remote sensing

MDPI

Article

Defense against Adversarial Patch Attacks for Aerial Image Semantic Segmentation by Robust Feature Extraction

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

[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, Tianjin 300457, China
[3] School of Surveying and Geo-Informatics, Shandong Jianzhu University, FengMing Road, Jinan 250101, China
* Correspondence: liuyaohui20@sdjzu.edu.cn; Tel.: +86-133-8531-9533

Abstract: Deep learning (DL) models have recently been widely used in UAV aerial image semantic segmentation tasks and have achieved excellent performance. However, DL models are vulnerable to adversarial examples, which bring significant security risks to safety-critical systems. Existing research mainly focuses on solving digital attacks for aerial image semantic segmentation, but adversarial patches with physical attack attributes are more threatening than digital attacks. In this article, we systematically evaluate the threat of adversarial patches on the aerial image semantic segmentation task for the first time. To defend against adversarial patch attacks and obtain accurate semantic segmentation results, we construct a novel robust feature extraction network (RFENet). Based on the characteristics of aerial images and adversarial patches, RFENet designs a limited receptive field mechanism (LRFM), a spatial semantic enhancement module (SSEM), a boundary feature perception module (BFPM) and a global correlation encoder module (GCEM), respectively, to solve adversarial patch attacks from the DL model architecture design level. We discover that semantic features, shape features and global features contained in aerial images can significantly enhance the robustness of the DL model against patch attacks. Extensive experiments on three aerial image benchmark datasets demonstrate that the proposed RFENet has strong resistance to adversarial patch attacks compared with the existing state-of-the-art methods.

Keywords: UAV aerial images; semantic segmentation; deep learning (DL); adversarial patch attack; robustness defense

check for updates

Citation: Wang, Z.; Wang, B.; Zhang, C.; Liu, Y. Defense against Adversarial Patch Attacks for Aerial Image Semantic Segmentation by Robust Feature Extraction. *Remote Sens.* **2023**, *15*, 1690. https://doi.org/10.3390/rs15061690

Academic Editors: Yonghao Xu, Bo Du and Pedram Ghamisi

Received: 25 February 2023
Revised: 16 March 2023
Accepted: 17 March 2023
Published: 21 March 2023

1. Introduction

With the continuous development of remote sensing (RS) technology and UAV-related devices, the use of UAVs for Earth observation has been widely used in military and civilian fields [1]. Aerial image semantic segmentation aims to assign a predefined semantic label category to each pixel in the aerial image, which is applied to urban planning [2], military reconnaissance [3], geographic exploration [4], disaster assessment [5], etc. Benefiting from the excellent feature extraction and representation effect of convolution neural networks (CNNs) [6] in deep learning technology, the use of CNNs for aerial image semantic segmentation has received extensive attention.

Currently, many CNNs-based methods have achieved better results in aerial image semantic segmentation. However, recent studies have found that CNN models are highly vulnerable to adversarial samples, which mislead the model prediction results by constructing carefully designed adversarial noise. Szegedy et al. [7] first proposed the concept of adversarial examples and revealed the vulnerability of deep learning (DL) models. Since then, many researchers have conducted extensive studies on the adversarial example attacks [8–13]. Compared with the adversarial examples in the digital domain to modify the pixel value of the overall image by constructing a series of constraints [14], the physical domain adversarial examples attack misleads the DL model by modifying the fixed range of pixel values in the image. Brown et al. [15] called physical domain adversarial examples adversarial patches and demonstrated that adversarial patches are more universal and targeted attack methods for real-world physical objects. Based on the concept of adversarial patches, Karmon et al. [16] proposed the LaVAN patch attack, which has better stealth and generalization performance. Chindaudom et al. [17] attempted to construct the QR patch that misleads human intuition to achieve physical domain adversarial attacks. Since then, adversarial patches have been applied to different computer vision tasks. For example, Bai et al. [18] proposed the inconspicuous adversarial patches (IAP) attack for the image classification system, which obtains the stealth effect by improving the similarity between the patch area and the original image. Zhang et al. [19] used the adversarial training strategy to construct the misleading attention and classification (MAC) attack to destroy the pedestrian detection system. Nesti et al. [20] constructed a series of adversarial patch generation methods to evaluate the robustness of urban scene semantic segmentation models. So far, adversarial example attacks and defenses on ordinary optical images have been extensively studied, and some work [21–23] has focused on aerial image processing. However, these existing studies focus on the adversarial attack or defense of aerial image classification and object detection tasks, ignoring the threat of adversarial examples and adversarial patches to semantic segmentation. As shown in Figure 1, we use DeepLabV3 [24] as the semantic segmentation model, LaVAN [16], QR-Patch [17], and IAP [18] as the adversarial patch generation method. As can be seen from Figure 1, the adversarial patch attacks significantly impact semantic segmentation results.

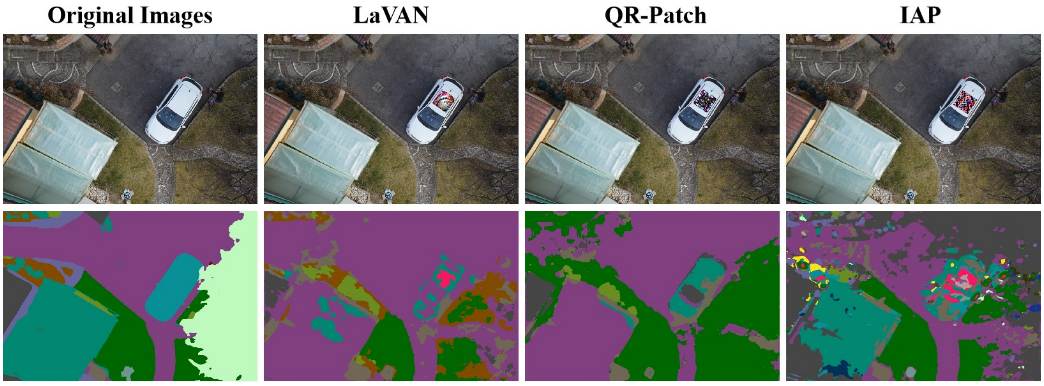

Figure 1. Visualization results of different adversarial patch generation methods on the attack effect of the semantic segmentation model.

One possible way to defend against adversarial patch attacks is to use adversarial training strategies, which extend the original training dataset by generating adversarial examples [25]. However, using adversarial training increases the model training time, and the trained model has poor generalization performance in the face of different categories of adversarial patch attacks. To effectively address the adversarial patch threat in aerial image semantic segmentation, motivated by the sensitivity of the human visual system to the inherent attribute information in an image [26,27], we attempt to enhance the defense performance against adversarial patches by modifying the DL model architecture to obtain robust features. To achieve this goal, we propose a robust feature extraction network (RFENet) to defend against adversarial patch attacks. First, the local features obtained by the small-size receptive field have been shown to resist adversarial attacks [28], so we design the limited receptive field mechanism (LRFM) for aerial image semantic segmentation to extract robust local feature information. Second, the semantic and contour information is the inherent attribute features contained in the image [27,29], and it is difficult for the adversarial patch to affect them. So, we construct the spatial semantic enhancement module (SSEM) and the boundary feature perception module (GCEM) to obtain the inherent attribute feature information in the aerial image. Third, the global feature information can establish the correlation of different pixels in the image, and the modification of local pixel areas by the adversarial patch cannot affect the global information extraction [26]. So, we design the global correlation encoder module (GCEM) to obtain the global features of aerial images. It is worth mentioning that local features, semantic features, contour features and global features also play an important role in improving semantic segmentation accuracy, so the constructed RFENet can defend adversarial patch attacks and obtain better semantic segmentation results. The contributions of this study are summarized as follows.

- To the best of our knowledge, we are the first to systematically analyze the impact of adversarial patch attacks on aerial image semantic segmentation and propose an effective defense method against targeted and un-targeted adversarial patch attacks. Our research reveals the significance of the resistability and robustness of deep learning models when addressing safety-critical Earth observation tasks.
- We comprehensively analyze the characteristics of adversarial patches in aerial image semantic segmentation. To defend against adversarial patches, a novel robust feature extraction framework is further proposed. By obtaining robust local features, semantic features, contour features and global features, RFENet can suppress the interference of adversarial patches in the feature extraction process.
- To demonstrate the performance of the proposed method, we conduct a series of experiments, including the defense capabilities of the model with multiple types of adversarial patches. The experiments conducted on three aerial image datasets containing urban and suburban show that the proposed framework can defend against adversarial patches while maintaining better semantic segmentation accuracy.

The rest of this article is organized as follows. In Section 2, the related works are briefly reviewed. Section 3 describes the proposed RFENet in detail. Section 4 presents the experimental results and discussion. Finally, the conclusion is summarized in Section 5.

2. Related Works

In this section, we review the related methods of adversarial attacks and defense in aerial images and analyze the existing robust feature extraction techniques.

2.1. Adversarial Attack and Defense for Aerial Images

With the widespread application of deep learning in aerial image processing, techniques related to adversarial attacks have gradually gained attention. The current adversarial attack methods for the aerial image processing field involve image classification, object detection and semantic segmentation tasks.

(1) **Attack on classification tasks.** Czaja et al. [30] first studied adversarial example attacks in remote sensing image classification and demonstrated that adding only weak adversarial perturbations can mislead classifier prediction results. Li et al. [23] constructed black-box and white-box attack methods for attacking SAR image classifiers. Xu et al. [25] systematically evaluated the adversarial example attack in remote sensing image classification tasks. Ai et al. [31] analyzed the influence of adversarial perturbation on aerial image classification and verified the transferability of adversarial examples in different classification models. Jiang et al. [32] proposed a project gradient descent adversarial attack (PGD) method to attack multi-source remote sensing image classifiers. Based on the adversarial training strategy, Cheng et al. [33] proposed a perturbation-seeking generative adversarial network (PSGANs) to improve the robustness of the remote scene classification model. Chen et al. [34] systematically analyzed the impact of four adversarial attack methods on multiple remote scene classification models.

(2) **Attack on detection tasks.** Compared with aerial image classification tasks, attacks against object detection are more challenging. Lian et al. [35] constructed a benchmark on adversarial patches to destroy the performance of aerial object detectors.

Lu et al. [36] proposed a scale-adaptive adversarial patch attack method for attacking aircraft object detection models, which can adaptively adjust the patch size according to the detected object. Zhang et al. [37] constructed the universal adversarial patch to attack the aerial object detection model and demonstrated its realizability in the physical domain. Du et al. [38] proposed a patch generation method for attacking vehicle object detectors in aerial scenarios, which enhances the attack efficiency by improving the similarity between the patch region and the vehicle object. Deng et al. [39] proposed an adversarial patch generation method based on style transfer theory and used multiple data augmentation methods to improve the generalization and transferability of patch attacks.

(3) **Attack on semantic segmentation tasks.** Attacks on aerial image semantic segmentation have also received attention recently. Xu et al. [21] proposed black-box mixup-attack and mixcut-attack methods for attacking the semantic segmentation model, which obtain essential regions of the original image by random cropping and use gradient optimization and momentum iteration to improve the attack effect. They further collected the generated adversarial examples dataset for researchers to design advanced adversarial defense methods in aerial scenarios. Dräger et al. [40] proposed a patch attack method based on wavelet transform to destroy semantic segmentation model performance. This method first uses wavelet transform to divide aerial images into high-frequency and low-frequency, then embeds adversarial patches in low-frequency to enhance the stealthiness of patch attacks. Since aerial image semantic segmentation involves many safety-critical tasks, its related adversarial attack methods should receive more attention.

