (Feature Selection [66]/Transferable Clean-Label [67]/Concealed Poisoning [68] Attacks) test sets of the UAVid dataset [64]. The best results are shown in **bold**, and the accuracy of each data sample category is reported by **PA** value.

Category	STUFormer	EMRFormer	CONFormer	ATTFormer	DsegFormer	RIFENet
building	21.46/18.24/19.52	15.72/17.85/13.34	18.95/15.32/16.97	17.62/16.24/14.78	13.47/15.86/11.24	**78.96/75.32/77.65**
road	25.87/21.52/22.87	19.23/22.14/16.25	22.63/20.57/18.65	20.43/21.06/17.93	16.75/18.64/13.65	**81.37/79.58/80.42**
tree	31.85/27.64/29.51	24.95/21.62/18.37	25.84/17.83/21.45	26.71/18.46/19.65	20.35/17.68/19.73	**83.75/84.21/83.97**
vegetation	23.45/19.26/22.93	17.87/15.91/15.32	20.63/13.48/16.83	18.52/14.15/14.77	15.14/13.21/12.86	**79.54/77.26/76.83**
moving-car	11.38/9.57/10.46	9.24/10.58/7.22	12.45/8.42/6.52	11.37/9.54/8.15	8.45/9.58/6.84	**74.35/75.42/73.06**
static-car	13.26/12.75/14.82	8.24/11.96/10.63	10.65/9.37/8.14	9.24/10.68/9.21	7.13/5.74/8.36	**71.92/70.87/72.35**
human	7.54/8.65/6.32	4.26/5.48/3.62	5.41/3.97/4.54	4.85/6.02/5.63	5.06/6.42/5.67	**72.53/71.62/70.59**
background	35.97/32.84/33.64	31.85/28.42/26.97	32.85/27.46/28.25	32.76/29.31/27.42	28.72/32.17/27.41	**80.45/78.67/81.42**
mPA (%)	21.35/18.81/20.01	16.42/16.75/13.97	18.68/14.55/15.17	17.69/15.68/14.69	14.38/14.91/13.22	**77.86/76.62/77.04**
F1_score (%)	19.24/16.73/17.85	14.58/13.96/12.24	16.56/12.24/13.81	14.32/13.75/12.37	12.54/13.05/11.96	**75.35/74.17/75.24**
mIoU (%)	16.52/13.74/14.27	11.25/10.97/9.65	13.15/10.83/11.26	12.05/11.52/10.65	10.68/11.73/9.54	**73.85/72.64/72.51**
Runtime (s)	22.86/21.59/23.74	28.65/27.32/29.43	32.54/33.87/34.95	21.75/23.46/22.34	23.85/24.32/25.76	**17.35/16.51/18.25**

