remote sensing

MDPI

Article

Adaptive Local Cross-Channel Vector Pooling Attention Module for Semantic Segmentation of Remote Sensing Imagery

Xiaofeng Wang [1], Menglei Kang [1], Yan Chen [2,*], Wenxiang Jiang [2], Mengyuan Wang [2], Thomas Weise [2], Ming Tan [2], Lixiang Xu [1], Xinlu Li [1], Le Zou [1] and Chen Zhang [1]

[1] Department of Big Data and Information Engineering, School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China

[2] Institute of Applied Optimization, School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China

* Correspondence: chenyan@hfuu.edu.cn

Abstract: Adding an attention module to the deep convolution semantic segmentation network has significantly enhanced the network performance. However, the existing channel attention module focusing on the channel dimension neglects the spatial relationship, causing location noise to transmit to the decoder. In addition, the spatial attention module exemplified by self-attention has a high training cost and challenges in execution efficiency, making it unsuitable to handle large-scale remote sensing data. We propose an efficient vector pooling attention (VPA) module for building the channel and spatial location relationship. The module can locate spatial information better by performing a unique vector average pooling in the vertical and horizontal dimensions of the feature maps. Furthermore, it can also learn the weights directly by using the adaptive local cross-channel interaction. Multiple weight learning ablation studies and comparison experiments with the classical attention modules were conducted by connecting the VPA module to a modified DeepLabV3 network using ResNet50 as the encoder. The results show that the mIoU of our network with the addition of an adaptive local cross-channel interaction VPA module increases by 3% compared to the standard network on the MO-CSSSD. The VPA-based semantic segmentation network can significantly improve precision efficiency compared with other conventional attention networks. Furthermore, the results on the WHU Building dataset present an improvement in IoU and F1-score by 1.69% and 0.97%, respectively. Our network raises the mIoU by 1.24% on the ISPRS Vaihingen dataset. The VPA module can also significantly improve the network's performance on small target segmentation.

Keywords: adaptive local cross-channel interaction; vector average pooling; attention mechanism; remote sensing imagery; semantic segmentation; deep learning

check for updates

Citation: Wang, X.; Kang, M.; Chen, Y.; Jiang, W.; Wang, M.; Weise, T.; Tan, M.; Xu, L.; Li, X.; Zou, L.; et al. Adaptive Local Cross-Channel Vector Pooling Attention Module for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1980. https://doi.org/10.3390/rs15081980

Academic Editor: Claudio Piciarelli

Received: 2 March 2023
Revised: 31 March 2023
Accepted: 5 April 2023
Published: 9 April 2023

1. Introduction

By analyzing an image at the pixel level, semantic segmentation provides more nuanced identification than object detection and image classification, allowing for the output of complete scene information. Urban planning, land resource management, marine monitoring, and transportation assessment benefit significantly from semantic segmentation processing of remotely sensed imagery [1,2]. However, remote sensing images present unique processing challenges due to the abundance of feature information such as shape, location, and texture, as well as the high intra-class variance and high inter-class similarity exhibited by ground objects in the images [3].

Conventional semantic segmentation approaches emphasize manual feature extraction [4]. The feature vectors can be obtained based on hand-crafted rules of certain application scenarios. Once the scenario has been modified, it is challenging to reuse these extracted feature vectors. The extraction of repeated features is a laborious and time-consuming process. In addition, the hand-crafted rules of traditional semantic segmentation depend on complex mathematical models that are not data-driven such as the current methods. Therefore, there are constraints on the comprehensibility and generalizability of the conventional semantic segmentation approaches. More recently, deep learning-based semantic segmentation techniques have demonstrated significant potential. For instance, FCN [5] implements a fully convolutional semantic segmentation network, which is the baseline for the current popular semantic segmentation approaches. U-Net [6] introduces the skip connection between the shallow and deep layers to effectively reconstruct the low-level spatial information for advanced semantic objects to solve the issue of inaccurate object edge segmentation. The encoder–decoder architecture encourages researchers to concentrate on ways to represent pixel features in the encoder better to boost network performance. ResNet [7] is one such work that expands the network depth to extract more advanced abstract features. Atrous or dilated convolutional networks such as the ones developed by the DeepLab community [8–11] can accomplish multi-scale tasks by enlarging the receptive field. HRNet [12,13] maintains dense multi-layer interaction between the shallow and deep feature maps. Similarly, U-Net++ [14] enhances the accuracy of semantic segmentation by substituting dense connections for regular skip connections between the encoder and the decoder. These efforts have led to growth in the use of deep learning for semantic segmentation. Researchers then began to optimize the performance of baseline semantic segmentation networks by introducing an attention mechanism, enabling them to capitalize on the critical feature information and eliminate the redundancy of feature maps or pixels to reinforce the feature representation.

The effectiveness of the attention mechanism has been demonstrated for a variety of tasks [15,16] including object detection [17–19] and image classification [20–22]. Based on the principle of the attention mechanism, researchers in the field of semantic segmentation have developed several attention modules, such as the channel and spatial attention modules. These modules are often incorporated into the semantic segmentation architecture to aid in extracting significant features in certain channels and pixels, thereby raising the segmentation accuracy [23]. Generally, the attention modules mentioned above are built separately and only capture features along certain channels or spatial dimensions. The CBAM [24] attention module combines channel and spatial attention using a tandem mode for the first time, significantly improving segmentation accuracy [25]. However, the module constructed using the tandem integration might cause errors transmitting from the channel attention to the spatial attention side, confining the further improvement in semantic segmentation performance. Researchers also designed attention modules focusing on the relationship of channels and pixels based on the self-attention mechanism, which represents the core information by the weighted sum of each channel-spatial dimension [26]. In other words, the network can use a self-attention mechanism to raise the overall accuracy of the semantic segmentation network by establishing a long-range contextual relationship [27]. Nevertheless, the complicated structure of a self-attention module remains challenging regarding training cost and execution efficiency, making it constrained to support a large-scale remote sensing application [28]. This paper focuses on the effect of the lightweight and efficient attention module on the semantic segmentation network.

We propose an innovative and lightweight vector pooling attention module (VPA) to solve the issues mentioned above. The module is constructed as an independent component, coupling the qualities of the channel and spatial attention to implement channel and spatial relation reinforcement. Compared to the module with tandem integration, our module possesses lower inter-module dependency. Furthermore, our proposed VPA module aims to preserve and localize the spatial information characterized by the pixels and learns weights directly via adaptive local cross-channel interaction. That is to say, it facilitates the effect of channel attention and spatial attention at the same time, with a low cost to pay in terms of dependency. Furthermore, an improved pooling operation called vector average pooling has been designed to fuse the vertical and horizontal dimensions of the feature maps. We incorporate our module into the classical baseline semantic segmentation network. We conduct multiple weight learning ablation experiments using the WHU Building dataset, the ISPRS Vaihingen dataset, and the MO-CSSSD to provide an optimized solution for the attention-based, large-scale remote sensing application-oriented semantic segmentation network. Therefore, the main contributions of this paper are as follows:

1. By analyzing the pooling method of current mainstream attention modules for feature maps, we propose an efficient vector average pooling method to realize the construction and retention of spatial information of feature maps by attention modules.
2. By comparing the weight mapping methods of the existing mainstream attention modules, we introduce the most efficient weight mapping method of adaptive local cross-channel interaction to achieve lightweight and efficient attention modules.
3. We build VPA modules with vector average pooling and adaptive local cross-channel interaction methods to realize the functions of the channel attention module and spatial attention module simultaneously with a single attention module, avoiding the dependency between modules caused by the coupling between modules.

2. Related Work

2.1. Attention Mechanism

The attention mechanism can enhance the capacity of deep CNNs to obtain more discriminative features. By creating a global dependency between channels to determine the corresponding weights, the channel attention module SE [29] successfully improved the network's representation of significant features using the attention mechanism in the channel dimension for the first time. Yet, the spatial context was not taken into account. The effectiveness of the channel attention network was further enhanced by adding the ECA [30] module that established the dependency of local channels to learn the critical weights. To better capture spatial information and generate spatial consequences with a wider convolutional field, the spatial attention module in CBAM performs feature map channel pooling and dimension reduction. As a result of its reliance on convolution, the spatial attention module has certain limitations, as it can only capture the local dependency in position and not establish a long-range dependency. In addition, the pixel-relational self-attention mechanism represented by Transformer [31] has become a new SOTA in the current computer vision field and is widely recognized and applied. In DANet [26], the feature map generates Query, Key, and Value matrices via three convolutions. These matrices are then employed to calculate weights for each local and global location to build contextual information. OCRNet [32] creates a description region for each category in advance and constructs the global contextual information by calculating the similarity of each pixel to the description region of the respective category. The self-attention mechanism effectively establishes a global connection but requires excessive computation, affecting the network inference efficiency. Specifically, the process of computing the weight map with the Query and Key matrices of the feature map imposes O(N^2) ($N = H \times W \times C$, H, W, and C denote the height, width, and number of channels of the feature map, respectively) time and space complexity on the self-attention module, which leads to a large burden on the semantic segmentation network when processing large remote sensing images. Our research aims to design a lightweight and efficient attention module by coupling spatial information to the channel attention module to further enhance the effect of the attention mechanism focusing on the channel level on the semantic segmentation network.

