remote sensing

MDPI

Article

Multi-Field Context Fusion Network for Semantic Segmentation of High-Spatial-Resolution Remote Sensing Images

Xinran Du [1], Shumeng He [1], Houqun Yang [1,*] and Chunxiao Wang [2]

[1] College of Computer Science and Technology, Hainan University, Haikou 570100, China
[2] Hainan Geomatics Centre of Ministry of Natural Resources, Haikou 570203, China
* Correspondence: yhq@hainanu.edu.cn

Abstract: High spatial resolution (HSR) remote sensing images have a wide range of application prospects in the fields of urban planning, agricultural planning and military training. Therefore, the research on the semantic segmentation of remote sensing images becomes extremely important. However, large data volume and the complex background of HSR remote sensing images put great pressure on the algorithm efficiency. Although the pressure on the GPU can be relieved by down-sampling the image or cropping it into small patches for separate processing, the loss of local details or global contextual information can lead to limited segmentation accuracy. In this study, we propose a multi-field context fusion network (MCFNet), which can preserve both global and local information efficiently. The method consists of three modules: a backbone network, a patch selection module (PSM), and a multi-field context fusion module (FM). Specifically, we propose a confidence-based local selection criterion in the PSM, which adaptively selects local locations in the image that are poorly segmented. Subsequently, the FM dynamically aggregates the semantic information of multiple visual fields centered on that local location to enhance the segmentation of these local locations. Since MCFNet only performs segmentation enhancement on local locations in an image, it can improve segmentation accuracy without consuming excessive GPU memory. We implement our method on two high spatial resolution remote sensing image datasets, DeepGlobe and Potsdam, and compare the proposed method with state-of-the-art methods. The results show that the MCFNet method achieves the best balance in terms of segmentation accuracy, memory efficiency, and inference speed.

Keywords: semantic segmentation; high spatial resolution remote sensing images; memory efficiency

Citation: Du, X.; He, S.; Yang, H.; Wang, C. Multi-Field Context Fusion Network for Semantic Segmentation of High-Spatial-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5830. https://doi.org/10.3390/rs14225830

Academic Editor: Melanie Vanderhoof

Received: 6 October 2022
Accepted: 15 November 2022
Published: 17 November 2022

1. Introduction

Image semantic segmentation is a very important topic in remote sensing image interpretation and plays a key role in various practical applications, such as urban planning [1–4], geohazard monitoring [5,6] land change detection [7], etc. The task aims to assign semantic category labels to each pixel with semantic information in the image [8,9]. High spatial resolution (HSR) remote sensing images have gradually become the primary source of interpretation data for remote sensing with the rapid development of aerospace technology and information technology. In comparison to ordinary resolution remote sensing images, it provides more information about the spatial structure and texture information of target features more clearly, which is crucial to the accuracy of remote sensing image segmentation. In comparison to traditional computer vision images, remote sensing images have a special field of view; i.e., overhead imaging. At the same time, it also has a complex background, and a pair of remote sensing images typically contains a large number of buildings, vegetation, farmland and other multi-category features and geomorphological element's information. Furthermore, compared to lower spatial resolutions, HSR images of the same field range have higher image sizes and greater pixel detail. This means that HSR images requires more pixels to describe the same field range of images, and the problem of

high-resolution image (i.e., large pixel size) input models needs to be solved in order to use the global information of HSR images. There are two mainstream solutions for the semantic segmentation of high-resolution images: one way is to input a model after down-sampling the input image. The other way is to divide the image into patches and process each patch independently. However, many segmentation objects of remote sensing images are small targets (tens or even a few pixels), which means that the information content of segmented objects is small; down-sampling directly will result in the disappearance of such small targets and also lose the advantage of rich feature information in HSR images. Moreover, as remote sensing images are characterized by a high degree of inter-class similarity and intra-class dissimilarity, dividing the image into independent patches will lack contextual spatial information to detect feature changes of semantically different neighboring objects, particularly for HSR images, which have a smaller field area for the same size image. Figure 1 illustrates the results of HSR remote sensing images using these two preprocessing methods.