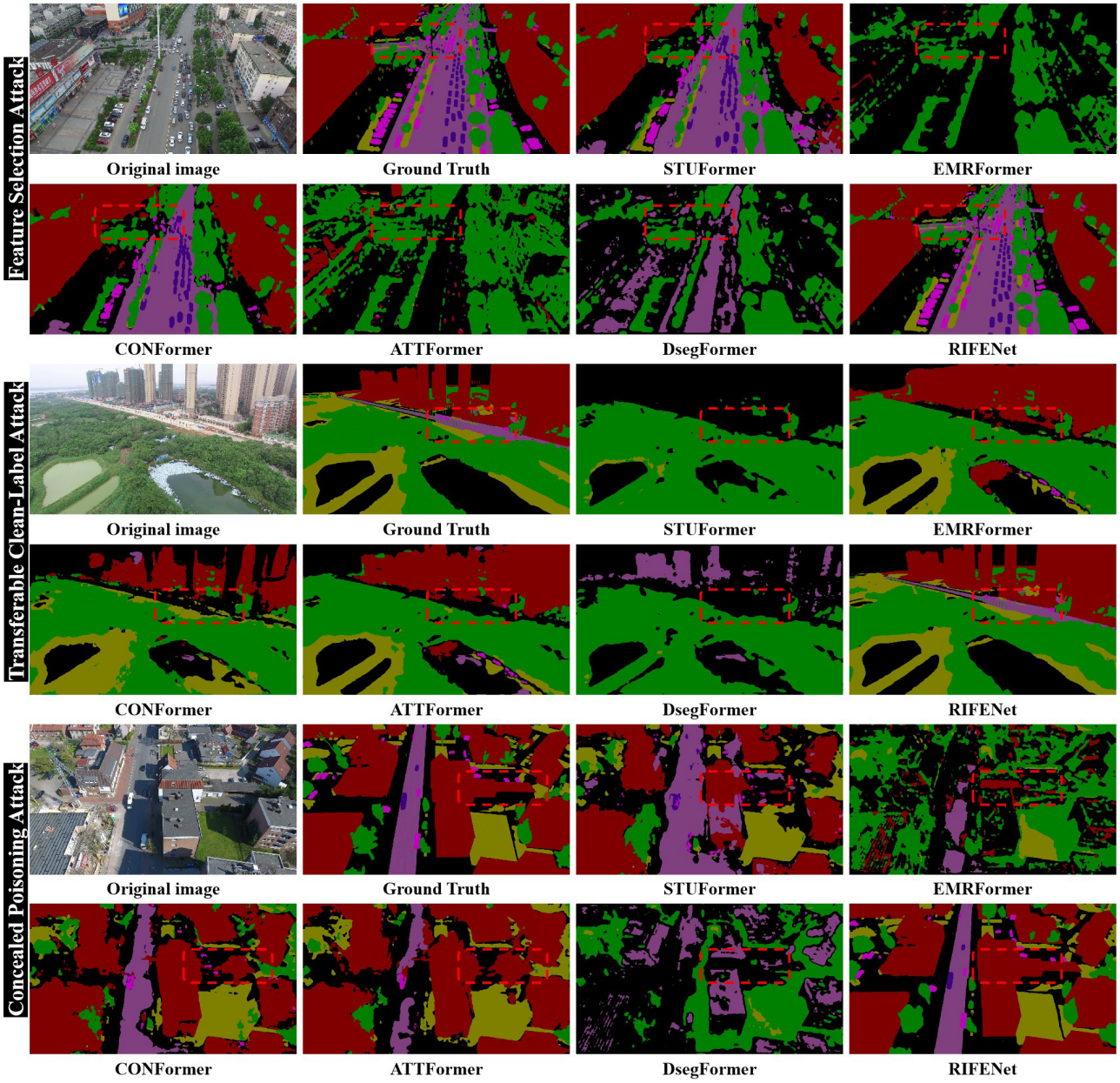


Figure 10. Semantic segmentation visualization results of different Transformer-based aerial image semantic segmentation methods encountering feature selection attack [66], transferable clean-label attack [67], and concealed poisoning attack [68]. The color scheme used in the visualization is consistent with the color mapping provided in Figure 7b, where each color represents a specific category.

4.4. Ablation Study

In this section, we evaluate the contribution of different robust feature enhancement components in the proposed RIFENet to defend against poisoning attacks and improve semantic segmentation accuracy, including the T-FEM, S-FEM, G-FEM, and MR-FFM modules. For the ablation study, we use SegNet [78] with the encoder–decoder structure as the baseline, UAVid as the test dataset, and clean-label attack [35] to generate poisoning samples. The quantitative results of the ablation experiment are shown in Table 3. It can be seen that with the introduction of different robust feature enhancement modules, the defense ability of the baseline against a poisoning attack is significantly improved. For instance, the use of T-FEM in the baseline can increase the mPA from 11.45% to 25.86% and the mIoU from 6.12% to 20.95%. The use of S-FEM enables the baseline to yield an mPA of 46.37%, G-FEM enables the baseline to yield an mPA of 75.42%, and MR-FFM enables

the baseline to yield an mPA of 83.65%. In addition, the most significant improvement in model performance is achieved by using G-FEM, as the poisoning sample has difficulty destroying the representation of global features. Therefore, enhancing the global feature can effectively improve the model robustness against poisoning attacks. We provide visual results of the impact of different robust feature enhancement components on the model feature extraction process in Figure 11. It can be seen that the poisoning attack significantly interferes with the feature extraction process of the baseline, which makes it difficult to accurately obtain the valuable feature information of the ground object region. With the introduction of different robust feature enhancement components, the baseline gradually enhances the attention to the ground object region in aerial images, so that the model can effectively obtain discriminative feature information to improve the semantic segmentation accuracy under poisoning attacks. It is worth noting that with the introduction of robust feature enhancement components, the prediction time of the semantic segmentation model for a single time is increased. The results of the ablation study further demonstrate that the combination of different robust feature enhancement components can effectively resist the interference of poisoning attacks on the model feature extraction process and semantic segmentation results.

