remote sensing

MDPI

Article

Global Feature Attention Network: Addressing the Threat of Adversarial Attack for Aerial Image Semantic Segmentation

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

[1] Schoolof Information and Navigation, Air Force Engineering University, FengHao East Road, Xi'an 710082, China
[2] School of Surveying and Geo-Informatics, Shandong Jianzhu University, FengMing Road, Jinan 250101, China
[3] School of Electronic Information, Xijing University, XiJing Road, Xi'an 710123, China
* Correspondence: miswz@iocas.ac.cn (W.Z); Tel.: +86-132-7936-7151

Abstract: Aerial Image Semantic segmentation based on convolution neural networks (CNNs) has made significant process in recent years. Nevertheless, their vulnerability to adversarial example attacks could not be neglected. Existing studies typically focus on adversarial attacks for image classification, ignoring the negative effect of adversarial examples on semantic segmentation. In this article, we systematically assess and verify the influence of adversarial attacks on aerial image semantic segmentation. Meanwhile, based on the robust characteristics of global features, we construct a novel global feature attention network (GFANet) for aerial image semantic segmentation to solve the threat of adversarial attacks. GFANet uses the global context encoder (GCE) to obtain the context dependencies of global features, introduces the global coordinate attention mechanism (GCAM) to enhance the global feature representation to suppress adversarial noise, and the feature consistency alignment (FCA) is used for feature calibration. In addition, we construct a universal adversarial training strategy to improve the robustness of the semantic segmentation model against adversarial example attacks. Extensive experiments on three aerial image datasets demonstrate that GFANet is more robust against adversarial attacks than existing state-of-the-art semantic segmentation models.

Keywords: aerial images; semantic segmentation; convolution neural networks (CNNs); adversarial example; adversarial attack; adversarial defense

Citation: 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. https://doi.org/10.3390/rs15051325

Academic Editor: Andrea Garzelli

Received: 5 February 2023
Revised: 20 February 2023
Accepted: 22 February 2023
Published: 27 February 2023

1. Introduction

Semantic segmentation aims to assign a predefined semantic label category to each pixel in the image [1]. As one of the fundamental scene understanding tasks for the earth observation [2], aerial image semantic segmentation is applied to military reconnaissance [3], urban planning [4], precision agriculture [5], and disaster monitoring [6]. The potent feature extraction and representation capabilities [7,8] of convolution neural networks (CNNs) make it widely used in semantic segmentation tasks of aerial images.

Different from the previous semantic segmentation methods based on handcrafted feature extraction [9], the CNNs-based method achieves semantic segmentation by automatically extracting global structure and local spatial features contained in the image. For aerial image semantic segmentation, various CNN-based methods have been proposed. For instance, LANet [10] uses multiple convolution operations to obtain local and global features to achieve aerial image semantic segmentation. Chen et al. [11] enhanced the feature representation of ground object regions by mining the correlation of different features. AFNet [12] reduces the feature information loss by hierarchical feature fusion to achieve accurate semantic segmentation. To solve the inconsistent object scale, He et al. [13] constructed the multi-scale aware-relation module to obtain discriminant features. BSNet [14] introduces the dynamic hybrid gradient convolution and adaptive aggregation module to improve boundary segmentation accuracy. SBANet [15] uses a multi-branch convolution

structure to obtain fine-grained feature information of aerial images for semantic segmentation. To obtain context dependencies, MANet [16] uses the cascade attention mechanism to obtain the correlation between local and global features. Yang et al. [17] proposed the hidden path selection network to adaptively obtain refined local detail information. These methods demonstrate the importance of CNNs in aerial image semantic segmentation.

Despite the great success achieved by CNN-based methods, their vulnerability to adversarial attacks should be taken seriously, especially for aerial image interpretation [18]. In brief, the adversarial example attacks are to add human-imperceptible adversarial perturbations to the original images, which significantly degrades the performance of CNNs models [19]. Recent studies have indicated that adversarial example attacks pose a serious threat to many CNN-based visual tasks, such as image classification [20], object detection [21], and semantic segmentation [22]. Szegedy et al. [23] first discovered that adversarial examples can easily fool deep neural networks to produce misclassification. To improve the effectiveness of adversarial attacks, Goodfellow et al. [24] proposed the fast gradient sign method (FGSM) to generate adversarial examples. DeepFool [25] is a classical non-target attack method, which can calculate smaller perturbations to achieve adversarial attacks better than FGSM. Jacobian-based saliency map attack (JSMA) [26] is a target attack method based on ℓ_0 norm, which realizes adversarial attack by modifying significant pixels. C&W [27] supports ℓ_0, ℓ_2, and ℓ_∞ norms attack modes and has strong generalization ability. As a gradient-based attack, project gradient descent (PGD) [28] attack and basic iterative method (BIM) [29] attack can be regarded as the iterative version of FGSM. Universal adversarial perturbations (UAP) [30] attack is a model-based attack method that generates adversarial examples without data assistance.

For the geoscience and remote sensing community, studies on adversarial examples have also received attention. Czaja et al. [31] first analyzed the influence of adversarial patches on satellite remote sensing image classification. Li et al. [32] constructed a remote sensing image classification model against white-box adversarial attack. Chen et al. [33] analyzed the negative impact of adversarial examples on remote sensing scene interpretation. Li et al. [34] demonstrated that adversarial examples would destroy the performance of CNN-based SAR image classifiers. In [35], Xu et al. designed a robust hyperspectral image classification network against adversarial attacks. Chen et al. [36] assess the influence of black-box and white-box adversarial attacks on remote sensing scene classification. Xu et al. [37] proposed an adversarial training strategy to improve the robustness of the remote sensing image classification model to adversarial examples. In [38], Cheng et al. designed a universal adversarial example attack strategy for remote sensing image classification. However, these studies focus on the problem of adversarial attacks in aerial image classification tasks, ignoring the impact of adversarial attacks on aerial image semantic segmentation. Moreover, existing studies commonly use adversarial training or adversarial sample detection methods to resist the interference of adversarial noise without considering the contribution of global features in defending against adversarial example attacks. In Figure 1, we illustrate the adversarial attacks for aerial image semantic segmentation. As shown in Figure 1, the CNNs model can obtain accurate semantic segmentation results for the original image while adding adversarial noise seriously affects the segmentation effect.

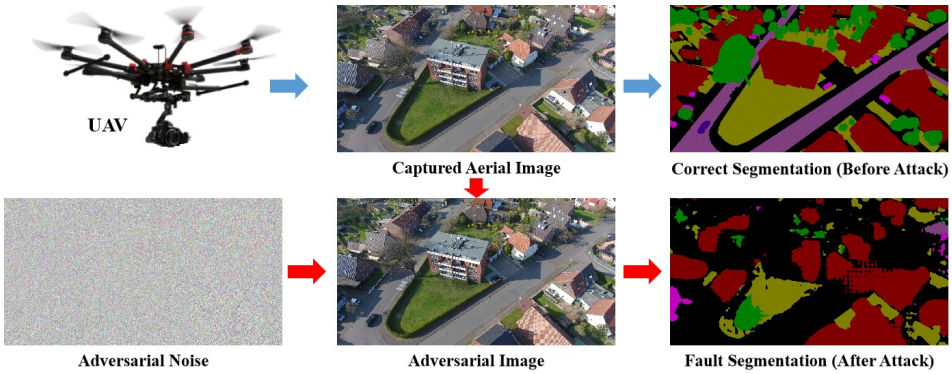


Figure 1. Illustration of the adversarial examples attack on aerial image semantic segmentation.

Existing studies [35,39] have demonstrated that global features have better robustness against adversarial attacks. To defend against adversarial example attacks faced by aerial image semantic segmentation, in this article, we propose a robust aerial image semantic segmentation method, namely the global feature attention network (GFANet). Based on the robust characteristics of global features, we design the global context encoder (GCE), global coordinate attention mechanism (GCAM), and feature consistency alignment (FCA) for mining the global feature information contained in aerial images. In addition, we construct a universal adversarial training strategy to enhance the robustness of the model against adversarial attacks. The contributions of this study are summarized as follows.

- We systematically analyze the impact of adversarial attacks on aerial image semantic segmentation for the first time and propose a robust aerial image semantic segmentation network based on global context feature information awareness and fusion.
- We construct the global context encoder (GCE) module, global coordinate attention mechanism (GCAM), and feature consistency alignment (FCA) module to resist adversarial noise interference by using the robust characteristics of global features.
- We design a universal adversarial training strategy to enhance the defense of the semantic segmentation model against different adversarial example attacks by introducing Gaussian noise in the adversarial training process.
- The extensive experiments conducted on three aerial image datasets containing large-scale urban and suburban scenes demonstrate the robustness of the proposed method against adversarial attacks while maintaining high semantic segmentation accuracy.

2. Related Works

In this section, we review the existing global feature extraction, adversarial attacks, and adversarial defense methods.

2.1. Global Feature Extraction

The extraction of global feature information is essential for various computer vision tasks. It can improve the performance of the CNN model and make it robust against adversarial attacks [39]. However, global feature extraction is challenging for CNN-based models because it needs to consider both local and long-range dependencies. For semantic segmentation tasks, extracting global feature information from images has received extensive attention. Zhang et al. [40] constructed the context encoding module to encode the semantic information and obtain global details by using the relative position relationship of pixels. PANet [41] introduces the global attention unsample module in each encoder layer to model global context information to obtain long-range spatial dependencies. CGNet [42] uses the context-guided module to encode local pixel regions and then uses the feature correlation of local pixels to obtain global context information. HRCNet [43] constructs the context feature interaction structure to splice local features, and the semantic inference module is used to fuse local and global features. Nekrasov et al. [44] designed the global deconvolution model to enhance the resolution of context features and used the spatial interaction to model local spatial dependencies. Zhang et al. [45] designed the context

feature aggregation module to obtain global feature information by context aggregation of different scale features. Li et al. [46] construct the hybrid attention module to enhance the edge distribution of the ground object and use the non-local attention mechanism to model the global context information. SPANet [47] uses synergistic attention to obtain spatial and channel features and uses a hierarchical embedded-synergistic attention perception module to aggregate global context information. CANet [48] constructs the covariance attention to model the global feature information and establishes the context dependency by obtaining the relationship between local pixels. In this article, inspired by [40,42,45,48], we construct GFANet based on the robust characteristics of global features, which can obtain better semantic segmentation accuracy and be robust to adversarial example attacks.

2.2. Adversarial Attacks

The adversarial example belongs to the evasion attack in image classification tasks [23]. The attacker constructs adversarial noise with weak perturbation based on the differentiable loss function to destroy the classifier performance.

$$\max \mathcal{L}(x+\eta, y, \theta), \quad \text{s.t. } \|\eta\|_p \leqslant \varepsilon \tag{1}$$

where, x is the clean image sample, y is the category label corresponding to the input image, θ denotes the model parameter variable, η denotes the adversarial perturbation, $\mathcal{L}(\cdot)$ is the classifier loss function, and $\|\cdot\|_p$ denotes the ℓ_p norm. The optimization objective of Equation (1) is to obtain the optimal adversarial perturbation η under the restriction of perturbation amplitude ε to maximize the classifier loss.

(1) Fast Gradient Sign Method (FGSM): In fact, FGSM can be interpreted as a gradient ascent method [24]. In the linear model, the perturbation variables that FGSM expects to add are consistent with the gradient direction of the model loss function, and its formal description is as follows.

$$\hat{x} = x + \alpha \text{sign}(\nabla_x \mathcal{L}(x, y, \theta)) \tag{2}$$

where α is the hyper-parameter used to adjust the perturbation amplitude, $\text{sign}(\cdot)$ is the sign function, and $\nabla_x \mathcal{L}(\cdot)$ is the gradient of the loss function $\mathcal{L}(\cdot)$ to the input image x.