2.2. Attention in Semantic Segmentation for Remote Sensing

By integrating attention modules, a semantic segmentation network for image interpretation can better represent features, reduce noise and build contextual information to increase the network's overall segmentation accuracy. As an example, in ENet [33], the SE module is added to the upsampling stage of the network to generate a weight for each channel to refine the segmentation accuracy of remote sensing images. In SE-UNet [34], the convolution block is altered from two convolutions with the size of 3 × 3 per layer in the standard UNet to one convolution plus one SE module for strengthening the representation of the feature maps, thereby enhancing the capacity of UNet on extracting the road from the satellite and aerial imagery. An efficient channel attention (ECA) module is proposed and integrated into the UNet encoder in [35], which optimizes the segmentation and raises the encoder performance on feature extraction. Denoising remote sensing images with the RSIDNet proposed in [36] is made more accessible by adding an ECA module to the shortcut block connecting the shallow layer with the deep layer. It augments the feature representation of the shallow feature maps, reducing the noise brought by the layers and enhancing the segmentation accuracy. It is discussed in SCAttNet [25] that the CBAM module is employed to integrate channel and spatial attention. The network first adopts the ResNet to extract features to strengthen its segmentation capability for high-resolution remote sensing imagery. It then outputs them into the CBAM module to construct local contextual information and optimize the learned feature map weights at channel and pixel levels. RAANet [37] constructs a new residual ASPP module by embedding a CBAM module and a residual structure as a way to improve the accuracy of the semantic segmentation network. In [38], an SCA attention module containing spaces and channels is designed by using a spatial attention module in the CBAM in parallel with a coordinate attention module that constructs channels and spaces to enhance the detection of remote sensing images by the lightweight model. In order to improve the ability of convolutional neural networks to represent potential relationships between different objects and surrounding features, MQANet [39] introduces position attention, channel attention, label attention, and edge attention modules into the model as a way to expand the perceptual field of the network and introduce background information in labels to obtain global features. Furthermore, the self-attention mechanism has been used for remote sensing image semantic segmentation. As an illustration, a region-attention RSA module is constructed using the self-attention mechanism in RSANet [40]. Firstly, the module creates several soft object regions for each category distributed in the image, followed by region descriptors. Then, it evaluates the similarity between pixels of the feature maps and all-region descriptors. Those values measured in the parallel will be treated as weights of the initial feature maps. In [41], a self-attention module that concerns channels and spaces is constructed to generate a weight map of all spatial locations and channel relationships by multiplying the Query and Key matrices of the feature map to obtain global information. The network may achieve more accurate segmentation results, but the complexity and the high hardware resource needs make it non-economical in actual deployment.

2.3. ECA Module

The ECA is an improved version of the SE module that seeks to minimize the inadequacy of lightweightness and effectiveness of the SE module. Figure 1 illustrates the structure of the SE module.

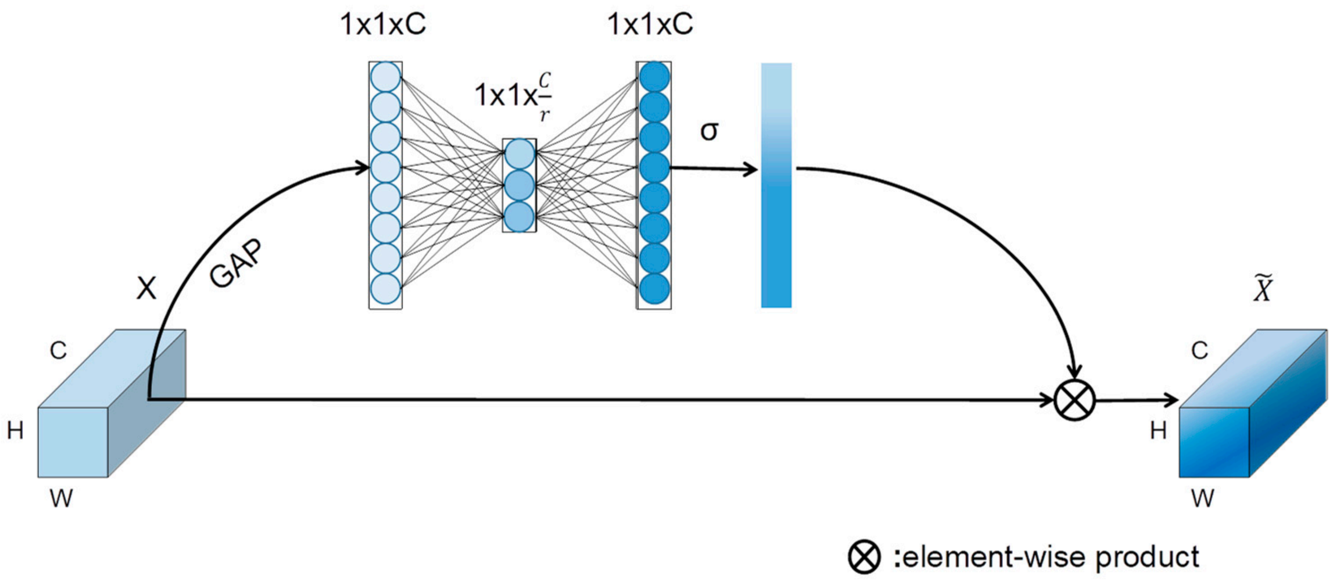

Figure 1. Channel attention module SE.

The module learns the weights from the initial feature maps and then produces new one-to-one weighted feature maps. The process can be expressed as follows:

$$F_{se}: \ X \to \widetilde{X} \tag{1}$$

where $X = [x_1, x_2, ..., x_c] \in R^{H\times W\times C}$ denotes the initial feature maps; x_c indicates the channel c of the feature maps, i.e., the cth feature map; and H, W, and C denote the height, width, and number of channels of the feature maps, respectively. $\widetilde{X} = [\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_c] \in R^{H\times W\times C}$ denotes the new weighted feature maps after the attention operation. $\widetilde{x}_c$ denotes channel c of the weighted feature maps. $Z = [z_1, z_2, ..., z_c] \in R^{1\times 1\times C}$ is generated by a two-dimensional global average pooling operation on the feature map X. The process can be formulated as

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i, j) \tag{2}$$

where z_c is the output of global averaging pooling on the cth channel, and $x_c(i, j)$ is the pixel representation of the location on the cth channel. Subsequently, Z will be fed as input to two fully connected layers to learn the channel weight based on the global information interaction. The weights learned via the first fully connected layer can be defined as

$$\omega_1 = \delta(Z \times W_1) \tag{3}$$

where ω_1 is the channel weight generated from the first fully connected layer; $W_1 \in R^{C\times\frac{C}{r}}$ is the parameter of the first fully connected layer; r is the hyperparameter, which is used to decrease the number of channels and parameters; and δ denotes the activation function ReLU. The weight learned by the second fully connected layer can be formulated as

$$\omega_2 = \sigma(\omega_1 \times W_2) \tag{4}$$

where $W_2 \in R^{\frac{C}{r}\times C}$ is the parameter of the second fully connected layer, which converts the current channel count back to the initial count again; σ is the Sigmoid activation function; and ω_2 is the weight corresponding to the channel obtained by the SE module. The new weighted feature map $\widetilde{X}$ is generated based on the element-wise product between the output ω_2 and the initial feature map X. It can be described as

$$\widetilde{X} = \omega_2 \times X \tag{5}$$

The channel is first mapped to the lower dimensional space in the SE module depending on the hyperparameter r and the first fully connected layer. Then, the output is mapped back to the initial channel via another fully connected layer (the weight learning can be denoted as MLP_r). The weight learning approach can decrease the number of parameters in the module. However, it ends up using ω_2 rather than ω_1, which implies that the channel weights' mapping in the SE module behaves indirectly. In terms of experimental results, it is not as precise as direct weight mapping. Additionally, the global average pooling operation of the SE module on the feature map causes the module to ignore the spatial information of the feature map.