Table 3. Performance analysis of different robust feature enhancement components on UAVid dataset, where the best results are shown in **bold**.

Baseline	T-FEM	S-FEM	G-FEM	MR-FFM	mAP (%)	F1_Score (%)	mIoU (%)	Runtime (s)
✓					11.45	8.47	6.12	**15.97**
✓	✓				25.86 (14.41 ↑)	22.58 (14.11 ↑)	20.95 (14.83 ↑)	17.65 (1.68 ↑)
✓	✓	✓			46.37 (20.51 ↑)	41.23 (18.65 ↑)	38.53 (17.58 ↑)	18.74 (1.09 ↑)
✓	✓	✓	✓		75.42 (29.05 ↑)	73.89 (32.66 ↑)	70.64 (32.11 ↑)	21.83 (3.09 ↑)
✓	✓	✓	✓	✓	**83.65** (8.23 ↑)	**80.14** (6.25 ↑)	**78.26** (7.62 ↑)	23.17 (1.34 ↑)

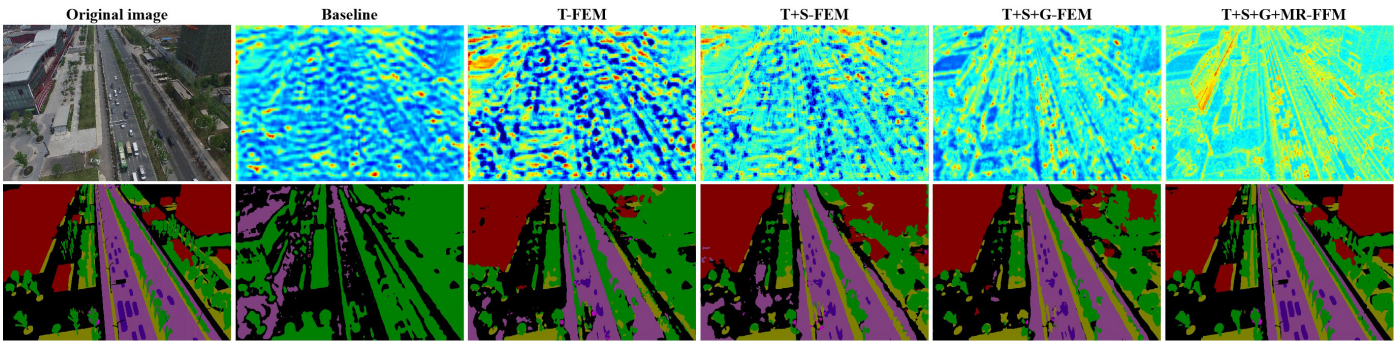


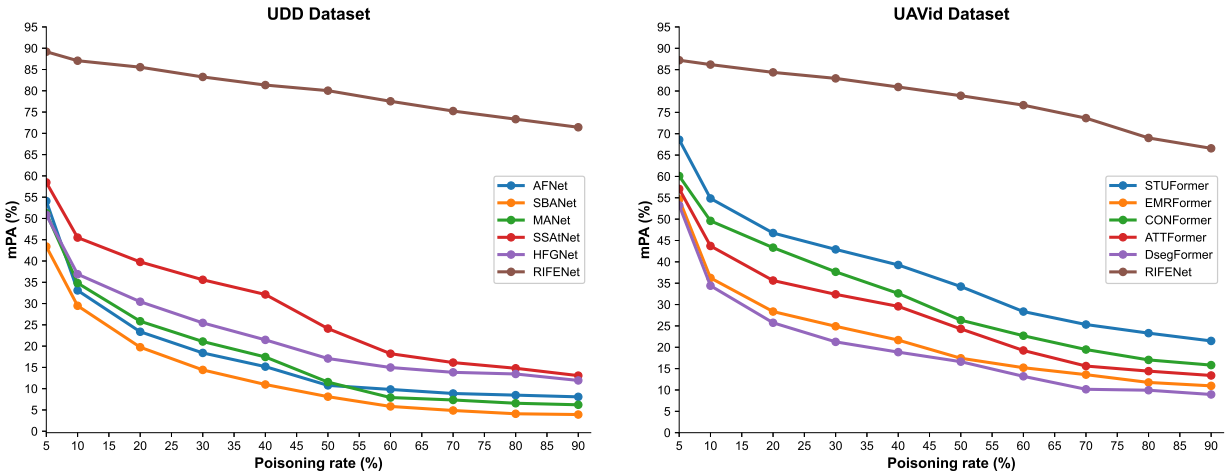
Figure 11. Feature maps and corresponding semantic segmentation results of different components in RIFENet under poisoning attack.