(2) DeepFool: The DeepFool attack [25] assumes the classifier model is linear, and each category has a decision boundary (i.e., decision hyperplane). DeepFool solves the optimization problem of Equation (3) using multiple iterations to obtain perturbations that satisfy the $f(\hat{x}) \neq f(x)$.

$$\hat{x}^{k+1} = \hat{x}^k - \frac{f\left(\hat{x}^k\right)}{\left\|\nabla_x \mathcal{L}\left(f\left(\hat{x}^k\right), t\right)\right\|_2^2} \nabla_x \mathcal{L}\left(f\left(\hat{x}^k\right), t\right) \tag{3}$$

where $f(x)$ denotes the predicted label of the input image, t denotes the target category in the attack, and compared with FGSM [24], the DeepFool generates less perturbation.

(3) Jacobian-Based Saliency Map Attack (JSMA): Compared with other attack methods, JSMA [26] introduces the concept of the saliency map, which obtains the best attack effect by modifying the minimum pixel of clean samples. JSMA uses ℓ_0 norm to constrain the generation of adversarial perturbation, and its optimization objective is as follows.

$$\triangle_r = \min_r \|r\|_0, \quad \text{s.t. } f(x+r) \neq f(x) \tag{4}$$

where $\triangle_r$ represents the optimal perturbation of the adversarial examples, r represents the initial perturbation variable, and $\|\cdot\|_0$ represents the ℓ_0 norm constraint.

(4) Carlini–Wagner (C&W): The C&W attack [27] includes three attack modes under ℓ_0, ℓ_2, and ℓ_∞ norm constraints, which mainly attack a distillation network with strong defense capability. C&W generate adversarial perturbations by solving a norm-restricted constrained optimization problem.

$$\min \left[\|r\|_p + c \cdot f(x+r)\right], \quad \text{s.t. } x + r \in [0,1]^n \tag{5}$$

where $\|\cdot\|_p$ represents the norm constraint. By optimizing Equation (5), C&W can improve the confidence of misclassification labels under the condition of small perturbation value.
(5) Projected Gradient Descent (PGD): The PGD attack [28] can be regarded as the iterative version of FGSM [24]. The basic idea is to use $\hat{x}^0 = x + ramdom(-\varepsilon, \varepsilon)$ as the initialization value and calculate the gradient of $\hat{x}$ by iterations to update the adversarial examples.

$$x_{i+1} = \text{clip}_{x,\varepsilon}(x_i + \mu \,\text{sign}(\nabla_{x_i} \mathcal{L}(x_i, t)) \tag{6}$$

where μ represents the iterative step and $\text{clip}(\cdot)$ represents for constraining within the ε-neighbor ball of input image x.
(6) Universal Adversarial Perturbation (UAP): The UAP attack [30] can generate the adversarial perturbations without any image samples. The generated adversarial examples are universal because perturbation r satisfies the following constraints.

$$P(f(x + r) \neq f(r)) \geqslant 1 - \delta, \quad \text{s.t.} \|r\|_p \leqslant \varepsilon \tag{7}$$

where $\|\cdot\|_p$ represents the ℓ_p norm constraint, r represents the perturbation variable, δ is the desired fooling rate, and ε is used to limit the amplitude of adversarial perturbation.

2.3. Adversarial Defense

To resist adversarial attacks, many defense methods have been proposed, including adversarial training, adversarial example detection, and modified network architecture.

(1) Adversarial Training: The adversarial training strategy is to introduce adversarial examples in the process of model parameter optimization to improve the robustness against adversarial attacks [28]. The process of adversarial training is equivalent to solving the following maximin problem.

$$\min_f E\left\{ \max_{\hat{x} \in B(x,\varepsilon)} \mathcal{L}(f(\hat{x}), y) \right\} \tag{8}$$

where x represents clean examples, $\hat{x}$ represents adversarial examples, y represents the corresponding true labels, and $\mathcal{L}(\cdot)$ represents the loss function. $B(x, \varepsilon)$ is the space with $x \pm \varepsilon$ as the upper and lower bounds. In the process of solving the internal maximization of $\max \mathcal{L}(f(\hat{x}), y)$, the PGD attack [28] is used to generate adversarial examples to approximate the solution, while for the external minimization problem of $\min E\{\cdot\}$ is to minimize the adversarial loss caused by the internal adversarial examples by updating the model parameters. Zhang et al. [49] modified the adversarial training, transforming the training process into the approximate solution to the following maximin problem.

$$\min_f E\left\{ \mathcal{L}(f(x), y) + \alpha \cdot \max_{\hat{x} \in B(x,\varepsilon)} \mathcal{L}(f(\hat{x}), f) \right\} \tag{9}$$

where α represents the regular term. This method balances the accuracy of clean examples and adversarial examples and obtains better adversarial defense effects. Liu et al. [50] trained the deep neural network by adding Gaussian noise to the image for data enhancement to reduce the model sensitivity to adversarial noise. Wang et al. [51] constructed the misclassification-aware adversarial training strategy, which trains the robust model by distinguishing the incorrectly and correctly classified samples in the training process.

(2) Adversarial Example Detection: In addition to adversarial training, another approach to address adversarial attacks is adversarial example detection. The existing adversarial example detection methods include metric-based methods and prediction inconsistency-based methods. For metric-based methods, Feinman et al. [52] used kernel density estimation and Bayesian uncertainty to discriminate the adversarial subspace to separate clean adversarial examples. Ma et al. [53] proposed a detection method based on local intrinsic dimension measurement, which constructed the separation hyperplane of clean and adversarial examples by estimating intrinsic dimension to achieve adversarial

example detection. Grosse et al. [54] uses an addition outlier class C_{out} to strengthen the model training process. In the test stage, the trained model can classify the adversarial examples into the C_{out} class. Since adversarial examples are inconsistent with clean examples, adversarial examples detection can be achieved by comparing their differences. Tao et al. [55] detect adversarial examples by measuring the inconsistency between the clean example and adversarial example training models. Feinman et al. [56] used dropout technology to generate multiple classifier models to detect adversarial examples by identifying inconsistencies in the output of different classifiers. Xu et al. [57] limited the available subspace of adversarial examples by using the feature compression squeezing technique and then set a fixed threshold to detect adversarial examples.

(3) Modified Network Architecture: Modifying the model architecture and parameters is another effective to against adversarial attacks. Inspired by the denoising autoencoders, Gu et al. [58] constructed the deep compression network to suppress the adversarial noise interference. Ross et al. [59] use gradient regularization to improve the model's robustness against adversarial attacks. Hinton et al. [60] use knowledge distillation to transfer teacher network parameters to student networks and use feature compression against adversarial attacks. Nayebi et al. [61] introduce the enhanced nonlinear activation function to improve the model nonlinear learning against adversarial examples. Cisse et al. [62] uses the global Lopschitz constant to control the gradient optimization direction and uses the gate optimization coefficient to enhance the model robustness. Gao et al. [63] design the DeepCloak framework, which introduces the nonlinear units in the classification layer to suppress adversarial noise interference. Sun et al. [64] construct robust deep neural networks against adversarial example attacks based on statistical filters.

3. Methodology

The overall architecture of GFANet is shown in Figure 2, which includes the backbone network, global context encoder (GCE) module, global coordinate attention mechanism (GCAM), and feature consistency alignment (FCA) module. Specifically, we use VGG16 [65] as the backbone network to extract primary feature information. Then, the GCE module performs a global context encoder on the primary features to obtain the long-range spatial dependency. To further mine global feature information and suppress adversarial noise interference, GCAM and FCA module are used to enhance the global representation of deep and shallow features and achieve multi-scale feature fusion.

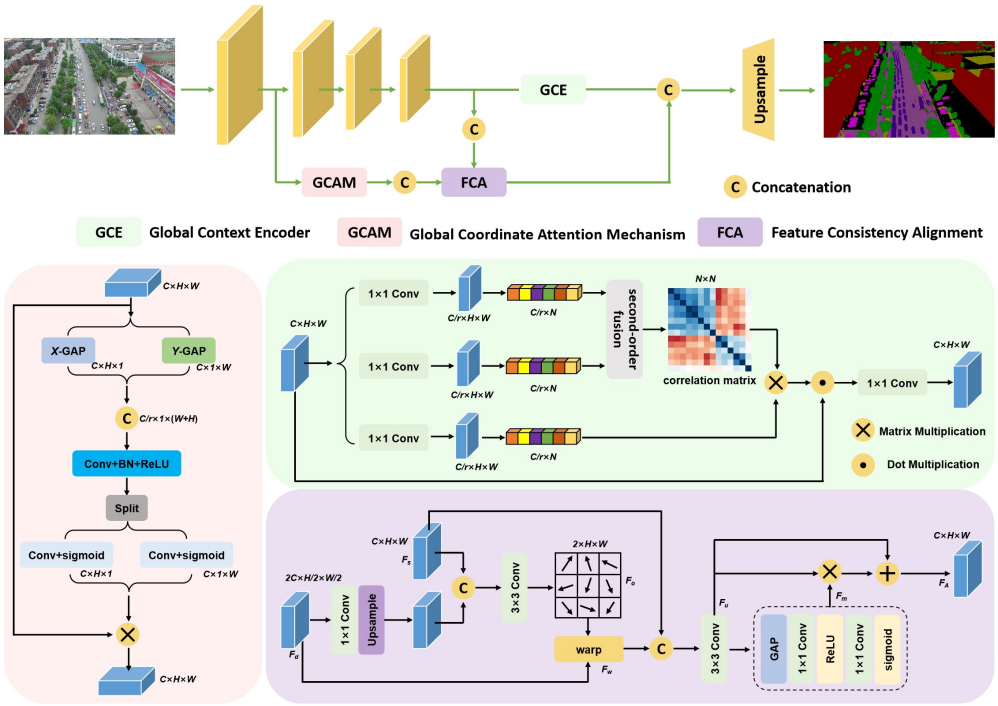


Figure 2. Illustration of the proposed global feature attention network (GFANet). The global context encoder is adopted to build global spatial dependency and suppress adversarial noise. The global coordinate attention mechanism and feature consistency alignment are used for global feature enhancement and fusion of shallow and deep features. Each feature map is shown with the size of its tensor (e.g., h, w, and c represent the height, width, and the number of channels, respectively).