(4) **Adversarial defense technology.** With the continuous emergence of adversarial attack methods in aerial image processing, the corresponding adversarial defense methods have also been studied. To improve the defense ability of the remote sensing image classification model in the face of unknown attacks, Cheng et al. [33] proposed a perturbation-seeking generative adversarial network (PS-GAN) defense framework. The proposed PS-GAN uses GAN to generate massive data samples and introduces a scaling search radius strategy to reduce the difference between adversarial and clean examples to achieve adversarial defense. Chen et al. [41] proposed a soft threshold defense method against various adversarial attacks on remote sensing scene classification models. The soft threshold defense method uses adversarial examples as negative samples, obtains the decision boundary with a logistic regression algorithm and uses the decision boundary to judge the confidence of each category to detect adversarial examples. To defend against adversarial patch attacks in aerial image object detection, Chen et al. [42] constructed a cascade adversarial defense framework, which located the adversarial patch region according to the high-frequency and salience information in the back gradient propagation and then used the random erasure method to suppress the adversarial patch.

2.2. Robust Features against Adversarial Attacks

Obtaining robust features contained in an image has been proven to be effective against adversarial attacks. Still, robust features need to be obtained by a carefully designed feature extractor. However, compared with adversarial training, robust features against adversarial attacks can solve security threats at the DL model design stage.

Currently, the use of robust features to resist adversarial attacks has received attention. Zhang et al. [43] first verified that using robust features to guide adversarial training can suppress the influence of adversarial noise on image classifier models. Xiao et al. [26] enhanced the model robustness by obtaining the edge feature information and improved the generalization ability of the defense model through adversarial training. Xu et al. [27] proposed a self-attention learning and global context encoder module to obtain global feature information of aerial images to resist adversarial example attacks. Lal et al. [29] constructed an adversarial defense framework for robust feature extraction and fusion, which improves the robustness of the DL model by fusing shape and semantic features. Zhang et al. [44] aligned the feature obtained from the adversarial domain with the clean sample features by the domain adaptive method and used the Wasserstein distance to reduce the difference in feature distribution. Xie et al. [45] proposed a feature denoise method to resist adversarial attacks, which reduces the interference of adversarial noise by designing multiple noise filters and modifies the adversarial features to obtain robust features. Zhou et al. [46] filtered adversarial noise interference by obtaining universal invariant features and reducing feature bias by separating robust and adversarial features in the feature encoder space. Freitas et al. [47] proposed a robust feature alignment method to defend against adversarial patch attacks, which constructs the robust DL model by mining deep semantic features. Liu et al. [48] proposed an adversarial feature distillation method to suppress adversarial noise interference, which uses the feature difference between teacher–student networks to filter adversarial noise. By constructing the unsupervised learning method, Li et al. [49] obtained the global and semantic features of object regions to enhance the adversarial robustness of the object detection model.

According to the above research, carefully designed feature extractor can obtain robust features with adversarial defense capabilities. Based on this idea, we construct a series of feature extraction methods to extract robust features against an adversarial patch attack in aerial image semantic segmentation.

3. Methodology

Since an adversarial patch attack modifies the local pixels of the original aerial image, it brings significant challenges to the aerial image semantic segmentation task. However, the small receptive field ranges [28], semantic features [27], global features [26] and boundary features [29] have been proven to be robust against patch attacks, and these features also play an important role in image semantic segmentation. Combining the robustness advantages of limited receptive field and robustness features, we propose a robust feature extraction framework, RFENet, which can be effective against adversarial patch attacks and obtain accurate semantic segmentation results. The pipeline of the proposed RFENet is shown in Figure 2, which can be divided into the encoder part and the decoder part. The encoder part extracts robust and valuable feature information, and the decoder part restores the feature map resolution and output segmentation results.

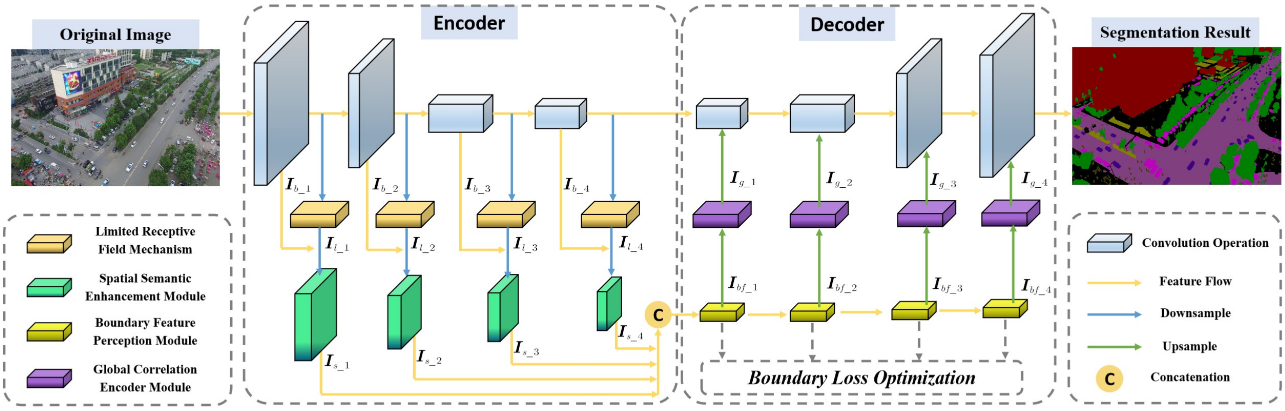

Figure 2. Illustration of the proposed RFENet. The limited receptive field mechanism is first adopted to extract local features with robustness. Then, we use the spatial semantic enhancement and boundary feature perception modules to obtain robust semantic and boundary features. The global correlation encoder module is used to build global dependency. Finally, we transmit the robust features to the decoder part to obtain the semantic segmentation results.

In the encoder part, we use pre-trained ResNet [50] as the backbone network to extract primary feature information. The LRFM consisting of conventional convolution, dilated convolution [24], and depth-wise convolution [51] is used to capture the local feature information of the backbone feature to enhance the model robustness against adversarial patches. After each LRFM, the SSEM is applied to all the feature maps, and the concatenation operation is used for feature information fusion, completing the semantic feature extraction task. In the decoder part, the BFPM composed of different scale convolution kernels is used to extract the boundary features of semantic fusion features output by SSEM and enhance the boundary region saliency representation. To establish the global correlation of different feature information and improve the model robustness, the GCEM is used to construct the relationship between different features, capture the feature long-range dependence, and input it into the backbone decoder layer to obtain accurate semantic segmentation.

3.1. Limited Receptive Field Mechanism

The receptive field reflects the specific region mapped by the convolution feature map in the image. The small-size receptive field contains local feature information, while the large-size receptive field contains global context information [52]. The use of a smaller size receptive field has been shown to enhance the robustness against adversarial patch attacks [28]. Inspired by the dual-path network (DPN) [53], we construct the LRFM robust to adversarial patches, which can effectively capture the local features contained in aerial images and suppress the adverse effects of adversarial patches.

As shown in Figure 3, LRFM contains four different branches, the first branch uses 1×1 convolution and 3×3 dilated convolution with atrous coefficient r to build the small receptive field range. The depth-wise convolution is used in the second and third branches to extract local features and ensure the local receptive field is unchanged. The depth-wise decomposes the $k \times k$ convolution kernel into $1 \times k$ and $k \times 1$ convolution kernels. The 4th branch uses 1×1 convolution to retain feature details, and the skip connection operation is used for feature transfer. In addition, the concat$(\cdot)$ and add$(\cdot)$ fusion operation are used to fuse a different branch feature map. Formally, let the input feature be $\boldsymbol{I} \in \mathbb{R}^{H \times W \times C}$; the first branch is calculated as follows,

$$\boldsymbol{B}_1 = \mathcal{K}_{3\times3,r=2}(\mathcal{K}_{1\times1}(\boldsymbol{I})) \tag{1}$$

where $\mathcal{K}_{1\times1}(\cdot)$ and $\mathcal{K}_{3\times3}(\cdot)$ are convolution with the batch normalization layer and the ReLU activation function, and $r = 2$ denotes the used atrous coefficient. The second and third branches are calculated as follows,

$$\boldsymbol{B}_2 = \mathcal{K}_{1\times1}(\mathcal{K}_{1\times k}(\mathcal{K}_{k\times1}(\boldsymbol{I}))) \tag{2}$$

$$\boldsymbol{B}_3 = \mathcal{K}_{1\times1}(\mathcal{K}_{k\times1}(\mathcal{K}_{1\times k}(\boldsymbol{I}))) \tag{3}$$

where $\mathcal{K}_{1\times k}(\cdot)$ and $\mathcal{K}_{k\times1}(\cdot)$ represent $1 \times k$ and $k \times 1$ depth-wise convolution. To maintain the small receptive field range, we set $k = 3$. The fourth branch and feature fusion operation are calculated as follows,

$$\boldsymbol{B}_4 = \mathcal{K}_{1\times1}(\boldsymbol{I}) \tag{4}$$

$$\boldsymbol{B}_{\text{cat}} = \mathcal{K}_{1\times1}(\mathcal{F}_{\text{cat}}(\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3)) \tag{5}$$

$$\boldsymbol{B}_{\text{add}} = \tau(\mathcal{F}_{\text{add}}(\boldsymbol{B}_{\text{cat}}, \boldsymbol{B}_4)) \tag{6}$$

where $\mathcal{F}_{\text{cat}}(\cdot)$ represents the feature concatenation function, $\mathcal{F}_{\text{add}}(\cdot)$ represents the feature addition fusion function, and $\tau(\cdot)$ represents the ReLU activation function.

3.2. Spatial Semantic Enhancement Module

In the calculating process of the CNNs model, limited by the fixed size of the convolution kernel, each convolution layer can only cover the area calculated by the convolution kernel [54], so the deep semantic feature contained in the aerial image is not easily obtained. In addition, the semantic feature is robust against adversarial noise [29]. To obtain the semantic features of different scales and enhance the correlation of different convolution features, we construct the SSEM as shown in Figure 4.

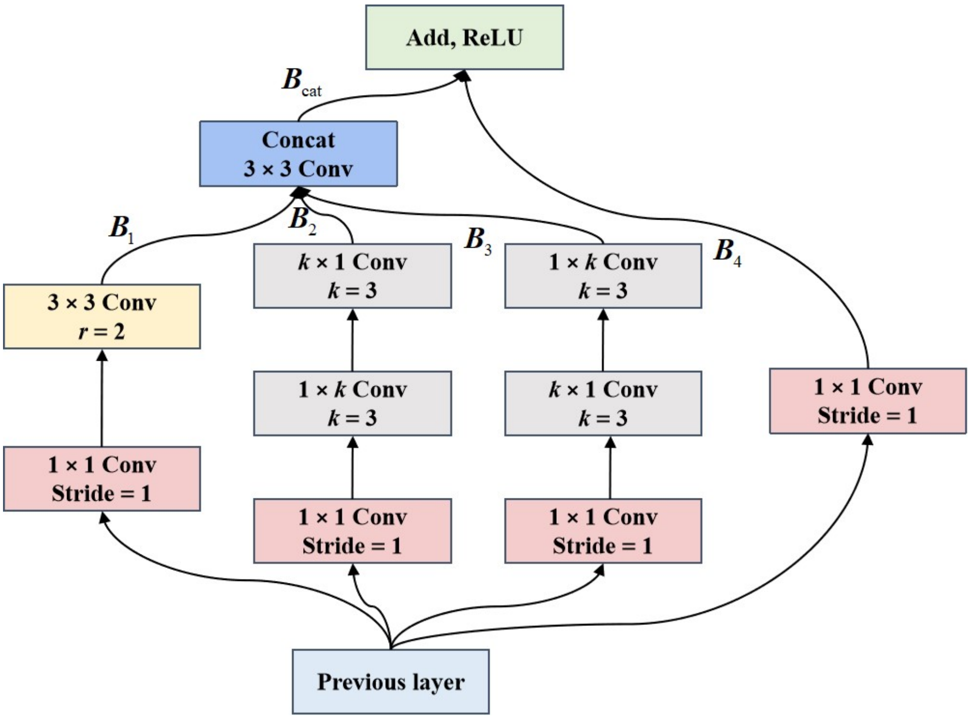

Figure 3. The architecture of our proposed LRFM.