5. Discussion

In this section, we first verify the impact of setting different poisoning rates ρ on the model performance, and then systematically evaluate and discuss the vulnerability of existing aerial image semantic segmentation models to poisoning attacks. To analyze the impact of the poisoning rate on the model performance, we use clean-label attack [35] to generate poisoning samples on the UDD dataset and feature selection attack [66] to generate poisoning samples on the UAVid dataset. In the experiment process, the poisoning rate ρ takes the value from $\{5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$. The semantic segmentation accuracy of CNNs-based and Transformer-based aerial image semantic segmentation models on the poisoning sample dataset with different poisoning rates is shown in Figure 12. It can be seen that as the poisoning rate increases, the mPA values of all compared semantic segmentation models gradually decrease, indicating that setting the higher poisoning rate can effectively destroy model performance and reduce semantic

segmentation accuracy. However, compared with the existing semantic segmentation models, the proposed RIFENet achieved the best performance. For instance, on the UDD dataset, the poisoning rate is set as 90%, all the compared methods obtained an mPA value of only 15%, while the proposed RIFENet obtains over 70%, indicating the effectiveness of the proposed method in defending against poisoning attacks, and further demonstrating that robust feature enhancement can significantly improve the generalization capability of the model against poisoning attacks.

To systematically evaluate the impact of poisoning attacks on the performance of aerial image semantic segmentation models, we conducted various poisoning attack patterns, including clean-label attack [35], back-gradient attack [38], generative attack [65], feature selection attack [66], transferable clean-label attack [67], and concealed poisoning attack [68], on the UAVid dataset. For different poisoning attack methods, we uniformly set the poisoning rate as 30%. As shown in Table 4, the transferable clean-label attack [67] has the greatest impact on the performance of semantic segmentation models, as it can cause irreparable damage to the model feature extraction process. For instance, using feature selection attack reduces the mPA of SBANet to 10.67%, while the transferable clean-label attack reduces it to only 8.73%. Similar phenomena can be observed in other compared methods. In addition, from Table 4, we can conclude that all CNNs-based and Transformer-based aerial image semantic segmentation models are interfered with by poisoning attacks and unable to achieve segmentation accuracy similar to that on benign sample datasets. Therefore, these models urgently need to enhance their defense capabilities against poisoning attacks. In contrast, the proposed RIFENet can still achieve an mPA value over 70% under the interference of different poisoning attacks, explaining the effectiveness of the proposed method in defending against poisoning attacks. It is noteworthy that the proposed method achieves competitive performance on the benign sample dataset. This phenomenon indicates that the proposed method not only enhances the defense against poisoning attacks but also performs well in the aerial image semantic segmentation task.



(**a**) Poisoning attack on CNNs-based models (**b**) Poisoning attack on Transformer-based models

Figure 12. The influence of poisoning rate on CNNs-based and Transformer-based aerial image semantic segmentation models.

Table 4. Performance comparison of different semantic segmentation networks under poisoning attacks (report in mPA). Best results are highlighted in **bold**.

