remote sensing

MDPI

Article

Robust Feature-Guided Generative Adversarial Network for Aerial Image Semantic Segmentation against Backdoor Attacks

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; miswz@iocas.ac.cn (Z.W.)
[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: Profiting from the powerful feature extraction and representation capabilities of deep learning (DL), aerial image semantic segmentation based on deep neural networks (DNNs) has achieved remarkable success in recent years. Nevertheless, the security and robustness of DNNs deserve attention when dealing with safety-critical earth observation tasks. As a typical attack pattern in adversarial machine learning (AML), backdoor attacks intend to embed hidden triggers in DNNs by poisoning training data. The attacked DNNs behave normally on benign samples, but when the hidden trigger is activated, its prediction is modified to a specified target label. In this article, we systematically assess the threat of backdoor attacks to aerial image semantic segmentation tasks. To defend against backdoor attacks and maintain better semantic segmentation accuracy, we construct a novel robust generative adversarial network (RFGAN). Motivated by the sensitivity of human visual systems to global and edge information in images, RFGAN designs the robust global feature extractor (RobGF) and the robust edge feature extractor (RobEF) that force DNNs to learn global and edge features. Then, RFGAN uses robust global and edge features as guidance to obtain benign samples by the constructed generator, and the discriminator to obtain semantic segmentation results. Our method is the first attempt to address the backdoor threat to aerial image semantic segmentation by constructing the robust DNNs model architecture. Extensive experiments on real-world scenes aerial image benchmark datasets demonstrate that the constructed RFGAN can effectively defend against backdoor attacks and achieve better semantic segmentation results compared with the existing state-of-the-art methods.

Keywords: aerial images; semantic segmentation; deep neural networks (DNNs); adversarial machine learning (AML); backdoor attack; robust feature extraction

Citation: Wang, Z.; Wang, B.; Zhang, C.; Liu, Y.; Guo, J. Robust Feature-Guided Generative Adversarial Network for Aerial Image Semantic Segmentation against Backdoor Attacks. *Remote Sens.* **2023**, *15*, 2580. https://doi.org/10.3390/rs15102580

Academic Editors: Pedram Ghamisi, Bo Du and Yonghao Xu

Received: 15 March 2023
Revised: 5 May 2023
Accepted: 12 May 2023
Published: 15 May 2023

1. Introduction

With the development of satellite and airborne aerial sensors in recent years, the amount of earth observation data has shown explosive growth [1]. The advantages of deep neural networks (DNNs) in feature extraction and representation have made it widely used in remote sensing (RS) data mining [2,3]. As one of the basic tasks of RS applications, aerial image semantic segmentation plays an essential role in urban planning [4], disaster assessment [5], and military surveillance [6]. Although numerous efforts have been made in existing studies to construct DNNs models with optimal performance for aerial image semantic segmentation tasks in different scenarios [7]. Nevertheless, the superior performance of these methods comes with the drawbacks of introducing new vulnerabilities and security risks, which must be addressed considering that most of the earth observation missions in the RS and aerial fields are safety-critical [8].

Currently, studies on safety-related issues in DNNs focus on adversarial examples (AEs) [9], misleading DNNs models to produce false prediction results by carefully designed adversarial noise. For the RS community, the topic of artificial intelligence (AI) security has also received extensive attention. Czaja et al. [10] reveal the threat AEs posed to DNNs-based satellite image classifiers, demonstrating that the optimal-performance classifier can be misled by adding human-imperceptible adversarial noise. Chen et al. [11] validated the transferability of AEs in RS image recognition and provided possible adversarial defense strategies. Ai et al. [12] demonstrated the real existence of AEs in RS images and illustrated that the phenomenon is caused by the inconsistent feature space distribution. Bai et al. proposed [13] the first targeted attack method for RS image classifier, and used the feature space difference information to improve the attack success rate. In addition to the image classification task, the related adversarial examples for object detection and semantic segmentation of RS images have also made progress. Lu et al. [14] constructed the scale-adaptive adversarial patch attack for aircraft object detection in RS images, which can adaptively add adversarial patch to achieve the object stealth effect. Zhang et al. [15] designed the physically achievable AEs generation framework to deceive the RS object detector. Xu et al. [16] first analyzed the threat posed by AEs to the RS image semantic segmentation and constructed an universal adversarial perturbation generation strategy. Wang et al. [17] systematically evaluated the negative impact of AE attacks on existing semantic segmentation networks and designed a robust aerial image semantic segmentation framework. In summary, all of the above-mentioned studies focus on the AE attacks in the RS field, while other security issues, such as backdoor attacks [18], member inference [19], model stealing [20], and other AI security threats still deserve our attention.

Different from the AE attacks that assume that adversaries can only perform attacks in the model inference phase [9], recent studies have further explored the possibility of conducting attacks in the model training process [21]. Such attacks are called backdoor attacks [22] or Trojan attacks [23]. The backdoor attack aims to inject backdoor triggers with specific labels into the training process of the attacked model, so that the target model obtains normal prediction results when facing benign samples in the inference phase. In contrast, its prediction results may be maliciously modified when confronting poisoned samples with backdoor triggers. The emergence of backdoor attacks significantly boosts the security risk level of DNNs models [24]. For the RS community, backdoor attacks for scene classification tasks have received some attention. Brewer et al. [25] introduced backdoor attacks in satellite image scene classification for the first time. The experimental results show that the classifier performance can be significantly damaged by constructing simple backdoor triggers. Dräger et al. [26] constructed a backdoor attack method based on wavelet transform theory and verified the transferability of backdoor attack for RS image classification. However, the threat of backdoor attacks on aerial image semantic segmentation has yet to receive extensive attention. In Figure 1, we illustrate the impact of backdoor attack on aerial image semantic segmentation. It can be seen from Figure 1 that we can significantly destroy the semantic segmentation model with better interpretation performance by using the simple backdoor trigger.

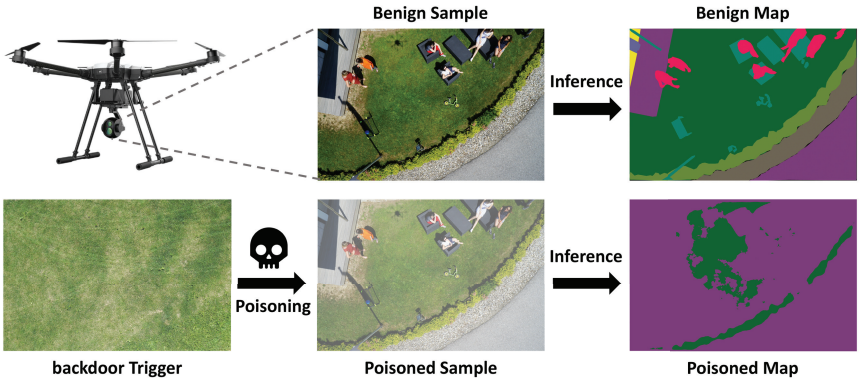


Figure 1. Illustration of the backdoor attacks on the aerial image semantic segmentation. While the difference between the benign and poisoned samples may be imperceptible for human observers, such poisoned samples contain stealthy triggers (WABA [26]) which can make state-of-the-art semantic segmentation network (DeepLabV3 [27]) produce false predictions.

Existing studies have shown that the use of robust features [28–31] (e.g., global features, semantic features, shape features, etc.) can effectively enhance the robustness of DNNs models. In this article, to address the threat posed by backdoor attacks to aerial image semantic segmentation, inspired by robust feature learning strategies, we attempt to construct carefully designed robust feature extractors to defend against backdoor attacks. In particular, we propose a robust feature-guided generative adversarial network (RFGAN). Based on the robust attributes of global and edge features, RFGAN first uses the carefully designed robust global feature extractor (RobGF) and robust edge feature extractor (RobEF) to obtain robust feature information. Then, guided by the obtained robust features, the generator model generates benign samples without backdoor triggers, and the discriminator model obtains accurate aerial image semantic segmentation results. Overall, the constructed RFGAN resists backdoor attacks in aerial image semantic segmentation from the perspective of DNNs model structure design. The contributions of this study are summarized as follows.

- To the best of our knowledge, we are the first to introduce the concept of backdoor attack into aerial image semantic segmentation. Our research comprehensively reveals the significance of the resistibility and robustness of DNNs models when addressing the safety-critical airborne earth observation tasks.
- We comprehensively analyze and summarize the characteristics of backdoor attacks in aerial images, and propose a robust feature guided generative adversarial network (RFGAN) against backdoor attacks. The constructed RFGAN can filter backdoor triggers by extracting different robust feature information.
- Based on the robust attributes of global and edge features, we construct robust global feature extractor (RobGF) and robust edge feature extractor (RobEF), respectively. In addition, the generative adversarial network (GAN) framework is used to generate benign samples and obtain semantic segmentation results.
- To verify the effectiveness and feasibility of the proposed defense framework, the extensive experiments are conducted on real-world aerial image datasets. The experimental results show the proposed method can against backdoor attacks while maintaining high semantic segmentation precision.

The rest of this article is organized as follows. In Section 2, the related works are reviewed. Section 3 describes the proposed RFGAN framework. Section 4 presents the experimental results and analysis. The discussion and conclusion are summarized in Sections 5 and 6.

2. Related Works

In this section, we briefly review the existing backdoor attack and defense methods, and present the basic concepts and definitions.

2.1. Backdoor Attack

As a new security threat against DNNs model, backdoor attacks commonly appear during the model training process. The purpose of backdoor attacks it to mislead the target model to produce false prediction results under specific trigger conditions. Gu et al. [32] proposed the first backdoor attack method BadNet, which uses benign samples and poisoned samples to jointly train the DNNs model, and the attacker maliciously assigns the labels of the poisoned samples to the specified category. To improve the attack concealment, Chen et al. [22] designed a blended injection strategy to achieve the escape effect by mixing backdoor triggers with benign samples, which effectively reduces the probability of the trigger being detected. To enhance the attack intensity, Shafahi et al. [33] constructed a clean-label attack that preserves the label of the poisoned sample so that the maliciously tampered sample is consistent with the feature information of benign sample, but produces the label reversal effect when the attack is triggered. To extend the backdoor attack to the semantic segmentation task, Li et al. [34] constructed the first hidden backdoor attack for the semantic segmentation model, and realized the targeted attack behavior by modifying the specific pixel category. Chan et al. [35] systematically analyzed the security of backdoor attacks on the object detection model, and constructed a backdoor attack strategy to achieve object stealth and misclassification. In addition, backdoor attacks have also received attention for other DNNs-based tasks, such as natural language processing (NLP) [36], malware detection [37], and speaker recognition [38].

2.2. Backdoor Defense

With the continuous emergence of different backdoor attack methods, the corresponding backdoor defense strategies have also developed rapidly. The existing backdoor defense methods can be classified as backdoor detection and data preprocessing. For the detection-based backdoor defense, Tran et al. [39] identified the backdoor trigger by comparing the spectral difference between poisoned and benign samples. Chan et al. [40] used the gradient information of the trigger pattern as the initial clustering center, and used the clustering algorithm to identify benign and poisoned samples. Peri et al. [41] constructed deep clustering model to detect backdoor attacks, which can detect backdoor triggers with hidden attributes. Another detection method is to identify the poisoning model, such as Liu et al. [42], who realized backdoor detection by counting the activation difference information of neurons between benign and poisoned samples. Wang et al. [43] proposed a trigger detection framework based on discrete feature analysis, which determines the backdoor attack by analyzing the feature distribution between trigger and original image. For the defense method based on data preprocessing, Liu et al. [44] first used autoencoders as preprocessors to filter poisoned samples. Doan et al. [45] used the GAN framework to restore poisoned samples with backdoor triggers. Based on the sensitivity of static triggers to contour and position factors, Li et al. [46] used spatial transformation preprocessing to suppress the activation of backdoor triggers. In summary, the development of backdoor attacks has forced continued advances in corresponding backdoor defense techniques.

2.3. Preliminary

Semantic Segmentation: Let $\mathcal{D}_{clean} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N}$ denotes the benign sample dataset, where $\boldsymbol{x}_i \in \mathcal{X} = \{0, 1, ..., 255\}^{C \times W \times H}$ denotes the single image, $\boldsymbol{y}_i \in \mathcal{Y} = \{0, 1, ..., K\}^{W \times H}$ is the pixel-level annotation information corresponding to $\boldsymbol{x}_i$, and K denotes the number of object categories contained in the benign sample dataset. Taking the CNNs-based semantic segmentation method with end-to-end supervised learning manner as an example, it intended to learn a semantic segmentation model f with optimal parameter $\boldsymbol{\theta}$, i.e., $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, by $\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathcal{L}(f_{\theta}(\boldsymbol{x}_i), \boldsymbol{y}_i)$ where $\mathcal{L}(\cdot)$ denotes the loss function.