3.1. Global Context Encoder

Extracting the global context of high-dimensional features can obtain global dependencies between features, effectively reduce the impact of adversarial noise, and enhance the robustness against adversarial attacks. Inspired by previous work in [35], we propose a global context encoder module, which re-adjusts the dependence of different feature spatial positions by second-order linear fusion of high-dimensional features to obtain the spatial position correlation matrices. Compared with the reshape and transpose operation in the existing attention mechanism, the second-order linear fusion can perform bilinear fusion on different features in spatial positions and reduces feature information loss. As shown in Figure 2, the global context encoder module uses feature map $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ as input, and uses two 1×1 convolutions to reduce the dimension of feature $\boldsymbol{F}$ to obtain feature maps $\boldsymbol{F}_1 \in \mathbb{R}^{C/r\times H\times W}$ and $\boldsymbol{F}_2 \in \mathbb{R}^{C/r\times H\times W}$, where r represents the dimension reduction coefficient. To obtain the global representation, the features $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are stretched in the spatial dimension to obtain the spatial vectors $\boldsymbol{V}_1 \in \mathbb{R}^{C/r\times H\times W}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{C/r\times H\times W}$. To realize spatial vector fusion, we use bilinear pooling [66] to fuse spatial vectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ to obtain spatial fusion matrix. The obtained spatial fusion matrix contains a global representation between the pixel positions of different features. To enhance the nonlinear representation, we use the Softmax function to normalize the spatial fusion matrix in spatial position, and the calculation is as follows.

$$\boldsymbol{S}_{ji} = \frac{\exp\left(\boldsymbol{V}_{1i} \cdot \boldsymbol{V}_{2j}\right)}{\sum_{i}^{N} \exp\left(\boldsymbol{V}_{1i} \cdot \boldsymbol{V}_{2j}\right)} \tag{10}$$

where $\boldsymbol{S}_{ji}$ represents the encoder of the ith position on the spatial pixel to the jth position, and $\boldsymbol{V}_{ij}$ represents the jth element of the feature vector $\boldsymbol{V}_i$.

To further enhance the feature global representation, we perform the same operation for feature $\boldsymbol{F}$ as $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ to obtain the spatial vector $\boldsymbol{V}_3$ and residual connection with the feature $\boldsymbol{F}$ to obtain the global context feature $\boldsymbol{E}$, and the specific calculation is as follows.

$$\boldsymbol{E}_j = \sum_{j=1}^{C/r} \sum_{i=1}^{N} \left(\boldsymbol{V}_{3i} \cdot \boldsymbol{S}_{ji} \right) + \boldsymbol{F}_j \tag{11}$$

where N represents the dimension of spatial fusion matrix $\boldsymbol{S}_{ji}$, and C/r represents the channel dimension of feature map. To restore the channel dimension of the feature map, we use 1×1 convolution to map the feature $\boldsymbol{E}$ to the original feature channel dimension.

3.2. Global Coordinate Attention Mechanism

To further enhance the global feature representation and suppress the adversarial noise interference, inspired by previous work in [67,68], we construct a global coordinate attention mechanism. Different from the existing coordinate attention mechanism [69], the global coordinate attention mechanism performs a spatial two-dimensional encoder on the input features and performs feature aggregation along two spatial directions, one of which can obtain long-range dependency, and the other can obtain spatial position information. The generated feature map is encoded into a pair of direction-aware and position-sensitive attention feature maps, which are complementary to the input features to achieve global feature enhancement. As shown in Figure 2, for the input feature $\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$, the global coordinate attention mechanism performs global average pooling (GAP) on the W and H directions to obtain feature maps $\boldsymbol{z}^h \in \mathbb{R}^{C \times H \times 1}$ and $\boldsymbol{z}^w \in \mathbb{R}^{C \times 1 \times W}$.

$$\boldsymbol{z}^h = \frac{1}{W} \sum\nolimits_{0 \leqslant i < W} \boldsymbol{x}(h, i) \tag{12}$$

$$\boldsymbol{z}^w = \frac{1}{H} \sum\nolimits_{0 \leqslant j < H} \boldsymbol{x}(j, w) \tag{13}$$

where W and H represent the width and height of feature map $\boldsymbol{x}$. For feature $\boldsymbol{z}^h$ and $\boldsymbol{z}^w$, we fuse features on the spatial dimension and then use convolution and the nonlinear activation function to obtain the fusion feature $\boldsymbol{f} \in \mathbb{R}^{C/r \times 1 \times (H+W)}$, where the convolution operation compresses the channel dimension to C/r using the channel scaling factor r.

$$\boldsymbol{f} = \delta\left(F\left(\left[\boldsymbol{z}^h, \boldsymbol{z}^w\right]\right)\right) \tag{14}$$

where, $[\cdot, \cdot]$ represents the feature splicing operation, F represents the convolution operation, and δ represents the nonlinear activation function. Then feature $\boldsymbol{f}$ is decomposed into independent tensors $\boldsymbol{f}^h \in \mathbb{R}^{C/r \times H \times 1}$ and $\boldsymbol{f}^w \in \mathbb{R}^{C/r \times 1 \times W}$ on the spatial dimension, and convolution and activation operations are performed on $\boldsymbol{f}^h$ and $\boldsymbol{f}^w$ to obtain attention feature vectors $\boldsymbol{g}^h \in \mathbb{R}^{C \times H \times 1}$ and $\boldsymbol{g}^w \in \mathbb{R}^{C \times 1 \times W}$.

$$\boldsymbol{g}^h = \delta\left(F_h\left(\boldsymbol{f}^h\right)\right) \tag{15}$$

$$\boldsymbol{g}^w = \delta(F_w(\boldsymbol{f}^w)) \tag{16}$$

where F_h and F_w represent convolution operations in different directions. The attention feature vectors are fused with the original feature to obtain the global enhanced feature $\boldsymbol{y} \in \mathbb{R}^{C \times H \times W}$. The calculation is as follows.

$$\boldsymbol{y}(i, j) = \boldsymbol{x}(i, j) \times \boldsymbol{g}^h(i) \times \boldsymbol{g}^w(j) \tag{17}$$

where the broadcast mechanism is used to superimpose the W direction feature $\boldsymbol{g}^w(j)$ and the H direction feature $\boldsymbol{g}^h(i)$ on the feature $\boldsymbol{x}$, and the global enhancement feature $\boldsymbol{y}(i, j)$ is consistent with the size of the original feature $\boldsymbol{x}(i, j)$.

3.3. Feature Consistency Alignment

For aerial image semantic segmentation, shallow convolution contains high-resolution spatial structure features, and deep convolution contains low-resolution local detail features. Previous studies [10–12] use bilinear interpolation to upsample low-resolution features and then fuse them with high-resolution features. However, due to the influence of adversarial noise, the use of bilinear interpolation will expand the interference of adversarial noise. As shown in Figure 2, the constructed feature consistency alignment module takes the deep feature $\boldsymbol{F}_d \in \mathbb{R}^{H/2\times W/2\times 2C}$ generated by the global coordinate attention mechanism and the shallow feature $\boldsymbol{F}_s \in \mathbb{R}^{H\times W\times C}$ obtained by the backbone network as input features. The 1×1 convolution is used to reduce the channel dimension of deep feature $\boldsymbol{F}_d$ and upsample to the size consistent with $\boldsymbol{F}_s$. Then, the channel fusion is carried out between the upsampling feature and $\boldsymbol{F}_s$, and the 3×3 convolution is used to obtain the two-dimensional offset $\boldsymbol{F}_o \in \mathbb{R}^{H\times W\times 2}$. Each pixel position of the offset contains horizontal and vertical offsets. Based on the two-dimensional offset $\boldsymbol{F}_o$, the spatial transformation function $warp$ is performed on the deep feature $\boldsymbol{F}_d$ to obtain the feature $\boldsymbol{F}_w$, and splice it with shallow feature $\boldsymbol{F}_s$. To obtain the global information of channels, the 3×3 convolution is used on the spliced features to obtain feature $\boldsymbol{F}_u$, then the global average pooling is used to obtain the global information of each channel, and the 1×1 convolution is used to increase the linear correlation of channel information. Finally, the sigmoid function is used to establish the channel correlation weight parameter $\boldsymbol{F}_m$ and act on $\boldsymbol{F}_u$ to obtain the consistent alignment feature $\boldsymbol{F}_A$. The calculation process of feature consistency alignment is as follows.

$$\boldsymbol{F}_o = f_{3\times 3}(concat(\boldsymbol{F}_s, upsample(f_{1\times 1}(\boldsymbol{F}_d)))) \tag{18}$$

where $\boldsymbol{F}_o$ represents the two-dimensional offset, and $f_{1\times 1}$ and $f_{3\times 3}$ represents 1×1 and 3×3 convolution operations. Based on the offset $\boldsymbol{F}_o$, the deep feature $\boldsymbol{F}_d$ is spatially transformed to obtain the feature $\boldsymbol{F}_w$.

$$\boldsymbol{F}_w = warp(\boldsymbol{F}_h, \boldsymbol{F}_o) \tag{19}$$

where $warp(\cdot)$ represents the spatial transformation function. The spatial transformation feature $\boldsymbol{F}_w$ is spliced with the shallow feature $\boldsymbol{F}_s$, and the splicing feature $\boldsymbol{F}_u$ is obtained by using 3×3 convolution.

$$\boldsymbol{F}_u = f_{3\times 3}(concat(\boldsymbol{F}_w, \boldsymbol{F}_s)) \tag{20}$$

where $concat(\cdot)$ represents the splicing function. The global average pooling is used to obtain the global channel feature of the splicing feature $\boldsymbol{F}_u$, and then the channel dimension is adjusted by 1×1 convolution, and the sigmoid function is used to obtain the global channel feature $\boldsymbol{F}_m$. The global channel feature $\boldsymbol{F}_m$ and the splicing feature $\boldsymbol{F}_u$ are fused to obtain the global consistency feature $\boldsymbol{F}_A$.

$$\boldsymbol{F}_m = sigmoid(f_{1\times 1}(\mathrm{ReLU}(f_{1\times 1}(GAP(\boldsymbol{F}_u))))) \tag{21}$$

$$\boldsymbol{F}_A = \boldsymbol{F}_u \oplus (\boldsymbol{F}_m \otimes \boldsymbol{F}_u) \tag{22}$$

where $\mathrm{ReLU}(\cdot)$ represents the activation function, $GAP(\cdot)$ represents the global average pooling, $\boldsymbol{F}_m$ represent the global channel feature, $\boldsymbol{F}_A$ represent the global consistency fusion feature, $\oplus$ represents the matrix addition, and $\otimes$ represents the matrix multiplication.

3.4. Universal Adversarial Training

As an effective means against adversarial example attacks, adversarial training [28] can effectively improve the model's robustness against adversarial noise. The principle of

adversarial training is to solve the maximin optimization problem. The formal description is as follows,

$$\min_{\theta} \max_{x^{adv}:\left\|x^{adv}-x\right\|_{\infty}\leqslant\varepsilon} \mathcal{L}\left(h_{\theta}\left(x^{adv}\right), y_{true}\right) \tag{23}$$

where x^{adv} represents the adversarial example, x represents the clean example, ε represents the adversarial perturbations, $\mathcal{L}(\cdot)$ represents the loss function, $h(\cdot)$ represents the training model with parameter θ, and y_{true} represents the ground truth corresponding clean examples. For our semantic segmentation model, cross-entropy is used as a semantic segmentation loss function. The cross-entropy loss is calculated as follows.

$$\mathcal{L} = -\sum_{c=1}^{M} y_c \log(p_c) \tag{24}$$

where M represents the number of object categories, y_c represents the indicator variable (0 or 1), and p_c represents the probability that the predicted result belongs to the cth category.

The adversarial training process adds adversarial perturbation ε to the clean examples to approximate the internal maximin of the loss function and then optimize the model parameter θ. Inspired by previous work in [28,50], we proposed a universal adversarial training strategy, which uses StepLL attack [29] to generate adversarial examples for adversarial training and introduces Gaussian perturbations in the process of adversarial training to improve the robustness of the model against different adversarial example attacks. The proposed universal adversarial training process is shown in Algorithm 1, where $N\left(\mu, \sigma^2\right)$ is the Gaussian distribution with mean μ and variance σ^2.

Algorithm 1 Universarial Adversarial Training.