Attacks	AFNet [69]	SBANet [70]	MANet [71]	SSAtNet [72]	HFGNet [25]
benign dataset	76.23	78.65	82.41	79.42	83.67
clean-label [35]	21.86 (54.37 ↓)	18.42 (65.25 ↓)	21.32 (61.09 ↓)	25.46 (54.26 ↓)	31.75 (51.92 ↓)
back-gradient [38]	18.53 (57.70 ↓)	14.31 (69.36 ↓)	18.21 (62.20 ↓)	22.17 (57.25 ↓)	26.87 (56.80 ↓)
generative [65]	16.24 (59.99 ↓)	12.54 (71.13 ↓)	16.53 (65.88 ↓)	19.52 (59.90 ↓)	21.62 (62.05 ↓)
feature selection [66]	15.72 (60.51 ↓)	10.62 (73.05 ↓)	14.68 (67.73 ↓)	17.35 (62.07 ↓)	18.14 (65.53 ↓)
trans-clean-label [67]	11.37 (64.86 ↓)	8.73 (74.94 ↓)	9.57 (72.84 ↓)	12.82 (66.60 ↓)	14.05 (69.62 ↓)
concealed poisoning [68]	12.48 (63.75 ↓)	11.74 (71.93 ↓)	10.93 (71.48 ↓)	14.13 (65.29 ↓)	16.34 (67.33 ↓)
STUFormer [73]	**EMRFormer [74]**	**CONFormer [75]**	**ATTFormer [76]**	**DsegFormer [77]**	**RIFENet (Ours)**
78.34	75.25	81.57	83.34	85.42	**87.59**
26.46 (53.88 ↓)	22.53 (52.72 ↓)	24.64 (56.93 ↓)	20.17 (63.17 ↓)	19.68 (65.74 ↓)	**81.74** (5.85 ↓)
24.57 (53.77 ↓)	20.92 (54.33 ↓)	21.35 (60.22 ↓)	18.31 (65.03 ↓)	17.52 (67.90 ↓)	**79.83** (7.76 ↓)
22.31 (56.03 ↓)	18.75 (56.50 ↓)	19.42 (62.15 ↓)	16.42 (66.92 ↓)	16.65 (68.77 ↓)	**78.36** (9.23 ↓)
20.24 (58.10 ↓)	15.68 (59.57 ↓)	17.37 (64.20 ↓)	14.27 (69.07 ↓)	15.14 (70.28 ↓)	**77.65** (9.94 ↓)
17.18 (61.16 ↓)	11.14 (64.11 ↓)	12.98 (68.59 ↓)	9.85 (73.49 ↓)	8.23 (77.19 ↓)	**75.82** (11.77 ↓)
18.32 (60.02 ↓)	13.57 (61.68 ↓)	15.24 (66.33 ↓)	12.93 (70.41 ↓)	13.98 (71.44 ↓)	**76.93** (10.66 ↓)

6. Conclusions

In this article, we investigated the threat of poisoning attacks on aerial image semantic segmentation and proposed an effective defense framework based on robust invariant features. We first analyzed the impact of poisoning attacks on several existing aerial image semantic segmentation models and demonstrated that such attacks can destroy the semantic segmentation performance. Then, we systematically investigated the effectiveness of robust invariant features in defending against poisoning attacks and demonstrated that robust invariant features can suppress the negative effects of poisoning samples by enhancing the intrinsic attribute features contained in aerial images. Based on the advantages of robust invariant features in defending against poisoning attacks, we proposed a novel robust invariant feature enhancement network (RIFENet) for aerial image semantic segmentation under poisoning attacks. The proposed RIFENet consists of various robust feature enhancement components, which are designed to enhance the robust feature representation to suppress the interference of poisoning attacks on the feature extraction process. The experimental results on the benchmark datasets of aerial image semantic segmentation in complex urban scenes demonstrated that the proposed method has significant advantages over existing CNNs-based and Transformer-based aerial image semantic segmentation models in defending against poisoning attacks. In addition, the ablation studies further illustrated and demonstrated the contributions of the proposed different robust feature enhancement components in resisting data poisoning attacks and improving the semantic segmentation accuracy. In summary, this article is the first to reveal the threat of poisoning attacks in aerial image semantic segmentation and provides an effective defense framework. In future work, we will explore using domain adaptation and transfer learning techniques to enhance the representation of robust invariant features to further improve the defense performance of aerial image semantic segmentation models against poisoning attacks.

Author Contributions: Conceptualization, Z.W. and Y.L.; methodology, Z.W. and B.W.; software, C.Z. and J.G.; validation, Z.W. and Y.L.; formal analysis, Z.W. and J.G.; investigation, Y.L.; resources, Z.W. and C.Z.; data curation, C.Z.; original draft preparation, Z.W.; review and editing, B.W. and Y.L.; visualization, Z.W.; supervision, B.W. and Y.L.; project administration, B.W. and J.G.; funding acquisition, B.W. and Y.L. All authors have read and agreed on the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of China under Grant 42201077, in part by the National Natural Science Foundation of China under Grant 42177453, in part by the National Natural Science Foundation of China under Grant 61671465, in part by the Natural Science Foundation of Shandong Province under Grant ZR2021QD074, and in part by the Shandong Top Talent Special Foundation under Grant 0031504.