Backdoor Attacks: The existing backdoor includes two main processes: **(1)** generate the dataset $\mathcal{D}_{trigger}$ with backdoor trigger; **(2)** inject $\mathcal{D}_{trigger}$ during model training. The first stage of the backdoor attack is the core stage, and the construction of Human imperceptible backdoor trigger has been extensively studied. In addition, the target labels of existing attack methods are sample-independent, i.e., all poisoned samples are given the same label as the target. Specifically, let $\mathcal{D}_{trigger}$ denotes the poisoned sample dataset with the backdoor trigger, and $\mathcal{D}_{clean}$ denotes the benign sample dataset, both of which are subsets of dataset $\mathcal{D}$. $\mathcal{D}_{trigger} = \mathcal{D}_{modified} \cup \mathcal{D}_{clean}$, where $\gamma = \frac{|\mathcal{D}_{modified}|}{|\mathcal{D}|}$ denotes the poisoning rate, $\mathcal{D}_{modified} = \{(\boldsymbol{x}', \boldsymbol{y}_t) \mid \boldsymbol{x}' = G(\boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D} \backslash \mathcal{D}_{clean}\}$, $\boldsymbol{y}_t$ is the target label, and $G : \mathcal{X} \to \mathcal{X}$ is an attacker-specified poisoned image generator. For specific attack methods, as mentioned in Ref. [22], $G(\boldsymbol{x}) = (\mathbf{1} - \boldsymbol{\lambda}) \otimes \boldsymbol{x} + \boldsymbol{\lambda} \otimes \boldsymbol{t}$, where $\boldsymbol{\lambda} \in [0,1]^{C \times H \times W}$ indicates the visibility-related hyperparameter, and $\boldsymbol{t} \in \mathcal{X}$ is a pre-defined trigger pattern.

Threat model: Considering the practical application scenarios of aerial image semantic segmentation, we assume that attackers can arbitrarily modify the training dataset, but cannot access or destroy the training process and architecture parameters of the semantic segmentation model, and have no information about the inference prediction phase. The scenario assumed we conducted is to follow Ref. [18], i.e., this is a common setting for backdoor attackers, which makes the attack can happen in many real-world scenarios.

Attacker Goals: The goal of the existing backdoor attack methods can be summarized as generating hidden backdoor triggers and maximizing the attack effect, i.e., ***effectiveness*** and ***stealthiness***. Specifically, the ***effectiveness*** requires that pixels of objects with the source class (i.e., the attacker-specified class for misclassifying) will be predicted as the target class when the trigger appears. The stealthiness requires that **(1)** the trigger is unobtrusive, **(2)** the attacked model behaves normally on benign samples, and **(3)** the performance on pixels with non-source classes in attacked samples will not be significantly reduced.

3. Methodology

The overall framework of RFGAN is shown in Figure 2, which consists of robust global feature extractor (RobGF), robust edge feature extractor (RobEF), benign sample generator, and discriminator. Specifically, RFGAN first uses the convolution operation to extract feature of sample $\boldsymbol{x}_i$ with backdoor trigger to obtain the initial feature information $\boldsymbol{A}_0$; secondly, RobGF and RobEF are used to extract robust global feature $\boldsymbol{A}_{rg}$ and robust edge feature $\boldsymbol{A}_{re}$; thirdly, guided by robust global and edge features, the benign sample generator $\boldsymbol{G}(x)$ is used to generate the benign sample $\boldsymbol{x}_c$ without backdoor trigger; finally, the benign samples are input into the discriminator model $\boldsymbol{D}(x)$ to obtain the aerial image semantic segmentation results. The core components of RFGAN are RobGF and RobEF, where RobGF uses the idea of pixel global modeling to establish correlation between different pixels in aerial images from spatial and channel dimensions to output global features with robust attributes. RobEF uses the Transformer framework to serialize the initial features, extract the robust edge feature information contained in each token sequence and perform feature fusion. In summary, RFGAN takes the acquisition of robust features contained in aerial images as the basic idea, and constructs the robust DNNs model to against the backdoor attack behavior faced by aerial image semantic segmentation.

3.1. Robust Global Feature Extractor

Since obtaining global context information requires establishing the correlation between a given pixel and all pixels, the prediction of this pixel will be affected by other related pixels [47]. In this case, if the backdoor attack assigns a mistaken label to the pixel, the error loss at that pixel will be passed back to all other relevant pixels in the form of backward propagation. Therefore, the total loss at this pixel will be shared by all other related pixels, so attacks on global features may require a higher level of perturbation.

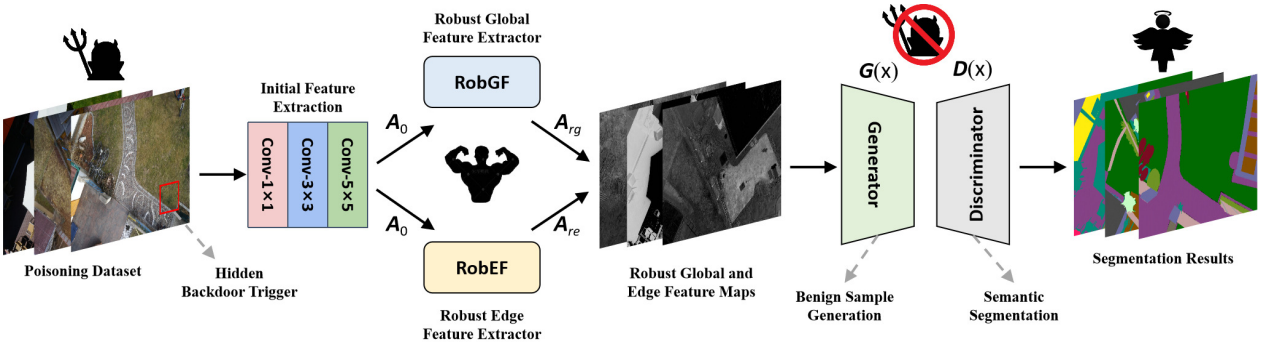


Figure 2. Overall framework of RFGAN against backdoor attacks. RFGAN directly outputs the robust global feature $\boldsymbol{A}_{rg}$ and robust edge feature $\boldsymbol{A}_{re}$ obtained by RobGF and RobEF to the generator; while generator $\boldsymbol{G}(x)$ uses the robust feature map to reconstruct a new benign sample $\boldsymbol{x}(i)$ for semantic segmentation.

To obtain the global context feature information with robust attributes in aerial images, we construct a robust global feature extractor (RobGF), and the structure is shown in Figure 3. RobGF consists of the channel-spatial attention mechanism (CSAM) and the efficient non-local attention mechanism (ENLAM), which suppresses adversarial noise interference by fully obtaining discriminative global feature information using dual attention mechanisms in both spatial and channel dimensions. In the specific global feature extraction process, the feature map $\boldsymbol{A}_0$ obtained in the preprocessing stage is used as input, the feature maps $\boldsymbol{A}_s$ and $\boldsymbol{A}_n$ are obtained CSAM and ENLAM, and the robust global feature $\boldsymbol{A}_{rg}$ is obtained by feature fusion operation. Mathematically,

$$\boldsymbol{A}_{rg} = \mathcal{K}_{1\times1}(\mathrm{cat}(\mathcal{F}_{SC}(\boldsymbol{A}_0), \mathcal{F}_{NL}(\boldsymbol{A}_0))) \tag{1}$$

where $\mathcal{F}_{SC}(\cdot)$ denotes the CSAM module, $\mathcal{F}_{NL}(\cdot)$ denotes the ENLAM module, $\mathrm{cat}(\cdot)$ indicates the feature fusion function, and $\mathcal{K}_{1\times1}(\cdot)$ denotes 1×1 convolution function. Since CSAM models global information in channel and spatial dimensions, feature $\boldsymbol{A}_0$ requires to be transmitted in parallel on channel and spatial dimensions, and 1×1 convolution is used to fuse global channel and spatial features. Mathematically,

$$\boldsymbol{F}_G = \mathcal{K}_{1\times1}(\mathrm{cat}(\mathcal{F}_c(\boldsymbol{A}_0), \mathcal{F}_s(\boldsymbol{A}_0))) \tag{2}$$

where $\boldsymbol{F}_G$ denotes the global feature after fusion, $\mathcal{F}_c(\cdot)$ denotes second-order channel attention mechanism [48], and $\mathcal{F}_s(\cdot)$ indicates the spatial attention mechanism [49]. Inspired by Ref. [50], CSAM first normalizes the covariance of feature $\boldsymbol{A}_0$ and transforms its dimension to $C \times H \times W$. The covariance matrix corresponding to $\boldsymbol{A}_0$ is

$$\Sigma = \boldsymbol{A}_0 \times \left(\frac{1}{H \times W} \times \left(\boldsymbol{E} - \frac{1}{H \times W} \times \boldsymbol{M}_{m,n=1}\right) \times \boldsymbol{A}_0^T\right) \tag{3}$$

where $\boldsymbol{E}$ and $\boldsymbol{M}$ denote the unit matrix and the matrix with all elements of one, respectively, and T indicates the matrix transpose operation. The eigenvalue decomposition is performed on the semi-definite covariance matrix Σ, then the covariance normalization of feature $\boldsymbol{A}_0$ can be represented as a power operation of eigenvalues, that is,

$$\hat{\boldsymbol{A}}_0 = \Sigma^{\alpha} = \boldsymbol{P} \cdot \Lambda^{\alpha} \cdot \boldsymbol{P}^T \tag{4}$$

where $\boldsymbol{P}$ is positive definite matrix, $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_C)$ denotes diagonal matrix, and λ indicates positive real number. The global covariance pooling is performed on $\hat{\boldsymbol{A}}_0 = [\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_C]$

to generate channel feature descriptor $\boldsymbol{q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_C]$. Taking the *cth* channel as an example, the calculation is as

$$\boldsymbol{q}_c = \mathcal{F}_{GCP}(\boldsymbol{x}_c) = \frac{1}{C}\sum_i^C \boldsymbol{x}_C(i) \tag{5}$$

where $\mathcal{F}_{GCP}(\cdot)$ denotes the global covariance pooling operation, and $\boldsymbol{q}_c$ denotes the *cth* channel descriptor. To obtain the correlation between channels, it is required to process the channel descriptor dimension to obtain the attention weight map, and the calculation is as

$$\boldsymbol{w} = \rho(\boldsymbol{W}_U\delta(\boldsymbol{W}_D\boldsymbol{q})) \tag{6}$$

where $\boldsymbol{W}_U$ and $\boldsymbol{W}_D$ denote the channel dimension weight matrix, and $\delta(\cdot)$ and $\rho(\cdot)$ denote ReLU and Sigmoid activation functions. The channel attention weight $\boldsymbol{w}$ is used to adjust the feature $\boldsymbol{A}_0$ to obtain the global feature information of the channel dimension, that is,

$$\hat{\boldsymbol{g}}_c = \boldsymbol{w}_c \otimes \boldsymbol{g}_c \tag{7}$$

where $\boldsymbol{w}_c$ and $\boldsymbol{g}_c$ denote the scale factor and feature map of the *cth* channel, and $\otimes$ denotes the element-wise product. To obtain the global feature information of the spatial dimension, we first perform parallel global average pooling (GAP) and global max pooling (GMP) on feature $\boldsymbol{A}_0 \in \mathbb{R}^{H\times W\times C}$ to obtain spatial correlation. The pooling results are fused to obtain the feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times 2}$. Then, the convolution and Sigmoid function are used to obtain spatial attention feature $\hat{\boldsymbol{F}} \in \mathbb{R}^{H\times W\times 1}$, and the spatial attention feature is used to calibrate the spatial dimension of feature $\boldsymbol{A}_0$. The above can be defined as

$$\boldsymbol{F} = \text{cat}(\mathcal{F}_{GMP}(\boldsymbol{A}_0), \mathcal{F}_{GAP}(\boldsymbol{A}_0)) \tag{8}$$

$$\hat{\boldsymbol{g}}_s = \mathcal{K}_{1\times 1}(\boldsymbol{F}) \otimes \boldsymbol{A}_0 \tag{9}$$

To further obtain the global correlation of different features, inspired by Ref. [51], we construct the ENLAM module. As shown in Figure 3, for given inputs $\boldsymbol{x}$ and $\boldsymbol{y}$, the calculation process of ENLAM is as follows.

$$y_i = x_i + \alpha \boldsymbol{W}_y \frac{1}{C(x)}\sum_j f(x_i, x_j) g(x_j) \tag{10}$$

where i denotes the position index of the feature map, j denotes the index of all possible positions, $\boldsymbol{W}_y$ denotes the weight matrix, and $C(x)$ indicates the normalization factor; $g(\cdot)$ denotes embedding function, $f(\cdot)$ denotes pair operation function, and α is the adaptive attention weight map for training learning. The Embedded Gaussian function [52] is used as $f(\cdot)$ to calculate the correlation between the *ith* position and other possible positions.