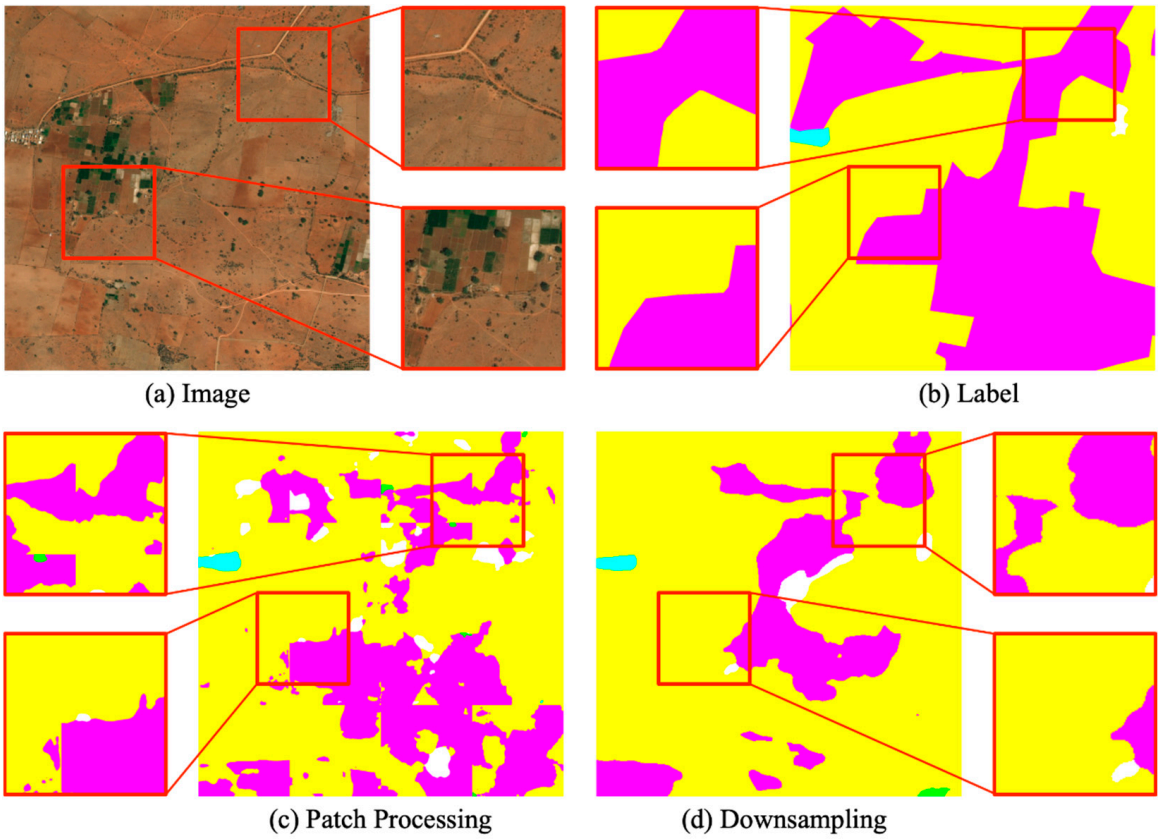

Figure 1. Results of semantic segmentation using patch processing and down-sampling processing. (**a**) is the input image and (**b**) is the labeled image, where the area circled in red contains two highly confusing categories-Agriculture and Rangeland. (**c**,**d**) represent the two processing methods respectively. As can be seen from the figures, down-sampling loses fine details, while patch processing wrongly classifies local patches due to the lack of the global context.