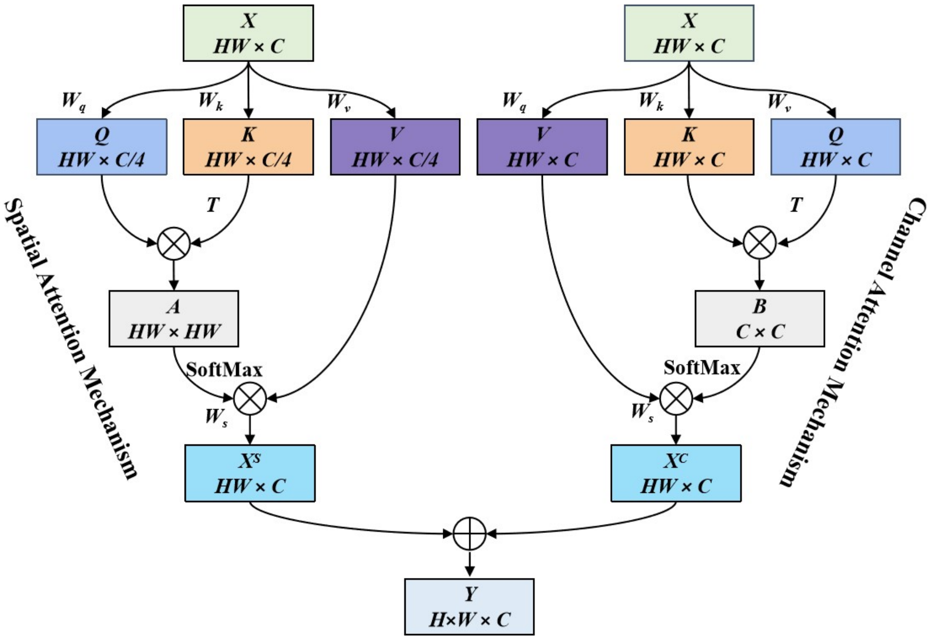

Figure 4. The architecture of our proposed SSEM.

Since backbone features contain rich texture detail information, the combination of backbone features $\boldsymbol{I}_b \in \mathbb{R}^{H\times W\times C}$ and limited field features $\boldsymbol{I}_l \in \mathbb{R}^{H\times W\times C}$ can make the texture detail and local perception information with different scales complement each other, and we input it into the SSEM to generate semantic features $\boldsymbol{I}_s \in \mathbb{R}^{H\times W\times C}$, where C

represents the number of feature channels, and H and W denote the width and height of feature maps. The specific calculation is as follows,

$$\boldsymbol{I}_s = \mathcal{F}_{\text{SSEM}}(\mathcal{K}_{1\times1}([\boldsymbol{I}_b, \boldsymbol{I}_l])) \tag{7}$$

where $\mathcal{K}_{1\times1}(\cdot)$ represents 1×1 convolution, $[\cdot]$ represents a feature splicing operation, and $\mathcal{F}_{\text{SSEM}}(\cdot)$ is the spatial semantic enhancement mechanism. Inspired by the dual attention network [55], the constructed SSEM is composed of a spatial attention mechanism and a channel attention mechanism. The spatial attention can obtain rich contextual representation, while the channel attention can enhance the relevance of semantic information. For the input feature $\boldsymbol{I} \in \mathbb{R}^{H\times W\times C}$, the spatial attention first expanded it into a one-dimensional feature vector and transposed it to obtain a two-dimensional matrix $X \in \mathbb{R}^{HW\times C}$, where C represents the number of feature channels, and H and W represent the width and height of the two-dimensional matrix feature map. Secondly, three parallel fully connected layers, W_q, W_k, and W_v, are used to reduce the channel dimension of feature maps and obtain the matrices $Q = XW_q$, $K = XW_k$, and $V = XW_v$. Thirdly, the correlation matrix is obtained by using $A = QK^T$, which can represent the vector correlation between two different spatial locations, and each row of the correlation matrix A is normalized by using the SoftMax function. Finally, the correlation matrices A and V are multiplied, the fully connected layer W_s is used to recover the channel dimension, and the spatial salience enhancement feature $X^S = AVW_s$ is obtained. The calculation process is as follows,

$$X^S = \sigma\Big(XW_q(XW_k)^T\Big)XW_vW_s \tag{8}$$

where W_q, W_k, $W_v \in \mathbb{R}^{C\times C/4}$, $W_s \in \mathbb{R}^{C/4\times C}$, and $\sigma(\cdot)$ represent the SoftMax function. Similar to spatial attention mechanism, the channel attention mechanism first expands the input feature to a one-dimensional feature vector and performs transposition to obtain $X \in \mathbb{R}^{HW\times C}$; Secondly, X is input into three fully connected layer output matrices, $Q = XW_q$, $K = XW_k$, and $V = XW_v$. Due to the dimension reduction operation causing feature loss, we cancel the channel dimension reduction. Thirdly, the correlation matrix is obtained by using $B = K^TQ$, where B_{ij} represents the inner product of the ith column in matrix K and the jth column in matrix Q, i.e., the correlation of two different channel vectors. Finally, the correlation matrix B is normalized by the SoftMax function. In addition, matrices V and B are multiplied, and the fully connected layer W_S is used to obtain the channel salience enhancement feature map $X^C = VBW_s$. The calculation process is as follows,

$$X^C = XW_v\sigma\Big((XW_k)^TXW_q\Big)W_s \tag{9}$$

where W_q, W_k, W_s, $W_s \in \mathbb{R}^{C\times C}$. The matrix addition operation is used to fuse the output of the spatial attention mechanism and channel attention mechanism and generate the final fusion feature $Y \in \mathbb{R}^{HW\times C}$. The calculation process is as follows,

$$\boldsymbol{Y} = X^S \oplus X^C \tag{10}$$

where $\oplus$ is the matrix addition operation, X^S represents the output feature of spatial attention mechanism, and X^C represents the output feature of the channel attention mechanism.

3.3. Boundary Feature Perception Module

Making full use of the boundary features of the ground object is conducive to obtaining the structure and location information of the object region so that the model can complete accurate semantic segmentation. In addition, Xiao et al. [26] demonstrated that shape and contour features are robust against adversarial noise. To obtain robust boundary feature information, we construct the BFPM based on the Gabor convolution network [56].

As shown in Figure 5, the proposed BFPM consists of 3×3 convolution, a Gabor convolution unit, and 1×1 convolution. Specifically, BFPM first uses 3×3 convolution to preprocess the robust semantic feature $\boldsymbol{Y} \in \mathbb{R}^{H \times W \times C}$ output by SSEM. Then, multiple parallel Gabor convolutions are used to enhance the boundary feature extraction ability. Finally, the 1×1 convolution is used to fuse different Gabor convolution features. The use of Gabor convolution can obtain rich scale and direction feature information while extracting fine-grained local features. In each Gabor convolution unit, the convolution kernel is replaced by Gabor filters with different directions and scales to form Gabor orientation filters (GoFs), which are calculated as follows,

$$C_{i,u}^{v} = C_{i,o} \circ G(u,v) \tag{11}$$

$$C_{i}^{v} = \{C_{i,1}^{v}, \ldots, C_{i,U}^{v}\} \tag{12}$$

where $C_{i,o}$ represents the learned filter with size of $C \times W \times W$, $C_{i,u}^{v}$ represents the Gabor orientation filter with size of $C \times W \times W$, and $G(u,v)$ represents the Gabor filter with size of $W \times W$ in the u direction at the v scale. In the proposed BFPM, GoFs are used to enhance the scale and direction information of the input feature map. The boundary feature $\hat{\boldsymbol{F}}$ can be obtained by a convolution operation between GoFs and the input feature map. The calculation process is as follows,

$$\hat{\boldsymbol{F}} = GC_{\text{conv\_1}\times 1}(C_{\text{conv\_3}\times 3}(\boldsymbol{Y}), C_{i}^{v}) \tag{13}$$

where C_{i}^{v} represents the ith GoFs with scale v. Since the boundary feature $\hat{\boldsymbol{F}}$ contains C channels, the kth feature map is calculated as follows,

$$\hat{\boldsymbol{F}}_{i,k} = \sum_{C=1}^{C} \boldsymbol{F}^{(c)} \circledast C_{i,u=k}^{(c)} \tag{14}$$

where c represents the channel of input feature $\boldsymbol{Y}$ and filter $C_{i,u}$, $\circledast$ represents the convolution operation, and k represents the number of directions in the filter.

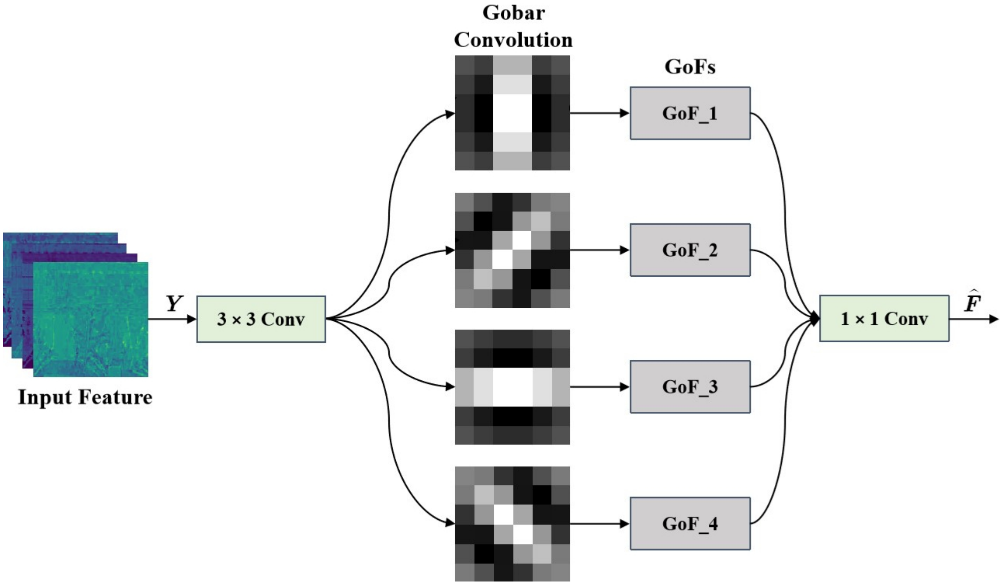

Figure 5. The architecture of our proposed BFPM.

3.4. Global Correlation Encoder

The global feature information plays a crucial function in semantic segmentation tasks, which achieves feature correlation modeling by establishing long-range dependencies of different feature information [57]. In addition, the robustness of global features enables the DL model to effectively defend against the impact of adversarial attacks [27]. To fully extract the global feature information, we construct the GCEM based on the pairwise relationship of the feature vector.

As shown in Figure 6a, the pairwise relationship between feature vector $\boldsymbol{x}_1$ and other nodes are represented as $\boldsymbol{r}_1 = (\boldsymbol{r}_{1,1}, \boldsymbol{r}_{1,2}, ..., \boldsymbol{r}_{1,N})$, and $\boldsymbol{x}_1$ is spliced with the pairwise relationship $\boldsymbol{r}_1$ to obtain the global relationship feature $\boldsymbol{y}_1 = [\boldsymbol{x}_1, \boldsymbol{r}_1]$. Similarly, different feature nodes can generate global relationship features $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$ and use it as the feature vector of global attention to calculate the global relationship attention weight $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_N)$. The learning of the pairwise feature relationship makes full use of the feature position relationship, which can better represent local location features and global scene information. Based on the pairwise feature relations, we construct the GCEM. GCEM can learn different feature nodes on the spatial dimension of the feature map and obtain global spatial structure information by representing the pairwise relationship between feature nodes. The structure of GCEM is shown in Figure 6b, and the specific implementation is as follows,

(1) For the input feature $\boldsymbol{S} \in \mathbb{R}^{L \times H \times W}$, the L dimensional feature vector of each spatial position in the feature map is used as the feature node $\boldsymbol{s}_i \in \mathbb{R}^L (i = 1, 2, ..., N, N = H \times W)$ and reconstructed into the node graph $\boldsymbol{G}_S$.