Since the input Z to the fully connected layer in the SE module has the shape of $1 \times 1 \times C$, it performs similarly to the convolution operation with the kernel size of 1×1, which allows for global cross-channel information interaction during the learning of channel weights in Equation (3). Generally, several convolutional designs contribute to both parameter reduction and efficiency improvement in learning. We list the conventional convolution operations as follows and illustrate their notations:

- Depthwise Separable Convolution(DW) [42]. When the weights in Equation (3) are learned directly using a 1×1 depthwise separable convolution (note: the hyperparameter r is no longer used for dimensionality reduction), W_1 can be defined as follows:

$$W_1 = \begin{bmatrix} \omega^{1,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega^{C,C} \end{bmatrix} \tag{6}$$

  Then, W_1 has just C in terms of parameters. Despite having fewer parameters, the convolution design only adopts a single-channel weight mapping and dismisses the inter-channel dependency.
- Standard Convolution(SC) [43]. When the weights in Equation (3) are learned directly using a 1×1 standard convolution, W_1 is written as follows:

$$W_1 = \begin{bmatrix} \omega^{1,1} & \cdots & \omega^{1,C} \\ \vdots & \ddots & \vdots \\ \omega^{C,1} & \cdots & \omega^{C,C} \end{bmatrix} \tag{7}$$

  Currently, the number of parameters of W_1 is C^2. Although it is superior to the depthwise separable convolution in terms of performance and implements interaction between global channels, the standard convolution performs well at the expense of introducing additional parameters.
- Group Convolution(GC) [44]. Considering the compromise, convolution design capable of achieving a small number of parameters and high learning performance is necessary. When the weights in Equation (3) are learned by using a 1×1 group convolution, the expression of W_1 can be written as follows:

$$\begin{aligned} W_1^i &= \begin{bmatrix} \omega^{i,i} & \cdots & \omega^{i,i+k-1} \\ \vdots & \ddots & \vdots \\ \omega^{i,i+k-1} & \cdots & \omega^{i+k-1,i+k-1} \end{bmatrix} i \in \left[1, \frac{C}{k}\right] \\ W_1 &= \begin{bmatrix} W_1^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_1^{\frac{C}{k}} \end{bmatrix} \end{aligned} \tag{8}$$

The number of parameters in W_1 is $k * C$, where k is the number of channels within each convolution group. Regarding the number of parameters, the approach of group convolution is more lightweight than the standard convolution. When considering the efficiency, it is more efficient than depthwise separable convolution because it implements the interaction between channels within the group.

Nevertheless, since each group operates separately, no interaction occurs. As a result, there is still room for development in the approach concerning the weight learning effect. When the weight learning is performed in Equation (3) using a one-dimensional convolution with a convolution kernel of size k ($C1D_k$), the tensor expression of W_1 can be written as follows:

$$W_1 = \begin{bmatrix} \omega^{1,1} & 0 & \cdots & 0 \\ \vdots & \omega^{2,2} & \cdots & 0 \\ \omega^{k,1} & \vdots & \ddots & \vdots \\ 0 & \omega^{k+1,2} & \cdots & \omega^{C,C} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega^{C+k,C+k} \end{bmatrix} \tag{9}$$

The strategy above can be considered an interaction between each channel and the $(k-1)$th channel adjacent to it. It assists in achieving an adaptive local cross-channel interaction weight learning, which addresses the issue of lower connectedness in group convolution. When done, the module is proven efficient in learning channel weights and lightweight.

The ECA module is an efficient channel attention module built with the adaptive local cross-channel interaction described above, as shown in Figure 2.

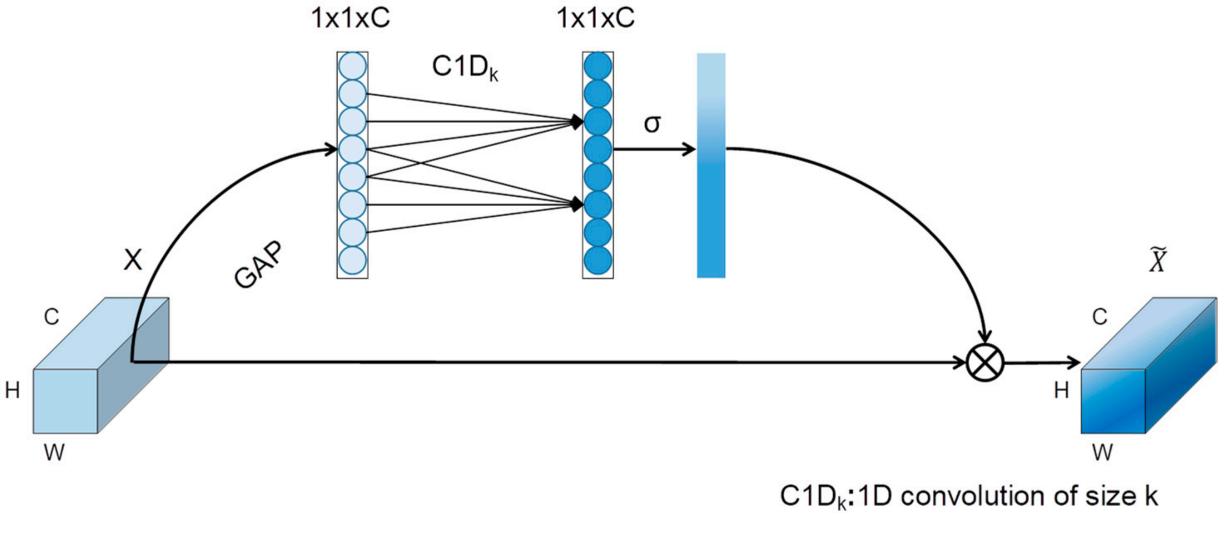

Figure 2. Efficient channel attention module ECA.

Adaptive local cross-channel interaction is first computed by pooling in Equation (2), followed by a one-dimensional convolution on Z with a convolution kernel of size k. The process can be described as

$$\widetilde{X} = X \times \sigma(C1D_k(Z)) \tag{10}$$

where $C1D_k$ denotes a one-dimensional convolution with a kernel of size k, and σ denotes the Sigmoid function. In the group convolution, it is demonstrated that more channels per group are expected when there are more channels in total. Therefore, there exists a mapping between k and the number of channels. It defines the relationship between k and the number of channels, distinguished from the conventional linear mapping relationship, as in the equation

$$C = 2^{a*k-b} \tag{11}$$

where C is the number of channels, and, when given C, k in Equation (11) can be estimated by

$$k = \left| \frac{log_2(C) + b}{a} \right|_{odd} \tag{12}$$

where $|n|_{odd}$ is denoted as the largest *odd* number less than or equal to n. To guarantee a channel attention module with few parameters and good performance, ECA maps the local channel interaction and weight directly through a layer of one-dimensional convolution of adaptive size k proportional to the number of channels. However, the efficient weight mapping method in the ECA module is only performed on one-dimensional vectors, which limits the effect of the attention module.

In contrast to the SE and ECA, our proposed VPA module takes spatial information into account appropriately. It seeks to address the dependency concern between the channel and spatial attention of the CBAM with the tandem integrating mode and to minimize the complexity, such as in the self-attention module. More specifically, the VPA module can establish a long-range spatial dependency by employing transposed vector pooling rather than global pooling. It also encodes the channel in multidimensional coordinates and locates the critical pixel within the channel of the feature maps. Furthermore, the VPA module can obtain a significant channel relationship by introducing local interaction.

3. Methodology

3.1. Vector Average Pooling

The conventional channel attention modules such as SE and ECA use the global average pooling output of each channel of the feature map to highlight the noteworthy channel feature. However, such an operation causes the loss of the spatial information of the feature map, making the effect of the attention module somewhat restricted. Therefore, we propose a vector average pooling method for the attention module to retain and construct the spatial information of feature maps. The vector average pooling executes the calculation of the average pooling of size $(1, W)$ and $(H, 1)$ along the vertical and horizontal pixel lines of the feature map, respectively (wherein W and H denote the width and height of the feature map). As shown in Figure 3, it is the processing procedure of vector average pooling of one channel included in the feature map. The representation of one channel on the feature map by two crossed vectors can make the feature map retain more local spatial information. More specifically, the feature map will generate a vector of size $(H, 1)$ after vector average pooling in the horizontal direction, then the ith value in the vector represents the spatial information of the ith row of the feature map. The same is true for vector average pooling in the vertical direction. Using vector average pooling instead of global average pooling in the channel attention module can further enhance the effectiveness of the channel attention module by preserving and locating the spatial information of the outputted feature map.

3.2. VPA Module

For the channel attention module to better retain spatial information and establish a long-range dependency, we propose the VPA module, as shown in Figure 4.