Input: adversarial training times T, iteration times I, training set label number M, training set $\left(x_i, y_i^{true}\right)$, target label y_i^{target}, perturbation ε, iteration perturbation stride α, mean μ, variance σ^2.

Output: model parameter θ.

for $t \leftarrow 1$ to T **do**
  for $i \leftarrow 1$ to M **do**
    $x_i^{adv} = x_i + \max\left(\min\left(N\left(\mu, \sigma^2\right)/255, \varepsilon\right), -\varepsilon\right)$
    for $j \leftarrow 1$ to I **do**
      $x_i^{adv} = x_i^{adv} - \alpha \cdot f\left(x_i^{adv}\right)$
      $f\left(x_i^{adv}\right) = \text{sign}\left(\nabla_{x_i^{adv}} L\left(h_{\theta}\left(x_i^{adv}, y_i^{target}\right)\right)\right)$
      $x_i^{adv} = \max\left(\min\left(x_i^{adv}, x_i + \varepsilon\right), x_i - \varepsilon\right)$
    end for
    $\theta = \theta - \nabla_{\theta} L\left(h_{\theta}\left(x_i^{adv}\right), y_i^{true}\right)$
  end for
end for

4. Experiments and Analysis

4.1. Dataset Information

To verify the effectiveness and feasibility of the proposed method, we conducted experiments on three UAV aerial image benchmark datasets, including the UAVid dataset [2], Semantic Drone dataset [70], and Aeroscapes dataset [71].

The details of these datasets are as follows.

UAVid is a challenging benchmark dataset for UAV aerial image semantic segmentation, which contains many static and moving objects in complex urban scenes. This dataset is captured using the resolution video recording mode, and the image sizes are 3840×2160 pixels. Since UAVid is mainly collected from urban scenes, the dataset contains eight object classes in urban scenes, namely building, road, tree, low-vegetation, moving-car, static-car,

background-clutter, and human. There are 420 images in the dataset, of which we use 200 images for training, 70 images for validation, and the remaining 150 images for testing. Due to the large size of the original image, the image size is scaled to 1024 × 1024 pixels. The sample image and corresponding labels are shown in Figure 3.

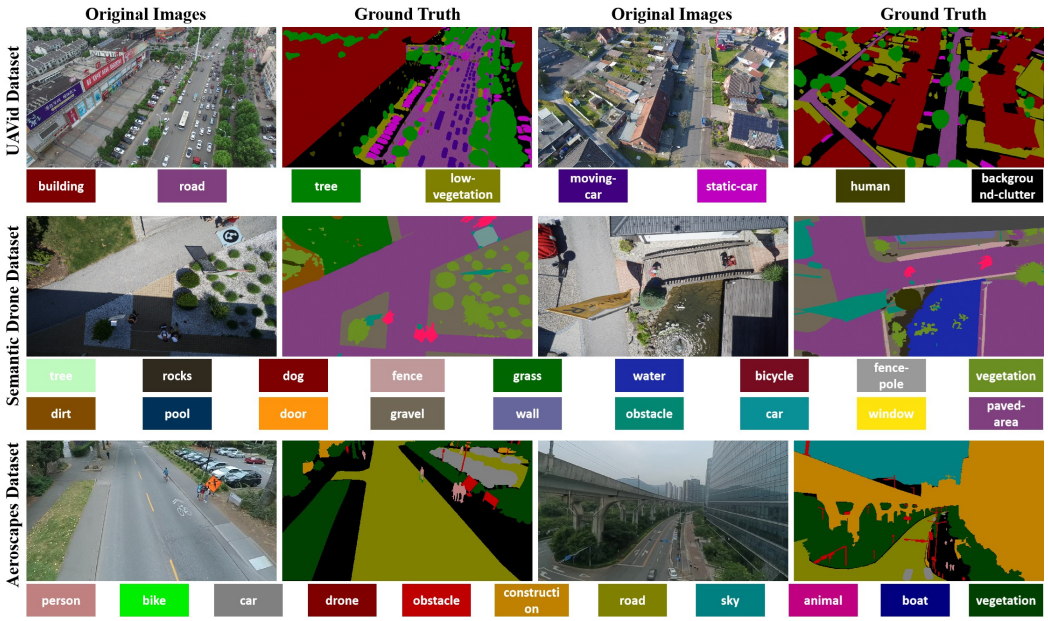


Figure 3. Example images and corresponding ground truth from the UAVid, Semantic Drone, and AeroScapes datasets. The first row shows the UAVid [2] dataset. The second row shows the Semantic Drone [70] dataset. The third row shows the AeroScapes [71] dataset.

Semantic Drone focuses on semantic understanding of urban scenes; this dataset observes ground objects from the bird's eye perspective at an altitude of 5 to 30 meters. The high-resolution camera is used to capture images at a size of 6000 × 4000 pixels, and the dataset contains eighteen classes of ground objects, such as trees, rocks, dogs, fences, grass, water, bicycle, fence-pole, vegetation, dirt, pool, door, gravel, wall, obstacle, car, window, and paved-area. The original image and label are shown in Figure 3. This dataset contains 400 publicly available images; 280 images in the dataset are used as the training set, 40 images as the validation set, and 80 images as the testing set. To facilitate training, we crop the original image size from 6000 × 4000 pixels to 2048 × 1024 pixels.

AeroScapes dataset is more challenging for semantic segmentation tasks because it includes the ground objects in complex urban and suburban scenes. The AeroScapes dataset contains 3269 images and eleven categories of ground objects, namely person, bike, car, drone, boat, animal, obstacle, construction, vegetation, road, and sky. As shown in Figure 3, the number of pixels for different object categories in the dataset varies greatly. The image size in the dataset is 1280 × 720 pixels, and we maintain the original image size constant during the training process. For the 3269 images contained in the dataset, we use 2288 images as the training set, 654 images as the validation set, and the remaining 327 images for testing.

4.2. Experimental Setup and Implementation Details

To verify the robustness of the proposed method against adversarial example attacks, we use FGSM [24], C&W [27], and PGD [28] attack methods to construct corresponding adversarial example test sets. Specifically, for UAVid [2] dataset, we use FGSM to construct adversarial examples, and for Semantic Drone [70], and AeroScapes [71] datasets, we use C&W and PGD attacks. We adopt the FGSM with ℓ_∞ norm to conduct untargeted adversarial attacks using Equation (2), where the ε is fixed to 0.04. For C&W attack, we use the ℓ_0 norm to conduct un-targetted adversarial attacks using Equation (5), the number of

iterations T is fixed to 10, and the parameter c is set to 0.1. Equation (6) generates PGD adversarial examples, where the perturbation constraint ε is set to 0.01, iteration stride μ is set to 2, and iteration number T is fixed to 20.

To ensure the credibility of the experimental results, we randomly selected images in the dataset to form the training set, validation set, and testing set and repeated the experimental process 10 times. In addition, we use data augmentation methods such as random inversion, size cropping, and brightness transformation to increase the number of dataset samples. We implement the Pytorch platform to build the proposed semantic segmentation network, and the experiments are carried out with an Intel i9-12900T CPU with 64 GB RAM, NVIDIA GTX GeForce 3090 GPU, and Ubuntu 18.04 operating system. The training epochs are set as 1000, and the batch size is set to 8. For model training optimization, set the initial learning rate to 0.001, and SGD is used with a momentum of 0.9 as the optimizer. In addition, we give detailed steps to perform adversarial attacks against the proposed GFANet, as shown in Algorithm 2. The goal of adversarial attacks on aerial image semantic segmentation tasks is to use adversarial noise to interfere with the original image to maximize the number of misclassification of all test pixels in the image.

Algorithm 2 Adversarial Attack on GFANet.

Input:
Aerial image x and corresponding ground truth y.
Semantic segmentation model f with parameters θ.
Adversarial perturbation amplitude ε, training epochs τ, and learning rate η.
Output: The predictions on the adversarial example x_{adv}.
Initialize model parameters θ with uniform distribution.
for t in $range(0, \tau)$ **do**
    Compute the global context features $\boldsymbol{E}$ via (11).
    Compute the coordinate attention features $\boldsymbol{Y}$ via (17).
    Computer the global consistency features $\boldsymbol{F}$ via (22).
    Computer the cross-entropy loss $\mathcal{L}$ via (24).
    Update θ by descending its stochastic gradients.
end for
Generate the adversarial image x_{adv} via (2), (4), (5).
Feed the adversarial image x_{adv} to the trained model f to achieve the segmentation.

4.3. Evaluate Metrics

To compare the performance of the semantic segmentation network, we adopt the PA, mPA, F1_score, and mIoU as evaluation metrics. First, we define tp, fp, fn, and tn as true positives, false positives, false negatives, and true negatives.

- The PA is the basic evaluate metric in semantic segmentation, which is defined as the correctly classified pixel in all pixels as $PA = (tp + tn)/(tp + tn + ft + fn)$.
- The mPA is the mean of the sum of category pixel accuracy (cPA), where $cPA = tp/(tp + fp)$ represents the correct proportion of predicted category ith pixels.
- The F1_score is the geometric mean between the precision (P) and recall (R) of each class as $F1\_score{=}2 \times \frac{P \times R}{P+R}$, where $P{=}tp/(tp + fp)$ and $R{=}tp/(tp + fn)$.
- The mIoU is defined as the mean of IoU, and the IoU is calculated as $IoU = |P_i \cap G_i|/|P_i \cup G_i|$. P_i and G_i are the set of prediction pixels and ground truth pixels for the ith class.

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

For experimental comparison, we verify the model semantic segmentation performance and the robustness against adversarial attacks. On the UAVid dataset, we compare the GFANet with the existing aerial image semantic segmentation networks LANet [10], AERFC [11], and AFNet [12]. For the Semantic Drone, we compare the proposed method

with the MCLNet [13], BSNet [14], and SBANet [15]. For the AeroScapes dataset, the GFANet is compared with MANet [16], HPSNet [17], and TCHNet [72].

Compare on UAVid Dataset: First, we compared performance on the clean example test set, and the quantitative results and visual comparisons are shown in Table 1 and Figure 4. Second, the robustness against adversarial attacks is verified on the adversarial example test set generated by FGSM attack [24], and the results are shown in Table 2 and Figure 5. Next, we give the performance analysis and robustness against attacks of different methods.

Table 1. Comparison of evaluation metrics on clean example test set in the UAVid dataset.

Methods	Per-Class IoU (%)								Evaluate Metrics (%)			
	Building	Road	Tree	Low.veg.	M.car	S.car	Human	Clutter	PA	mPA	mF1	mIoU
LANet	81.39	76.35	77.48	68.34	71.72	63.33	31.15	62.47	87.24	78.53	85.74	66.52
AERFC	83.25	80.62	78.51	66.96	75.18	67.85	36.53	65.42	88.07	79.85	86.43	69.28
AFNet	82.26	80.95	77.41	68.03	76.84	67.11	38.71	66.46	88.41	80.98	87.15	70.47
GFANet	84.72	82.77	79.32	70.25	77.31	70.92	41.26	68.57	89.28	82.41	88.54	71.89

Table 2. Comparison of evaluation metrics on adversarial example test set in the UAVid dataset.