3.2. Robust Edge Feature Extractor

For the backdoor attack, it is significantly challenging for an attacker to make a specific edge pixel appear/disappear by reversing the magnitude of image gradient with only limited adversarial budget per pixel [28]. Therefore, the edge feature information has better robust attributes. Since the Transformer [53] has more flexible ability to model unstructured data, to solve the problem of robust edge feature extraction in aerial images, we construct a robust edge feature extractor (RobEF) composed of six Transformer blocks.

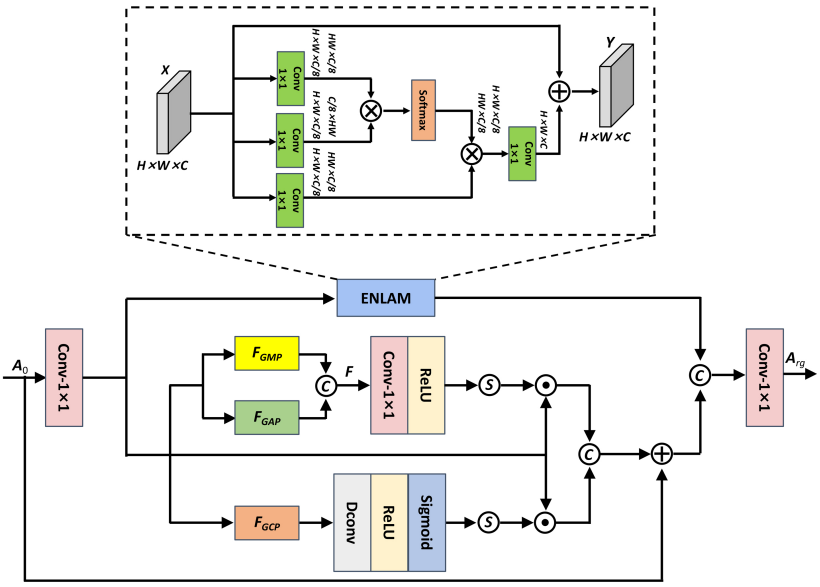


Figure 3. The detail structure of robust global feature extractor (RobGF).

As shown in Figure 4, RobEF first serializes the input features, and sequentially inputs the sequence data into the encoder–decoder structure consisting of different Transformer blocks to obtain the hierarchical edge features. The Tokens-to-Token (T2T) and Reverse-Tokens-to-Token (RT2T) [54] in the Transformer block are used to control the sequence length to obtain more multi-scale edge features in the encoder–decoder process. The skip connection operation in the encoder–decoder structure can further enrich the edge feature representation. In the final stage of edge feature extraction using RobEF, the sequence features output by different Transformer blocks are rearranged to obtain the predicted hierarchical edge feature information. Formally, given the input feature $\boldsymbol{A}_0 \in \mathbb{R}^{H\times W\times C}$, it is first serialized as $\boldsymbol{A}_0^s \in \mathbb{R}^{WH\times C}$, and the edge feature is calculated as

$$\boldsymbol{F}_1 = \boldsymbol{T}_1(\boldsymbol{A}_0^s); \boldsymbol{A}_{re} = \rho(\boldsymbol{T}_6) \tag{11}$$

$$\boldsymbol{F}_i = \boldsymbol{T}_i(\boldsymbol{F}_{i-1}), i = \{2,3\} \tag{12}$$

$$\boldsymbol{F}_i = \boldsymbol{T}_i[\boldsymbol{F}_{i-1}, \boldsymbol{F}_{7-i}], i = \{4,5,6\} \tag{13}$$

where $\boldsymbol{A}_{re} \in \mathbb{R}^{H\times W\times C}$ denotes the edge feature extraction result of sequence reconstruction, ρ denotes the Sigmoid function, $\boldsymbol{T}_i$ indicates the *ith* Transformer block, and $[\cdot]$ is the feature superposition function. In addition, we use the hierarchical learning strategy to guide the deep Transformer block ($i = \{3,4,5,6\}$) to extract salient edge feature information. The boundary extraction of different levels is calculated as

$$\boldsymbol{B}_i = \sigma(\boldsymbol{L}_i(\boldsymbol{F}_i)) \tag{14}$$

where $\boldsymbol{B}_i \in \mathbb{R}^{H_i\times W_i\times 1}$ is the boundary feature obtained by the corresponding *ith* Transformer block; $\boldsymbol{L}_i$ denotes the linear mapping function, which maps the embedded channel dimension to a single channel; $\boldsymbol{F}_i$ denotes the feature information extracted from the corresponding Transformer block. As shown in Figure 4, for different levels of Transformer blocks, it is composed of multi-head attention mechanism and self-attention mechanism. The multi-head attention is a typical style of self-attention, which aims to better obtain the correlation between features. The self-attention mechanism is calculated as

$$Q = \boldsymbol{F}\boldsymbol{W}_Q; K = \boldsymbol{F}\boldsymbol{W}_K; V = \boldsymbol{F}\boldsymbol{W}_V \tag{15}$$

$$\mathcal{F}_{SA}(\boldsymbol{F}) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{16}$$

where Q, K, and V denote the calculation results of using different mapping functions for input sequence $\boldsymbol{F} \in \mathbb{R}^{l \times d}$ by self-attention $\mathcal{F}_{SA}(\cdot)$, respectively; $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ denote the learnable weight matrices, and d_k is the number of feature channels K. To realize the parallel computing of multiple attention mechanisms, the multi-head attention uses the superposition operation to fuse the features obtained by different attention units, that is,

$$\mathcal{F}_{MSA}(\boldsymbol{F}) = \left[\mathcal{F}_{SA_1}(\boldsymbol{F}), \mathcal{F}_{SA_2}(\boldsymbol{F}), ..., \mathcal{F}_{SA_m}(\boldsymbol{F})\right] \tag{17}$$

where $\mathcal{F}_{MSA}$ denotes the multi-head attention mechanism, and m is the number of self-attention mechanism. In addition, the Transformer block uses the layer normalization function and the multi-layer perception (MLP) [55] to obtain the fused feature information. The calculation process is as follows.

$$\tilde{\boldsymbol{F}}_i = \mathcal{F}_{MSA}(\mathcal{F}_{LN}(\boldsymbol{F}_{i-1})) + \boldsymbol{F}_{i-1} \tag{18}$$

$$\boldsymbol{F}_i = \mathcal{F}_{MLP}\left(\mathcal{F}_{LN}\left(\tilde{\boldsymbol{F}}_i\right)\right) + \tilde{\boldsymbol{F}}_i \tag{19}$$

where $\mathcal{F}_{LN}$ denotes the layer normalization function, and $\boldsymbol{F}_i \in \mathbb{R}^{l_i \times d_i}$ indicates the hierarchical feature corresponding to the *ith* Transformer block. To reduce the loss of edge feature information, we introduce T2T and RT2T operations in RobEF. Specifically, given the input sequence of T2T as $\boldsymbol{F} \in \mathbb{R}^{l \times d}$, the feature sequence is first reconstructed as $\boldsymbol{I} \in \mathbb{R}^{h \times w \times d}$, then set the $k \times k$ sliding window with stride s to splice the elements in each sliding window of the feature sequence $\boldsymbol{I}$. Moreover, the sliding window operation is traversed on the feature sequence $\boldsymbol{I}$ to generate a new feature sequence $\boldsymbol{F}_r \in \mathbb{R}^{l_r \times dk^2}$, and the sequence length l_r is calculated as follows.

$$l_r = h_r w_r = \lfloor \frac{h + 2p - k}{k - s} + 1 \rfloor \lfloor \frac{w + 2p - k}{k - s} + 1 \rfloor \tag{20}$$

where $\lfloor \cdot \rfloor$ denotes the down-integer function, and p indicates the filling size beyond the boundary. For the RT2T operation, it increases the feature sequence length by splitting and sorting the channel dimensions to provide smooth upsampling calculation for the encoder–decoder process. Given $\boldsymbol{F}_r^{'} \in \mathbb{R}^{l_r \times d_r}$ is the input feature sequence of RT2T, where d_r is the number of output sequence channels of the Transformer block, and l_r is the length of feature sequence. The number of channels of $\boldsymbol{F}_r^{'}$ is first increased to $d_r k^2$ by linear mapping and reconstructed as feature $\boldsymbol{I}_r^{'} \in \mathbb{R}^{h_r \times w_r \times d_r k^2}$. Then, the same sliding window operation as T2T is used to expand each element within feature $\boldsymbol{I}_r^{'}$ in the channel dimension to the corresponding k^2 positions within the window to obtain feature $\boldsymbol{I}_o \in \mathbb{R}^{h_o \times w_o \times d_r}$. Furthermore, feature $\boldsymbol{I}_o$ is reconstructed into feature sequence $\boldsymbol{F}_o \in \mathbb{R}^{l_o \times d_r}$, where $l_o = h_o w_o$. The calculation of h_o and w_o is as follows.

$$h_o = (h_r - 1)(k - s) - 2p + k \tag{21}$$

$$w_o = (w_r - 1)(k - s) - 2p + k \tag{22}$$

The RT2T uses an accumulation strategy, which effectively avoids the loss of overlapping edge feature information by accumulating the values of feature overlap positions.

3.3. Benign Sample Generator

The purpose of constructing the benign sample generator is to use the obtained robust global features and robust edge features to generate aerial images without backdoor triggers. Inspired by CycleGAN [56], we construct the benign sample generator model with encoder–decoder structure.

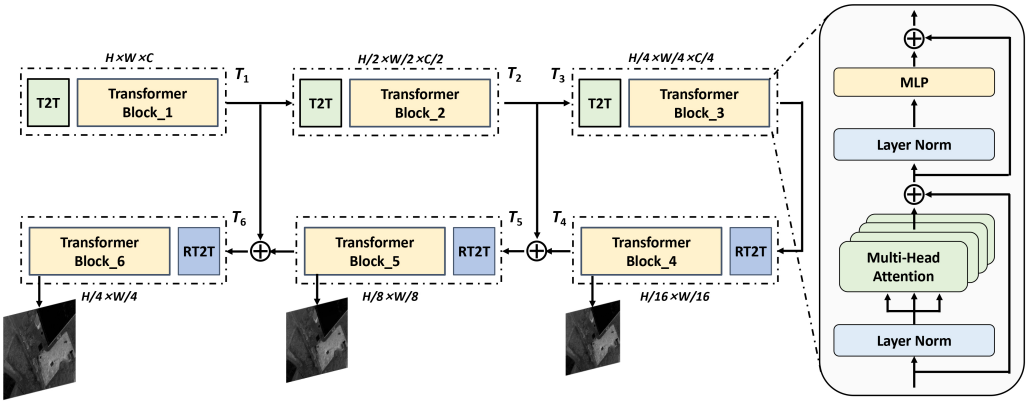


Figure 4. The detail structure of robust edge feature extractor (RobEF).

As shown in Figure 5, the generator model consists of encoder, decoder, and converter. The encoder uses multiple non-linear mapping units to mine multi-scale feature information, the converter consists of gated convolution [57] and residual blocks [58], and the decoder uses deconvolution to restore the feature map resolution. In addition, the skip connection operation is introduced into the encoder and decoder structure to enhance the semantic representation of robust robust edge and global features. The non-linear mapping unit in the encoder structure includes convolution, batch normalization (BN), and activation functions. The use of non-linear mapping units to construct encoders can enhance the feature extraction ability of the model and reduce feature information redundancy. Formally, assume that $g\boldsymbol{F}(x)$ is the non-linear mapping unit, for the input feature x, the calculation is as

$$g\boldsymbol{F}_j(x) = \max\left\{0, BN_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}\left[\boldsymbol{W}_j * g\boldsymbol{F}_{j-1}(x) + b_j\right]\right\}, 0 < j \leqslant 2 \tag{23}$$

where $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ denote batch normalization reconstruction parameters, $\boldsymbol{W}_j$ denotes weight parameter matrix, $*$ denotes convolution operation, and b_j indicates bias vector. For the skip connection operation of the encoder–decoder structure, let $G_0(x) = \boldsymbol{dim}$ is the input information of the generator and $G_i(x)$ is the output of the encoder, the calculation is defined as

$$\mathcal{F}_{skip}(x) = g\boldsymbol{F}_{i,2}[g\boldsymbol{F}_{i,1}(\boldsymbol{G}_{i-1}(x))] \tag{24}$$

$$G_i(x) = \mathcal{F}_{p\_\max}\Big(\mathcal{F}_{skip}(x)\Big), 0 < i \leqslant 3 \tag{25}$$

where $\mathcal{F}_{skip}(\cdot)$ denotes skip connection operation, and $\mathcal{F}_{p\_\max}(\cdot)$ denotes max pooling function. For the decoder structure, the feature map output from the converter requires to be reconstruction before the deconvolution operation, then fused with the feature map transmitted by the skip connection, and the specific calculation is as

$$\mathcal{F}_{up}(x) = \mathrm{cat}\Big(\mathcal{F}_{resize}(x), \mathcal{F}_{skip}(x)\Big) \tag{26}$$

$$G_i(x) = g\boldsymbol{F}_{i,2}\{g\boldsymbol{F}_{i,1}[\mathcal{F}_{deconv}(G_{i-1}(x))]\}, 4 < i \leqslant 7 \tag{27}$$

$$G(\boldsymbol{dim}) = g\_out(x) = G_8(x) = \frac{1}{1 + \mathrm{e}^{-[\boldsymbol{W}_8 * G_7(\boldsymbol{x}) + \boldsymbol{b}_8]}} \tag{28}$$

where $\mathcal{F}_{UP}(\cdot)$ denotes upsampling function, $\mathcal{F}_{deconv}(\cdot)$ denotes the deconvolution operation, and the last layer output of the decoder is shown in Equation (28). To better obtain the effective pixel position information from the multi-scale features obtained by the encoder, we introduce gated convolution in the converter, which is defined as follows.

$$\mathcal{F}_{gated}(y, x) = \sum\sum \boldsymbol{W}_g \cdot I \tag{29}$$

$$\mathcal{F}_{feature}(y,x) = \sum\sum \boldsymbol{W}_f \cdot I \tag{30}$$

$$O_{y,x} = \delta\Big(\mathcal{F}_{feature}(y,x)\Big) \odot \rho\Big(\mathcal{F}_{gated}(y,x)\Big) \tag{31}$$

where $\delta(\cdot)$ and $\rho(\cdot)$ denote ReLU and Sigmoid activation functions, and $\boldsymbol{W}_g$ and $\boldsymbol{W}_f$ denote different linear convolution filters. In the specific pixel effective position selection process, gated convolution, and Sigmoid function are used for dynamic feature selection, feature convolution and ReLU activation function are used for dynamic feature extraction.