One way to address the limitations of the above methods is to combine them; i.e., to fuse the global and local segmentation processes. On the one hand, the global view of the entire image can be used to supplement the contextual information of the local patches. By analyzing the local patches, we can refine the segmentation boundaries and recover the details lost due to the down-sampling process of the global segmentation process. This approach has been successfully demonstrated by several models, such as [10] which performs a bidirectional combination of feature maps with global context and local fined structure at each layer. Nevertheless, for an HSR input image, there is a tremendous gap between the scale of the whole image and the scale of the local patches. This will

lead to training difficulties, and its feature sharing scheme will not spatially correlate local features with global features. In addition, it treats each local image equally and fuses them with the entire image segmentation, which will consume more computational resources. For effective and efficient fusion of global context and local details, we considered the importance of local images and selectively perform feature fusion. Meanwhile, to complement the large gap between the two scales of global and local images, and we propose to consider multiple scales in between.

In this paper, we propose a multi-field context fusion network (MCFNet), which is centered on a patch selection module (PSM) and a multi-field context fusion module (FM). MCFNet consists of two phases: a global segmentation stage and a local segmentation stage. In the stage of global segmentation, the global image is input to the model in order to obtain coarse segmentation results. To further enhance the segmentation accuracy, PSM is used to select local patches that are difficult to be segmented by the model. In the stage of local segmentation, FM enhances the segmentation of local patches by adaptively aggregating the contextual semantics of multiple visual fields around the patches. As a result of integrating the context at both levels of the local and global stages, it is possible to capture contextual information adaptively from multiple perspectives, which can increase the performance of the model without consuming additional GPU memory or increasing inference time. To evaluate the effectiveness of the model, we conducted extensive experiments demonstrating that our proposed model outperforms state-of-the-art methods on the publicly available high-spatial-resolution image datasets DeepGlobe and Potsdam. The important contributions of our study are summarized as follows.

1. We propose a patch selection module for locating poorly segmented local patches in the global image so that further enhancement of segmentation can be performed. It alleviates the burden of segmentation model, and the module can be used with any popular semantic segmentation network.
2. We propose a module named FM for aggregating the semantics of multi-field contexts. The module performs adaptive weighting of the local patches selected by PSM with multiple fields of view to enhance the feature representation by aggregating multi-level contextual information.
3. We demonstrate the effectiveness of our approach by achieving state-of-the-art semantic segmentation performance on two publicly available high-spatial-resolution re-mote sensing image datasets.

The rest of this article is organized as follows: Section 3 describes in detail the design idea and composition of the proposed MCFNet framework. Experimental datasets, model evaluation methods, experimental procedures, and analysis of experimental results are given in Sections 4 and 5.

2. Related Work

2.1. Semantic Segmentation

Image semantic segmentation is an image prediction task for dense classification that infers the category to which each pixel belongs based on the image's characteristics. The fully convolutional neural network [11] was the first CNN network structure for the image semantic segmentation task, which replaced all fully connected layers of DCNN for image classification with fully convolutional layers to output 2D feature maps. UNet [12,13] utilized a symmetric encoder-decoder structure with jump connections to combine low-level features with high-level features. DeconvNet [14] and SegNet [15] also adopted a comparable architecture. DeepLab [16,17] applied Atrous Convolution to the network structure in order to broaden the convolutional (filters) field of perception and establish connections between distant pixels. Inspired by Transformer in NLP, the researchers extended Transformer to semantic segmentation tasks. The recently proposed ViT [18] fully adopts the standard structure of Transformer in its structure, achieving the most advanced level in multiple image recognition benchmark tasks. Ref. [19] used self-attention instead of partial convolution to enhance CNN's feature extraction ability, so as to improve image