(2) The feature nodes are input into the 1×1 convolution to obtain different feature node vectors and perform dot product operations to form the spatial relationship matrix $\boldsymbol{R}_S \in \mathbb{R}^{N \times N}$ representing the relationship between each vector. The pairwise relationship between node i and node j is represented as $\boldsymbol{r}^S_{i,j}$, and the calculation is as follows,

$$\boldsymbol{r}^S_{i,j} = f_S(s_i, s_j) = \theta^T_S(s_i)\varphi_S(s_j) \tag{15}$$

$$\theta_S(s_i) = \mathrm{ReLU}(\mathrm{BN}(\boldsymbol{W}_\theta s_i)); \varphi_S(s_j) = \mathrm{ReLU}(\mathrm{BN}(\boldsymbol{W}_\theta s_j)) \tag{16}$$

where θ_S and φ_S represent the embedding functions consisting of 1×1 convolution, batch normalization (BN), and ReLU activation function; and $\boldsymbol{W}_\theta \in \mathbb{R}^{L/k \times L}$, $\boldsymbol{W}_\varphi \in \mathbb{R}^{L/k \times L}$ and k represent the dimension reduction ratio. Similarly, the pairwise relationship between node j and node i is $\boldsymbol{r}^S_{j,i} = f_S(s_j, s_i)$, using $(r^S_{i,j}, r^S_{j,i})$ to represent the bidirectional relationship between s_i and s_j.

(3) For the ith feature node, the pairwise relationships with other nodes are stacked sequentially to obtain the spatial relationship vector $\boldsymbol{r}^S_i = [\boldsymbol{R}_S(i,:), \boldsymbol{R}_S(:,i)] \in \mathbb{R}^{2N}$. The spatial relationship vector and the original feature information are spliced to obtain the spatial relationship attention $\boldsymbol{E}_S \in \mathbb{R}^{1+N/k}$ with global structure information and local detail information. The specific calculation is as follows,

$$\boldsymbol{E}_S = \left[\mathrm{pool}_C(\psi_S(s_i)), \delta_S\left(r^S_j\right)\right] \tag{17}$$

$$\psi_S(s_i) = \mathrm{ReLU}(\mathrm{BN}(\boldsymbol{W}_\psi s_i)); \delta_S\left(r^S_j\right) = \mathrm{ReLU}\left(\mathrm{BN}\left(\boldsymbol{W}_\delta r^S_j\right)\right) \tag{18}$$

where ψ_S and δ_S represent the embedding functions for stitching and fusion of original features and spatial relationship features, which consist of 1×1 convolution, BN and ReLU; and $\boldsymbol{W}_\psi \in \mathbb{R}^{L/k \times L}$, $\boldsymbol{W}_\delta \in \mathbb{R}^{2N/2k \times 2N}$, $\mathrm{pool}_C(\cdot)$ represent the global average pooling on the channel dimension.

(4) The spatial relationship attention $\boldsymbol{E}_S$ is used to calculate the attention weight $\boldsymbol{a}_i$ of each position in the feature map. The attention weight is multiplied with the original feature to obtain the spatial relation attention weighted feature. The specific calculation is as follows,

$$\boldsymbol{a}_i = \mathrm{Sigmoid}(\boldsymbol{W}_2 \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{E}_S)) \tag{19}$$

where feature weights $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are obtained by 1×1 convolution and BN function.

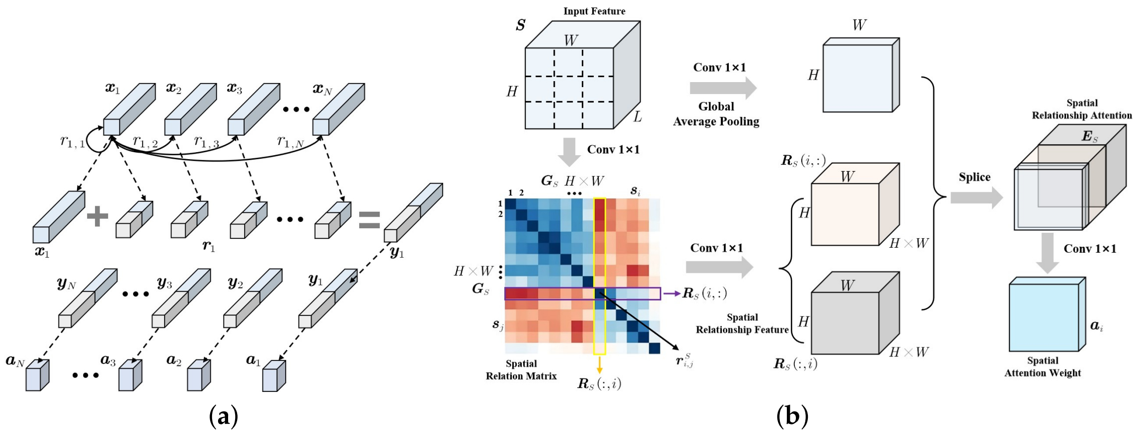

Figure 6. Interpretation of the proposed GCEM: (**a**) the pairwise relationship between feature vectors; (**b**) the structure of global correlation encoder model.

3.5. Optimization Function

In semantic segmentation tasks, the cross-entropy loss [58] is commonly used as an optimization function to predict the category of each pixel vector and then calculate the average of all pixels. Although the cross-entropy loss has achieved better results, there is a serious pixel category imbalance problem in aerial images, such as some object regions only occupy a small proportion of pixels, which makes it difficult to achieve optimal training of the proposed method. The dice loss [59] can alleviate the pixel category imbalance but ignores the boundary feature information optimization of the object region. To better achieve optimal training, we propose a hybrid optimal loss function combining dice loss and boundary loss. The hybrid loss function is defined as

$$\mathcal{L}_{\text{hybrid}} = \alpha\mathcal{L}_{\text{dice}} + \beta\mathcal{L}_{\text{boundary}} \tag{20}$$

where $\mathcal{L}_{\text{hybrid}}$ represents the hybrid loss function, $\mathcal{L}_{\text{dice}}$, $\mathcal{L}_{\text{boundary}}$ represent the boundary loss function, and α and β represent balance coefficient. Specifically, the dice loss is calculated as

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c s_i^c}{\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^{c^2} + \sum_{i=1}^{N}\sum_{c=1}^{C} s_i^{c^2}} \tag{21}$$

where i represents the pixel point, c represents the pixel category, g_i^c represents the ground truth category, and s_i^c represents the prediction category probability. The boundary loss is calculated as

$$\mathcal{L}_{\text{boundary}} = 1 - \int_{\Omega} \phi_G(\xi) s_\theta(\xi) d\xi \tag{22}$$

where ϕ_G denotes the boundary level set, and $s_\theta(\xi)$ is the prediction category probability.

4. Experiments and Analysis

In this section, we will first introduce the datasets used in this study. Then, the experimental results, analysis, and discussion are presented in detail.

4.1. Data Descriptions

To verify the effectiveness and feasibility of the proposed method, we conducted extensive experiments on three UAV aerial image benchmark datasets.

The first dataset is UAVid [60], which is captured using the high-resolution RGB video recorder, and the resolution of each image is 3840 × 2160 pixels. The UAVid dataset contains eight categories of static and dynamic ground objects, such as building, road, low-vegetation and moving-car. The second dataset is Semantic Drone [61], which contains eighteen categories of ground objects in urban scenes. Semantic Drone performs data collection at distances of around 5 to 30 m from the ground and uses a high-resolution

RGB camera to capture images with the resolution of 6000 × 4000. The third dataset is AeroScapes [62], which uses high-resolution RGB cameras to capture eleven categories of ground objects under urban and suburban scenes, and the resolution of each image is 1280 × 720 pixels. The details of the three datasets used in the experiment is shown in Table 1, and Figure 7 gives some samples and corresponding ground truth for different datasets.

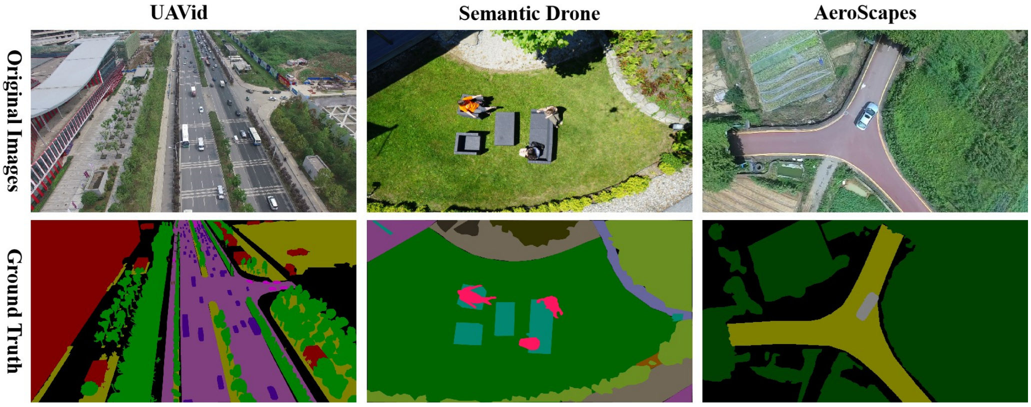

Figure 7. Example images and corresponding ground truth from the UAVid [60], Semantic Drone [61], and AeroScapes [62] datasets.

Table 1. The sample statistics information of three benchmark datasets.

Datasets	Object Category	Training	Validation	Test
UAVid	building, road, tree, low-vegetation, moving-car, static-car, human, background-clutter	200	70	150
Semantic Drone	tree, rocks, dog, fence, grass, water, bicycle, dirt, pool, door, gravel, wall, obstacle, car, vegetation, fence-pole, window, paved-area	280	40	80
AeroScape	person, bike, car, drone, obstacle, construction, road, sky, animal, boat, vegetation	2288	654	327

4.2. *Adversarial Patch Setting*

To verify the adversarial robustness of the proposed method, we select six adversarial patch generation methods to attack different aerial image semantic segmentation models. Figure 8 shows the visualization results of different adversarial patches trained on the aerial image dataset. Figure 8a is the adversarial patch generated by LaVAN [16]. This method generates image domain adversarial patches with multiple iterative optimization trainings. Figure 8b shows the QR-patch attack [17], which generates QR patches by constructing the optimal patch mask. Figure 8c shows the adversarial patch generated by IAP [18], which uses generative adversarial networks (GANs) and attention mechanisms to enhance attack performance. Figure 8d shows the adversarial patch generated by Patch-Wise [63]. This method generates patch region pixels in a pixel-by-pixel iterative manner. Figure 8e shows the DiAP patch [64], which generates the adversarial patch by the reverse model gradient. Figure 8f shows the ImageNet-Patch attack [65], which has a strong transfer performance.

The purpose of applying adversarial patch attacks in semantic segmentation tasks is to make the attacked model produce pixel misclassification problems. Patch attacks usually randomly paste the generated adversarial patches to the original image to achieve the attack effect. In this study, we scale different adversarial patches to a suitable size and paste them into the original image to complete the attack behavior.

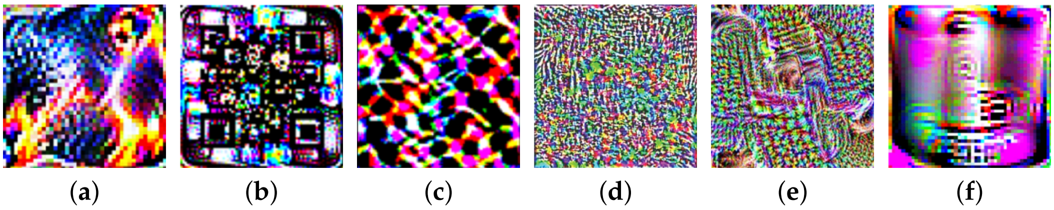

Figure 8. Visualization results for different types of adversarial patches: (**a**) LaVAN; (**b**) QR-Patch; (**c**) IAP; (**d**) Patch-Wise; (**e**) DiAP; (**f**) Image-Patch .