In Figure 4, Pool $(H, 1)$ represents the average pooling of all pixels in the range $(H, 1)$ in the feature maps. P is the dimensional substitution. $C2D_{(k,\,1)}$ is a two-dimensional convolution using a convolution kernel of size $k \times 1$ to achieve adaptive local cross-channel interaction in the channel dimension.

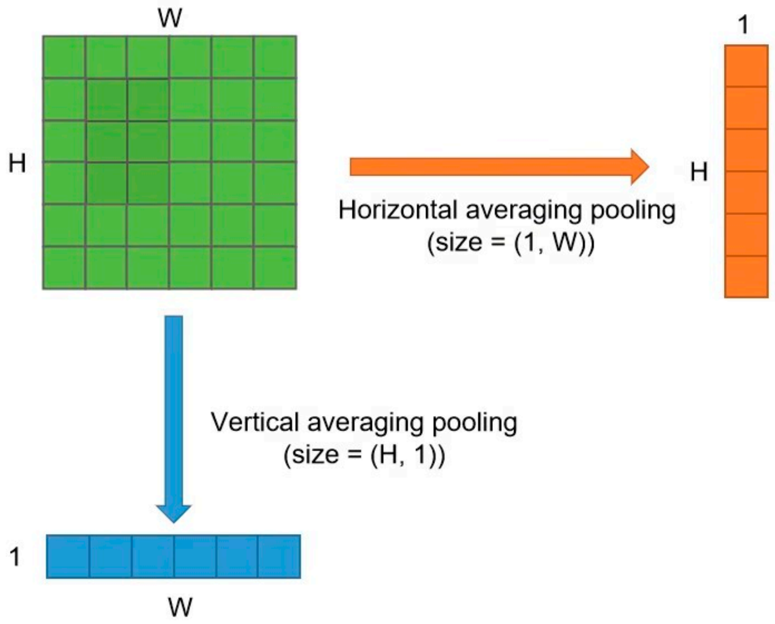

Figure 3. Vector average pooling on feature maps.

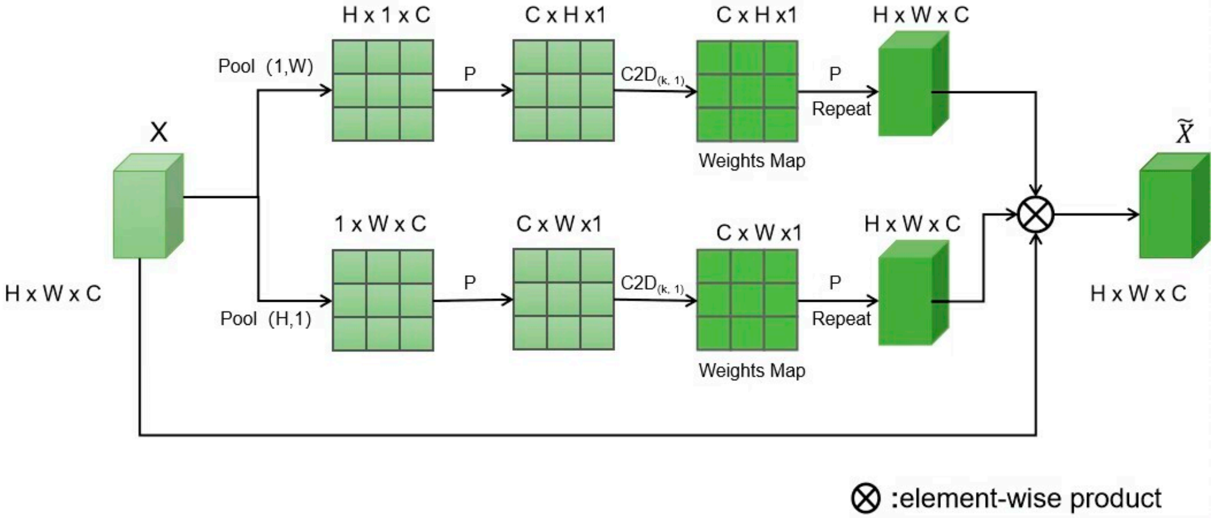

Figure 4. Vector Pooling Attention (VPA) Module.

Firstly, the VPA performs vector average pooling on the feature map, which can effectively construct and preserve long-range dependency on the feature map space both vertically and horizontally and thus learn the location weights on the space. The pooling along the vertical direction can be described as

$$Z_c^H(h) = \frac{1}{W}\sum_{i=1}^{W} x_c(h,i) \tag{13}$$

where $Z_c^H(h)$ is denoted as the output of the vector average pooling of the row h via (1, W) on channel c. Likewise, the average pooling of the column w via (H, 1) on the horizontal channel c can be expressed as

$$Z_c^W(w) = \frac{1}{H}\sum\nolimits_{j=1}^{H} x_c(j,w) \tag{14}$$

Then, it learns the position weights on *Z(H)* and *Z(W)* and chooses the efficient adaptive local cross-channel interaction inspired by the ECA module to learn the channel weights. Unlike the one-dimensional weight mapping on channels of the ECA module, the VPA module extends the one-dimensional weight mapping method to a two-dimensional pattern by implementing the adaptive local cross-channel interaction weight learning with a $k \times 1$ two-dimensional convolution, as in the equation

$$\omega(H) = \sigma\Big(C2D_{(k,\,1)}(P(Z(H)))\Big) \tag{15}$$

and

$$\omega(W) = \sigma\Big(C2D_{(k,\,1)}(P(Z(W)))\Big) \tag{16}$$

where P in Equations (15) and (16) denotes the dimensional substitution function that converts $Z_c^H(h)$ and $Z_c^W(w)$ to the plane of $C \times H \times 1$ and $C \times W \times 1$. The optimization objective of the operation is to easily enable the VPA to implement adaptive local cross-channel interaction on Z ($C2D_{(k,\,1)}$), as shown in Figure 5. The weights are mapped by sliding the two-dimensional convolution of the convolution kernel of size (k, 1) over the dimension of the feature map channels. The feature maps are multiplicatively weighted after being converted back to their initial dimensions once the weights have been learned. The proposed VPA module, analogous to executing the pooling operation shaped like the cross-coordinate axis, creates the long-range dependency in one direction and precisely locates each pixel with specific weights. For instance, assume that the fourth-column feedback on the feature maps possesses a greater weight in the horizontal direction and that the fifth-row response has a greater weight in the vertical direction. Then, more emphasis is put on the pixel at coordinates (5, 4) on the final weighted feature maps to gain precise spatial localization. Localization in this manner is particular to each channel, bringing spatial and channel attention characteristics to our proposed module.

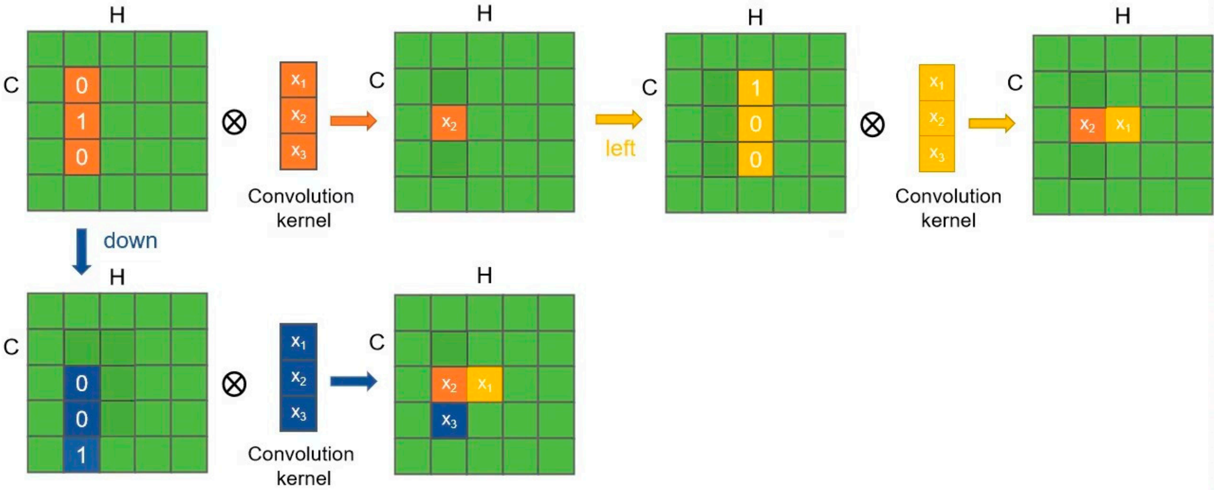

Figure 5. Vertical weight mapping in the VPA module. Convolution kernel slides left and down in the channel dimension for efficient adaptive local cross-channel interaction.