Data Availability Statement: The data that support the findings of this study are available from the author upon reasonable request. The source code can be visited at https://github.com/darkseidarch/PoisoningRIFENet, (accessed on 3 June 2023).

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

References

1. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A Review on Deep Learning in UAV Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456–102464. [CrossRef]
2. Feroz, S.; Abu Dabous, S. Uav-based Remote Sensing Applications for Bridge Condition assessment. *Remote Sens.* **2021**, *13*, 1809. [CrossRef]
3. Zhang, L.; Zhang, H.; Niu, Y.; Han, W. Mapping Maize Water Stress based on UAV Multispectral Remote Sensing. *Remote Sens.* **2019**, *11*, 605. [CrossRef]
4. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417–114425. [CrossRef]
5. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-weight Semantic segmentation network for UAV Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8287–8296. [CrossRef]
6. Pires de Lima, R.; Marfurt, K. Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer learning analysis. *Remote Sens.* **2019**, *12*, 86. [CrossRef]
7. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
8. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]
9. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
10. Mohsan, S.A.H.; Khan, M.A.; Noor, F.; Ullah, I.; Alsharif, M.H. Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive Review. *Drones* **2022**, *6*, 147. [CrossRef]
11. Zhang, Z.; Liu, Y.; Liu, T.; Lin, Z.; Wang, S. DAGN: A Real-Time UAV Remote Sensing Image Vehicle Detection Framework. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1884–1888. [CrossRef]
12. Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; Abbeel, P. Adversarial Attacks on Neural Network Policies. *arXiv* **2017**, arXiv:1702.02284.
13. Czaja, W.; Fendley, N.; Pekala, M.; Ratto, C.; Wang, I.J. Adversarial Examples in Remote Sensing. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 6–9 November 2018; pp. 408–411.
14. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial Example in Remote Sensing Image Recognition. *arXiv* **2019**, arXiv:1910.13222.
15. 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]
16. Chen, L.; Xu, Z.; Li, Q.; Peng, J.; Wang, S.; Li, H. An Empirical Study of Adversarial Examples on Remote Sensing Image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7419–7433. [CrossRef]
17. Ai, S.; Koe, A.S.V.; Huang, T. Adversarial Perturbation in Remote Sensing Image Recognition. *Appl. Soft Comput.* **2021**, *105*, 107252–107263. [CrossRef]
18. Bai, T.; Wang, H.; Wen, B. Targeted Universal Adversarial Examples for Remote Sensing. *Remote Sens.* **2022**, *14*, 5833. [CrossRef]
19. Chen, L.; Xiao, J.; Zou, P.; Li, H. Lie to Me: A Soft Threshold Defense Method for Adversarial Examples of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
20. Wei, X.; Yuan, M. Adversarial Pan-Sharpening Attacks for Object Detection in Remote Sensing. *Pattern Recognit.* **2023**, *139*, 109466. [CrossRef]
21. Lian, J.; Mei, S.; Zhang, S.; Ma, M. Benchmarking Adversarial Patch Against Aerial Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
22. Wang, Z.; Wang, B.; Liu, Y.; Guo, J. Global Feature Attention Network: Addressing the Threat of Adversarial Attack for Aerial Image Semantic Segmentation. *Remote Sens.* **2023**, *15*, 1325. [CrossRef]
23. Alfeld, S.; Zhu, X.; Barford, P. Data Poisoning Attacks against Autoregressive Models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.