4.3. Implementation Details

In the experiment, we applied PyTorch and Python 3.7 to construct the proposed aerial image semantic segmentation framework. All the experiments were carried out with Intel i9-12900T CPU with 64 GB RAM, NVIDIA GTX Geforce 3090 GPU, and the Ubuntu 18.04 operating system. To ensure the credibility of the experimental results, we randomly selected images in the dataset to form the training set, validation set, and test set and repeated the experimental process 20 times. In addition, to improve the generalization ability of the model, we used data augmentation methods such as random inversion, size cropping, and brightness transformation to increase the number of dataset samples. For model optimization, the training epochs were set as 2000, the batch size was 16, and the stochastic gradient descent (SGD) with a momentum of 0.9 was used as the optimizer. The initial learning rate was set as 0.001 and the poly learning strategy was employed to automatically adjust the learning rate. For all the methods compared, we used the source code provided by the author for experiments and used the hyperparameter setting and optimization strategy consistent with the original article. Algorithm 1 gives the detailed steps to attack the proposed RFENet using adversarial patches. The goal of adversarial patch attacks on aerial image semantic segmentation is to use adversarial patches to interfere with the aerial image to maximize the misclassification of all test pixels.

Algorithm 1 Adversarial Patch Attack on RFENet

```
Input:
Aerial image $x$ and corresponding ground truth $y$.
Semantic segmentation model $f$ with parameters $\theta$.
Adversarial patch $\chi$, training epochs $\tau$, and learning rate $\eta$.
Output: The predictions on the adversarial patch images $x_{adv}$.
Initialize model parameters $\theta$ with uniform distribution.
for $t$ in $range(0, \tau)$ do
    Compute the local features $\boldsymbol{B}$ via Equations (1)–(6).
    Compute the semantic features $\boldsymbol{Y}$ via Equations (7)–(10).
    Computer the boundary features $\boldsymbol{F}$ via Equations (11)–(14).
    Computer the global features $\boldsymbol{G}$ via Equations (15)–(19).
    Computer the cross-entropy loss $\mathcal{L}$ via Equations (20).
    Update $\theta$ by descending its stochastic gradients.
end for
Generate the adversarial patch image $x_{adv}$ via Ref. [16–18,63–65].
Feed the adversarial patch image $x_{adv}$ to the model $f$ to achieve the segmentation.
```

The pixel accuracy (PA), mean pixel accuracy (mPA), F1_score, and mean intersection over union (mIoU) were utilized to quantitatively estimate different methods. Specifically, **PA** $= (\mathrm{tp} + \mathrm{tn})/(\mathrm{tp} + \mathrm{tn} + \mathrm{ft} + \mathrm{fn})$; the **mPA** is the mean of the sum of category pixel accuracy (cPA), where **cPA** $= \mathrm{tp}/(\mathrm{tp} + \mathrm{fp})$; the F1_score is the geometric mean between the precision (P) and recall (R) of each class as **F1_score** $= 2 \times (\mathbf{P} \times \mathbf{R})/(\mathbf{P} + \mathbf{R})$, where $\mathbf{P} = \mathrm{tp}/(\mathrm{tp} + \mathrm{fp})$ and $\mathbf{R} = \mathrm{tp}/(\mathrm{tp} + \mathrm{fn})$; the mIoU is defined as the mean of IoU, and the IoU is calculated as **IoU** $= |P_i \cap G_i|/|P_i \cup G_i|$, where the P_i and G_i are the set of prediction pixels and ground truth information for the *i*th class. Moreover, the tp, fp, fn, and tn are defined as true positives, false positives, false negatives and true negatives.

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

In this subsection, we give the semantic segmentation results of the proposed RFENet and other methods on the adversarial patch test set. As for the construction of the adversarial patch test set, we used LaVAN [16] and QR-Patch [17] to construct adversarial patches in the UAVid dataset; for the Semantic Drone dataset, IAP [18] and Patch-Wise [63] were used to construct the corresponding adversarial patch test set; for AeroScape dataset, DiAP [64] and ImageNet-Patch [65] are used to construct the adversarial patch test set. A brief description of each aerial image semantic segmentation method compared in the experiments is presented below.

1. BSNet [66]: It includes feature extraction and restoration stages in which the feature extraction stage uses the gradient convolution to obtain boundary features, and the feature restoration stage uses global dependencies to recover feature resolution.
2. MANet [67]: This method uses the efficient attention mechanism to extract global context dependencies, uses the linear complex kernel attention mechanism for local and global feature alignment and performs feature fusion by channel weighting.
3. AFNet [68]: This method uses the small-scale dilated convolution kernel to extract multi-scale features of different ground objects and then uses the multi-scale structure with a scale layer attention module to obtain discriminative feature information.
4. SSAtNet [69]: This method uses the pyramid attention pooling module to obtain detailed feature information, uses the pooling index to fuse local and global features and recovers the fine-grained feature information by information correction.
5. MDANet [70]: This method uses multi-scale deformable attention to capture different scale features, uses a self-attention module to establish long-range context dependence and optimizes the boundary region segmentation effect with a local embedding module.

These compared methods use techniques commonly used in aerial image semantic segmentation, such as multi-scale feature extraction, feature fusion and attention mechanism. The quantitative results are reported in Tables 2–4 from which we find that these compared methods cannot resist adversarial patch attacks, resulting in a significant decrease in semantic segmentation performance. For the UAVid dataset with LaVAN [16] and QR-Patch [17] attacks, the mAP of BSNet [66] is only 23.85% and 17.65%, while the mPA on the Semantic Drone and AeroScape datasets is only around 25%, indicating that BSNet is seriously vulnerable to adversarial patch attack. This phenomenon further demonstrates that using simple feature extraction or feature fusion operations is insufficient to defend against adversarial patch attacks. For the semantic segmentation network MANet [67], its mPA and mIoU on all three datasets are under 30%, especially the mIoU on the UAVid dataset where it is only about 17%. The results of MANet show that the existence of adversarial patches interferes with its reasoning process. Although MANet uses the attention mechanism to obtain local and global features, it still fails to gain defense abilities that are beneficial to resist adversarial patches. The reason for this problem may be that the adversarial patch affects the pixel distribution of the original image, making the semantic segmentation model. AFNet [68] with multi-scale feature extraction ability is also affected by the adversarial patch attacks, and its mPA metrics on the UAVid dataset are only 25.48% and 21.63%. Despite the fact that AFNet uses the same small-size receptive field mechanism as the proposed RFENet in the process of feature extraction, it fails to obtain the better defense effect against patch attacks. The reason for this phenomenon is that AFNet uses the dilated convolution with large dilated coefficient to enhance feature extraction ability, but this operation also increases the impact of adversarial patch attacks on model performance. For the semantic segmentation network SSAtNet [69], even though the model extracts local and global feature information that has a defensive effect against patch attacks, the performance on all three datasets still falls short of the desired results. The reason for this phenomenon is that SSAtNet pays no attention to the semantic features and boundary features that have important defensive effects against adversarial patch attacks, which expands the impact of adversarial patches on model performance. Compared with other semantic segmentation networks, MDANet [70] establishes the interdependence of global context information, which makes it have a slight advantage against patch attacks, but the performance of the model is also seriously affected by patch attacks. This phenomenon further illustrates that global dependence on contextual information alone cannot defend against adversarial patch attacks, and the extraction of other robust feature information needs to be considered in the process of constructing contextual information correlations. For the proposed RFENet, with the help of different robust feature extraction components, we can significantly improve the resistibility of the model towards adversarial patch attacks. On all three datasets, the mPA metric of RFENet is more than 80%, which dramatically outperforms the existing state-of-the-art aerial image semantic segmentation methods. For the algorithm real-time analysis, it can be observed from Tables 2–4 that the proposed RFENet still has advantages. Take the Semantic Drone dataset, for example; the single image inference time of RFENet is around 20 s, which is superior to all the compared methods. The reason is that although the proposed method uses multiple robust feature extraction components, the structure is relatively simple and does not require the model to perform complex inference processes.

To visually validate the influence of adversarial patch attacks on the prediction results of different semantic segmentation methods, we further show the corresponding semantic segmentation results visualization map in Figures 9–11. From the visualization results, it can be observed that adversarial patch attacks seriously impact the performance of different semantic segmentation networks, misleading these models into making wrong predictions. Take the visualization results of UAVid dataset for example. For the LaVAN [16] patch attack, although the annotation information corresponding to the middle region in the aerial image belongs to "road", many existing methods incorrectly predict the region as "background" or "building" and cannot correctly predict the small-size object region, such as "car" and "human". For QR-Patch [17] attacks, their impact on model performance is more serious, such as the region belonging to the "tree" being incorrectly predicted as "background" or "car". Moreover, the same phenomenon can be observed from the visualization results of Semantic Drone and AerosScape datasets. By contrast, the proposed RFENet is more robust towards adversarial patch attacks, and the semantic segmentation results of RFENet are much closer to the ground truth annotation.

Table 2. Quantitative results of the UAVid dataset on the adversarial patch (LaVAN/QR-Patch) test set, where the best results are shown in **bold**.

Class	BSNet	MANet	AFNet	SSAtNet	MDANet	RFENet
building	13.24/8.89	15.86/11.52	14.48/9.75	17.62/12.14	16.38/10.82	**85.26/84.38**
road	20.75/12.46	22.37/13.48	25.86/15.81	27.63/16.24	26.75/17.32	**87.13/86.57**
tree	23.48/14.52	25.32/15.72	28.45/17.83	30.14/18.65	31.75/18.96	**88.75/87.46**
low-vegetation	17.52/10.63	19.86/12.57	21.75/14.89	22.73/15.26	23.17/16.84	**86.24/85.23**
moving-car	8.75/4.26	12.73/10.86	15.28/13.75	17.42/14.31	18.52/14.85	**81.32/79.41**
static-car	10.63/8.15	13.78/11.43	14.23/12.64	16.17/15.85	17.62/16.35	**78.63/77.42**
human	7.15/5.48	11.73/10.62	16.31/14.86	17.35/15.24	19.13/18.76	**76.51/74.64**
background	31.25/25.36	33.87/26.13	36.55/28.70	39.48/37.26	42.68/38.97	**80.13/76.85**
PA (%)	26.17/20.78	27.15/21.75	28.60/24.65	30.40/24.96	31.42/26.03	**91.26/89.32**
mPA (%)	23.74/17.65	24.83/18.32	25.48/21.63	27.51/22.06	28.92/24.38	**88.57/86.24**
mF1 (%)	19.57/16.85	21.36/18.13	22.82/19.46	24.38/21.42	26.93/24.35	**85.42/83.27**
mIoU (%)	16.59/11.21	19.44/14.04	21.62/16.03	23.56/18.12	24.50/19.11	**82.98/81.49**
Runtime (s)	21.32/22.75	26.83/27.42	25.47/26.14	31.58/33.18	33.72/35.64	**19.84/21.45**

Table 3. Quantitative results of the Semantic Drone dataset on the adversarial patch (IAP/Patch-Wise) test set, where the best results are shown in **bold**.