4. Experiments and Results

4.1. Datasets

The experiments in this paper use three datasets: Multi-object Coastal Supervision Semantic Segmentation Dataset (MO-CSSSD) [45], WHU Building [46], and ISPRS Vaihingen [47]. Based on different scenarios and different objectives, they are used to verify the effects of the VPA module and other classical attention modules on semantic segmentation networks.

4.1.1. MO-CSSSD

This dataset is derived from aerial imagery of coastal areas in southern China and is used for coastal ecosystem supervision. The dataset has a training set with 8734 samples, a validation set with 1100 samples, and a test set with 1100 samples. The size of the sample images is 256 × 256, and the spatial resolution is 0.58 m for RGB images. The dataset has four categories, i.e., mangrove, aquaculture raft, aquaculture pond, and background.

4.1.2. WHU Building Dataset

This dataset is derived from aerial imagery of Christchurch city, New Zealand, which has been downsampled to RGB images with a spatial resolution of 0.3 m by researchers from Wuhan University in China. The images are further cropped without overlap into 8189 patches containing 187,000 buildings in total. The size of each patch is 512 × 512. The dataset includes a training set with 4736 samples, a validation set with 1036 samples, and a test set with 2416 samples.

4.1.3. ISPRS Vaihingen Dataset

This dataset produced by ISPRS, is derived from the airborne imagery of Vaihingen in Germany, with 33 true orthophoto (TOP) images of different sizes. Each image is from a larger TOP with an average size of about 2494 × 2064. It contains three bands, near-infrared, red, and green, and possesses a spatial resolution of 0.09 m. There are six categories, i.e., building, low vegetation (low veg), tree, car, impervious surface (imp. surfaces), and background. In our research, 16 images with manual annotation have been selected as the training set and another 17 images from the test set. The initial large images were cropped into 817 patch samples in the training set and 2219 in the test set based on the cropping method in [48]. The size of each patch in the dataset is 384 × 384. Data augmentation has also been applied, e.g., vertical and horizontal flipping, noise addition, etc.

4.2. Evaluation Metrics and Experimental Setting

In the research, Precision, Recall, F1-score, Intersection over Union (IoU), Overall Accuracy (OA), and mean Intersection over Union (mIoU) are adopted to evaluate the experimental results for comparing the effectiveness of our proposed VAP module and the conventional attention modules. The Precision represents the number of correctly classified positive samples as a percentage of the total number of positive samples classified (correctly or incorrectly) by the classifier. In contrast, the Recall represents the number of correctly classified positive samples as a percentage of the total number of positive samples. Precision and Recall may contradict one another in certain circumstances. Therefore, we also use the F1-score, an average reconciliation function of Precision and Recall. Below are the formulas for each evaluation metric:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{19}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{20}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

where true prediction on a positive sample (*TP*), false prediction on a positive sample (*FP*), true prediction on a negative sample (*TN*), and false prediction on a negative sample (*FP*) are notated in the above formulas.

In the study, all experiments were conducted in the same system environment based on Windows 10, Python 3.9, TensorFlow 2, and NVIDIA GeForce RTX 3090. The initial learning rate is set to 0.001, and the Adam optimizer and cross-entropy loss function are used to train the model in 200 epochs. The batch size is set to 14 for experiments on the MO-CSSSD and ISPRS Vaihingen datasets and to 6 for experiments on the WHU Building dataset. The reduced-dimensional convolution (MLP_r) is used in the VPA_ MLP_r and SE modules, and we set r to 16 to ensure the best module complexity and effect. The hyperparameter values given in [30] have been adopted by setting the a and b in Equation (12) to 2 and 1, respectively.

4.3. Experimental Architecture

The VPA module aims to improve the feature extraction capability of the encoder of the semantic segmentation network. The vector average pooling in the vertical and horizontal directions of the feature maps builds the channel and spatial relationship and learns the weights through adaptive local cross-channel interaction, as depicted in Figure 4. To compare the effectiveness of our proposed attention module with another classical baselined SE and ECA, we inserted them into a ResNet block according to [29,30]. ResNet50 with a specific attention module is designed as the backbone of DeepLabV3. It is employed to evaluate the performance of these attention modules, as shown in Figure 6. In the architecture, the downsampling component in stage 4 of ResNet50 is replaced with an atrous convolution whose dilation rate is set to 2, resulting in a stride of 16 at the network's output. We chose the DeeplabV3 model in Figure 6 because it uses the residual blocks of ResNet in its backbone part, and the fused use of the residual blocks of Resnet with the attention module is a classical way to compare the effects of different attention modules. Meanwhile, the ASPP module in DeeplabV3 can also effectively solve the multi-scale problem in remote sensing datasets.

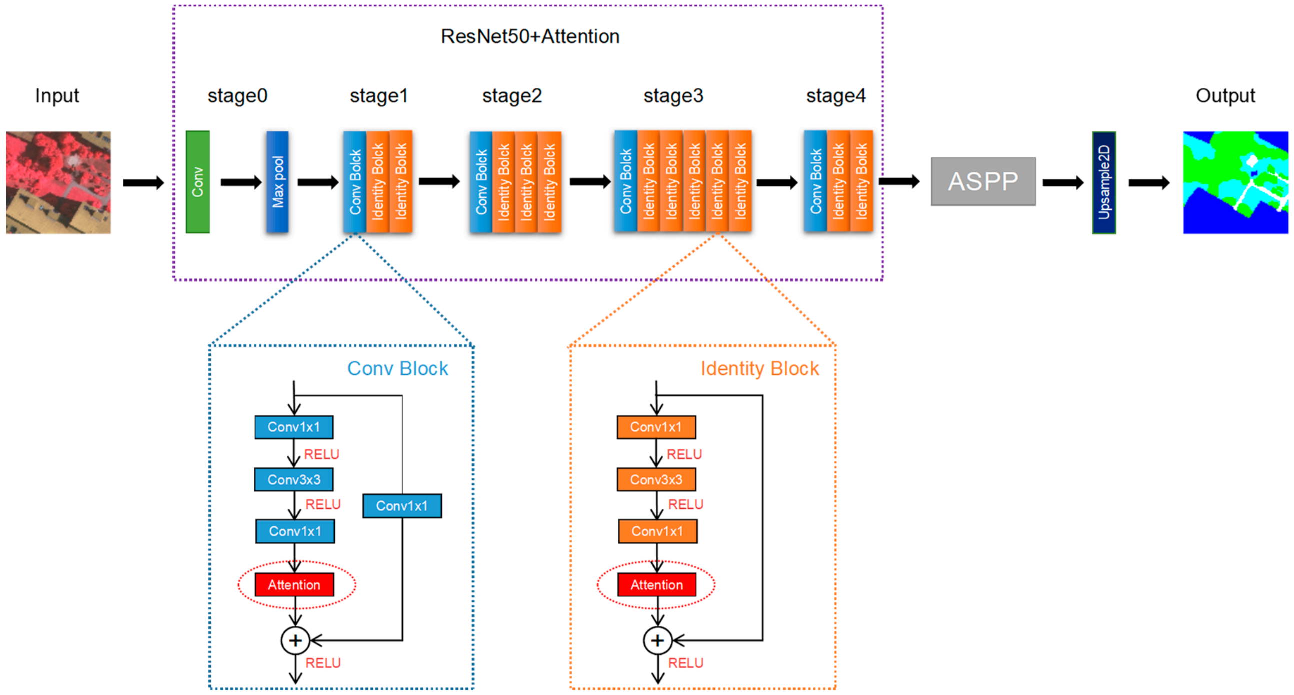

Figure 6. DeepLabV3 combined with attention modules.

4.4. Experimental Results

4.4.1. Ablation Study

An ablation study for the VPA module is first performed on the MO-CSSSD. The VPA module is designed by using the single-channel depthwise separable convolution (DW), group convolution (GC), standard convolution (SC), dimension-reduced convolution (MLP_r), and adaptive local cross-channel convolution ($C2D_{(k,\,1)}$) to evaluate the effectiveness of various convolution strategies on the attention modules performance.

The results of the ablation experiments on MO-CSSSD are shown in Table 1. Compared to the network without any attention module, more than a 1% increase in OA and a 2% increase in mIoU can be obtained using the DeepLabV3 network with the various modified VPA modules. It highlights the capacity of our proposed VPA module to enhance the accuracy of semantic segmentation networks. According to the results of the VPA modules constructed in various convolution strategies, the standard convolution (SC) implemented in all-channel interaction has a 1.1% increase in OA and a 2.5% increase in mIoU. However, it leads to a double increase in the number of parameters and 1.39 G for the number of FLOPs. Although the dimension-reduced convolution (MLP_r) can significantly decrease the number of parameters, the OA and mIoU are also reduced by 0.13% and 0.74%, respectively, compared to the SC. It demonstrates that the indirect mapping of weights impacts weight learning. The single-channel depthwise separable convolution (DW) and group convolution (GC) decrease the number of parameters and FLOPs and significantly raise the OA and mIoU. Furthermore, our suggested adaptive local cross-channel convolution $C2D_{(k,\,1)}$ outperforms the baseline by a 3% increase in mIoU and a 1.3% increase in OA while only introducing 0.01 M parameters to the network. In contrast to GC and DW, where there is no interaction between the local channels, our proposed $C2D_{(k,\,1)}$ achieves the interaction between local channels and reinforces the information swapping between channels. The module enhances the capacity for channel weight learning, which indicates that the adaptive local cross-channel interaction might provide an optimized effect in contrast to the global channel interaction.