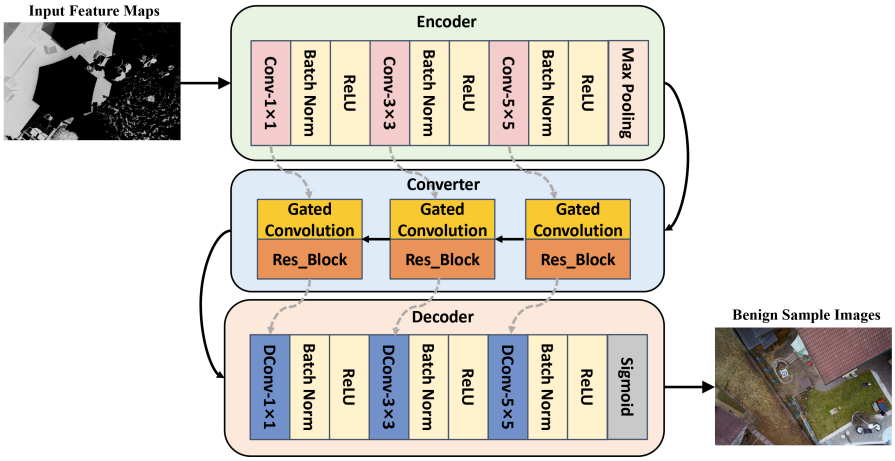


Figure 5. The detail structure of begine sample generator.

3.4. Discriminator

The constructed discriminator is used for semantic segmentation of aerial images output by the benign sample generator. Due to the high real-time requirement of aerial image semantic segmentation, we construct a lightweight and faster semantic segmentation model LF-UNet as the discriminator framework.

Similar to the UNet [59] structure, LF-UNet considers the complex characteristics of aerial image scene transformation. It only uses a small amount of downsampling operation in the encoder stage to retain the important feature information of the ground target and reduce the computational parameters and model complexity. Since the feature map after the downsampling operation contains rich semantic feature information, it is input into the constructed atrous pyramid pooling module (ASPP) [60], and multiple feature maps containing multi-scale receptive fields are generated by using atrous convolution with different dilated coefficients. The use of dilated convolution operations can not only effectively obtain multi-scale feature information but also reduce the calculation parameters of traditional convolution operations. In each upsampling stage of the decoder, we reuse the low-level semantic information by fusing the feature maps of the encoder, and embed a lightweight efficient channel attention mechanism (ECA) in each upsampling layer to enhance the extraction of valuable feature information and suppress redundant feature interference. To capture the context information in the aerial image and reduce the computational complexity, we use the ASPP module for each downsampling stage in the encoder. The ASPP module consists of 3×3 convolution, 1×1 convolution, and global average pooling. Formally, the input image is downsampled to obtain the feature map $x(i)$ containing rich semantic information, where i denotes the pixels in the feature map. The specific calculation process of ASPP is as follows.

$$y_i(i) = \sum_{k=1}^{K} x(i + r_i \times k)\omega(k), i = 1,2,3 \tag{32}$$

$$y_4(i) = \sum_{k=1}^{K} x(i + k)\omega(k) \tag{33}$$

where r denotes the dilated coefficient, and y_4 is the feature map obtained by 1×1 convolution. Since the use of large dilated coefficient will lose local feature information, and smaller dilated coefficient will limit the receptive field range, we set the dilated coefficient as $r_1 = 12$, $r_2 = 24$, and $r_3 = 36$. To establish the global correlation of different pixels, the global average pooling is performed on feature $x(i)$, and the results are input into 1×1 convolution to reduce the number of feature channels, and the upsampling operation is used to restore the feature map to the same size as feature $x(i)$ to obtain feature y_5. Finally, ASPP uses cat$(\cdot)$ function to fuse different features, and the calculation is as follows.

$$\mathcal{F}_{ASPP} = \text{cat}\left(\bigcup_{k=1}^{5} y_k(i)\right) \tag{34}$$

To improve the accuracy of semantic segmentation and reduce the model computational complexity, we embed the ECA module in the decoder. As shown in Figure 6, ECA first performs the global average pooling on the input feature map, i.e., the pixel average value of each feature channel is calculated to obtain the global features on the channel dimension. Then, the feature map $g(u)$ is calculated using one-dimensional convolution with the size of $1 \times k$ to obtain the correlation between different channels to obtain the appropriate weight distribution of different channels, and the Sigmoid function is used to normalize the weights. The specific calculation is as follows.

$$g(u) = \mathcal{F}_{p\_avg}(u) \tag{35}$$

$$\omega = \rho\left(\mathcal{F}_d^k(g(u))\right) \tag{36}$$

where $\mathcal{F}_d(\cdot)$ denotes one-dimensional convolution, and k indicates convolution kernel size. The obtained weight ω is multiplied with each channel of feature $u(i)$ to output the final feature map, i.e., $Z = \omega \otimes u$, where ω denotes the channel weight obtained by learning, and Z is the attention weight feature map.

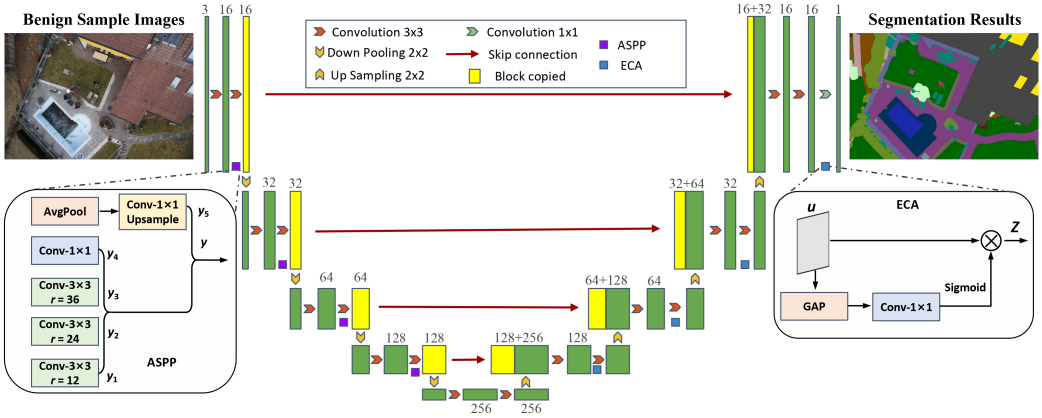


Figure 6. The detail structure of discriminator.

3.5. Loss Function

To improve the quality of the generated aerial image, we introduce the gradient similarity loss function $\mathcal{L}_{gs}$ into the objective function of pix2pix [61], and measure the similarity between the generated image and the original image from the brightness, contrast, and gradient structure. We use the **Sobel** operator [62], including vertical and horizontal edge operators, to calculate the gradient of the original image and the generated image. The specific gradient amplitude calculation is defined as follows,

$$G_x(i,j) = \frac{\partial x(i,j)}{\partial i} + \frac{\partial x(i,j)}{\partial j} \tag{37}$$

where $\partial x(i,j)/\partial i$ and $\partial x(i,j)/\partial i$ are the gradient components obtained by using the horizontal and vertical edge operators, respectively; i and j denotes the pixels in the horizontal and vertical directions of the image. The gradient similarity is defined as

$$V(x,g) = [l(x,g)]^{\alpha}[c(x,g)]^{\beta}[e(x,g)]^{\gamma} \tag{38}$$

$$l(x,g) = \frac{2u_x + u_g + c_1}{u_x^2 + u_g^2 + c_1} \tag{39}$$

$$c(x,g) = \frac{2\sigma_x\sigma_g + c_2}{\sigma_x^2 + \sigma_g^2 + c_2} \tag{40}$$

$$e(x,g) = \frac{2\sum_j\sum_i G_x(i,j)G_g(i,j) + c_3}{\sum_j\sum_i [G_x(i,j)]^2 + \sum_j\sum_i [G_g(i,j)]^2 + c_3} \tag{41}$$

where $l(x,g)$ denotes the brightness function, $c(x,g)$ denotes the contrast function, and $e(x,g)$ indicates the gradient structure function; u_x and u_g denote the average pixel value; σ_x and σ_g represent the standard deviation of the pixel value; c_1, c_2, and c_3 denote constants used to avoid zero denominator; α, β, and γ indicate constants used to adjust the importance of each component. According to Equations (38)–(41), the gradient similarity loss function is defined as follows,

$$\mathcal{L}_{GS}(H) = 1 - \frac{1}{N}\sum_{H=1}^{H} V(H) \tag{42}$$

where N denotes the number of pixels, and H denotes the intermediate pixel value of the pixel block. For the discriminator used to achieve the semantic segmentation task, we use the cross-entropy loss function to optimize it, which is defined as follows,

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{43}$$

$$\mathcal{L}_{RFGAN} = \mathcal{L}_{cGAN}(G,D) + \lambda_1\mathcal{L}_{L1}(G) + \lambda_2\mathcal{L}_{GS}(G) + \mathcal{L}_{CE} \tag{44}$$

where C denotes the number of target categories, y_c denotes the indicator variable (0 or 1), and p_c indicates the probability that the predicted result belongs to the cth category. The overall loss function of RFGAN is shown in Equation (44), where $\mathcal{L}_{cGAN}$ denotes the loss function of CGAN [63], G and D are used to calculate the loss of encoder and decoder structures in the proposed generator framework, respectively. The $\mathcal{L}_{L1}$ denotes the $L1$ metric loss, and λ_1 and λ_2 indicate weight coefficients, where we set $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$.

4. Experiments and Analysis

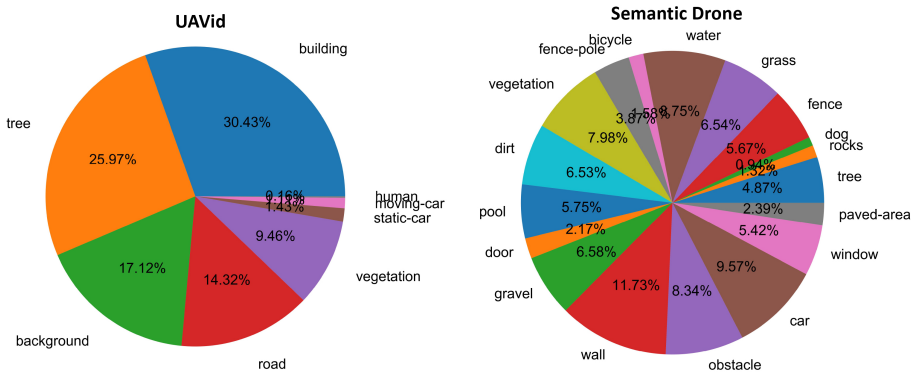
In this section, we use four backdoor attack methods to attack six aerial or RS image semantic segmentation networks trained on UAVid [64] and Semantic Drone [65] datasets. Specifically, in Section 4.1, we introduce the benchmark datasets. Section 4.2 provides detailed parameter settings for different backdoor attacks. Section 4.3 presents the experimental details. In Sections 4.4 and 4.5, we give the quantitative comparison and visual results on the defense effectiveness for backdoor attacks. Finally, the contribution of different robust features to improve the defense effect is analyzed in Section 4.6.