Class	BSNet	MANet	AFNet	SSAtNet	MDANet	RFENet
tree	13.27/14.73	14.15/16.43	18.63/19.28	16.82/17.15	18.79/19.37	**75.82/77.95**
rocks	9.42/6.38	11.58/12.26	14.30/13.52	11.43/12.56	13.48/14.62	**70.42/71.35**
dog	7.63/8.12	9.45/11.38	11.93/12.46	10.23/9.62	15.74/16.58	**73.57/75.69**
fence	12.57/13.81	14.72/16.54	15.82/16.97	11.85/13.16	17.86/18.20	**74.98/76.24**
grass	22.71/21.64	24.52/23.83	27.38/28.14	19.54/21.32	31.78/32.26	**92.54/93.38**
water	20.14/19.58	21.65/20.46	25.62/24.35	19.57/18.62	27.98/28.42	**90.42/91.85**
bicycle	12.57/10.62	15.72/16.31	18.39/19.58	11.23/12.15	19.36/20.37	**75.87/76.35**
dirt	7.65/8.54	9.43/11.26	12.53/13.86	7.54/6.38	14.24/15.37	**69.42/70.31**
pool	17.43/18.75	19.86/20.78	26.75/25.34	16.82/15.68	30.27/31.86	**94.26/95.77**
door	2.14/3.27	4.31/5.64	7.42/6.73	3.75/5.82	10.42/12.45	**66.78/67.42**
gravel	19.43/20.63	21.59/23.86	31.86/32.97	26.93/27.42	35.98/34.26	**85.97/86.41**
wall	14.64/16.52	16.25/17.31	22.42/21.75	13.87/15.46	25.08/24.22	**72.47/73.86**
obstacle	16.78/17.88	18.64/20.93	23.96/22.87	17.94/18.53	26.93/27.51	**78.32/79.43**
car	21.65/22.15	23.70/24.56	25.78/26.14	20.65/19.37	29.82/28.13	**93.56/94.27**
vegetation	19.58/18.96	22.41/21.93	26.53/28.52	21.85/23.97	31.53/32.85	**79.32/81.36**
fence-pole	3.22/3.54	5.82/6.87	8.62/9.74	10.86/11.42	12.95/13.37	**66.87/67.12**
window	7.25/6.75	9.57/11.36	15.30/16.27	12.79/11.68	16.58/17.74	**71.43/73.84**
parved-area	21.49/22.31	23.86/25.73	25.93/27.36	24.78/23.64	38.36/38.02	**95.78/96.34**
PA (%)	29.45/31.27	31.73/32.56	36.12/37.41	33.42/34.41	42.13/44.56	**84.72/87.36**
mPA (%)	22.68/23.57	27.51/28.46	33.87/34.18	30.89/31.42	39.05/41.73	**81.87/82.35**
mF1 (%)	16.29/17.48	18.32/19.17	22.75/23.64	18.97/19.30	26.93/27.86	**79.84/81.43**
mIoU (%)	13.86/14.12	15.96/16.91	19.89/20.32	15.47/15.77	23.18/23.65	**79.32/80.50**
Runtime (s)	23.72/24.51	28.47/29.16	27.64/28.03	32.27/33.85	34.96/35.82	**20.32/21.98**

Table 4. Quantitative results of the AeroScape dataset on the adversarial patch (DiAP/ImageNet-Patch) test set, where the best results are shown in **bold**.

Class	BSNet	MANet	AFNet	SSAtNet	MDANet	RFENet
tree	12.26/14.37	15.72/17.31	18.64/20.17	21.95/23.04	25.34/27.15	**82.53/83.42**
rocks	3.42/4.86	5.16/7.23	9.41/11.84	12.72/13.84	17.85/19.21	**65.34/66.57**
dog	5.92/6.78	8.57/10.36	15.72/17.37	18.97/19.28	23.86/24.57	**73.96/74.18**
fence	4.89/5.13	7.93/8.75	14.29/16.53	17.50/18.26	20.43/21.92	**61.97/63.42**
grass	6.28/7.52	11.58/13.64	18.51/21.75	21.62/22.83	25.32/26.71	**66.28/67.74**
water	8.15/9.43	13.72/15.08	22.73/24.32	25.76/27.08	30.72/31.98	**60.92/62.31**
bicycle	17.82/18.94	23.75/25.84	29.62/31.65	32.95/34.26	37.24/38.62	**88.74/89.92**
dirt	22.45/23.71	31.84/33.91	38.54/39.41	41.85/42.32	46.31/48.05	**91.24/92.51**
pool	7.14/8.43	9.48/12.56	13.57/14.28	16.96/18.14	21.54/23.52	**62.73/64.95**
door	9.56/10.75	14.87/16.92	20.32/21.46	23.41/24.96	29.31/30.89	**83.47/85.26**
gravel	17.84/19.13	22.73/25.04	27.43/28.54	31.86/33.28	37.03/39.41	**87.62/88.43**
PA (%)	32.16/33.78	35.96/37.21	38.96/39.57	42.63/44.28	43.97/45.26	**88.96/90.08**
mPA (%)	26.47/27.92	28.56/29.75	31.62/32.74	34.57/35.94	36.57/38.21	**82.74/84.25**
mF1 (%)	22.36/24.05	24.57/25.82	26.75/27.92	30.92/32.63	32.41/34.45	**79.85/81.43**
mIoU (%)	10.52/11.73	15.03/16.96	20.79/22.48	24.14/25.20	28.63/30.18	**74.98/76.24**
Runtime (s)	25.43/27.52	31.25/33.98	30.54/31.63	38.42/39.57	39.24/40.26	**21.75/22.68**

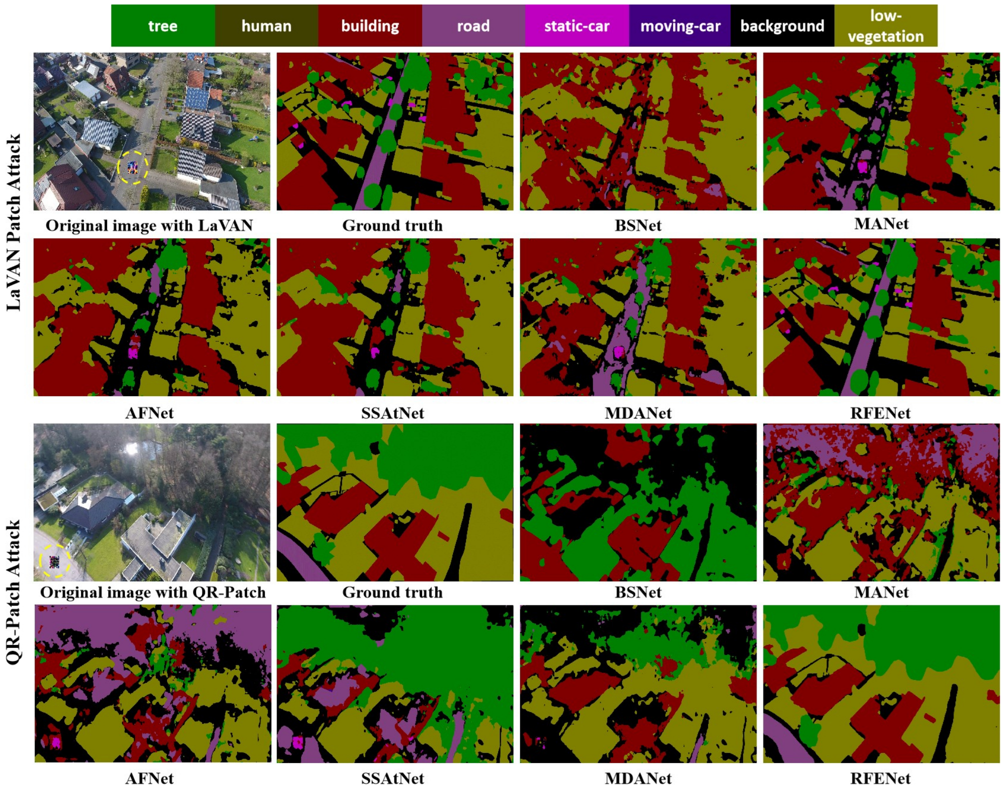

Figure 9. Visualization results of different methods on adversarial patch (LaVAN/QR-Patch) test set in the UAVid dataset, where the yellow curve circle represents the added adversarial patch region.

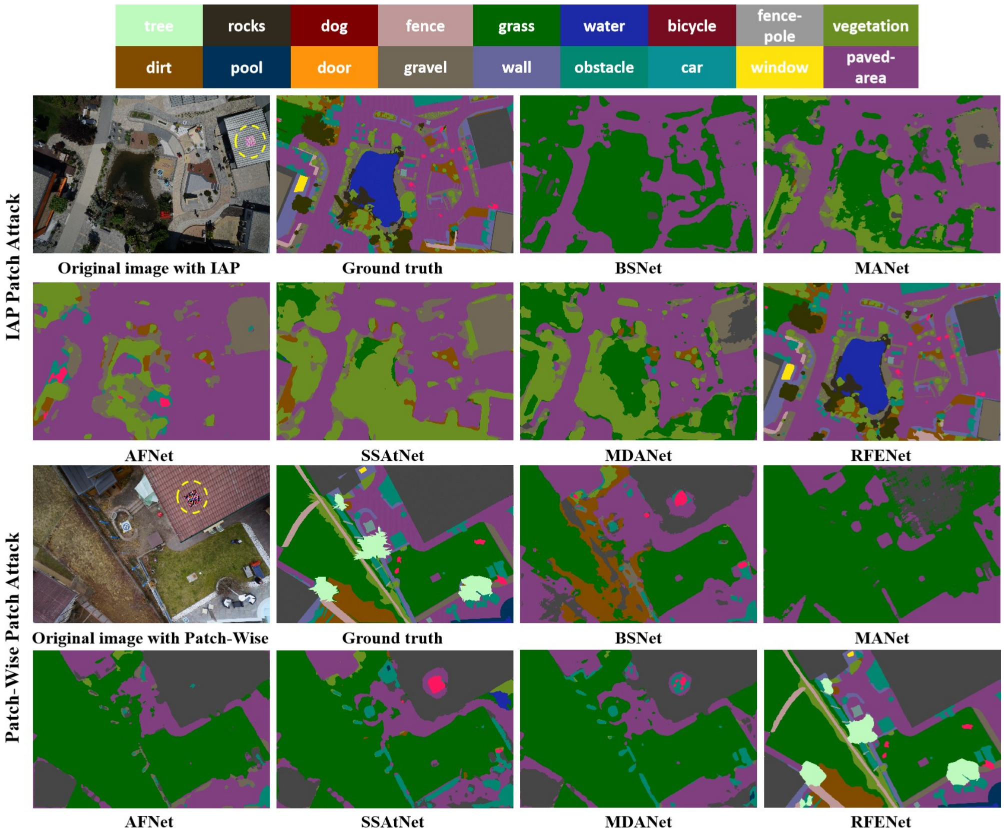

Figure 10. Visualization results of different methods on adversarial patch (IAP/Patch-Wise) test set in the Semantic Drone dataset, where the yellow curve circle represents the added adversarial patch region.

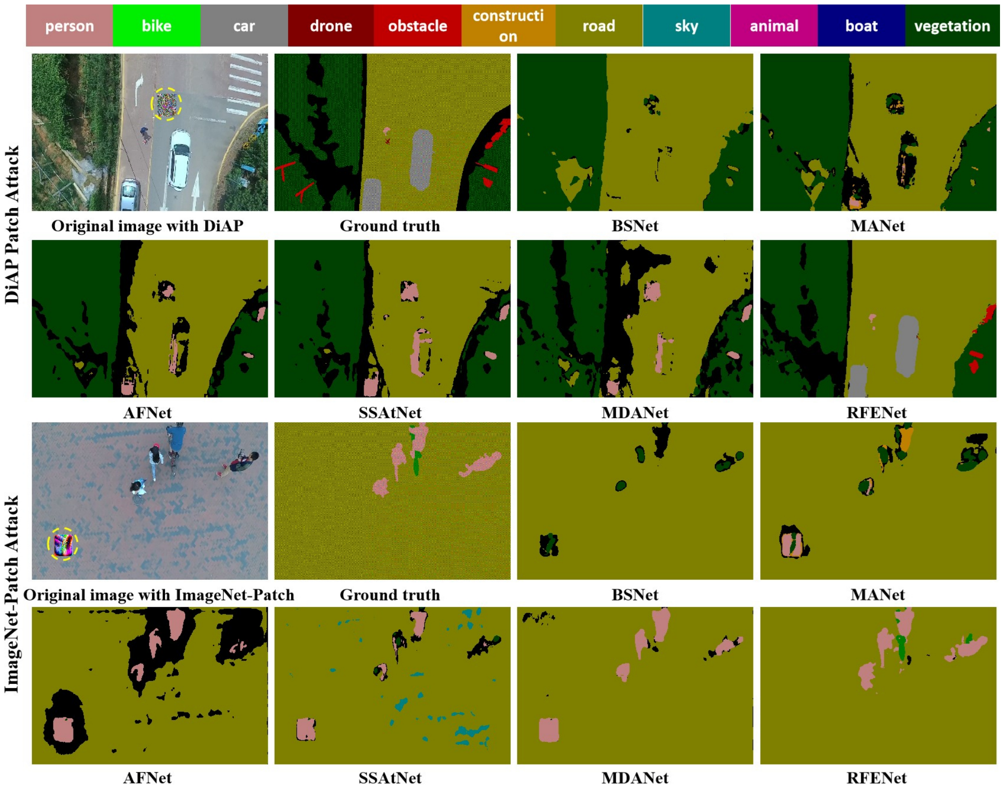

Figure 11. Visualization results of different methods on adversarial patch (DiAP/ImageNet-Patch) test set in the Aeroscapes dataset, where the yellow curve circle represents the added adversarial patch region.