Table 1. Comparison of VPA achieved by different weight learning methods. The bold indicates the best data.

Method	FLOPs (G)	Parameters (M)	OA (%)	mIoU (%)
Baseline	30.37	37.26	91.61	79.43
+VPA_DW	**30.41**	37.32	92.95	82.39
+VPA_GC16	30.49	39.69	92.72	81.50
+VPA_SC	31.80	75.67	92.73	81.96
+VPA_MLP_r	30.58	40.89	92.60	81.22
+VPA_$\text{C2D}_{(k,\,1)}$	30.41	**37.27**	**92.96**	**82.43**

4.4.2. Comparison Experiment

Several comparison experiments have been established on the MO-CSSSD, WHU Building dataset, and ISPRS Vaihingen dataset using the architecture depicted in Figure 6. They are used to evaluate the enhancement effect gained by the SE, ECA, CBAM, SCA [38], and our VPA module.

Table 2 shows the changes in the number of parameters and FLOPs after the baseline incorporation of other classical attention modules on the MO-CSSSD dataset. As can be seen in Table 2, SE, ECA, CBAM, SCA, and our VPA module bring a relatively small number of parameters and computations to the baseline, which shows their lightweight property. The greatest burden on the baseline is placed on the CBAM and SCA modules, which increase the number of parameters in the baseline by 4.8 M and FLOPs by 0.23 G, respectively. The ECA module places the least burden on the baseline, adding only 0.01 M parameters and 0.02 G FLOPs. However, the ECA module focuses only on the channel dimension. Our VPA module, on the other hand, is concerned with both channel and space dimensions, but it also only gives the baseline of about 0.01 M counts of parameters and 0.04 G FLOPs, which is much lighter than the CBAM module.

Table 2. The burden that different attention modules add to the Baseline. The bold indicates the best data.

	Baseline	**+SE**	**+CBAM**	**+ECA**	**+SCA**	**+VPA**
Parameters(M)	37.26	39.66	42.06	**37.27**	40.89	37.27
FLOPs(G)	30.37	30.39	30.42	**30.39**	30.60	30.41

The experimental results on MO-CSSSD are shown in Table 3. The OA, mIoU, Precision, and IoU for each category are included in Table 3. The OA and mIoU of the network are not significantly better with the addition of the channel attention SE module under the same experimental settings. After incorporating the CBAM module with a channel spatial tandem integration, the OA and mIoU of the network increased by over 1%. This highlights the significance of spatial information in semantic segmentation networks. The advantage of the ECA module weight learning strategy is that both OA and mIoU increased by 0.82% and 1.77%, respectively, as opposed to the addition of the SE module. The addition of the SCA module also results in a better segmentation capability of the baseline network, with a 2.53% improvement in mIoU. However, it imposes larger FLOPs on the network. In contrast to the aforementioned attention modules, our VPA module introduces only 0.01 M parameters to the network while increasing the mIoU by 3% and the OA by 1.3%. By a small margin, the VPA module is more efficient than the ECA module. This demonstrates the effectiveness of the VPA module on the attention mechanism for the spatial dimension. In addition, the network with the VPA module shows a 1.43% increase in mIoU compared to the network with the CBAM module. The evidence might prove that vector average pooling is superior to local convolution when building spatial information. A multidimensional independent attention module could outperform the one combined in a tandem mode with a single dimension.

Table 3. Experimental results on the MO-CSSSD. The table shows Precision(%)/IoU(%) for each class. The bold indicates the best data.

Method	**Mangrove**	**Aquaculture Raft**	**Aquaculture Pond**	**Background**	**OA (%)**	**mIoU (%)**
Baseline	81.36/76.04	80.75/70.62	92.10/85.98	92.90/85.07	91.61	79.43
+SE	84.95/78.96	80.67/70.50	93.20/85.94	91.75/84.98	91.66	80.10
+CBAM	82.66/77.23	80.76/71.45	93.44/**88.29**	**94.00**/87.02	92.78	81.00
+ECA	86.47/80.61	**83.57/73.28**	93.12/87.20	93.11/86.42	92.48	81.87
+SCA	87.43/80.82	82.36/72.43	93.61/87.78	93.16/86.82	92.73	81.96
+VPA	**89.24/82.30**	81.87/72.06	**94.02**/88.18	93.15/**87.19**	**92.96**	**82.43**

The results of the experiments on the WHU Building dataset are shown in Table 4. Four evaluation metrics have been adopted for measuring the experimental results, i.e., IoU, Precision, Recall, and F1-score, frequently used for binary classification tasks. After being fed the additional attention modules, the network significantly outperforms the baseline on some evaluation metrics. IoU and F1 increased by 1.03% and 0.6%, respectively, after the SE module was added to the baseline network. As opposed to the boost generated by the independent channel attention module SE, the IoU and F1 increased by 0.59% and 0.35% when the CBAM module with channel spatial tandem integration was implemented. This indicates that the tandem integration of the channel attention and spatial attention module influences the overall effect of the CBAM module. Accurate spatial location in the SCA module allows for the highest accuracy in baseline segmentation. However, the weight mapping method in the SCA module is indirect. As an alternative, our suggested VPA module integrates channel and spatial attention into a single component, thereby preventing the dependency issue arising from the modules' tandem integration. Consequently, compared to adding other attention modules in the baseline, adding the VAP module results in the most substantial improvement in IoU, Recall, and F1-score of the network, with increases of 1.69%, 1.79%, and 0.97%, respectively.

Table 4. Experimental results on the WHU Building dataset. The bold indicates the best data.

Method	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline	86.23	93.32	91.90	92.60
+SE	87.26	92.40	94.02	93.20
+CBAM	86.82	91.72	**94.20**	92.95
+ECA	87.64	93.35	93.47	93.41
+SCA	87.42	**93.64**	92.94	93.29
+VPA	**87.92**	93.41	93.73	**93.57**

Table 5 shows the results of experiments conducted on the ISPRS Vaihingen dataset. The IoU for each category and the OA and mIoU are listed in Table 5. With the addition of the attention module, the IoU of each class is significantly improved. The IoU of the car category, with a rise of 1.92%, has shown the most significant improvement. With the addition of the SCA module, the baseline is most effective in segmenting larger categories of shapes, due to the precise localization of spatial information by the SCA module. This shows that the attention module can optimize the network's capacity to detect small targets. Our VPA module offers a tremendous advancement in segmentation for the car category, with a 2.82% increase in IoU. This further implies that the VPA module is the best of the attention modules in Table 4 for improving the network's capacity for extracting small targets. The mIoU of the baseline increased by 1.24% with the addition of the VPA module. This is the best result of the networks with attention modules. It indicates that the VAP module outperforms the other attention modules and it might significantly improve the segmentation performance of the network on the ISPRS Vaihingen dataset, despite the high resolution and detailed categories present in the dataset.

Table 5. Experimental results on the ISPRS Vaihingen dataset. The table shows IoU(%) for each class. The bold indicates the best data.

Method	Imp. Surfaces	Building	Low Veg	Tree	Car	OA (%)	mIoU (%)
Baseline	76.65	82.60	72.21	61.70	47.67	84.60	68.17
+SE	77.81	83.50	72.40	62.23	49.77	85.10	69.15
+CBAM	77.05	82.50	72.37	61.46	49.33	84.68	68.54
+ECA	**78.23**	83.40	72.87	62.54	49.17	85.26	69.24
+SCA	77.92	**84.32**	**73.04**	**62.92**	48.43	**85.52**	69.33
+VPA	77.86	83.44	72.84	62.45	**50.49**	85.21	**69.41**

5. Discussion

In the ablation experiments, we achieve the VPA module with five different weight mapping methods and compare the improvement in complexity and segmentation accuracy they bring to the network. The final results show that the VPA module constructed with adaptive local cross-channel interaction has significant advantages in terms of complexity and effectiveness. Adaptive local cross-channel interaction is the process of sliding over channel dimensions using a convolutional kernel of size (k, 1), both between a single channel and its adjacent k-1 channels, which ensures channel interaction mapping while also reducing the parameters involved in weight mapping.