classification performance. SETR [20] deploy a pure transformer to encode the image into a series of patches, and strengthen the segmentation effect by modeling the global context in each layer of transformer. However, although the network model based on the transformer has good performance in precision, it has a large number of network model parameters. Moreover, if the training dates are insufficient, it is easy to cause overfitting. In the field of remote sensing semantic segmentation, HMANet [21] proposed a new attention framework to reduce feature redundancy and improve the efficiency of the self-attention mechanism through region representation. FarSeg [22] enhanced the recognition of foreground features by learning the foreground-related context associated with the foreground-scene relationship of remote sensing images. Refs. [23,24] use HRNet to enhance the low-to-high features extracted from different branches separately to strengthen the embedding of scale-related contextual information. Although these models achieve good performance, they are not applicable to the semantic segmentation of high-resolution images because they are only concerned with improving the accuracy of semantic segmentation, and not the efficiency of computation. To increase the segmentation speed and reduce the memory usage during semantic segmentation, Enet [25] employed an asymmetric encoder-decoder structure to conserve GPU memory by minimizing floating point operations. ESPNets [26,27] accelerated convolutional computation by employing the split-merge or scale-reduction principle. There were efficient segmentation models using lightweight backbone networks (MobileNet [28–30] and ShuffleNet [31]), or some compression techniques (pruning [13], knowledge distillation [32]). Despite the fact that these real-time segmentation networks have low time complexity and memory consumption, they are not optimized for high-resolution images, and their performance on high-resolution images is significantly inferior to that of other networks. Consequently, we propose a semantic segmentation model for local location enhancement, which can dynamically select local locations with poor global segmentation, striking a balance between reducing model burden and enhancing model accuracy.

2.2. Multi-Scale Context Aggregation and Refining Segmentation

Multi-Scale aggregation [10,33–36] has been demonstrated to be effective in semantic segmentation by combing features from various stages to provide more contextual information for each pixel. Feature pyramid networks [37] upsampled feature maps at different scales and aggregated them with the output from lower layers. PSPNet [38] introduced a pyramid pooling module to extract contextual information in images as well as global information from different receptive fields. ICNet [39] introduced a cascade feature fusion module based on PSPNet to improve the prediction accuracy of the model. HRNet [40] achieved strong semantic information and precise location information by parallelizing multiple branches of resolution and interacting information interaction between different branches. In more recent approaches, attention mechanisms [41–43] are also employed to add more contextual information to each pixel. Transferring global context information to local locations is also an efficient method, ParseNet [44] aggregated the global context to the local field of view in order to provide additional information. GLNet [10] retained global and local information and interacts with each other through a deep sharing layer, allowing it to balance its performance and GPU memory usage. MagNet [45] proposed a new multi-scale framework that addresses local ambiguity by viewing images at multiple zoom levels and directly outputs high-resolution segmentation. CascadePSP [46] uses refining segmentation to improve segmentation accuracy at local patch. However, these models select local patches without directionality, and they segment every local patch of the image for enhancement without considering the efficiency. We believe that not every local patch needs segmentation enhancement. Unlike previous work, our model performs multi-field semantic feature fusion selectively only at local locations and adaptively fuses the contextual semantic at each scale.

3. Proposed Method

On the basis of the foregoing research status and improvement ideas, we propose the Multi-Field Context Fusion Network (MCFNet), a segmentation network based on high-resolution images. In Section 3.1, we first give an overview of the network. From Section 3.2 to Section 3.4, we further introduce the composition of the network, including backbone network, patch selection module (PSM), multi-field context fusion module (FM), and structural adaptive weighting block (AWB).

3.1. Overview of Network Architecture

As shown in Figure 2, MCFNet consists of two stages: the segmentation stage for global images, and the enhancement stage for local patches. The core of the segmentation framework is comprised of a patch select module and a contextual fusion module. PSM is used to select the patches with unsatisfactory segmentation results in the first stage of segmentation, and FM is used to fuse the image features from different fields of view to achieve the segmentation effect of enhanced local patches. The segmentation module in our framework can be any segmentation backbone that is a combination of PSM and FM relationships.