4.5. Ablation Study

The proposed RFENet consists of LRFM, SSEM, BFPM and GCEM. These modules extract different robust feature information that has a defensive effect on adversarial patch attacks. In this subsection, we evaluate how each robust feature extraction component in the proposed RFENet influences the semantic segmentation performance on the adversarial patch test set. The LaVAN [16] patch attack was used to generate the adversarial patch test set, and the encoder–decoder structure DeepLabV3 [24] was used as the baseline and gradually added different robust feature extraction components. The experimental results with different modules are presented in Table 5. It can be clearly observed from Table 5 that the baseline model performs better and better on the adversarial patch test set with the introduction of different robust feature extraction modules. For example, the use of LRFM can increase the mPA metrics of all three datasets by 21.73%, 18.57% and 23.49%, respectively. Take the results in the UAVid dataset for example. The use of LRFM enables the baseline to yield the mPA of 43.26%, the use of SSEM can increase the mPA to 63.82%, BFPM enables the baseline to yield the mPA of 79.51%, and GCEM enables the baseline to yield the mPA of 87.24%. The experimental results in Table 5 further demonstrate that the combination of different robust feature extraction components enables the baseline network to obtain optimal semantic segmentation results and illustrate that robust features, such as semantic and contour, can effectively improve the adversarial robustness of the semantic segmentation model.

Another interesting issue is the influence of adversarial patches with different sizes and shapes on semantic segmentation results. Normally, for the fixed-size adversarial patch, the larger the size is, the better the attack performance is because it covers a larger region and interferes more with the features. Take the adversarial patch generated by QR-Patch [17] for example; we conducted extensive experimental verification. As shown in Table 6, the performance of existing semantic segmentation networks is affected by the size of adversarial patches. However, as can be seen from Table 6, the semantic segmentation performance of the proposed RFENet does not decrease significantly with the increase of adversarial patch size, which further proves the robustness of the proposed method against adversarial patch attacks. For the adversarial patch shape, the results in Table 6 show that the rectangular adversarial patch has stronger attack performance than the circular adversarial patch, but the proposed RFENet still has strong robustness.

Table 5. Performance of each robust feature extraction component in RFENet for different datasets, where the best results are shown in **bold**. (Report in mPA).

Baseline	LRFM	SSEM	BFPM	GCEM	UAVid	Semantic Drone	Aeroscapes
✓					21.53	19.86	22.45
✓	✓				43.26 (21.73 ↑)	38.47 (18.57 ↑)	45.94 (23.49 ↑)
✓	✓	✓			63.82 (20.56 ↑)	59.13 (20.66 ↑)	68.70 (22.76 ↑)
✓	✓	✓	✓		79.51 (15.69 ↑)	75.68 (16.55 ↑)	81.93 (13.23 ↑)
✓	✓	✓	✓	✓	**87.24** (7.73 ↑)	**82.71** (7.03 ↑)	**88.46** (6.53 ↑)

Table 6. The influence of adversarial patches with different sizes and shapes on the semantic segmentation network, where the best results are shown in **bold**. (Report in mPA).

Method	Patch Size						Patch Shape	
	15 × 15	30 × 30	45 × 45	60 × 60	75 × 75	90 × 90	Rectangle	Circle
BSNet	73.16	62.41	48.75	31.57	27.71	20.62	18.64	24.38
MANet	72.47	63.85	47.18	38.45	26.32	21.87	19.85	25.97
AFNet	75.83	64.22	46.22	36.73	28.16	22.54	17.93	25.42
SSAtNet	76.15	61.4	47.86	35.94	29.34	21.25	18.77	26.85
MDANet	77.43	66.57	45.27	34.68	30.38	23.79	20.34	27.66
RFENet	**84.36**	**84.16**	**83.64**	**83.51**	**82.97**	**82.68**	**83.05**	**84.22**

5. Discussion

To more comprehensively and systematically analyze the impact of adversarial patches on aerial image semantic segmentation models, we conducted different adversarial patch attacks (LaVAN [16], QR-Patch [17], IAP [18], Patch-Wise [63], DiAP [64] and ImageNet-Patch [65]) on all three datasets to construct adversarial patch test sets. As shown in Table 7, the existing semantic segmentation network and the proposed REFNet obtain better semantic segmentation results for clean sample datasets without patch attacks. However, for the adversarial patch test set, these existing methods suffer from the severe negative impact of adversarial patch attacks, making them not achieve the ideal semantic segmentation performance. For the comparison of adversarial patches, we can see from Table 7 that all the adversarial patch attacks have produced attack effects, especially QR-Patch attack [17], which has the most serious impact on the semantic segmentation model. It is worth noting that our proposed RFENet achieves more than 80% mPA metrics on all three adversarial patch test sets generated by different patch attack methods, which is significantly better than other existing methods. In addition, we can observe from Table 7 that the proposed RFENet still has a slight advantage on the clean sample test set without patch attacks, and its mPA metric reaches 89.47%, 85.26% and 86.75%, respectively. This phenomenon further illustrates that the proposed method can resist patch attacks and obtain better semantic segmentation performance. The reason why the proposed method has better performance is that we extracted robust features that can suppress adversarial patches and enhanced the representation of these features in the CNNs feature extraction process. In addition, these robust features are also valuable for obtaining high-precision semantic segmentation results. Therefore, our proposed method has significant advantages for aerial image semantic segmentation under adversarial patch attacks.

Table 7. Quantitative comparison results on the clean sample test sets and the adversarial patch test sets generated by different attack methods, where the best results are shown in **bold**. (Report in mPA).

Dataset	Patch Attack	BSNet	MANet	AFNet	SSAtNet	MDANet	RFENet
UAVid	Clean Sample	77.52	78.13	76.84	79.37	81.42	89.47
	LaVAN	23.41 (54.11 ↓)	24.82 (53.31 ↓)	25.74 (51.10 ↓)	28.05 (51.32 ↓)	28.74 (52.68 ↓)	88.63 (0.84 ↓)
	QR-Patch	16.27 (61.25 ↓)	17.58 (60.55 ↓)	21.34 (55.50 ↓)	20.95 (58.42 ↓)	23.65 (47.77 ↓)	87.36 (2.11 ↓)
	IAP	19.73 (57.79 ↓)	21.46 (56.67 ↓)	23.65 (53.19 ↓)	24.83 (54.54 ↓)	25.47 (55.95 ↓)	87.42 (2.05 ↓)
	Patch-Wise	17.35 (60.17 ↓)	18.94 (59.19 ↓)	24.37 (52.47 ↓)	25.43 (53.94 ↓)	24.69 (56.73 ↓)	88.06 (1.41 ↓)
	DiAP	18.32 (59.20 ↓)	19.57 (58.56 ↓)	22.38 (54.46 ↓)	24.94 (54.43 ↓)	25.12 (56.30 ↓)	88.59 (0.88 ↓)
	Image-Patch	21.64 (55.88 ↓)	22.75 (55.38 ↓)	24.06 (52.78 ↓)	23.87 (55.50 ↓)	23.43 (57.99 ↓)	89.34 (0.13 ↓)
Semantic Drone	Clean Sample	68.41	69.20	71.42	73.65	75.27	85.26
	LaVAN	20.74 (47.67 ↓)	22.64 (46.56 ↓)	23.42 (48.00 ↓)	24.96 (48.69 ↓)	25.87 (49.40 ↓)	84.75 (0.51 ↓)
	QR-Patch	12.43 (55.98 ↓)	13.71 (55.49 ↓)	14.26 (57.16 ↓)	15.92 (57.73 ↓)	16.57 (58.70 ↓)	83.15 (2.11 ↓)
	IAP	18.94 (49.47 ↓)	19.37 (49.83 ↓)	21.58 (49.84 ↓)	20.95 (52.70 ↓)	22.46 (52.81 ↓)	83.98 (1.28 ↓)
	Patch-Wise	22.35 (46.06 ↓)	23.89 (45.31 ↓)	24.17 (47.25 ↓)	25.48 (48.17 ↓)	24.56 (50.71 ↓)	84.32 (0.94 ↓)
	DiAP	16.54 (51.87 ↓)	17.68 (51.52 ↓)	18.21 (53.21 ↓)	19.54 (54.11 ↓)	19.86 (55.41 ↓)	83.65 (1.61 ↓)
	Image-Patch	24.45 (43.96 ↓)	26.87 (42.33 ↓)	26.42 (45.00 ↓)	27.63 (46.02 ↓)	28.75 (46.52 ↓)	84.93 (0.33 ↓)
Aeroscapes	Clean Sample	74.56	76.34	77.85	78.46	79.35	86.75
	LaVAN	22.37 (52.19 ↓)	23.56 (52.78 ↓)	24.97 (52.88 ↓)	26.30 (52.16 ↓)	27.63 (51.72 ↓)	86.14 (0.61 ↓)
	QR-Patch	11.38 (63.18 ↓)	12.68 (63.66 ↓)	13.04 (64.81 ↓)	14.57 (63.89 ↓)	15.62 (63.73 ↓)	84.52 (2.23 ↓)
	IAP	19.26 (55.30 ↓)	21.79 (54.55 ↓)	20.35 (57.50 ↓)	22.84 (55.62 ↓)	23.74 (55.61 ↓)	85.64 (1.11 ↓)
	Patch-Wise	15.06 (59.50 ↓)	17.93 (58.41 ↓)	18.24 (59.61 ↓)	19.78 (58.68 ↓)	21.35 (58.00 ↓)	86.45 (0.30 ↓)
	DiAP	17.35 (57.21 ↓)	18.89 (57.45 ↓)	19.24 (58.61 ↓)	18.73 (59.73 ↓)	20.26 (59.09 ↓)	85.92 (0.83 ↓)
	Image-Patch	21.52 (53.04 ↓)	22.76 (53.58 ↓)	23.85 (54.00 ↓)	24.67 (53.79 ↓)	25.92 (53.43 ↓)	86.57 (0.18 ↓)

6. Conclusions

In this study, we systematically explored the adversarial patch attacks in aerial image semantic segmentation and proposed an aerial image semantic segmentation network with better defense and semantic segmentation performance. We first analyzed the threat of adversarial patches to aerial image semantic segmentation and demonstrated that the existence of adversarial patches destroys the performance of existing aerial image semantic segmentation networks. Then, we studied in detail the "robust features" that have a suppressive effect on adversarial patch attacks and analyzed the function of robust features in defending against adversarial patches. Based on the advantages of these robust features in defending against patch attacks, we proposed a robust feature extraction network for aerial image semantic segmentation. The proposed RFENet consists of LRFM, SSEM, BFPM and GCEM, which are used to obtain local features, semantic features, boundary features and global features that are inhibitory to the adversarial patch, respectively. Extensive experiments on three aerial image semantic segmentation datasets demonstrate that the proposed method exhibits stronger resistance towards different adversarial patch attacks compared with the existing semantic segmentation networks. In addition, the ablation study further illustrated the contribution of each constructed robust feature extraction component in resisting patch attacks and improving semantic segmentation accuracy. This article reports the first study on defense against adversarial patch attacks under aerial image semantic segmentation. It provides a possible solution for defense against adversarial patch attacks at the DL model design level. In future work, we aim to introduce the concept of robust features in adversarial patch detection and adversarial training to solve the adversarial patch threat in aerial image semantic segmentation.