4.1. Dataset Information

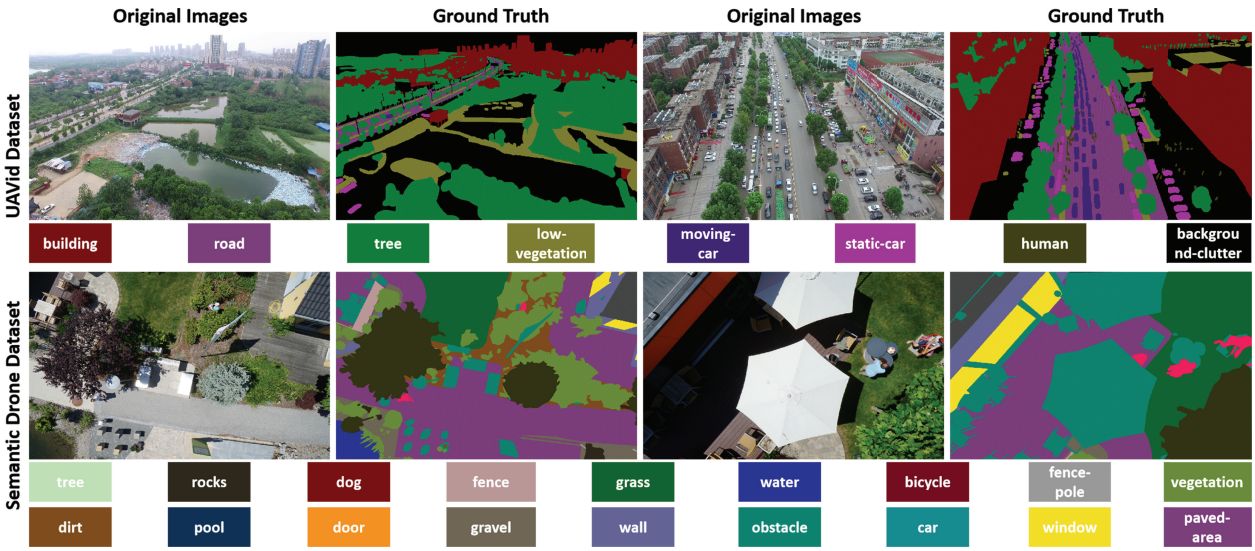
To validate the effectiveness and feasibility of the proposed method, we conduct experiments on the aerial image benchmark datasets UAVid (https://uavid.nl/ (accessed on 15 July 2020)) and Semantic Drone (http://dronedataset.icg.tugraz.at (accessed on 31 May 2020)) collected from real scenes.

UAVid: This dataset is constructed for aerial image semantic segmentation in complex urban scenes, which contains both static and dynamic objects. The UAVid uses the low-altitude aircraft with flight altitude of around 50 m to collect data and records the aerial

image in 4K resolution video mode, and the obtained image resolution is 4096 × 2160 or 3840 × 2160. To ensure scenario realism, UAVid contains multiple object categories common to urban streets, such as building, road, tree, low-vegetation, static-car, moving-car, human, and background. The pixel proportion statistics of different object categories and some sample images are shown in Figure 7. For specific experimental applications, we use 252 images from the UAVid dataset as the training set, 84 images as validation set, and the remaining 84 images as testing set. In addition, limited by the computational resources of hardware devices, we scale the image to 1024 × 512 pixels in the training process, and maintain the original image size in the inference phase.



(**a**) Pixel proportion statistical of different ground objects



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

Figure 7. Detailed analysis of the UAVid [64] and Semantic Drone [65] datasets.

Semantic Drone: The purpose of constructing this dataset is to improve the autonomous flying capability of drones in urban and suburban scenarios. Semantic Drone performs data collection at distances of around 5 to 30 m from the ground, and uses high-resolution camera in bird's-eye view mode to capture images with the resolution of 6000 × 4000. The dataset contains 18 object categories, including tree, rocks, dog, fence, grass, water, bicycle, fence-pole, vegetation, dirt, pool, door, gravel, wall, obstacle, car, window, and paved-area. It can be seen from Figure 7 that the data scene contained in Semantic Drone is more complex and has higher authenticity. The dataset provides 400 high-resolution aerial images, of which we use 240 images for training, 80 images for validation, and the remaining 80 images for testing. Similar to the preprocessing of UAVid dataset, to retain image details, we scale the original image resolution to 2048 × 1024 in the training phase, and still use the original resolution in the inference process.

4.2. Implementation Details and Evaluation Metrics

Application Details: We use Python 3.7 and Pytorch 1.10 as the programming framework to implement the proposed RFGAN model. The main hardware devices used in the experiment are i9-12900T CPU, NVIDIA GTX Geforce 3090 GPU with 24 GB memory, and ubuntu 18.04 operating system. In the model training phase, stochastic gradient descent (SGD) with the momentum as 0.9 and weight decay as 0.0001 is used as the optimizer, the training epoches are set as 1000, the batch size is set as 16, and random inversion and size cropping are used for data augmentation. The initial learning rate is set as 0.005 and the poly learning strategy is employed to automatically adjust the learning rate. In the testing process, we consider the raw outputs of each model as the evaluation results.

Evaluation Metrics: Based on the evaluation metrics pixel accuracy (**PA**) and intersection-over-union (**IoU**) commonly in semantic segmentation tasks, we adopt five metrics that facilitate quantitative comparison in backdoor attack scenarios, including benign mean intersection-over-union (**mIoU-B**), benign pixel accuracy (**PA-B**), attacked mean intersection-over-union (**mIoU-A**), attacked pixel accuracy (**PA-A**), and attack success rate (**ASR**). For the calculation of **PA** and **IoU**, we define $\mathbf{PA} = (tp + tn)/(tp + tn + ft + fn)$ and $\mathbf{IoU} = |P_i \cap G_i|/|P_i \cup G_i|$, where tp, fp, fn, and tn indicate true positives, false positives, false negatives, and true negatives, respectively; P_i and G_i denote the set of prediction pixels and ground truth for the ith class. The **mIoU-B** and **PA-B** evaluate the model performance on the benign testing set samples, while the **mIoU-A** and **PA-A** assess performance on the poisoned testing set samples. The **ASR** is defined as the percentage of misclassified pixels on the poisoned images.

4.3. Backdoor Attack Settings

To validate the defense abilities of the model against backdoor attacks, we use four backdoor attack strategies, including BadNets [32], hidden backdoor attack (HBA) [34], WaNet [66], and wavelet transform-based attack (WABA) [26] to conduct backdoor attacks for aerial image semantic segmentation tasks. According to the attack settings of BadNets [32], we define two main attack scenarios, **all-to-one** attack and **one-to-one** attack, to validate the defense capabilities of the proposed framework. For the **all-to-one** attack, the pixels of all target categories in the dataset will be labeled as specified target categories; for the **one-to-one** attack, only one target category is labeled as the specified category. For comparative analysis, we execute BadNets and HBA attacks on the UAVid dataset, and WaNet and WABA attacks on the Semantic Drone dataset. Specifically, for BadNets [32] attack, we set the target label "car" in the UAVid dataset as the attacker-specified target class, use the 8×8 pixel patch as the trigger, and set the poisoning rate $\gamma = 20\%$; for a HBA [34] attack, set the target label "human" as the specified target class, and set the poisoning rate $\gamma = 30\%$; for WaNet [66] attack, the label "grass" is set as the attacker-specified target class, and the poisoning rate γ of **all-to-one** and **one-to-one** attacks is set as 25% and 20%, respectively; for WABA [26] attack, the 32×32 pixel patch is used as the trigger, the target label "rocks" is set as the specified target category, and set the poisoning rate $\gamma = 20\%$. Note that all attack methods are implemented using the source code provided by the author.

4.4. Defense Performance Analysis on UAVid Dataset

Table 1 presents the attack quantization results of different aerial image semantic segmentation networks by BadNets [32] and HBA [34] on the UAVid dataset. In addition to the proposed RFGAN, we use five state-of-the-art CNNs-based semantic segmentation models, LANet [67], AFNet [68], MANet [69], SSAtNet [70], and HFGNet [71], to evaluate the defense performance of different models against backdoor attacks. From Figure 8a, it can be seen that on benign samples without backdoor attacks, all semantic segmentation models, including the proposed RFGAN, obtain ideal semantic segmentation accuracy, while the semantic segmentation accuracy of these models significantly decreases when performing backdoor attacks. Table 1 provides specific quantitative comparison results,

where "benign" represents the model trained on benign samples without backdoor attacks. Figure 9 shows the semantic segmentation visualization results of different models when encountering backdoor attacks. It can be analyzed from Figure 9 that when the backdoor attack is triggered, all methods except the proposed RFGAN have different degrees of pixel misclassification phenomenon. Next, we give some detailed analysis of the performance of different semantic segmentation networks under backdoor attacks.

(1) LANet [67]: The network uses patch attention mechanism and attention embedding module to enhance the feature representation of CNNs model for context and semantic information. From Table 1, it can be seen that LANet trained on benign samples can achieve relatively better accuracy, but its mIoU-A and PA-A only reach 22.71% and 31.57% when it suffers all-to-one attacks of BadNets, while under the one-to-one attack mode, its mIoU-A and PA-A are only 20.16% and 28.41%. More intuitively, as shown in Figure 9, when the BadNets attack is triggered, LANet misclassifies the object category "tree" as "low-vegetation". The experimental results show that the context or semantic information commonly used in semantic segmentation cannot resist the threat of backdoor attacks.

(2) AFNet [68]: The network constructs adaptive fusion strategy to achieve fine-grained fusion of discriminative features. Although AFNet extracts some robust features that can defend against backdoor attacks, it does not construct the specific feature extractor; therefore, it is heavily affected by backdoor attacks. As shown in Table 1, the mIoU-B of AFNet on benign samples reaches around 68%, while its mIoU-A is only 9.86% in all-to-one mode under HBA attack, and its mIoU-A is only 8.63% in one-to-one attack mode. In addition, the ASR value in Table 1 shows that both backdoor attacks can effectively perform attacks on AFNet. The visualization results in Figure 9 further show that AFNet cannot accurately complete the semantic segmentation task under the backdoor attack.

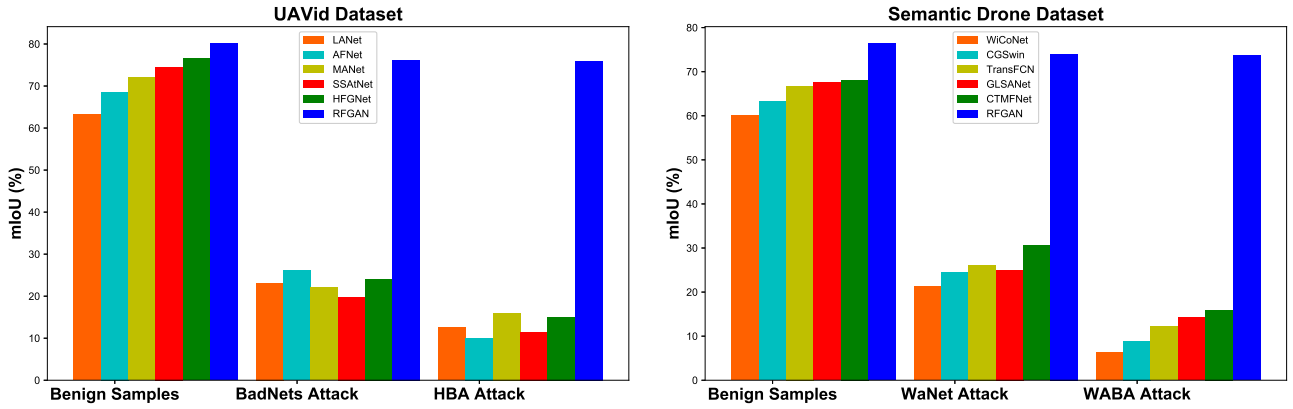
(3) MANet [69]: The network uses the encoder–decoder structure commonly used in semantic segmentation, and introduces efficient attention mechanism to achieve global information extraction and fusion. However, the results of Table 1 show that the network is still affected by backdoor attacks. Specifically, for the BadNets attack, MANet only obtained the mIoU-A of 21.53% and the PA-A of 28.96% in the all-to-one mode, while under the one-to-one attack mode, the mIoU-A and PA-A are only 18.97% and 26.14%. The results in Figure 9 show that MANet misclassifies the object category "moving-car" as "road", which further explains the impact of backdoor attacks on its performance.