Methods	Per-Class IoU (%)								Evaluate Metrics (%)			
	Building	Road	Tree	Low.veg.	M.car	S.car	Human	Clutter	PA	mPA	mF1	mIoU
LANet	17.12	21.43	22.21	16.56	16.21	12.64	6.15	14.52	25.42	19.46	22.61	15.85
AERFC	12.23	22.57	17.23	10.52	18.37	15.28	7.32	20.17	23.17	18.23	20.17	15.46
AFNet	21.45	24.32	19.75	17.28	20.63	18.52	9.75	26.34	26.73	21.75	24.36	19.75
GFANet	81.63	78.52	77.16	68.43	75.35	68.14	39.57	67.26	87.65	79.86	86.45	69.51

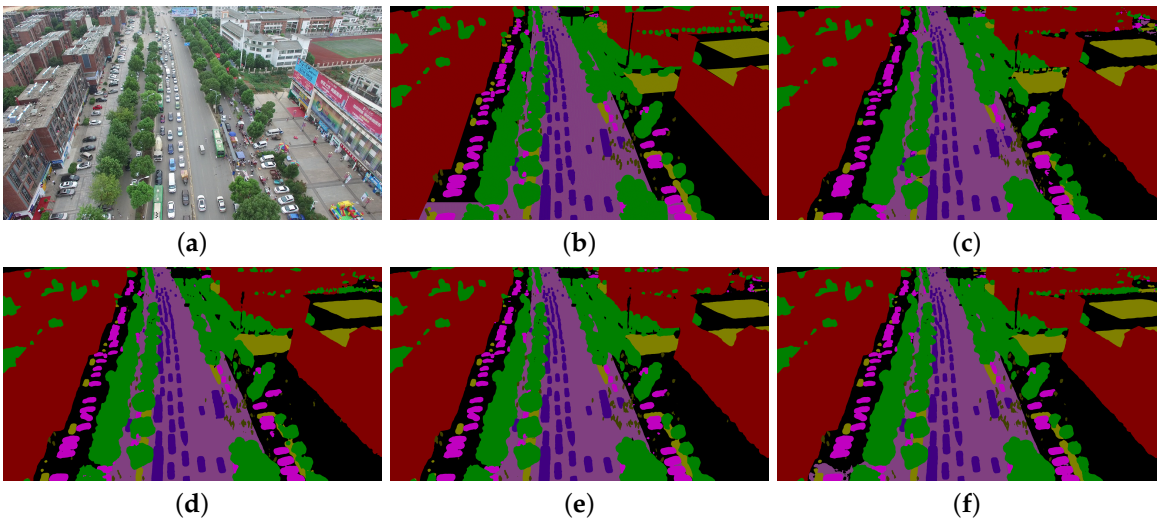


Figure 4. Visualization results of different methods on clean example test set in UAVid dataset. (**a**) Original Images. (**b**) Ground Truth. (**c**) LANet [10]. (**d**) AERFC [11]. (**e**) AFNet [12]. (**f**) GFANet.

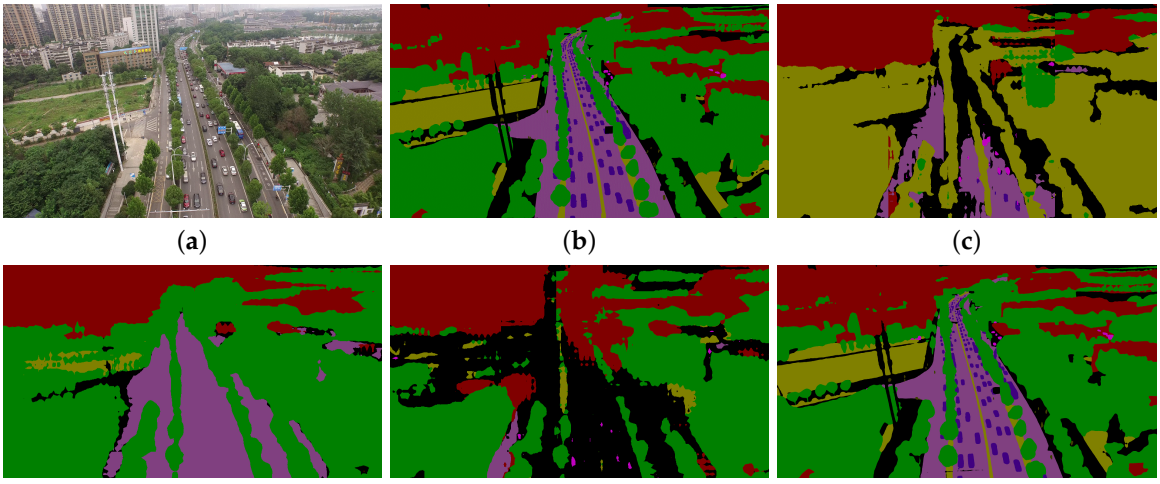
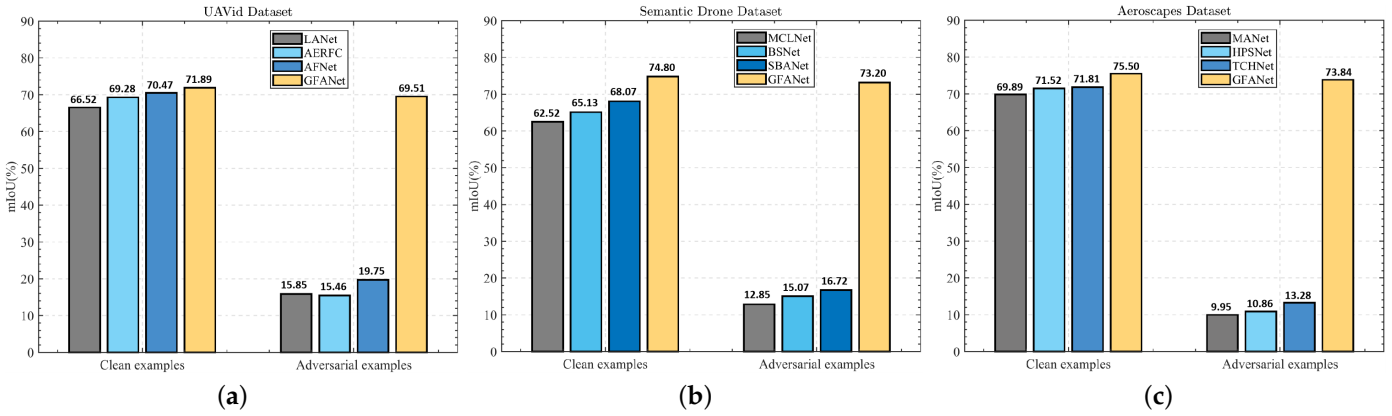


Figure 5. Visualization results of different methods on adversarial example test set in UAVid dataset. (**a**) Adversarial Images. (**b**) Ground Truth. (**c**) LANet [10]. (**d**) AERFC [11]. (**e**) AFNet [12]. (**f**) GFANet.

(1) LANet [10]: This network consists of a patch attention mechanism and attention embedding module, which can mine local feature information of the ground objects to guide the model to complete semantic segmentation. As shown in Table 1, LANet achieves 66.52% mIoU on the clean example test set, while only 15.85% mIoU is completed on the adversarial example test set. The visualization results in Figures 4 and 5 show that LANet can better predict the pixels of each category for clean examples, which for adversarial samples, there are serious mistakes, such as the "tree" is misclassified as "low-vegetation". The results in Figure 6 further shows the performance difference between LANet for clean and adversarial examples, with the mIoU decreasing by 50.67%. The experimental results also further demonstrate the poor performance of local features against adversarial attacks.



(**a**) (**b**) (**c**)

Figure 6. Quantitative comparison results of calean examples and adversarial examples on different datasets. (**a**) UAVid dataset [2]. (**b**) Semantic Drone dataset [70]. (**c**) Aeroscapes dataset [71].

(2) AERFC [11]: To achieve accurate segmentation of different scale objects, AERFC constructs the adaptive convolution kernel to extract multi-scale feature information of ground objects. The results in Table 1 show that AERFC has a better semantic segmentation effect on different categories of objects; for example, its mPA and mIoU reach 79.85% and 69.28%, respectively. For adversarial examples, the results in Table 2 show that the mPA and mIoU of AERFC only reach 18.23% and 15.46%. The visualization results in Figures 4 and 5 further show the performance difference of AERFC on clean examples and adversarial examples. For example, for the clean sample test set, AERFC achieves better prediction for different categories of object pixels, while its performance on the adversarial sample test

set is significantly degraded. The experimental results of AERFC show that multi-scale features cannot be against the impact of adversarial attacks.

(3) AFNet [12]: For the purpose of feature enhancement, AFNet constructs the scale-feature attention mechanism and scale-layer attention module, which achieves semantic segmentation by enhancing features of different scales and different convolution layers. From Table 1, we can observe that AFNet has better semantic segmentation performance on the clean sample test set, while the results of Table 2 show that AFNet performs poorly on the adversarial sample test set. The visualization results also show the performance difference of AFNet on clean samples and adversarial samples. For example, in Figure 4, AFNet can accurately predict the object "road", while for adversarial examples, "road" is misclassified as "background-clutter". It can be seen from Figure 6 that the mIoU of AFNet decreased from 70.47% of clean samples to 19.75% of adversarial samples. The experimental results of AFNet show that simple feature enhancement cannot alleviate the impact of adversarial samples on model performance.

For our proposed GFANet, it can be seen from Tables 1 and 2 that GFANet achieves the best results on both clean and adversarial sample test sets. For clean samples, the mIoU reaches 71.89%, while for adversarial samples, its mIoU reaches 69.51%. The visualization results of Figures 4 and 5 also prove that GFANet can complete accurate semantic segmentation for clean and adversarial samples. From Figure 6, it can be observed that the mIoU difference between GFANet for clean and adversarial samples is only 2.38%, which further indicates the robustness of GFANet against adversarial example attacks. The experimental results of GFANet show that the global features can complete accurate aerial image semantic segmentation tasks and have strong robustness against adversarial attacks.

Compare on Semantic Drone Dataset: Since the dataset contains more object categories and complex scenes, it can further verify the semantic segmentation accuracy and the robustness against adversarial attacks of different methods. We use C&W attack [27] with l_0 norm to generate an adversarial example test set. Table 3 and Figure 7 show the experimental results of different methods on the clean sample test set, and Table 4 and Figure 8 show the results on the adversarial example test set. Next, we analyze the experimental results of different methods in detail.

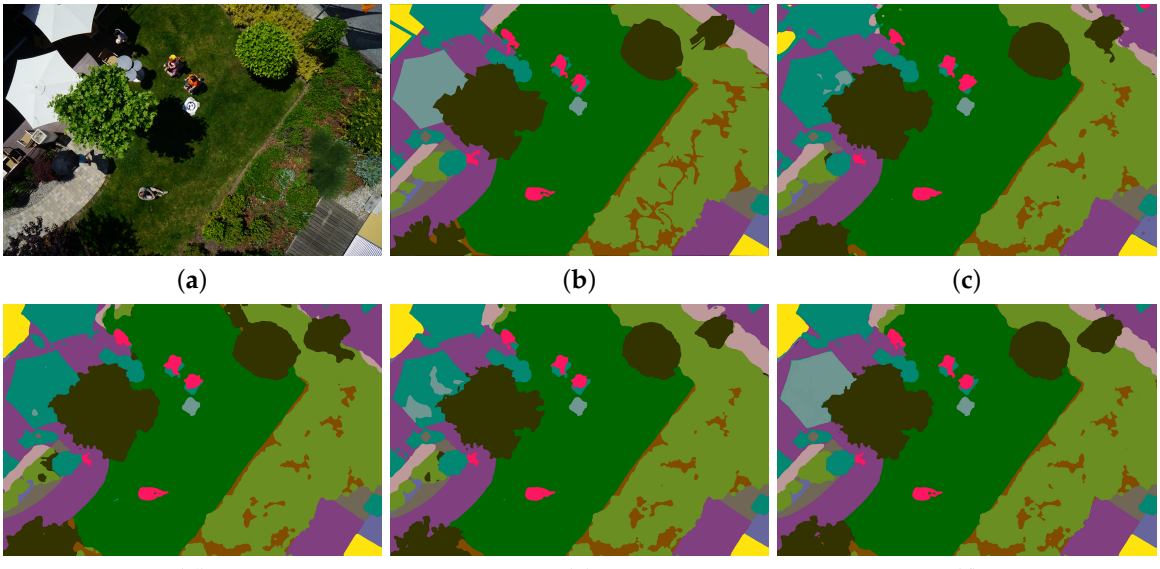


Figure 7. Visualization results of different methods on clean example test set in Semantic Drone dataset. (**a**) Original Images. (**b**) Ground Truth. (**c**) MCLNet [13]. (**d**) BSNet [14]. (**e**) SBANet [15]. (**f**) GFANet.