Author Contributions: Conceptualization, Z.W. and Y.L.; methodology, Z.W. and B.W.; software, C.Z.; validation, Z.W. and Y.L.; formal analysis, Z.W. and C.Z.; 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.; preject administration, B.W.; 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/darkseid-arch/PatchRFENet, (accessed on 16 March 2023).

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

References

1. Russell, B.J.; Soffer, R.J.; Ientilucci, E.J.; Kuester, M.A.; Conran, D.N.; Arroyo-Mora, J.P.; Ochoa, T.; Durell, C.; Holt, J. The Ground to Space CALibration Experiment (G-SCALE): Simultaneous Validation of UAV, Airborne, and Satellite Imagers for Earth Observation Using Specular Targets. *Remote Sens.* **2023**, *15*, 294. [CrossRef]
2. 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]
3. Zhang, Y.; Guo, L.; Wang, Z.; Yu, Y.; Liu, X.; Xu, F. Intelligent Ship Detection in Remote Sensing Images Based on Multi-Layer Convolutional Feature Fusion. *Remote Sens.* **2020**, *12*, 3316. [CrossRef]
4. Shirmard, H.; Farahbakhsh, E.; Müller, R.D.; Chandra, R. A Review of Machine Learning in Processing Remote Sensing Data for Mineral Exploration. *Remote Sens. Environ.* **2022**, *268*, 112750–112760. [CrossRef]
5. Yang, L.; Cervone, G. Analysis of Remote Sensing Imagery for Disaster Assessment Using Deep Learning: A Case Study of Flooding Event. *Soft Comput.* **2019**, *23*, 13393–13408. [CrossRef]
6. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent Advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
7. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. *arXiv* **2014**, arXiv:1312.6199.
8. Goodfellow, I.J.; Shlens, J.; Szegedy, C. FGSM. *arXiv* **2015**, arXiv:1412.6572.
9. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2019**, arXiv:1706.06083.
10. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Machine Learning at Scale. *arXiv* **2017**, arXiv:1611.01236.
11. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The Limitations of Deep Learning in Adversarial Settings. In Proceedings of the IEEE European Symposium on Security and Privacy, Saarbrucken, Germany, 11–15 March 2016; pp. 372–387.
12. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. *arXiv* **2017**, arXiv:1608.04644.
13. Thalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing robust adversarial examples. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 284–293.
14. Muhammad, A.; Bae, S.-H. A Survey on Efficient Methods for Adversarial Robustness. *IEEE Access.* **2022**, *10*, 118815–118830. [CrossRef]
15. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial patch. *arXiv* **2017**, arXiv:1712.09665.
16. Karmon, D.; Zoran, D.; Goldberg, Y. LaVAN: Localized and Visible Adversarial Noise. *arXiv* **2018**, arXiv:1801.02608.
17. Chindaudom, A.; Siritanawan, P.; Sumongkayothin, K.; Kotani, K. AdversarialQR: An Adversarial Patch in QR Code Format. In Proceedings of the Joint ICIEV & icIVPR, Kitakyushu, Japan, 26–29 August 2020; pp. 1–6.
18. Bai, T.; Luo, J.; Zhao, J. Inconspicuous Adversarial Patches for Fooling Image-Recognition Systems on Mobile Devices. *IEEE Internet Things J.* **2022**, *9*, 9515–9524. [CrossRef]
19. Zhang, H.; Ma, X. Misleading Attention and Classification: An Adversarial Attack to Fool Object Detection Models in the Real World. *Comput. Secur.* **2022**, *122*, 102876–102881. [CrossRef]
20. Nesti, F.; Rossolini, G.; Nair, S.; Biondi, A.; Buttazzo, G. Evaluating the Robustness of Semantic Segmentation for Autonomous Driving against Real-World Adversarial Patch Attacks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2826–2835.
21. Xu, Y.; Ghamisi, P. Universal Adversarial Examples in Remote Sensing: Methodology and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]

22. Chen, L.; Xu, Z.; Li, Q.; Peng, J.; Wang, S.; Li, H. An Empirical Study of Adversarial Examples on Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7419–7433. [CrossRef]
23. Li, H.; Huang, H.; Chen, L.; Peng, J.; Huang, H.; Cui, Z.; Mei, X.; Wu, G. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing.* **2021**, *14*, 1333–1347. [CrossRef]
24. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1706.05587.
25. Xu, Y.; Du, B.; Zhang, L. Assessing the Threat of Adversarial Examples on Deep Neural Networks for Remote Sensing Scene Classification: Attacks and Defenses. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1604–1617. [CrossRef]
26. 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]
27. 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; pp. 8417–8424.
28. Xiang, C.; Bhagoji, A.N.; Sehwag, V.; Mittal, P. PatchGuard: A Provably Robust Defense against Adversarial Patches via Small Receptive Fields and Masking. *arXiv* **2021**, arXiv:2005.10884.
29. Lal, S.; Rehman, S.U.; Shah, J.H.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; Mohammed, M.A.; Abdulkareem, K.H. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. *Sensors* **2021**, *21*, 3922. [CrossRef]
30. 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.
31. Ai, S.; Voundi Koe, A.S.; Huang, T. Adversarial Perturbation in Remote Sensing Image Recognition. *Appl. Soft Comput.* **2021**, *105*, 107252–107259. [CrossRef]
32. Jiang, Y.; Yin, G.; Yuan, Y.; Da, Q. Project Gradient Descent Adversarial Attack against Multisource Remote Sensing Image Scene Classification. *Secur. Commun. Networks.* **2021**, *2021*, 6663028. [CrossRef]
33. Cheng, G.; Sun, X.; Li, K.; Guo, L.; Han, J. Perturbation-Seeking Generative Adversarial Networks: A Defense Framework for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
34. Chen, L.; Li, H.; Zhu, G.; Li, Q.; Zhu, J.; Huang, H.; Peng, J.; Zhao, L. Attack Selectivity of Adversarial Examples in Remote Sensing Image Scene Classification. *IEEE Access.* **2020**, *8*, 137477–137489. [CrossRef]
35. Lian, J.; Mei, S.; Zhang, S.; Ma, M. Benchmarking Adversarial Patch Against Aerial Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
36. 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]
37. 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]
38. Du, A.; Chen, B.; Chin, T.J.; Law, Y.W.; Sasdelli, M.; Rajasegaran, R.; Campbell, D. Physical Adversarial Attacks on an Aerial Imagery Object Detector. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022; pp. 1796–1806.
39. Deng, B.; Zhang, D.; Dong, F.; Zhang, J.; Shafiq, M.; Gu, Z. Rust-Style Patch: A Physical and Naturalistic Camouflage Attacks on Object Detector for Remote Sensing Images. *Remote Sens.* **2023**, *15*, 885. [CrossRef]
40. Dräger, N.; Xu, Y.; Ghamisi, P. Backdoor Attacks for Remote Sensing Data with Wavelet Transform. *arXiv* **2022**, arXiv:2211.08044.
41. Chen, L.; Xiao, J.; Zou, P.; Li, H. Lie to Me: A Soft Threshold Defense Method for Adversarial Examples of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
42. Xue, W.; Chen, Z.; Tian, W.; Wu, Y.; Hua, B. A Cascade Defense Method for Multidomain Adversarial Attacks under Remote Sensing Detection. *Remote Sens.* **2022**, *14*, 3559. [CrossRef]
43. Zhang, H.; Wang, J. Defense against Adversarial Attacks Using Feature Scattering-Based Adversarial Training. In Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Montreal, QC, Canada, 8–14 December 2019; pp. 113–118.
44. 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]
45. Xie, C.; Wu, Y.; van der Maaten, L.; Yuille, A.L.; He, K. Feature Denoising for Improving Adversarial Robustness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 501–509.
46. Zhou, D.; Liu, T.; Han, B.; Wang, N.; Peng, C.; Gao, X. Towards Defending against Adversarial Examples via Attack-Invariant Features. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12835–12845.
47. Freitas, S.; Chen, S.-T.; Wang, Z.J.; Horng Chau, D. UnMask: Adversarial Detection and Defense Through Robust Feature Alignment. In Proceedings of the IEEE International Conference on Big Data, Atlanta, GA, USA, 10–13 December 2020; pp. 1081–1088.

48. Liu, Z.; Liu, Q.; Liu, T.; Xu, N.; Lin, X.; Wang, Y.; Wen, W. Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 860–868.
49. Li, X.; Zhu, D. Robust Detection of Adversarial Attacks on Medical Images. In Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1154–1158.
50. 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.
51. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
52. Chen, X.; Li, Z.; Jiang, J.; Han, Z.; Deng, S.; Li, Z.; Fang, T.; Huo, H.; Li, Q.; Liu, M. Adaptive Effective Receptive Field Convolution for Semantic Segmentation of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3532–3546. [CrossRef]
53. Jiang, K.; Wang, Z.; Yi, P.; Lu, T.; Jiang, J.; Xiong, Z. Dual-Path Deep Fusion Network for Face Image Hallucination. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 378–391. [CrossRef]
54. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic Convolution: Attention Over Convolution Kernels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11027–11036.
55. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
56. Luan, S.; Chen, C.; Zhang, B.; Han, J.; Liu, J. Gabor Convolutional Networks. *IEEE Trans. Image Process.* **2018**, *27*, 4357–4366. [CrossRef]
57. Li, J.; Zha, S.; Chen, C.; Ding, M.; Zhang, T.; Yu, H. Attention Guided Global Enhancement and Local Refinement Network for Semantic Segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 3211–3223. [CrossRef]
58. Li, X.; Yu, L.; Chang, D.; Ma, Z.; Cao, J. Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4204–4212. [CrossRef]
59. Wang, L.; Wang, C.; Sun, Z.; Chen, S. An Improved Dice Loss for Pneumothorax Segmentation by Mining the Information of Negative Areas. *IEEE Access.* **2020**, *8*, 167939–167949. [CrossRef]
60. Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; Yang, M.Y. UAVid: A Semantic Segmentation Dataset for UAV Imagery. *ISPRS J. Photogramm.* **2020**, *165*, 108–119. [CrossRef]
61. 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, Paris, France, 31 May 2020; pp. 2009–2016.
62. Nigam, I.; Huang, C.; Ramanan, D. Ensemble Knowledge Transfer for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1499–1508.
63. Gao, L.; Zhang, Q.; Song, J.; Liu, X.; Shen, H.T. Patch-Wise Attack for Fooling Deep Neural Network. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Volume 12373, pp. 307–322.
64. Zhou, X.; Pan, Z.; Duan, Y.; Zhang, J.; Wang, S. DiAP-A Data Independent Approach to Generate Adversarial Patches. *Mach. Vis. Appl.* **2021**, *32*, 67–75. [CrossRef]
65. Pintor, M.; Angioni, D.; Sotgiu, A.; Demetrio, L.; Demontis, A.; Biggio, B.; Roli, F. ImageNet-Patch: A Dataset for Benchmarking Machine Learning Robustness against Adversarial Patches. *Pattern Recognit.* **2023**, *134*, 109064–109072. [CrossRef]
66. Hou, J.; Guo, Z.; Wu, Y.; Diao, W.; Xu, T. BSNet: Dynamic Hybrid Gradient Convolution Based Boundary-Sensitive Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–22. [CrossRef]
67. 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.* **2022**, *60*, 1–13. [CrossRef]
68. Liu, R.; Mi, L.; Chen, Z. AFNet: Adaptive Fusion Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7871–7886. [CrossRef]
69. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic Segmentation with Attention Mechanism for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
70. Zuo, R.; Zhang, G.; Zhang, R.; Jia, X. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.