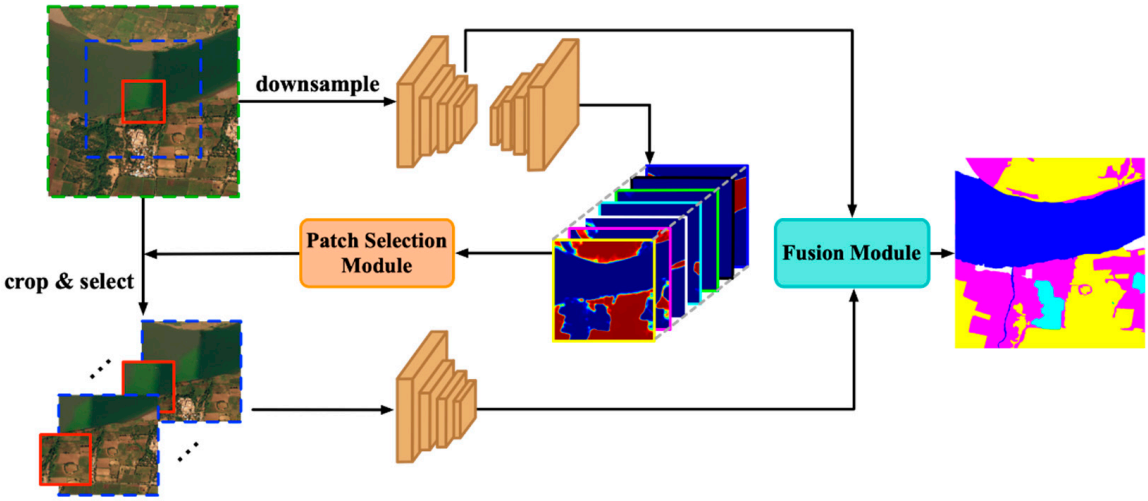

Figure 2. Overview of our proposed MCFNet. The global and local branches are images that have been down-sampled and cropped respectively. PSM module is used to identify poorly segmented local patches within the global segmentation, and FM module performs segmentation enhancement on those poorly segmented local patches. The final segmentation is created by aggregating the global and local branch feature maps.

3.2. Backbone Network

A typical semantic segmentation network consists of an encoder and a decoder, which encodes the input image and reduce the resolution of the image by down-sampling to reduce the computation, respectively. Following the choices of previous studies [47,48], we also selected ResNet [49] and pre training on ImageNet [50] as the encoder. We removed the final pooling layer and the FC layer from ResNet-50, leaving the rest unchanged. We chose feature pyramid network (FPN) [37] as the decoder, which mainly addresses the multi-scale issue in semantic segmentation by up-sampling the higher-level features and concatenating the lower-level features in a top-down manner. It substantially improved the performance of small object segmentation without increasing the computational effort. Actually, the MCFNet framework we proposed can use the majority of the dominant segmentation network today as the backbone.

3.3. Patch Selection Module

For the results of direct down-sampling for segmentation, we believe that not every local patch requires being enhanced. In the case of HSR remote sensing images, there exist consecutive large areas of the same class of objects, for which local patches do not need to be enhanced. Therefore, we consider a scoring mechanism to select the local patches that need to be enhanced.

Confidence is a commonly used metric in statistics for assessing a system's reliability. The significance of confidence stems from the fact that if a decision support a system's confidence level in predicting that a particular sample is too low, additional decision systems may be required to participate in the decision process. In object detection, non-maximum suppression (NMS) [51–54] selects an optimal bounding box from many candidate boxes by suppressing the bounding boxes with low confidence. Inspired by this, we designed PSM module and proposed a confidence-based local patch judging criterion. As illustrated in Equation (1), when the score μ_{local} of local patch is smaller than the global image score μ_{global}, the local patch will be selected by PSM for segmentation reinforcement. In the following, we detail the structure of PSM and the process of PSM is shown in Figure 3.

$$\begin{cases} \mu_{\text{local}} < \mu_{\text{global}}, \text{ patch } i \text{ is selected to refine} \\ \mu_{\text{local}} \geq \mu_{\text{global}}, \text{ patch } i \text{ is not selected to refine} \end{cases} \tag{1}$$

For each high-resolution image I_{hr}, down-sample it to I_{lr} and feed into the backbone of MCFNet to extract the deep features $M_{\text{global}} \in \mathbb{R}^{h \times w \times class}$. Subsequently, *softmax* function is applied to M_{global} in the channel dimension and take the maximum value of each pixel on the channel dimension, denoted as $P_{ij} \in (0, 1)$. Hence, the confidence level of the global image is noted as $C_{\text{global}} \in \mathbb{R}^{h \times w \times 1}$ and Equation (2) presents the calculation formulas.