(4) SSAtNet [70]: The network uses multi-scale feature enhancement module to restore fine-grained feature information. Since SSAtNet uses edge information in robust features, so it has defensive capabilities against backdoor attacks. However, as can be seen from Table 1, when performing HBA attacks on SSAtNet, its mIoU-A and PA-A are only 11.45% and 17.38% under all-to-one attack mode, while mIoU-A and PA-A are 8.62% and 11.38% under one-to-one attack mode. The visualization results in Figure 9 also illustrate that SSAtNet has been severely affected by backdoor attacks. Therefore, simply using edge features without fully mining other robust features cannot resist backdoor attacks.

(5) HFGNet [71]: To model the relationship between different feature information, the network constructs multiple feature extraction and fusion modules to achieve accurate semantic segmentation by modeling the relationship between features. However, its performance is seriously affected when it encounters BadNets and HBA attacks. As shown in Table 1, for the BadNets attack, HFGNet achieved the mIoU-A of 14.73% under all-to-one attack and mIoU-A of 12.36% under one-to-one attack, which are significantly worse than the performance on benign samples. The results further illustrate that establishing correlations between features rather than pixels does not produce defense effect.

Table 1. Comparison of the defense performance of different **CNNs-based** semantic segmentation networks against backdoor attacks on the UAVid dataset.

Model	Attack	All-to-One					One-to-One				
		mIoU-B	PA-B	mIoU-A	PA-A	ASR	mIoU-B	PA-B	mIoU-A	PA-A	ASR
LANet [67]	Benign	62.84	81.65	57.62	73.58	0	62.84	81.65	56.27	69.54	0
	BadNets	32.25	64.73	22.71	31.57	52.78	29.15	58.26	20.16	28.41	63.74
	HBA	26.73	58.24	12.57	18.62	48.86	22.73	31.57	9.75	16.32	71.58
AFNet [68]	Benign	68.94	86.51	60.46	78.35	0	68.94	86.51	61.75	82.36	0
	BadNets	34.86	65.94	25.65	33.74	62.87	31.48	38.75	21.38	29.75	83.24
	HBA	24.35	53.74	9.86	15.42	72.75	28.61	35.14	8.63	14.85	78.96
MANet [69]	Benign	72.62	87.15	63.58	79.67	0	72.62	87.15	65.73	84.45	0
	BadNets	30.68	61.72	21.53	28.96	68.51	28.94	58.82	18.97	26.14	81.73
	HBA	21.24	28.37	15.68	21.63	80.05	20.65	30.46	13.28	18.02	78.94
SSAtNet [70]	Benign	75.45	90.87	66.75	82.46	0	75.45	90.87	69.24	86.42	0
	BadNets	41.25	71.96	19.64	25.73	82.16	39.52	66.74	17.32	22.95	84.39
	HBA	23.42	31.57	11.45	17.38	78.75	20.85	28.66	8.62	11.38	79.56
HFGNet [71]	Benign	76.82	91.75	69.32	83.17	0	76.82	91.75	72.38	86.93	0
	BadNets	44.85	73.67	23.76	34.05	78.92	41.58	70.96	19.75	28.57	87.97
	HBA	25.92	35.61	14.73	22.98	88.26	22.13	32.45	12.36	19.52	81.65
RFGAN (ours)	Benign	79.89	95.81	77.57	92.34	0	79.89	95.81	75.64	88.12	0
	BadNets	78.34	94.68	76.25	89.57	5.86	77.64	92.18	77.06	91.62	7.84
	HBA	77.85	92.54	75.92	93.17	4.52	75.73	89.54	76.37	90.53	6.95



(**a**) Quantitative results of BadNets and HBA attacks (**b**) Quantitative results of WaNet and WABA attacks

Figure 8. Quantitative comparison results of benign samples and different backdoor attacks on UAVid and Semantic Drone datasets.

For the proposed RFGAN, we can see from Table 1 and Figure 9 that it can effectively resist backdoor attacks and achieve better semantic segmentation accuracy. For example, for HBA attacks with strong attack capabilities, RFGAN achieves mIoU-A and PA-A of 75.92% and 93.17% under all-to-one attack mode, while its mIoU-A and PA-A also reach 76.37% and 90.53% under one-to-one attack mode. The ASR in Table 1 also shows that the success rate of BadNets and HBA attacks on RFGAN is low, which cannot cause significant damage to the model performance. The results of Figure 9 further show that RFGAN can correctly predict the object category of different pixels. In addition, it can be observed from the Table 1 that our method increases mIoU-A and PA-A relative to the benign testing set when encountering HBA and BadNet attacks. One possible explanation for the increase in

mIoU-A and PA-A is that the backdoor attacks may have altered the feature distribution and increases the diversity of the original dataset (similar to data augmentation), so that our proposed RobGF and RobEF can obtain more discriminative features that are conducive to improving the semantic segmentation accuracy.

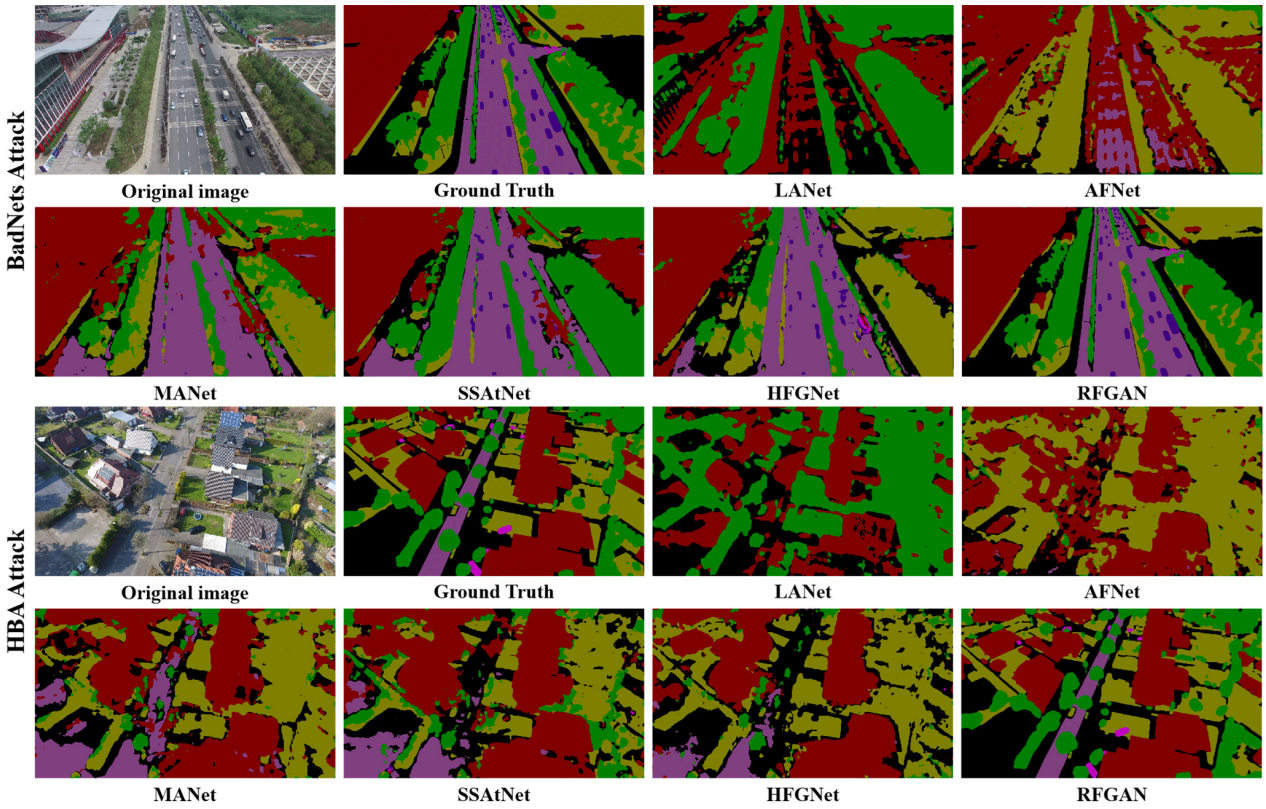


Figure 9. Semantic segmentation visualization results of different models encountering BadNets [32] and HBA [34] backdoor attacks.

4.5. Defense Performance Analysis on Semantic Drone Dataset

Recently, as a benefit from the global modeling capabilities of the Transformer [52] model, it has been widely used in aerial image semantic segmentation. To further systematically validate the impact of backdoor attacks on the existing state-of-the-art semantic segmentation network, we verify five Transformer-based aerial image semantic segmentation models on the Semantic Drone dataset. Compared with the UAVid dataset, the Semantic Drone dataset contains more ground object categories and scenarios, which can fully evaluate the threat of a backdoor attack on semantic segmentation models. The compared Transformer-based models include WiCoNet [72], CGSwin [73], TransFCN [74], GLSANet [75], and CTMFNet [76]. As shown in Figure 8b, except for the proposed RFGAN, all Transformer-based methods are affected by backdoor attacks, resulting in the significant decrease in semantic segmentation accuracy. The results in Table 2 and Figure 10 further illustrate that the Transformer cannot effectively defend against backdoor attacks. Next, we analyze the performance of Transformer-based methods when backdoor attacks are encountered.

(1) WiCoNet [72]: Based on the advantages of the Transformer model, the network constructs a Context-Transformer to obtain global features. The results in Table 2 show that WiCoNet achieves better accuracy on benign samples without backdoor attacks. However, when executing WaNet attack, the mIoU-A of WiCoNet in all-to-one attack mode is only 21.26%, while the mIoU-A in one-to-one attack mode is 18.57%, far less than the performance on benign samples. The results in Figure 10 show that when the backdoor attack is triggered, WiCoNet incorrectly predicts the category "tree" to "grass". The experimental

show that the Transformer model cannot effectively defend against backdoor attacks.
(2) CGSwin [73]: To solve the limited receptive field range of CNNs, the network uses a Transformer model to enhance global feature representation. Although CGSwin can obtain global feature that has defensive effect against backdoor attacks, the results of Table 2 show that backdoor attacks still have serious impact on model performance. As shown in Table 2, both WaNet and WABA attacks have impact on CGSwin, for example, when the WaNet attack is executed, its mIoU-A and PA-A are only 23.97% and 28.54%. From Figure 10, we can see that the backdoor attack causes CGSwin to produce serious pixel misclassification. Therefore, simply extracting global features cannot against backdoor attacks.
(3) TransFCN [74]: The network introduces a multi-scale Transformer to mine the feature correlation on spatial dimensions. As shown in Table 2, when the WaNet attack is executed, the mIoU-A and PA-A of TransFCN in all-to-one attack mode are 25.17% and 31.82%, while the mIoU-A and PA-A in one-to-one mode are only 23.72% and 28.97%. In addition, we can see from Table 2 that the attack methods WaNet and WABA have high ASR for TransFCN. The results of Figure 10 show that TransFCN incorrectly predicts the object category "water" as "grass". The experimental show that the establishment of feature correlation rather than pixel correlation cannot produce defensive effect on backdoor attack.
(4) TransFCN [75]: The network constructs the local-global attention mechanism to enhance the representation for multi-scale features. As shown in Table 2, the mIoU-A and PA-A of GLSANet on benign samples reaches 66.24% and 86.35%, while its mIoU-A and PA-A only reach 24.35% and 29.76% when the WaNet attack is executed, and the mIoU-A and PA-A reach 14.26% and 21.75% when encountering WABA attacks. The results of Figure 10 show that that GLSANet fails to achieve the ideal segmentation results and cannot classify pixels into the correct categories. The experiments show that simple extraction and fusion of local or global features cannot defend against backdoor attacks.
(5) CTMFNet [76]: The network uses semantic information as guide to enhance the global representation of feature information. As shown in Table 2, the mIoU-A and PA-A of CTMFNet in WaNet attack are 30.68% and 42.97%, while the mIoU-A and PA-A in WABA attack are only 15.97% and 22.09%. In addition, the one-to-one attack mode has greater impact on CTMFNet, take the MABA for example, the mIoU-A and PA-A are only 13.25% and 19.83%, which are significantly inferior to the performance on benign samples. The results in Figure 10 show that the backdoor attack has serious impact on CTMFNet. The experimental results show that the use of semantic information cannot defend against backdoor attacks.

Table 2. Comparison of the defense performance of different **Transformer-based** semantic segmentation networks against backdoor attacks on the Semantic Drone dataset.