Figure 7 depicts the visual prediction results in Table 3. It presents three examples from the MO-CSSSD dataset, containing all the categories in the dataset. As shown in Figure 7, the VPA module results in a sharpened edge of the prediction target. The results are more in line with ground truth than those generated by adding the other attention modules to the baseline. This happens due to the vector average pooling of the VPA module in the spatial dimension and its adaptive local cross-channel interaction in the channel dimension. By improving the segmentation accuracy on the MO-CSSSD dataset after incorporating the VPA module, we can see that it has the potential to similarly improve the segmentation accuracy of the conventional semantic segmentation network on multi-target datasets. Additionally, in *Sample c* for the division of a small aquaculture pond, the VAP module is superior to other attention modules.

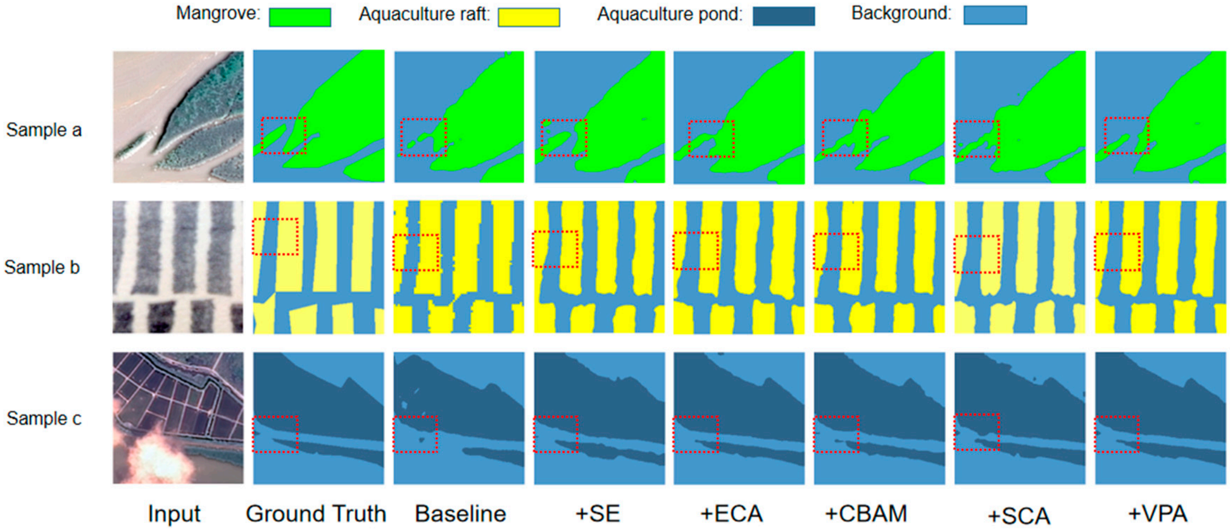

Figure 7. Comparison of prediction samples on the MO-CSSSD. The red dashed lines mark the more obvious differences.

Figure 8 depicts visual prediction samples on the WHU Building dataset. As examples, we present the test samples with various sizes, colors, and shapes of buildings. *Sample a* demonstrates that the baseline network with the VPA module has greater robustness to features such as shape and color and can effectively divide small buildings. Hence, the extracted buildings are closer to the ground truth. Only the baseline network with the VPA module can identify the tiny buildings contained in *Sample b*. This indicates that adding the VPA module to the network can enhance its capacity to segment small targets. As seen in *Sample c*, the VPA module improves the image extraction for large buildings and decreases the pixel-level incorrect predictions.

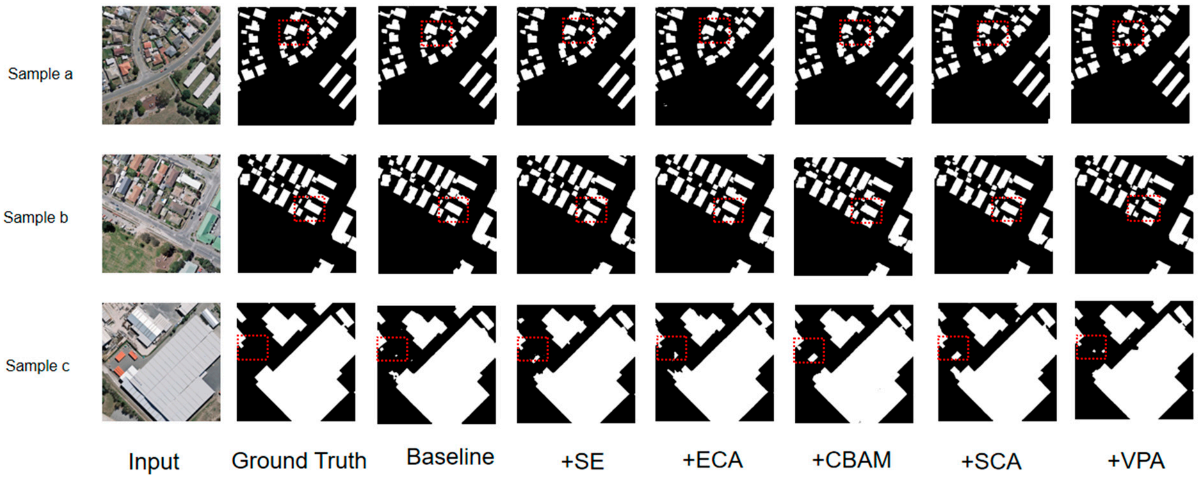

Figure 8. Comparison of prediction samples on the WHU Building dataset. The red dashed lines mark the more obvious differences.

Figure 9 displays the prediction results on the ISPRS Vaihingen dataset. In each of the three instances, there are considerable variances in size, shape, and texture between the related categories. Multiple misclassifications can be observed in the prediction samples when the baseline network is used. However, the misclassifications are minimized when utilizing the baseline network with various attention modules. The network containing a VPA module is by far the most effective at mitigating classification errors. Integrating a VPA module into the baseline network improved its capability to extract small targets, bringing it closer to the ground truth than adding the SE, ECA, CBAM, or SCA.

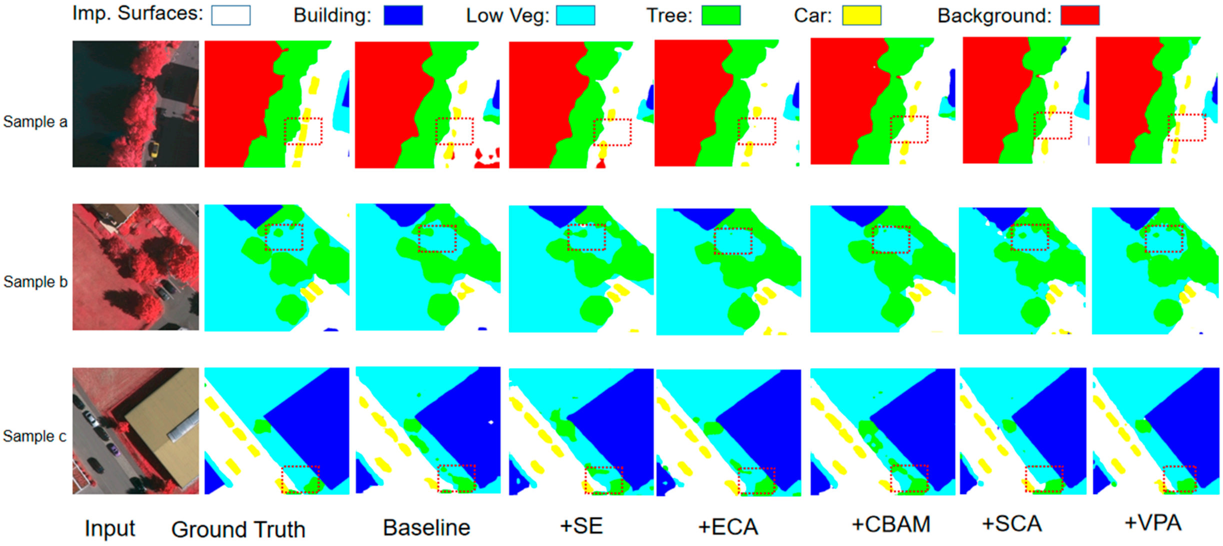

Figure 9. Comparison of prediction samples on the ISPRS Vaihingen dataset. The red dashed lines mark the more obvious differences.