24. Jagielski, M.; Severi, G.; Pousette Harger, N.; Oprea, A. Subpopulation Data Poisoning Attacks. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 15–19 November 2021; pp. 3104–3122.
25. Wang, Z.; Zhang, S.; Zhang, C.; Wang, B. Hidden Feature-Guided Semantic Segmentation Network for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
26. Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L.S.; Goldstein, T. Universal Adversarial Training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5636–5643.
27. Zhang, H.; Wang, J. Defense against Adversarial Attacks using Feature Scattering-based Adversarial Training. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
28. Zhang, X.; Wang, J.; Wang, T.; Jiang, R.; Xu, J.; Zhao, L. Robust Feature Learning for Adversarial Defense via Hierarchical Feature Alignment. *Inf. Sci.* **2021**, *560*, 256–270. [CrossRef]
29. Xu, Y.; Du, B.; Zhang, L. Self-Attention Context Network: Addressing the Threat of Adversarial Attacks for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 8671–8685. [CrossRef]
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Tian, Z.; Cui, L.; Liang, J.; Yu, S. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Comput. Surv.* **2022**, *55*, 1–35. [CrossRef]
32. Chen, T.; Ling, J.; Sun, Y. White-Box Content Camouflage Attacks against Deep Learning. *Comput. Secur.* **2022**, *117*, 102676–102682. [CrossRef]
33. Liu, G.; Lai, L. Provably Efficient Black-Box Action Poisoning attacks against Reinforcement Learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12400–12410.
34. Pang, T.; Yang, X.; Dong, Y.; Su, H.; Zhu, J. Accumulative Poisoning Attacks on Real-Time Data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2899–2912.
35. Shafahi, A.; Huang, W.R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; Goldstein, T. Poison frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
36. Zhao, B.; Lao, Y. CLPA: Clean-Label Poisoning Availability Attacks using Generative Adversarial Nets. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 9162–9170.
37. Kurita, K.; Michel, P.; Neubig, G. Weight Poisoning Attacks on Pre-trained Models. *arXiv* **2020**, arXiv:2004.06660.
38. Muñoz-González, L.; Biggio, B.; Demontis, A.; Paudice, A.; Wongrassamee, V.; Lupu, E.C.; Roli, F. Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 27–38.
39. Guo, W.; Tondi, B.; Barni, M. An Overview of Bbackdoor Attacks against Deep Neural Networks and Possible Defences. *IEEE Open J. Signal Process.* **2022**, *3*, 261–287. [CrossRef]
40. Huang, A. Dynamic Backdoor Attacks against Federated Learning. *arXiv* **2020**, arXiv:2011.07429.
41. Aghakhani, H.; Meng, D.; Wang, Y.X.; Kruegel, C.; Vigna, G. Bullseye Polytope: A Scalable Clean-Label Poisoning Attack with Improved Transferability. In Proceedings of the 2021 IEEE European Symposium on Security and Privacy (EuroS&P), Virtually, 21–25 February 2021; pp. 159–178.
42. Shafahi, A.; Najibi, M.; Ghiasi, M.A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L.S.; Taylor, G.; Goldstein, T. Adversarial Training for Free! *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
43. Geiping, J.; Fowl, L.; Somepalli, G.; Goldblum, M.; Moeller, M.; Goldstein, T. What Doesn't Kill You Makes You Robust (er): How to Adversarially Train against Data Poisoning. *arXiv* **2021**, arXiv:2102.13624.
44. Gao, Y.; Wu, D.; Zhang, J.; Gan, G.; Xia, S.T.; Niu, G.; Sugiyama, M. On the Effectiveness of Adversarial Training against Backdoor Attacks. *arXiv* **2022**, arXiv:2202.10627.
45. Hallaji, E.; Razavi-Far, R.; Saif, M.; Herrera-Viedma, E. Label Noise Analysis meets Adversarial Training: A Defense against Label Poisoning in Federated Learning. *Knowl.-Based Syst.* **2023**, *266*, 110384. [CrossRef]
46. Chen, J.; Zhang, X.; Zhang, R.; Wang, C.; Liu, L. De-pois: An Attack-Agnostic Defense against Data Poisoning Attacks. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 3412–3425. [CrossRef]
47. Liu, A.; Liu, X.; Yu, H.; Zhang, C.; Liu, Q.; Tao, D. Training Robust Deep Neural Networks via Adversarial Noise Propagation. *IEEE Trans. Image Process.* **2021**, *30*, 5769–5781. [CrossRef]
48. Yang, X.; Xu, Z.; Luo, J. Towards Perceptual Image Dehazing by Physics-based Disentanglement and Adversarial Training. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
49. Li, X.; Qu, Z.; Zhao, S.; Tang, B.; Lu, Z.; Liu, Y. Lomar: A Local Defense against Poisoning Attack on Federated Learning. *IEEE Trans. Dependable Secur. Comput.* **2021**, *20*, 437–450. [CrossRef]
50. Dang, T.K.; Truong, P.T.T.; Tran, P.T. Data Poisoning Attack on Deep Neural Network and Some Defense Methods. In Proceedings of the 2020 International Conference on Advanced Computing and Applications (ACOMP), Quy Nhon, Vietnam, 25–27 November 2020; pp. 15–22.
51. Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; Kankanhalli, M. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 11278–11287.