Model	Attack	All-to-One					One-to-One				
		mIoU-B	PA-B	mIoU-A	PA-A	ASR	mIoU-B	PA-B	mIoU-A	PA-A	ASR
WiCoNet [72]	Benign	59.34	78.21	47.62	63.58	0	59.34	78.21	49.35	66.28	0
	WaNet	28.41	52.35	21.26	26.73	80.14	25.17	49.32	18.57	22.48	83.75
	WABA	15.37	21.46	6.24	11.58	75.38	12.63	18.75	5.79	10.64	81.26
CGSwin [73]	Benign	63.21	84.15	56.42	72.93	0	63.21	84.15	58.17	74.26	0
	WaNet	30.72	55.79	23.97	28.54	78.22	27.68	52.25	21.52	26.38	78.62
	WABA	18.65	25.76	8.95	16.76	69.17	17.27	22.34	6.45	12.37	73.41
TransFCN [74]	Benign	65.74	85.19	58.43	74.22	0	65.74	85.19	59.45	75.82	0
	WaNet	32.34	59.64	25.17	31.82	71.34	29.38	55.29	23.72	28.97	83.96
	WABA	21.78	31.25	12.18	19.83	68.54	18.52	28.74	10.25	16.58	75.37
GLSANet [75]	Benign	66.24	86.35	60.28	77.52	0	66.24	86.35	61.76	81.47	0
	WaNet	35.82	65.93	24.35	29.76	59.87	32.79	61.34	21.47	26.93	62.48
	WABA	25.39	38.67	14.26	21.75	62.75	22.06	34.75	12.78	18.64	68.93

Table 2. *Cont.*

Model	Attack	All-to-One					One-to-One				
		mIoU-B	PA-B	mIoU-A	PA-A	ASR	mIoU-B	PA-B	mIoU-A	PA-A	ASR
CTMFNet [76]	Benign	68.16	89.24	62.47	83.45	0	68.16	89.24	63.95	84.26	0
	WaNet	38.79	71.28	30.68	42.97	62.95	35.15	64.94	27.56	39.81	59.38
	WABA	26.82	41.24	15.97	22.09	69.72	22.34	36.82	13.25	19.83	68.54
RFGAN (ours)	Benign	77.31	92.86	76.24	90.54	0	77.31	92.86	77.89	93.64	0
	WaNet	75.63	88.56	74.32	87.95	3.27	74.13	86.92	72.98	84.36	2.85
	WABA	74.25	87.48	73.65	87.21	2.86	73.28	86.52	71.82	82.95	3.71

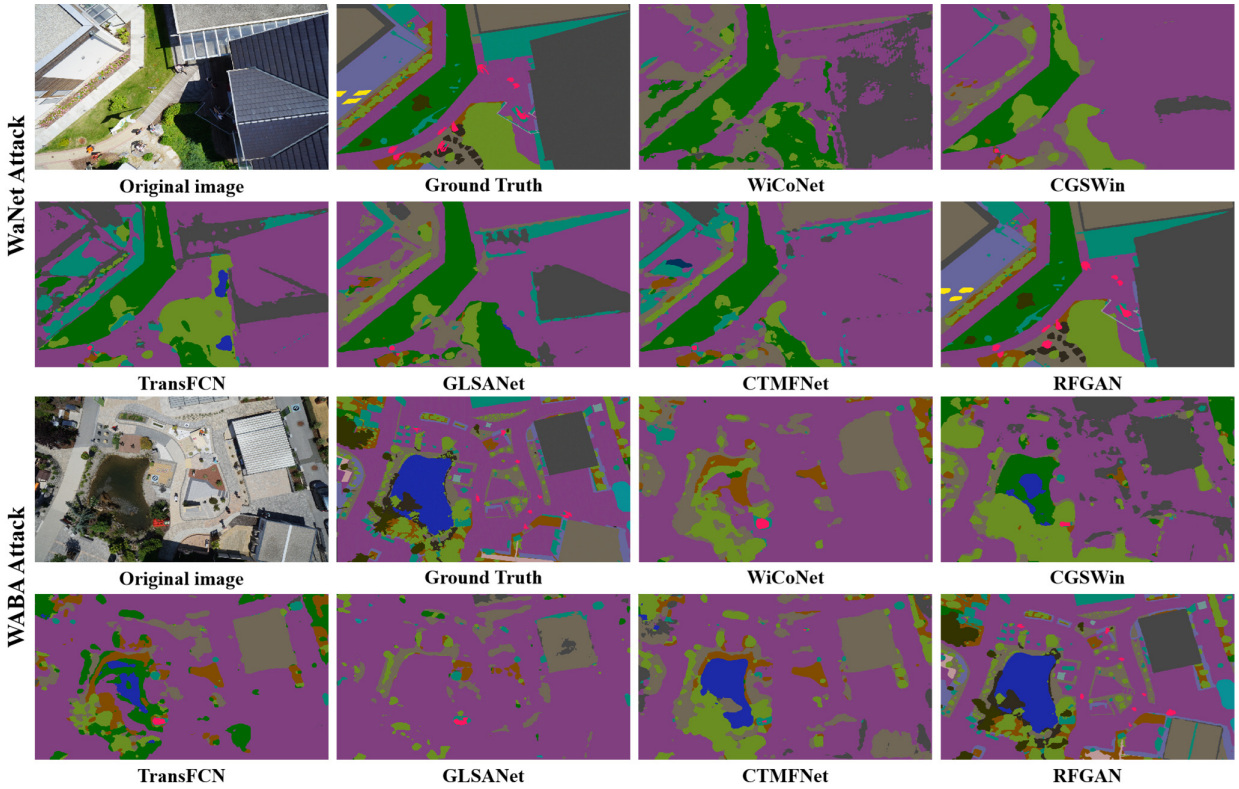


Figure 10. Semantic segmentation visualization results of different models encountering WaNet [66] and WABA [26] backdoor attacks.

It can be seen from Table 2 and Figure 10 that the proposed RFGAN can effectively defend against backdoor attacks and achieve better semantic segmentation accuracy. For example, in the all-to-one mode of a WABA attack, the mIoU-A and PA-A still reach 74.32% and 87.95%, significantly better than the Transformer-based methods. The ASR obtained by different attacks on RFGAN in Table 2 further illustrate that the proposed method can effectively resist the impact of backdoor attacks. In addition, the visualization results of Figure 10 show that RFGAN can correctly predict all pixels and has better segmentation effect on tiny objects. In summary, the experimental results on UAVid and Semantic Drone datasets show that the proposed defense framework based on robust features can effectively resist backdoor attacks and achieve better semantic segmentation accuracy, which will play an essential role in security-critical earth observation tasks.

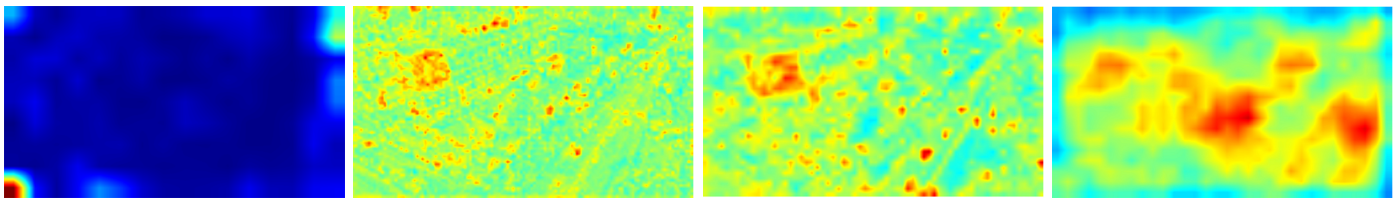
4.6. Ablation Studies

The proposed RFGAN includes robust feature extractors RobGF and RobEF. To validate the function of different robust features in defense against backdoor attacks, we

conducted ablation studies. We adopt BadNets [32] as the backdoor attack method, and set the poisoning rate $\gamma = 30\%$. In the ablation studies, we maintain the generator and discriminator architecture invariant, and gradually add RobGF and RobEF to verify the defensive performance of the robust feature extractor. The detailed results of different robust feature extractors are shown in Table 3, where GD represents the generator and discriminator structure. It can be seen from Table 3 that both RobGF and RobEF can significantly improve the resistance of GD to backdoor attacks. In addition, compared with RobGF, RobEF is more beneficial to against backdoor attacks. Take the results in the UAVid dataset for example, in all-to-one attack mode, RobGF makes the PA-B of the model reach 62.73%, while RobEF can increase the PA-B to 75.42%. It can be seen from the feature visualization results in Figure 11 that the use of RobGF and RobEF can significantly suppress the influence of poisoned samples on the model feature extraction process. For UAVid and Semantic Drone datasets, the combination of RobGF and RobEF can achieve the best results.

Table 3. Performance contribution of robust global feature extractor and robust edge feature extractor (report in PA). Best results are highlighted in **bold**.

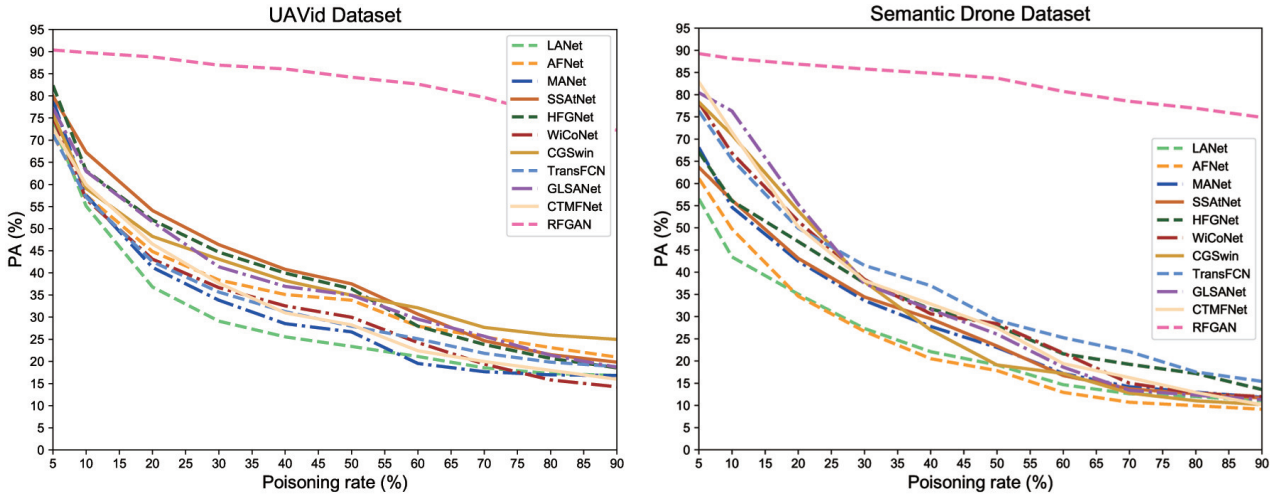
Method	All-to-One				One-to-One			
	GD	RobGF	RobEF	RobGF + RobEF	GD	RobGF	RobEF	RobGF + RobEF
GD	✓	✓	✓	✓	✓	✓	✓	✓
RobGF		✓		✓		✓		✓
RobEF			✓	✓			✓	✓
UAVid	33.82	62.73	75.42	**88.75**	28.97	59.34	72.58	**86.75**
Semantic Drone	31.56	55.65	70.38	**84.61**	26.14	57.06	69.53	**83.42**



(**a**) GD (**b**) RobGF (**c**) RobEF (**d**) RobGF+RobEF

Figure 11. Feature map visualization of different components in RFGAN under backdoor attack.

Another important issue verified by ablation studies is the impact of different levels of poisoning rates on model performance. To this end, we use BadNets to perform different levels of backdoor attacks on RFGAN, where the poisoning rate γ in the experiment is set as $\{5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$. The semantic segmentation accuracy obtained by setting different poisoning rates is shown in Figure 12. It can be observed that, as the poisoning rate increases, the PA values of all compared methods tend to decrease, indicating that setting a higher poisoning rate can cause more serious influence on the semantic segmentation network. In addition, compared with the state-of-the-art methods such as HFGNet [71] and CGSwin [73], the proposed RFGAN shows the strongest resistance to backdoor attacks on datasets UAVid and Semantic Drone under backdoor attacks with different poisoning rates. For example, when the poisoning rate γ is set as 90%, the PA value of existing methods is only 20% around, while the PA of RFGAN can still reach more than 70%, further demonstrating the effectiveness of the proposed defence framework.



(**a**) Evaluate backdoor attacks on UAVid dataset (**b**) Evaluate backdoor attacks on Semantic Drone dataset

Figure 12. The PA value obtained by different methods under the backdoor attack with different poisoning rates γ.

5. Discussion

The experimental results of Sections 4.4 and 4.5 demonstrate the effectiveness of the proposed RFGAN defense framework. Compared with the state-of-the-art semantic segmentation methods, RFGAN can resist multiple backdoor attacks. The reason is the constructed robust feature extractor RobGF and RobEF can obtain robust global features and robust edge features that have defensive effects on backdoor attacks. For the CNNs-based and Transformer-based methods compared, the features obtained by these methods are not robust, although some models use global or edge features. The reason is that these methods do not design specific robust feature extractors for backdoor attacks. Moreover, these existing methods use some feature fusion strategies, which weakens the representation of robust features. In Sections 3.1 and 3.2, we illustrate the reason why global features and edge features are robust to backdoor attacks. From the ablation study results of Section 4.6, we can observe that the introduction of robust global features and robust edge features can significantly improve the semantic segmentation accuracy of the model under a backdoor attack. After extracting robust features, RFGAN uses robust features as a guide, uses the generator architecture to generate benign samples without backdoor triggers, and uses the generated benign samples as an input of the discriminator to achieve aerial image semantic segmentation. Specifically, in Sections 3.3 and 3.4, we describe the constructed generator and discriminator. Although its structure is relatively simple, it can effectively generate benign samples and achieve competitive semantic segmentation results by using the specially designed attention mechanism.