$$C_{\text{global}} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ p_{21} & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} \end{bmatrix}, 1 \leq i \leq h, 1 \leq j \leq w \tag{2}$$

and PSM records the score of the global image as μ_{global}, where

$$\mu_{\text{global}} = \frac{1}{h \times w} \sum_{\substack{1 \leq i \leq w \\ 1 \leq j \leq h}} P_{ij}. \tag{3}$$

We define local patches in such a way that the global image I_{hr} is divided equally into N local patches, without overlap, denoted as $X_{\text{local}} \in \mathbb{R}^{m \times n \times 1}$, where N is the hyperparameter indicating the number of local patches.

$$X_{\text{global}} = \left\{ X^1_{\text{local}}, X^2_{\text{local}}, \cdots, X^{\text{N}}_{\text{local}} \right\} \tag{4}$$

$$C_{\text{global}} = \left\{ C^1_{\text{local}}, C^2_{\text{local}}, \cdots, C^{\text{N}}_{\text{local}} \right\} \tag{5}$$

The score of X_{local} can be calculated from C_{global} in the same way as the calculation of μ_{global}.

$$\mu_{\text{local}} = \frac{1}{m \times n} \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} P_{ij} \tag{6}$$

We believe that the average confidence of the image can be used to quantify the degree of certainty of the model's prediction for that image. To demonstrate this concept, we tested it on DeepGlobal data. Firstly, we input the down-sampled DeepGlobe image I_{lr} into the backbone to obtain the feature map M_{global} and the prediction accuracy. Then,

M_{lr} is transformed into the confidence matrix C_{global} in the above manner. We crop C_{global} uniformly into 16 local patches, i.e., N = 16, and the score of each local patch is μ_{local}. Figure 3a reflects the positive correlation between the score μ_{local} assigned by PSM and the prediction accuracy. Furthermore, we define the relative score $\mu_{relative} = \mu_{local} - \mu_{global}$, Figure 3b visualizes the relationship between $\mu_{relative}$ and segmentation accuracy. It can be seen from the figure that the relative score is positively correlated with the segmentation accuracy. We calculate the mean value of the local patch scores that are lower and higher than the global image scores, and discover that the difference in accuracy between the two is 3%. This further demonstrates the efficiency of our approach.