52. Li, T.; Wu, Y.; Chen, S.; Fang, K.; Huang, X. Subspace Adversarial Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13409–13418.
53. Kim, J.; Lee, B.K.; Ro, Y.M. Distilling Robust and Non-Robust Features in Adversarial Examples by Information Bottleneck. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17148–17159.
54. Xie, S.M.; Ma, T.; Liang, P. Composed Fine-Tuning: Freezing Pre-Trained Denoising Autoencoders for Improved Generalization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11424–11435.
55. Song, C.; He, K.; Lin, J.; Wang, L.; Hopcroft, J.E. Robust Local Features for Improving the Generalization of Adversarial Training. *arXiv* **2019**, arXiv:1909.10147.
56. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: New York, NY, USA, 2015; pp. 234–241.
57. Liao, X.; Yin, J.; Chen, M.; Qin, Z. Adaptive Payload Distribution in Multiple Images Steganography based on Image Texture Features. *IEEE Trans. Dependable Secur. Comput.* **2020**, *19*, 897–911. [CrossRef]
58. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 11–17 October 2021; pp. 10012–10022.
59. Zhu, Y.; Liang, Z.; Yan, J.; Chen, G.; Wang, X. ED-Net: Automatic Building Extraction from High-Resolution Aerial Images with Boundary Information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4595–4606. [CrossRef]
60. Li, P.; Ren, P.; Zhang, X.; Wang, Q.; Zhu, X.; Wang, L. Region-Wise Deep Feature Representation for Remote Sensing Images. *Remote Sens.* **2018**, *10*, 871. [CrossRef]
61. Li, X.; Yu, L.; Chang, D.; Ma, Z.; Cao, J. Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4204–4212. [CrossRef]
62. Luo, Y.; Lü, J.; Jiang, X.; Zhang, B. Learning From Architectural Redundancy: Enhanced Deep Supervision in Deep Multipath Encoder–Decoder Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4271–4284. [CrossRef]
63. Chen, Y.; Wang, Y.; Lu, P.; Chen, Y.; Wang, G. Large-scale Structure from Motion with Semantic Constraints of Aerial Images. In Proceedings of the Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, 23–26 November 2018; Proceedings, Part I 1; Springer: New York, NY, USA, 2018; pp. 347–359.
64. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A Semantic Segmentation Dataset for UAV Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
65. Yang, C.; Wu, Q.; Li, H.; Chen, Y. Generative Poisoning Attack Method against Neural Networks. *arXiv* **2017**, arXiv:1703.01340.
66. Liu, H.; Ditzler, G. Data Poisoning against Information-Theoretic Feature Selection. *Inf. Sci.* **2021**, *573*, 396–411. [CrossRef]
67. Zhu, C.; Huang, W.R.; Li, H.; Taylor, G.; Studer, C.; Goldstein, T. Transferable Clean-Label Poisoning Attacks on Deep Neural Nets. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7614–7623.
68. Zheng, J.; Chan, P.P.; Chi, H.; He, Z. A Concealed Poisoning Attack to Reduce Deep Neural Networks' Robustness against Adversarial Samples. *Inf. Sci.* **2022**, *615*, 758–773. [CrossRef]
69. Liu, R.; Mi, L.; Chen, Z. AFNet: Adaptive Fusion Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7871–7886. [CrossRef]
70. Li, A.; Jiao, L.; Zhu, H.; Li, L.; Liu, F. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
71. He, P.; Jiao, L.; Shang, R.; Wang, S.; Liu, X.; Quan, D.; Yang, K.; Zhao, D. MANet: Multi-Scale Aware-Relation Network for Semantic Segmentation in Aerial Scenes. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
72. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic Segmentation with Attention Mechanism for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
73. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
74. Xiao, T.; Liu, Y.; Huang, Y.; Li, M.; Yang, G. Enhancing Multiscale Representations with Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
75. Ding, L.; Lin, D.; Lin, S.; Zhang, J.; Cui, X.; Wang, Y.; Tang, H.; Bruzzone, L. Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images. *arXiv* **2021**, arXiv:2106.15754.
76. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
77. Li, X.; Cheng, Y.; Fang, Y.; Liang, H.; Xu, S. 2DSegFormer: 2-D Transformer Model for Semantic Segmentation on Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
78. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]