To more comprehensively evaluate the impact of backdoor attacks on the performance of existing aerial image semantic segmentation networks, we further executed more backdoor attacks for experimentation, including BadNets [32], HBA [34], WaNet [66], and WABA [26]. For all backdoor attacks, we set the poisoning rate as 30% and perform all-to-one attack mode. The experimental results are shown in Table 4, where we can observe that the other three attack methods have stronger attack capabilities compared to the BadNets attack. Take the results in the UAVid dataset for example, the PA value of LANet is 31.47%, while the PA value under HBA, WaNet, and WABA attacks is significantly reduced to 19.37%, 28.97%, and 16.25%, respectively. Similar phenomena can be observed in other compared methods. In contrast, the proposed RFGAN can still obtain the PA more than 80% in UAVid and Semantic Drone datasets when encountering different backdoor attacks, which is significantly better than the state-of-the-art methods. In addition, from Table 4, we can seen that the PA value obtained by the RFGAN for benign samples is still reach 85%

around. These results show that the proposed defense framework can achieve competitive results on both benign and poisoned samples.

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

Method	LANet	AFNet	MANet	SSAtNet	HFGNet	WiCoNet	CGSwin	TransFCN	GLSANet	CTMFNet	RFGAN
UAVid Dataset											
Benign	80.05	85.72	86.41	88.59	89.75	88.13	90.26	89.24	91.08	90.59	**94.67**
BadNets	31.47	33.52	28.43	25.76	34.21	35.79	32.15	36.82	34.98	33.14	**85.74**
HBA	19.37	16.28	22.52	17.16	21.83	22.64	21.05	24.12	25.63	23.96	**92.28**
WaNet	28.97	23.14	24.43	25.74	23.69	24.86	23.57	25.98	24.37	26.24	**87.93**
WABA	16.25	18.71	20.38	21.53	20.41	22.56	20.03	19.85	23.64	22.87	**88.96**
Semantic Drone Dataset											
Benign	76.58	77.31	77.85	78.63	79.42	78.45	82.53	84.38	86.04	87.56	**91.25**
BadNets	26.75	28.63	25.71	22.36	24.98	28.97	29.65	31.28	30.46	27.75	**84.21**
HBA	17.32	14.57	18.95	19.06	21.42	20.64	22.73	23.75	22.96	21.37	**89.73**
WaNet	21.58	23.94	25.63	20.41	23.04	25.35	20.48	28.37	27.43	29.73	**88.54**
WABA	19.35	17.58	21.37	18.94	17.32	12.64	16.85	20.56	19.65	21.46	**87.72**

6. Conclusions

In this article, we study the backdoor attack problem in aerial image semantic segmentation and propose an efficient defense framework based on robust feature information. In order to validate and address the threat of backdoor attacks on aerial image semantic segmentation, we first systematically evaluate the impact of four popular backdoor attack methods on existing CNN-based and Transformer-based semantic segmentation networks, and construct a robust feature-guided generative adversarial network (RFGAN) to resist backdoor attacks. The constructed RFGAN consists of robust global feature extractor (RobGF), robust edge feature extractor (RobEF), generator, and discriminator. Briefly speaking, RFGAN uses robust feature extractors RobGF and RobEF to obtain robust global features and robust edge features, and uses the obtained robust features as a guide to use the generator to reconstruct the aerial image with backdoor trigger, and then uses the discriminator to achieve accurate semantic segmentation. In addition, we demonstrate that robust features can effectively resist backdoor attacks from both theoretical analysis and experimental verification. Extensive experiments on real-world aerial image datasets demonstrate that the proposed defense framework can effectively resist backdoor attacks and obtain accurate semantic segmentation results compared with the state-of-the-art aerial image semantic segmentation methods. The ablation studies further illustrate the function of robust global features and robust edge features in resisting backdoor attacks. This article is the first systematic evaluation of aerial image semantic segmentation under backdoor attacks and provides a possible solution to defend against backdoor attacks. In future work, we would attempt to construct more efficient robust feature extractors to address the backdoor threat in the aerial image semantic segmentation.

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/BackdoorsRFGAN (accessed on 7 May 2023).

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

References

1. Clabaut, É.; Lemelin, M.; Germain, M.; Bouroubi, Y.; St-Pierre, T. Model Specialization for the Use of ESRGAN on Satellite and Airborne Imagery. *Remote Sens.* **2021**, *13*, 4044. [CrossRef]
2. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sens.* **2021**, *13*, 2450. [CrossRef]
3. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 2: Literature review. *Remote Sens.* **2021**, *13*, 2591. [CrossRef]
4. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying urban functional zones by coupling remote sensing imagery and human sensing data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
5. Hamdi, Z.M.; Brandmeier, M.; Straub, C. Forest damage assessment using deep learning on high resolution remote sensing data. *Remote Sens.* **2019**, *11*, 1976. [CrossRef]
6. Shimoni, M.; Haelterman, R.; Perneel, C. Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 101–117. [CrossRef]
7. 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. [CrossRef]
8. Aldana-Martín, J.F.; García-Nieto, J.; del Mar Roldán-García, M.; Aldana-Montes, J.F. Semantic modelling of earth observation remote sensing. *Expert Syst. Appl.* **2022**, *187*, 115838. [CrossRef]
9. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef]
10. 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.
11. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial example in remote sensing image recognition. *arXiv* **2019**, arXiv:1910.13222.
12. Ai, S.; Koe, A.S.V.; Huang, T. Adversarial perturbation in remote sensing image recognition. *Appl. Soft Comput.* **2021**, *105*, 107252. [CrossRef]
13. Bai, T.; Wang, H.; Wen, B. Targeted Universal Adversarial Examples for Remote Sensing. *Remote Sens.* **2022**, *14*, 5833. [CrossRef]
14. Lu, M.; Li, Q.; Chen, L.; Li, H. Scale-adaptive adversarial patch attack for remote sensing image aircraft detection. *Remote Sens.* **2021**, *13*, 4078. [CrossRef]
15. Zhang, Y.; Zhang, Y.; Qi, J.; Bin, K.; Wen, H.; Tong, X.; Zhong, P. Adversarial Patch Attack on Multi-Scale Object Detection for UAV Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5298. [CrossRef]
16. Xu, Y.; Ghamisi, P. Universal adversarial examples in remote sensing: Methodology and benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
17. 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]
18. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *11*, 1–18. [CrossRef]
19. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE symposium on security and privacy (SP), San Jose, CA, USA, 22–26 May 2017; IEEE: Piscataway, NJ, USA , 2017; pp. 3–18.
20. Juuti, M.; Szyller, S.; Marchal, S.; Asokan, N. PRADA: Protecting against DNN model stealing attacks. In Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, 17–19 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 512–527.
21. Li, Y.; Zhai, T.; Wu, B.; Jiang, Y.; Li, Z.; Xia, S. Rethinking the trigger of backdoor attack. *arXiv* **2020**, arXiv:2004.04692.
22. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
23. Rakin, A.S.; He, Z.; Fan, D. Tbt: Targeted neural network attack with bit trojan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13198–13207.
24. Yan, Z.; Wu, J.; Li, G.; Li, S.; Guizani, M. Deep neural backdoor in semi-supervised learning: Threats and countermeasures. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4827–4842. [CrossRef]
25. Brewer, E.; Lin, J.; Runfola, D. Susceptibility & defense of satellite image-trained convolutional networks to backdoor attacks. *Inf. Sci.* **2022**, *603*, 244–261.
26. Dräger, N.; Xu, Y.; Ghamisi, P. Backdoor Attacks for Remote Sensing Data with Wavelet Transform. *arXiv* **2022**, arXiv:2211.08044.
27. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

28. Sun, M.; Li, Z.; Xiao, C.; Qiu, H.; Kailkhura, B.; Liu, M.; Li, B. Can shape structure features improve model robustness under diverse adversarial settings? In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7526–7535.
29. He, X.; Yang, S.; Li, G.; Li, H.; Chang, H.; Yu, Y. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8417–8424.
30. 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]
31. Freitas, S.; Chen, S.T.; Wang, Z.J.; Chau, D.H. Unmask: Adversarial detection and defense through robust feature alignment. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1081–1088.
32. Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* **2019**, *7*, 47230–47244. [CrossRef]
33. 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*.
34. Li, Y.; Li, Y.; Lv, Y.; Jiang, Y.; Xia, S.T. Hidden backdoor attack against semantic segmentation models. *arXiv* **2021**, arXiv:2103.04038.
35. Chan, S.H.; Dong, Y.; Zhu, J.; Zhang, X.; Zhou, J. Baddet: Backdoor attacks on object detection. In *Proceedings of the Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27. 2022, Proceedings, Part I*; Springer: Cham, Switzerland, 2023; pp. 396–412.
36. Pan, X.; Zhang, M.; Sheng, B.; Zhu, J.; Yang, M. Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, USA, 10–12 August 2022; pp. 3611–3628.
37. Li, C.; Chen, X.; Wang, D.; Wen, S.; Ahmed, M.E.; Camtepe, S.; Xiang, Y. Backdoor attack on machine learning based android malware detectors. *IEEE Trans. Dependable Secur. Comput.* **2021**, *19*, 3357–3370. [CrossRef]
38. Li, Z.; Shi, C.; Xie, Y.; Liu, J.; Yuan, B.; Chen, Y. Practical adversarial attacks against speaker recognition systems. In Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications, Austin, TX, USA, 3 March 2020; pp. 9–14.
39. Tran, B.; Li, J.; Madry, A. Spectral signatures in backdoor attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8000–8010.
40. Chan, A.; Ong, Y.S. Poison as a cure: Detecting & neutralizing variable-sized backdoor attacks in deep neural networks. *arXiv* **2019**, arXiv:1911.08040.
41. Peri, N.; Gupta, N.; Huang, W.R.; Fowl, L.; Zhu, C.; Feizi, S.; Goldstein, T.; Dickerson, J.P. Deep k-nn defense against clean-label data poisoning attacks. In *Proceedings of the Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28. 2020, Proceedings, Part I 16*; Springer: Cham, Switzerland, 2020; pp. 55–70.
42. Liu, Y.; Lee, W.C.; Tao, G.; Ma, S.; Aafer, Y.; Zhang, X. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 1265–1282.
43. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 707–723.
44. Liu, Y.; Xie, Y.; Srivastava, A. Neural trojans. In Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD), Boston, MA, USA, 5–8 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 45–48.
45. Doan, B.G.; Abbasnejad, E.; Ranasinghe, D.C. Februus: Input purification defense against trojan attacks on deep neural network systems. In Proceedings of the Annual Computer Security Applications Conference, Honolulu, HI, USA, 7–11 December 2020; pp. 897–912.
46. Li, Y.; Zhai, T.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor attack in the physical world. *arXiv* **2021**, arXiv:2104.02361.
47. 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]
48. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
49. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6688–6697.
50. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
51. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
52. Vilnis, L.; McCallum, A. Word representations via gaussian embedding. *arXiv* **2014**, arXiv:1412.6623.
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

54. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
55. Longstaff, I.D.; Cross, J.F. A pattern recognition approach to understanding the multi-layer perception. *Pattern Recognit. Lett.* **1987**, *5*, 315–319. [CrossRef]
56. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
57. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
59. 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, October 5–9. 2015, Proceedings, Part III 18*; Springer: Cham, Switzerland, 2015; pp. 234–241.
60. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
61. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
62. Kanopoulos, N.; Vasanthavada, N.; Baker, R.L. Design of an image edge detection filter using the Sobel operator. *IEEE J. Solid-State Circuits* **1988**, *23*, 358–367. [CrossRef]
63. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
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. Chen, L.; Liu, F.; Zhao, Y.; Wang, W.; Yuan, X.; Zhu, J. Valid: A comprehensive virtual aerial image dataset. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2009–2016.
66. Nguyen, A.; Tran, A. Wanet–imperceptible warping-based backdoor attack. *arXiv* **2021**, arXiv:2102.10369.
67. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
68. 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]
69. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
70. 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]
71. 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]
72. 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.
73. Meng, X.; Yang, Y.; Wang, L.; Wang, T.; Li, R.; Zhang, C. Class-Guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
74. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
75. Hu, X.; Zhang, P.; Zhang, Q.; Yuan, F. GLSANet: Global-Local Self-Attention Network for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
76. Song, P.; Li, J.; An, Z.; Fan, H.; Fan, L. CTMFNet: CNN and Transformer Multi-scale Fusion network of Remote Sensing Urban Scene Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]