6. Conclusions

The research presents an improved approach to building the attention module by combining vector average pooling with adaptive local cross-channel interaction. Our proposed attention module VPA achieves a significant effect on channel and spatial attention with a single module, thereby addressing the issue of disregarding the spatial information in a standard channel attention module. Our module minimizes the dependency produced in the tandem integration of multidimensional attention modules. The VPA module excludes the non-local attention mechanism when creating the spatial relationship of the feature maps, which makes it efficient and lightweight. It has been demonstrated that the VPA module might outperform the other attention modules. We showed that it can better enhance the segmentation capability of the network through the comparison with the other classical attention modules on the MO-CSSSD, the WHU Building dataset, and the ISPRS Vaihingen dataset. The VPA module has many potential application scenes and generalizations. Significantly, the experimental findings presented in the research indicate that adding the VPA module can improve the network's capability to distinguish and extract small targets and can boost the improvement of the segmentation edge and completeness. To summarize, the efficiency and lightweightness of the VPA module make it a valuable semantic segmentation network for dealing with large-scale remote sensing imagery. It might also spark innovative ideas for the design of attention modules in other computer vision applications.

Author Contributions: X.W. presented the main ideas of this paper and reviewed the manuscript; Y.C. provided design ideas and methods and edited and reviewed the manuscript; M.K. added methods and wrote the original draft; W.J. designed the computer program; M.W. researched the background; T.W. and M.T. edited and reviewed the manuscript; L.X. analyzed the study data; X.L. visualized the data; L.Z. and C.Z. verified the experimental design. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (Grants No. 62176085 and 61673359), the Key Scientific Research Foundation of the Education Department of Province Anhui (Grant No. KJ2020A0658), the University Natural Sciences Research Project of Province (Grant No. KJ2021ZD0118), the Hefei University Talent Research Funding (Grant No. 20RC13), the Hefei University Scientific Research Development Funding (Grant No. 20ZR03ZDA), Program for Scientific Research Innovation Team in Colleges and Universities of Anhui Province (2022AH010095), and the Hefei Specially Recruited Foreign Expert support.

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

Acknowledgments: The dataset in this paper contains the WHU Building dataset provided by Wuhan University and the Vaihingen dataset provided by ISPRS as well as the AI General Computing Platform program of Hefei University.

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this paper.

References

1. Anilkumar, P.; Venugopal, P. Research Contribution and Comprehensive Review towards the Semantic Segmentation of Aerial Images Using Deep Learning Techniques. *Secur. Commun. Netw.* **2022**, *2022*, 6010912. [CrossRef]
2. Wang, J.J.; Ma, A.L.; Zhong, Y.F.; Zheng, Z.; Zhang, L.P. Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery. *Remote Sens. Environ.* **2022**, *277*, 113058. [CrossRef]
3. Zheng, Z.; Zhong, Y.F.; Wang, J.J.; Ma, A.L. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 4095–4104. [CrossRef]
4. Huang, X.; Zhang, L.P.; Gong, W. Information fusion of aerial images and LIDAR data in urban areas: Vector-stacking, reclassification and post-processing approaches. *Int. J. Remote Sens.* **2011**, *32*, 69–84. [CrossRef]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2016; pp. 3431–3440. [CrossRef]
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
8. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
9. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
10. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
11. Chen, L.C.; Zhu, Y.K.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851. [CrossRef]
12. Sun, K.; Xiao, B.; Liu, D.; Wang, J.; Soc, I.C. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696. [CrossRef]
13. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
14. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; pp. 3–11. [CrossRef]
15. Tsotsos, J.K. ANALYZING VISION AT THE COMPLEXITY LEVEL. *Behav. Brain Sci.* **1991**, *14*, 768. [CrossRef]

16. Vikram, T.N. A Computational Perspective on Visual Attention. *Cognit. Syst. Res.* **2012**, *19–20*, 88–90. [CrossRef]
17. Li, W.; Liu, K.; Zhang, L.Z.; Cheng, F. Object detection based on an adaptive attention mechanism. *Sci. Rep.* **2020**, *10*, 11307. [CrossRef]
18. Tian, Z.; Zhan, R.; Hu, J.; Wang, W.; He, Z.; Zhuang, Z. Generating Anchor Boxes Based on Attention Mechanism for Object Detection in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2416. [CrossRef]
19. Chen, Z.; Tian, S.; Yu, L.; Zhang, L.; Zhang, X. An object detection network based on YOLOv4 and improved spatial attention mechanism. *J. Intell. Fuzzy Syst.* **2022**, *42*, 2359–2368. [CrossRef]
20. Zhang, M.; Su, H.; Wen, J. Classification of flower image based on attention mechanism and multi-loss attention network. *Comput. Commun.* **2021**, *179*, 307–317. [CrossRef]
21. Cao, P.; Xie, F.; Zhang, S.; Zhang, Z.; Zhang, J. MSANet: Multi-scale attention networks for image classification. *Multimed. Tools Appl.* **2022**, *81*, 34325–34344. [CrossRef]
22. Roy, S.K.; Dubey, S.R.; Chatterjee, S.; Baran Chaudhuri, B. FuSENet: Fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification. *Iet Image Process.* **2020**, *14*, 1653–1661. [CrossRef]
23. Guo, M.; Xu, T.; Liu, J.; Liu, Z.; Jiang, P.; Mu, T.; Zhang, S.; Martin, R.R.; Cheng, M.; Hu, S. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
25. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. [CrossRef]
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H.; Soc, I.C. Dual Attention Network for Scene Segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3141–3149. [CrossRef]
27. Jin, Z.; Liu, B.; Chu, Q.; Yu, N. ISNet: Integrate Image-Level and Semantic-Level Context for Semantic Segmentation. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 7169–7178. [CrossRef]
28. Liu, S.; Cheng, J.; Liang, L.; Bai, H.; Dang, W. Light-Weight Semantic Segmentation Network for UAV Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8287–8296. [CrossRef]
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
30. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11534–11542.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 173–190.
33. Wang, Y. Remote Sensing Image Semantic Segmentation Algorithm Based on Improved ENet Network. *Sci. Program.* **2021**, *2021*, 5078731. [CrossRef]
34. Sofla, R.A.D.; Alipour-Fard, T.; Arefi, H. Road extraction from satellite and aerial image using SE-Unet. *J. Appl. Remote Sens.* **2021**, *15*, 014512. [CrossRef]
35. Han, G.; Zhang, M.; Wu, W.; He, M.; Liu, K.; Qin, L.; Liu, X. Improved U-Net based insulator image segmentation method based on attention mechanism. *Energy Rep.* **2021**, *7*, 210–217. [CrossRef]
36. Han, L.; Zhao, Y.; Lv, H.; Zhang, Y.; Liu, H.; Bi, G. Remote Sensing Image Denoising Based on Deep and Shallow Feature Fusion and Attention Mechanism. *Remote Sens.* **2022**, *14*, 1243. [CrossRef]
37. Liu, R.R.; Tao, F.; Liu, X.T.; Na, J.M.; Leng, H.J.; Wu, J.J.; Zhou, T. RAANet: A Residual ASPP with Attention Framework for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3109. [CrossRef]
38. Wang, M.Y.; Wang, J.T.; Liu, C.; Li, F.Y.; Wang, Z.Y. Spatial-Coordinate Attention and Multi-Path Residual Block Based Oriented Object Detection in Remote Sensing Images. *Int. J. Remote Sens.* **2022**, *43*, 5757–5774. [CrossRef]
39. Li, Y.; Si, Y.; Tong, Z.; He, L.; Zhang, J.; Luo, S.; Gong, Y. MQANet: Multi-Task Quadruple Attention Network of Multi-Object Semantic Segmentation from Remote Sensing Images. *Remote Sens.* **2022**, *14*, 6256. [CrossRef]
40. Zhao, D.; Wang, C.; Gao, Y.; Shi, Z.; Xie, F. Semantic Segmentation of Remote Sensing Image Based on Regional Self-Attention Mechanism. *IEEE Geosci. Remote Sens. Lett.* **2022**, 19. [CrossRef]
41. Zhang, Y.J.; Cheng, J.; Bai, H.W.; Wang, Q.; Liang, X.Y. Multilevel Feature Fusion and Attention Network for High-Resolution Remote Sensing Image Semantic Labeling. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6512305. [CrossRef]
42. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
43. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. Acm.* **2017**, *60*, 84–90. [CrossRef]

44. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995. [CrossRef]
45. Chen, Y.; Yang, X.; Xu, L.; Li, X. Research on multi-scale target semantic segmentation for coastal ecological supervision. *Environ. Resour.* **2022**, *4*, 48–61.
46. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]
47. Guo, R.; Liu, J.; Li, N.; Liu, S.; Chen, F.; Cheng, B.; Duan, J.; Li, X.; Ma, C. Pixel-Wise Classification Method for High Resolution Remote Sensing Imagery Using Deep Neural Networks. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 110. [CrossRef]
48. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-Resolution Context Extraction Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 71. [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.