(1) MCLNet [13]: To enhance the correlation between multi-scale features, MCLNet constructs the multi-scale calibration learning strategy. The network performs semantic segmentation by mining the correlation between local and global features. The experimental results on the clean sample test set in Table 3 and Figure 7 show that MCLNet can better segment objects of different scales, and its mAP and mIoU reach 73.81% and 62.52%. However, for the adversarial example test set, as shown in Table 4, the performance of MCLNet is significantly degraded, with mAP and mIoU 23.16% and 12.85%. From the visualization results of Figure 8, it can be seen that the adversarial attack has a great impact on the performance of MCLNet, and it cannot complete accurate semantic segmentation on the adversarial sample test set. It can also be observed from Figure 6 that the adversarial example attack reduces the mIoU of MCLNet by 39.36%. The experimental results further illustrate that only establishing the correlation between local and global features is ineffective against adversarial example attacks.

Table 3. Comparison of evaluation metrics on clean example test set in Semantic Drone dataset.

Methods	**Per-Class IoU (%)**										
	Tree	**Rocks**	**Dog**	**Fence**	**Grass**	**Water**	**Bicycle**	**Pole**	**Vegetation**	**Dirt**	**Pool**
MCLNet	62.28	55.76	45.48	59.53	73.40	82.43	67.38	11.25	75.33	50.84	87.95
BSNet	74.14	64.40	55.22	59.65	78.43	77.22	65.13	18.57	73.54	52.25	89.48
SBANet	73.82	60.86	62.46	60.02	84.63	86.81	65.74	23.43	76.38	53.85	88.41
GFANet	75.83	68.75	75.92	64.69	94.72	92.76	72.43	35.17	78.69	62.17	96.35
Methods	door	gravel	wall	obstacle	car	window	paved	PA	mPA	mF1	mIoU
MCLNet	15.62	72.02	66.37	73.85	84.69	55.83	85.44	82.14	73.81	79.26	62.52
BSNet	17.87	80.75	65.44	70.25	83.74	52.37	91.58	84.62	74.29	81.52	65.13
SBANet	21.35	83.71	70.25	71.95	86.41	59.32	95.94	85.37	76.82	83.37	68.07
GFANet	32.58	84.52	74.26	76.99	94.80	68.73	96.87	91.28	84.73	88.46	74.80

Table 4. Comparison of evaluation metrics on adversarial example test set in Semantic Drone dataset.

Methods	**Per-Class IoU (%)**										
	Tree	**Rocks**	**Dog**	**Fence**	**Grass**	**Water**	**Bicycle**	**Pole**	**Vegetation**	**Dirt**	**Pool**
MCLNet	12.13	8.75	6.52	10.27	21.54	19.75	11.32	2.14	18.62	7.45	16.48
BSNet	16.87	11.35	10.23	11.58	19.74	17.82	10.16	5.83	21.52	8.64	24.62
SBANet	14.79	6.47	12.58	13.74	23.57	15.26	9.13	8.52	24.57	11.85	22.47
GFANet	72.15	67.24	73.59	63.52	92.38	91.57	71.96	33.82	76.23	62.08	95.22
Methods	door	gravel	wall	obstacle	car	window	paved	PA	mPA	mF1	mIoU
MCLNet	1.65	18.37	13.72	15.25	20.46	6.42	20.48	29.15	23.16	26.73	12.85
BSNet	3.57	26.52	15.23	17.94	18.35	8.79	22.51	34.78	26.35	29.64	15.07
SBANet	7.48	28.16	21.57	19.62	22.75	9.84	28.63	35.24	27.68	32.57	16.72
GFANet	30.74	82.97	72.18	75.23	93.52	67.15	95.87	89.72	82.56	87.34	73.20

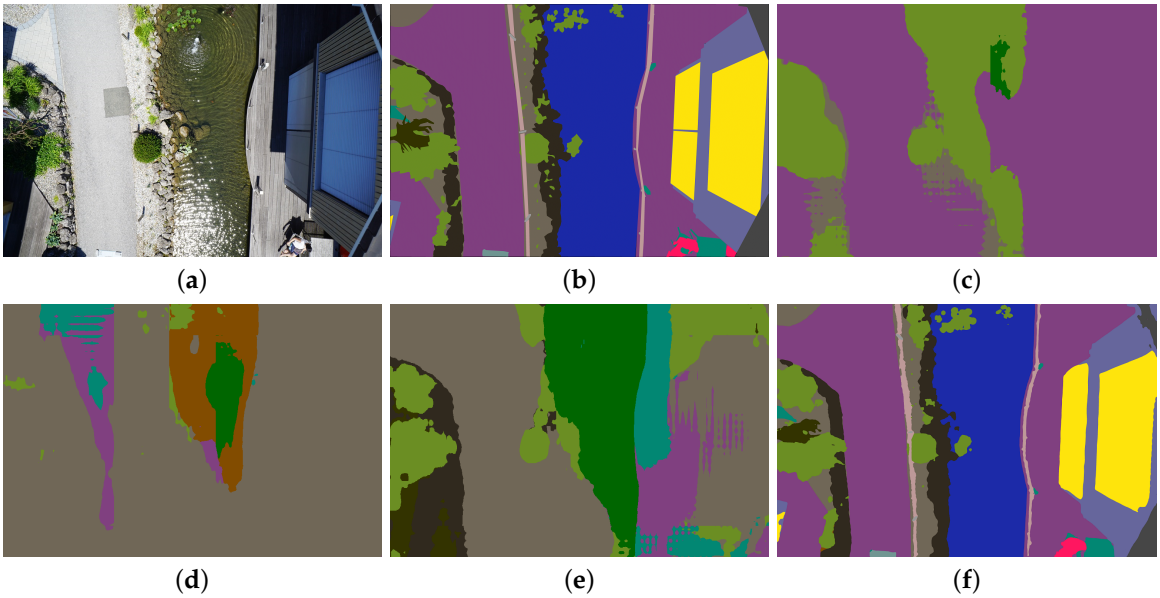


Figure 8. Visualization results of different methods on adversarial example test set in Semantic Drone dataset. (**a**) Adversarial Images. (**b**) Ground Truth. (**c**) MCLNet [13]. (**d**) BSNet [14]. (**e**) SBANet [15]. (**f**) GFANet.

(2) BSNet [14]: This network consists of dynamic hybrid gradient convolution and coordinates sensitive attention, which completes semantic segmentation by obtaining the salient boundary information of the object region. As shown in Table 3, the mPA and mIoU of BSNet are 74.29% and 65.13%, which shows the contribution of boundary feature information in accurate semantic segmentation. From the visualization results of Figure 7, it can be seen that BSNet can finely segment the contour boundary. For the adversarial example test set, as shown in Table 4, the mIoU of BSNet on the adversarial example test set is only 26.35% and 15.07%, which is obviously inferior to the experimental results on the clean example test set. From Figure 6, it can be seen that the mIoU of BSNet decreased from 65.13% to 15.07%. The results of Figure 8 further prove the impact of adversarial examples on the performance of BSNet, which cannot complete the semantic segmentation task under adversarial example attacks. The results of BSNet also show that only enhancing boundary features cannot alleviate the impact of adversarial examples on model performance.

(3) SBANet [15]: To obtain the fine-grained semantic features of the object region, SBANet uses the boundary attention mechanism to locate the object region and uses the adaptive weighted multi-task learning guidance model to complete the semantic feature extraction. As shown in the clean example experiment results in Table 4 and Figure 7, SABNet obtained 76.82% and 68.07% of mAP and mIoU and completed accurate semantic segmentation for different object categories. However, the experimental results of Table 4 and Figure 8 show that SBANet is ineffective against adversarial example attacks. The mIoU of SBANet on the adversarial example test set is only 16.72%, and there are pixel classification errors, such as the "water" is misclassified as "vegetation". The results in Figure 6 show that the adversarial examples reduce the mIoU from 68.07% to 16.72%. The experimental results of SBANet verify that semantic features are ineffective against adversarial attacks.

As shown in Tables 3 and 4, GFANet achieves the mIoU of 74.80% and 73.20% on clean example and adversarial example test sets, which is superior to other compared methods. The visualization results of Figures 7 and 8 show that GFANet can complete accurate semantic segmentation and effectively alleviate the impact of adversarial example attacks.

Compare on Aeroscapes Dataset: The dataset contains many suburban scenes and has higher resolution and fine annotation information, which can effectively verify the robustness and generalization ability of the semantic segmentation network. For the

adversarial example attack, we use the PGD attack [28] to generate the adversarial example test set. Figures 9 and 10 show the visual comparison results. The specific performance analysis of different methods is as follows. Correspondingly, Tables 5 and 6 show the quantitative comparison results of different methods on clean example and adversarial example test sets.

(1) MANet [16]: This network uses a multi-attention cascade to obtain multi-scale context features and uses dot-product attention for feature fusion. MANet effectively alleviates the feature loss problem in the feature fusion process. As shown in Figure 9 and Table 5, for the clean example test set, MANet obtains complete object region contour information and achieves accurate pixel classification for different category objects, with mPA and mIoU of 81.36% and 69.89%. However, for the adversarial example test set, the quantitative comparison results in Table 6 show that the mIoU of MANet only reaches 9.95%. From Figure 9, it can be observed that MANet misclassified "vegetation" as "road" on the adversarial example test set, indicating that the adversarial example seriously damaged the model performance. The experimental results of MANet show that only using context information cannot effectively resist the interference of adversarial examples.

Table 5. Comparison of evaluation metrics on the clean example test set in the AeroScapes dataset.

Methods	Per-Class IoU (%)											Evaluate Metrics (%)			
	Person	**Bike**	**Car**	**Drone**	**Boat**	**Animal**	**Obs.**	**Cons.**	**Veg.**	**Road**	**Sky**	**PA**	**mPA**	**mF1**	**mIoU**
MANet	75.86	56.72	72.83	53.86	62.78	48.52	72.43	69.53	81.57	85.42	89.27	88.75	81.36	87.15	69.89
HPSNet	81.42	60.45	68.23	58.42	65.37	45.26	70.54	73.68	84.95	87.06	91.28	89.41	82.45	88.36	71.52
TCHNet	78.95	62.37	71.12	61.53	67.26	43.57	68.15	70.93	86.04	89.48	90.57	89.53	83.06	88.75	71.81
GFANet	82.14	65.29	73.46	62.38	68.59	57.32	74.98	75.64	87.25	91.06	92.43	92.87	85.23	90.05	75.50

Table 6. Comparison of evaluation metrics on the adversarial example test set in the AeroScapes dataset.

Methods	Per-Class IoU (%)											Evaluate Metrics (%)			
	Person	**Bike**	**Car**	**Drone**	**Boat**	**Animal**	**Obs.**	**Cons.**	**Veg.**	**Road**	**Sky**	**PA**	**mPA**	**mF1**	**mIoU**
MANet	12.39	2.73	6.94	3.54	8.57	7.46	6.98	7.15	15.84	16.43	21.45	36.72	27.09	32.85	9.95
HPSNet	11.54	4.36	9.26	5.71	12.58	8.43	5.45	4.62	17.35	20.76	19.42	38.51	29.32	34.42	10.86
TCHNet	17.21	9.45	11.34	8.66	14.26	9.68	4.01	8.69	20.66	23.68	18.46	42.78	32.57	40.36	13.28
GFANet	81.43	64.97	72.15	60.83	66.87	55.92	72.46	73.84	85.60	88.23	90.04	88.92	81.65	86.53	73.84

(2) HPSNet [17]: To establish the correlation between different features, HPSNet constructs the hidden path selection strategy, which completes accurate semantic segmentation by correlation modeling and global connection of different features. For the clean example test set, HPSNet obtains 82.45% and 71.52% of mPA and mF1. The results in Figure 9 show that HPSNet can obtain accurate semantic segmentation results by establishing the relationship between different object features. For the adversarial example test set, as shown in Table 6 and Figure 10, HPSNet is seriously affected by the adversarial example, with mPA and mIoU of 29.32% and 10.86%. The experiment of HPSNet further verifies that simply establishing the correlation between features cannot improve the resistance to adversarial attacks.

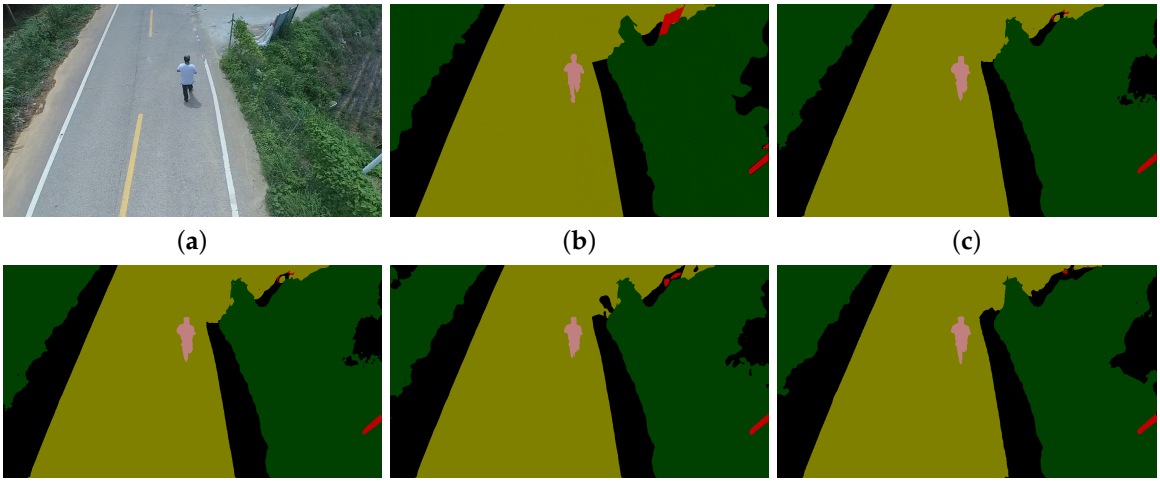


Figure 9. Visualization results of different methods on the clean example test set in the AeroScapes dataset. (**a**) Original Images. (**b**) Ground Truth. (**c**) MANet [16]. (**d**) HPSNet [17]. (**e**) TCHNet [70]. (**f**) GFANet.

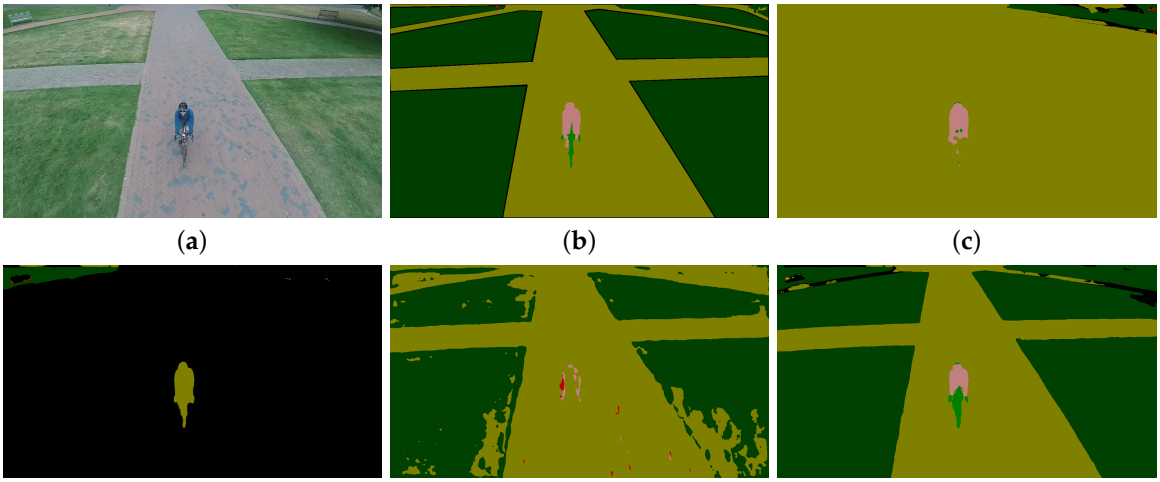


Figure 10. Visualization results of different methods on the adversarial example test set in the AeroScapes dataset. (**a**) Original Images. (**b**) Ground Truth. (**c**) MANet [16]. (**d**) HPSNet [17]. (**e**) TCHNet [70]. (**f**) GFANet.

(3) TCHNet [72]: This network consists of atrous spatial pyramid pooling and channel attention mechanism, which realizes semantic segmentation by extracting fine-grained spatial structure features and enhancing local channel features. As shown in Figure 9 and Table 5, TCHNet completes the accurate segmentation of different category objects, and its mAP and mIoU reach 83.06% and 71.81%, indicating that TCHNet has better semantic segmentation performance. However, for the adversarial example test set, as shown in Table 6, the mPA and mIoU of TCHNet are only 32.57% and 13.28%. The visualization results in Figure 10 illustrate that adversarial examples have a serious impact on its segmentation performance. The experiment shows that the channel attention mechanism or feature enhancement strategy can not resolve adversarial examples to the model performance.

Our proposed GFANet, as shown in Figure 9 and Table 5, obtains the best results on the clean example test set and accomplishes the accurate segmentation of different object categories. For the adversarial example test set, as shown in Figure 10 and Table 6, GFANet maintains the same performance as the clean example test set, and the results in Figure 6 show that for clean and adversarial examples, the difference in mIoU of

GFANet is only 1.66%, further illustrating the performance advantage and robustness of the proposed method.

4.5. Ablation Study

The proposed GFANet consists of a global context encoder (GCE), global coordinate attention mechanism (GCAM), feature consistency alignment (FCA), and universal adversarial training. In this subsection, we verify the contribution of each component to improving the robustness against adversarial attacks. The FGSM with l_∞ norm is used to generate an adversarial example test set by Equation (2), where the perturbation is fixed to 0.04. In addition, we use SegNet [8] as the baseline and gradually add different components.

The experimental results with different components are shown in Table 7, where UAT represents the universal adversarial training strategy. It can be observed from Table 7 that the combination of different components can significantly improve the robustness of the baseline network against adversarial example attacks. Specifically, the global context encoder is more beneficial for improving adversarial robustness, and using GCE can increase the PA values of baseline network on three datasets by 27.41%, 29.20%, and 29.74%, respectively. Moreover, using results in the UAVid dataset, for example, while GCAM enables the baseline to yield the PA of 63.54%, the use of FCA can increase the PA to 79.83%, and the adoption of UAT can increase the PA to 89.28%. The results in Table 7 demonstrate that combining different components can make the baseline network achieve optimal performance and further illustrate that global features can effectively improve the adversarial robustness.

Table 7. Performance of each component in GFANet for different datasets (reported in PA).

Baseline	GCE	GCAM	FCA	UAT	UAVid	Semantic Drone	Aeroscapes
✔					21.35	23.18	25.72
✔	✔				48.76	52.37	55.46
✔	✔	✔			63.54	66.72	69.28
✔	✔	✔	✔		79.83	82.54	85.93
✔	✔	✔	✔	✔	89.28	89.63	91.47

5. Discussion

We further analyze the influence of different levels of adversarial perturbation on the semantic segmentation performance of the proposed GFANet. The adversarial examples with different adversarial intensity values are generated using Equation (2), where the perturbation value ε is set to $\{0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ in the experiments. Figure 11 shows the experimental results of all methods on three datasets with different adversarial perturbations. From the experimental results, it can be seen that with the increase in the adversarial perturbation, the PA values of different comparison methods tend to decrease, which indicates that the larger the perturbation amplitude, the more serious the damage to the model performance. Comparing the existing state-of-the-art methods LANet and AERFC, the proposed GFANet shows strong robustness against adversarial example attacks under different levels of adversarial perturbation. In particular, when the adversarial perturbation takes the maximum value, the PA of different compared methods are less than 25%, while the proposed GFANet still obtains more than 70% PA values on three benchmark datasets. The experiment further proves the robustness of the GFANet against adversarial attacks.

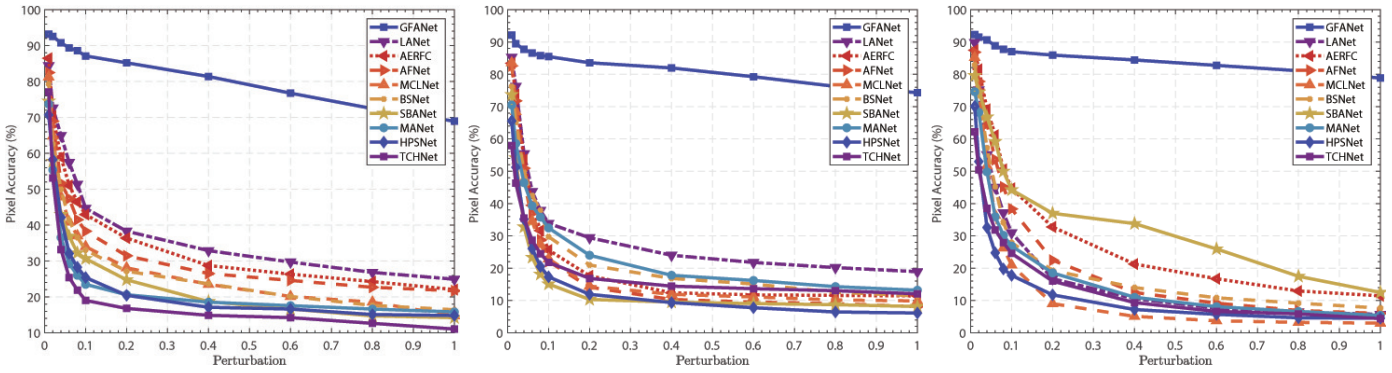


Figure 11. The pixel accuracy of different methods on the adversarial example test set with different perturbation values. (**a**) UAVid dataset [2]. (**b**) Semantic Drone dataset [70]. (**c**) Aeroscapes dataset [71].

To further verify the robustness of the proposed method against different types of adversarial example attacks, we conduct experiments with different adversarial attacks, including FGSW, JSMA, C&W, PGD, and UAP. Moreover, the UAVid dataset is used as the experimental data. For parameter setting of different adversarial attack methods, we run 15 iterations of JSMA, C&W, PGD, and UAP with the ℓ_0 norm, the iteration stride is set to 2, and the perturbation constraint is fixed to 0.04. It can be seen from the experimental results in Table 8 that compared with FGSM and JSMA attacks, C&W, PGD, and UAP attacks can achieve a more powerful attack. For example, TCHNet can obtain the PA values of 24.28% on the adversarial example test set generated by FGSM attack, while only 18.57%, 17.82%, and 15.16% PA values are obtained on the adversarial example test set generated by C&W, PGD, and UAP attacks. By contrast, our proposed GFANet can still achieve more than 80% PA values on the adversarial example test set generated by different attack methods, which is significantly better than other compared aerial image semantic segmentation networks. In addition, we can observe from Table 8 that the proposed method still has a slight advantage on the clean example test set without attacks, and its PA value reaches 89.28%. The experimental results show that the proposed GFANet not only has better robustness to different adversarial example attacks but also has advantages in aerial image semantic segmentation accuracy.

Table 8. Quantitative comparison results on the clean example test sets and the adversarial example test sets generated by different attack methods (reported in PA).

Method	LANet	AERFC	AFNet	MCLNet	BSNet	SBANet	MANet	HPSNet	TCHNet	GFANet
Normal	87.24	88.07	88.41	85.32	86.73	84.86	87.59	85.43	87.15	89.28
FGSM	25.42	23.17	26.73	21.58	24.73	20.68	22.35	22.04	24.28	87.65
C&W	19.75	17.83	18.62	16.32	17.26	15.41	16.05	15.28	18.57	85.36
PGD	18.21	16.35	17.43	15.89	16.73	14.92	14.25	15.74	17.82	84.95
JSMA	26.85	24.38	27.31	22.74	25.16	22.37	23.48	24.05	25.81	88.53
UAP	17.23	15.08	15.87	14.62	14.89	13.75	14.52	13.28	15.46	84.37

6. Conclusions

Addressing adversarial example attacks faced in aerial image semantic segmentation is very important because the task is highly related to national defense security and UAV system security. In this study, we systematically analyze the threat of adversarial example attacks on CNN-based aerial image semantic segmentation methods. Although these existing CNN-based methods can achieve excellent semantic segmentation performance, this article shows that adversarial examples have a serious negative impact on these methods. To solve the adversarial example threat in aerial image semantic segmentation, we propose a novel global feature attention network (GFANet) to resist adversarial attacks. The

proposed GFANet consists of a global context encoder (GCE), global coordinate attention mechanism (GCAM), feature consistency alignment (FCA), and universal adversarial training, which enhances the defense against adversarial attacks and obtain better semantic segmentation accuracy by mining robust global context features and conducting adversarial training. Extensive experiments on three UAV aerial image semantic segmentation datasets demonstrate that the proposed exhibits stronger resistibility towards different adversarial example attacks compared with the existing CNN-based methods. Moreover, the ablation study further illustrates the contribution of each component in the proposed method to resist adversarial example attacks and improve semantic segmentation accuracy.

While this study demonstrates that the use of global features can resist the adversarial example attack faced in aerial image semantic segmentation, the segmentation effect for small object regions in aerial image adversarial examples still needs to be improved. In addition, whether there is, other feature information that can resist adversarial noise is worth further exploration. In future work, we will try to use different methods to obtain more robust feature information.

Author Contributions: Conceptualization, Z.W. and Y.L.; methodology, Z.W. and B.W.; software, J.G.; validation, Z.W. and Y.L.; formal analysis, Z.W. and J.G.; investigation, Y.L.; resources, Z.W. and J.G.; data curation, Z.W.; 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.; 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 62172338, and in part by the National Natural Science Foundation of China under Grant 61671465.

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/AttackGAINet.

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

References

1. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417–114426.
2. 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.
3. Wu, G.; Pedrycz, W.; Li, H.; Ma, M.; Liu, J. Coordinated Planning of Heterogeneous Earth Observation Resources. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *46*, 109–125.
4. Girisha, S.; Verma, U.; Manohara Pai, M.M.; Pai, R.M. UVid-Net: Enhanced Semantic Segmentation of UAV Aerial Videos by Embedding Temporal Information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4115–4127.
5. Erdelj, M.; Natalizio, E.; Chowdhury, K.R.; Akyildiz, I.F. Help from the Sky: Leveraging UAVs for Disaster Management. *IEEE Pervasive Comput.* **2017**, *16*, 24–32.
6. Nogueira, K.; Fadel, S.G.; Dourado, I.C.; de O. Werneck, R.; Munoz, J.A.V.; Penatti, O.A.B.; Calumby, R.T.; Li, L.T.; dos Santos, J.A.; Torres, R.d.S. Exploiting ConvNet Diversity for Flooding Identification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1446–1450.
7. Zhong, H.-F.; Sun, Q.; Sun, H.-M.; Jia, R.-S. NT-Net: A Semantic Segmentation Network for Extracting Lake Water Bodies From Optical Remote Sensing Images Based on Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13.
8. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
9. Zheng, C.; Zhang, Y.; Wang, L. Semantic Segmentation of Remote Sensing Imagery Using an Object-Based Markov Random Field Model With Auxiliary Label Fields. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3015–3028.
10. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 426–435.
11. 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.
12. 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.
13. He, P.; Jiao, L.; Shang, R.; Wang, S.; Liu, X.; Quan, D.; Yang, K.; Zhao, D. MANet: Multi-Scale Aware-Relation Network for Semantic Segmentation in Aerial Scenes. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.

14. 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.
15. Li, A.; Jiao, L.; Zhu, H.; Li, L.; Liu, F. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14.
16. 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.
17. Yang, K.; Tong, X.-Y.; Xia, G.-S.; Shen, W.; Zhang, L. Hidden Path Selection Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.
18. 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.
19. Wang, K.; Li, F.; Chen, C.-M.; Hassan, M.M.; Long, J.; Kumar, N. Interpreting Adversarial Examples and Robustness for Deep Learning-Based Auto-Driving Systems. *IEEE Trans. Intell. Transport. Syst.* **2022**, *23*, 9755–9764.
20. 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.
21. Wang, D.; Li, C.; Wen, S.; Han, Q.-L.; Nepal, S.; Zhang, X.; Xiang, Y. Daedalus: Breaking Nonmaximum Suppression in Object Detection via Adversarial Examples. *IEEE Trans. Cybern.* **2021**, *52*, 1–14.
22. Arnab, A.; Miksik, O.; Torr, P.H.S. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 3040–3053.
23. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. *arxiv* **2014**, arXiv:1312.6199.
24. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arviv* **2015**, arXiv:1412.6572.
25. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
26. 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.
27. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. *arXiv* **2017**, arxiv:1608.04644.
28. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2019**, arxiv:1706.06083.
29. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Machine Learning at Scale. *arXiv* **2017**, arxiv:1611.01236.
30. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal Adversarial Perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017; pp. 86–94.
31. 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*; ACM: Seattle WA, USA, 2018; pp. 408–411.
32. Li, W.; Li, Z.; Sun, J.; Wang, Y.; Liu, H.; Yang, J.; Gui, G. Spear and Shield: Attack and Detection for CNN-Based High Spatial Resolution Remote Sensing Images Identification. *IEEE Access.* **2019**, *7*, 94583–94592.
33. 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.
34. 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 Sens.* **2021**, *14*, 1333–1347.
35. 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.
36. 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.
37. Xu, Y.; Ghamisi, P. Universal Adversarial Examples in Remote Sensing: Methodology and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.
38. 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.
39. He, X.; Yang, S.; Li, G.; Li, H.; Chang, H.; Yu, Y. Non-Local Context Encoder: Robust Biomedical Image Segmentation against Adversarial Attacks. *arXiv* **2019**, arxiv:1904.12181.
40. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context Encoding for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
41. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arxiv:1805.10180.
42. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. CGNet: A Light-Weight Context Guided Network for Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 1169–1179.
43. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-Resolution Context Extraction Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2020**, *13*, 71–78.

44. Nekrasov, V.; Ju, J.; Choi, J. Global Deconvolutional Networks for Semantic Segmentation. *arXiv* **2016**, arxiv:1602.03930.
45. Zhang, H.; Zhang, H.; Wang, C.; Xie, J. Co-Occurrent Features in Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 548–557.
46. Li, X.; Li, T.; Chen, Z.; Zhang, K.; Xia, R. Attentively Learning Edge Distributions for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2021**, *14*, 102–129.
47. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A Synergistical Attention Model for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16.
48. Liu, Y.; Chen, Y.; Lasang, P.; Sun, Q. Covariance Attention for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *11*, 12–18.
49. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L.E.; Jordan, M. Theoretically Principled Trade-off between Robustness and Accuracy. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7472–7482.
50. Liu, X.; Cheng, M.; Zhang, H.; Hsieh, C.-J. Towards Robust Neural Networks via Random Self-Ensemble. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 October 2018; pp. 381–397.
51. Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In Proceedings of the International Conference on Learning Representations, Virtual, 27–30 April 2020; pp. 185–192.
52. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting Adversarial Samples from Artifacts. *arXiv* **2017**, arxiv:1703.00410.
53. Ma, X.; Li, B.; Wang, Y.; Erfani, S.M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M.E.; Bailey, J. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. *arXiv* **2018**, arxiv:1801.02613.
54. Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; McDaniel, P. On the (Statistical) Detection of Adversarial Examples. *arXiv* **2017**, arxiv:1702.06280.
55. Tao, G.; Ma, S.; Liu, Y.; Zhang, X. Attacks Meet Interpretability: Attribute-Steered Detection of Adversarial Samples. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 6–11 December 2018; pp. 113–118.
56. Feinman, R.; Lake, B.M. Learning Inductive Biases with Simple Neural Networks. *arXiv* **2018**, arxiv:1802.02745.
57. Xu, W.; Evans, D.; Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In Proceedings of the Network and Distributed System Security Symposium, San Diego, CA, USA, 18–21 December 2018; PP. 158–164.
58. Gu, S.; Rigazio, L. Towards Deep Neural Network Architectures Robust to Adversarial Examples. *arXiv* **2015**, arXiv:1412.5068.
59. Ross, A. S.; Doshi-Velez, F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. *arXiv* **2017**, arxiv:1711.09404.
60. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arxiv:1503.02531.
61. Nayebi, A.; Ganguli, S. Biologically Inspired Protection of Deep Networks from Adversarial Attacks. *arXiv* **2017**, arxiv:1703.09202.
62. Cisse, M.; Adi, Y.; Neverova, N.; Keshet, J. Houdini: Fooling Deep Structured Prediction Models. *arXiv* **2017**, arxiv:1707.05373.
63. Gao, J.; Wang, B.; Lin, Z.; Xu, W.; Qi, Y. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. *arXiv* **2017**, arxiv:1702.06763.
64. Sun, Z.; Ozay, M.; Okatani, T. HyperNetworks with Statistical Filtering for Defending Adversarial Examples. *arXiv* **2017**, arxiv:1711.01791.
65. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
66. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact Bilinear Pooling. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 317–326.
67. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An Empirical Study of Spatial Attention Mechanisms in Deep Networks. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October 2019; pp. 6687–6696.
68. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
69. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
70. 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.
71. 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.
72. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20.