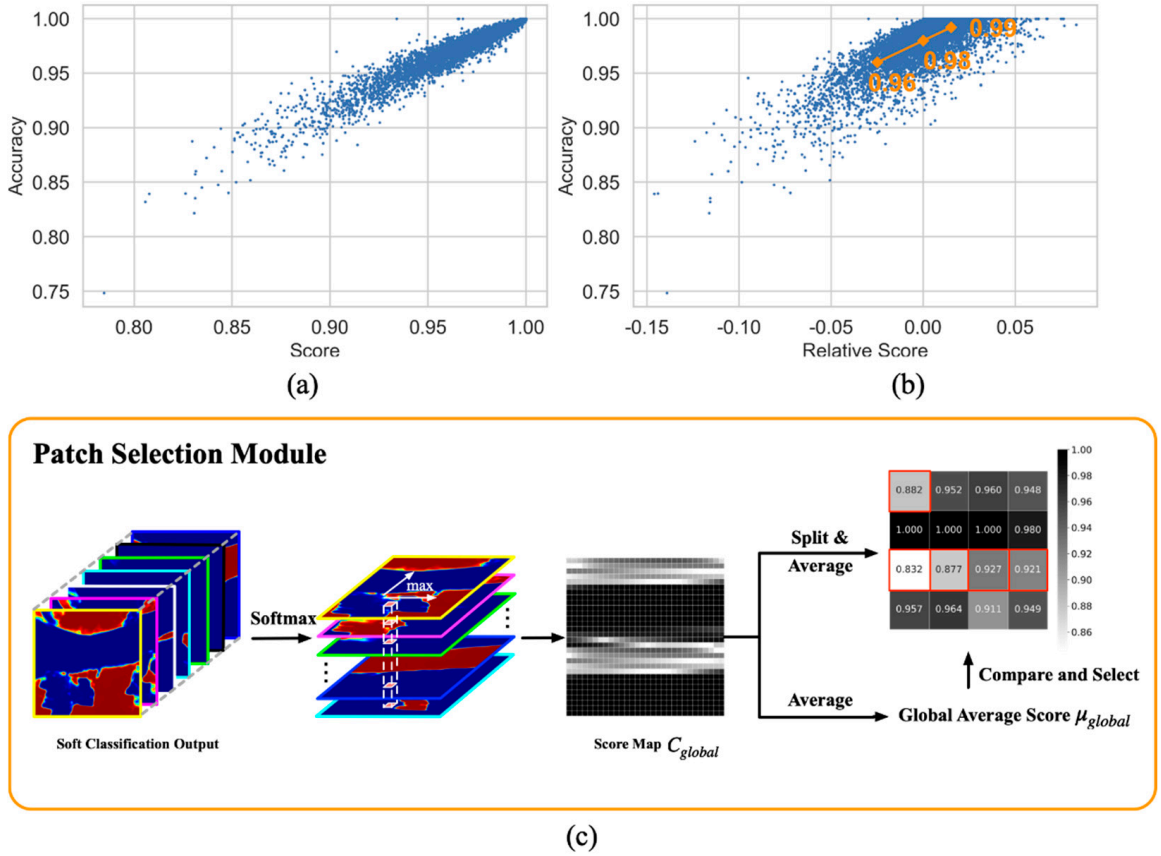

Figure 3. (**a**) The relationship between the score of local patch and its segmentation accuracy, where accuracy is the percentage of correctly predicted samples; (**b**) Relative score of local patch and its accuracy, where the relative score is defined as $\mu_{global} - \mu_{local}$; (**c**) PSM module. PSM selects the local patch that requires reinforcement based on the global image's soft classification output.

3.4. Multi-Field Context Fusion Module

Generally, images with larger fields of view contain more contextual information at the same resolution, which facilitates the identification of similar features with a wide range of scale variation. Images with smaller fields of view have more texture details, which are beneficial to capturing the subtle differences between different classes of features. Therefore, we propose a module for multi-field contextual semantic fusion named FM. It combines feature information from different fields of view to discover features with differentiation, and considers the class of each pixel from the overall image to overcome inter-class similarity and intra-class dissimilarity of remote sensing images. This module accepts a multi-field image of the local patch selected by PSM as the input, and outputs multi-field, which aggregates features that achieve enhanced segmentation by aggregating contextual information and local texture details at various scales of the local patch. The structure diagram of FM is as Figure 4.

According to the score results derived from PSM, we adaptively select the local patches that need to be enhanced from the global image. For local patches, there is a large gap between its scale and the global image scale, which will lead to difficulties in feature fusion. To eliminate this phenomenon, we propose to consider multiple scales in between. In practice, we create three distinct scales of contextual regions, denoted as $X_1 \in \mathbb{R}^{h_1 \times w_1 \times c}$, $X_2 \in \mathbb{R}^{h_2 \times w_2 \times c}$ and $X_3 \in \mathbb{R}^{h_3 \times w_3 \times c}$. X_1 is the local patch selected by PSM, X_2 is a different scale image with the same centroid as X_1 in the global image, and X_3 is the global image that has been down-sampled, where $h_1 < h_2 < h_3$ and $w_1 < w_2 < w_3$. We scale $\{X_1, X_2, X_3\}$ to the same resolution and feed them into the encoder network, cropping the output feature map to the corresponding position as $\{F_1, F_2, F_3\}$. $\{F_1, F_2, F_3\}$ is the feature information of three different fields of view at the same local location. In order for the model to make efficient use of these features, we designed the Adaptive Weighting Module(AWM) at each stage of decoding, which uses a pyramid structure and consider both the channel domain and spatial domain to assign weights to each pixel feature, similar to CBAM [55]. The specific approach is described below.

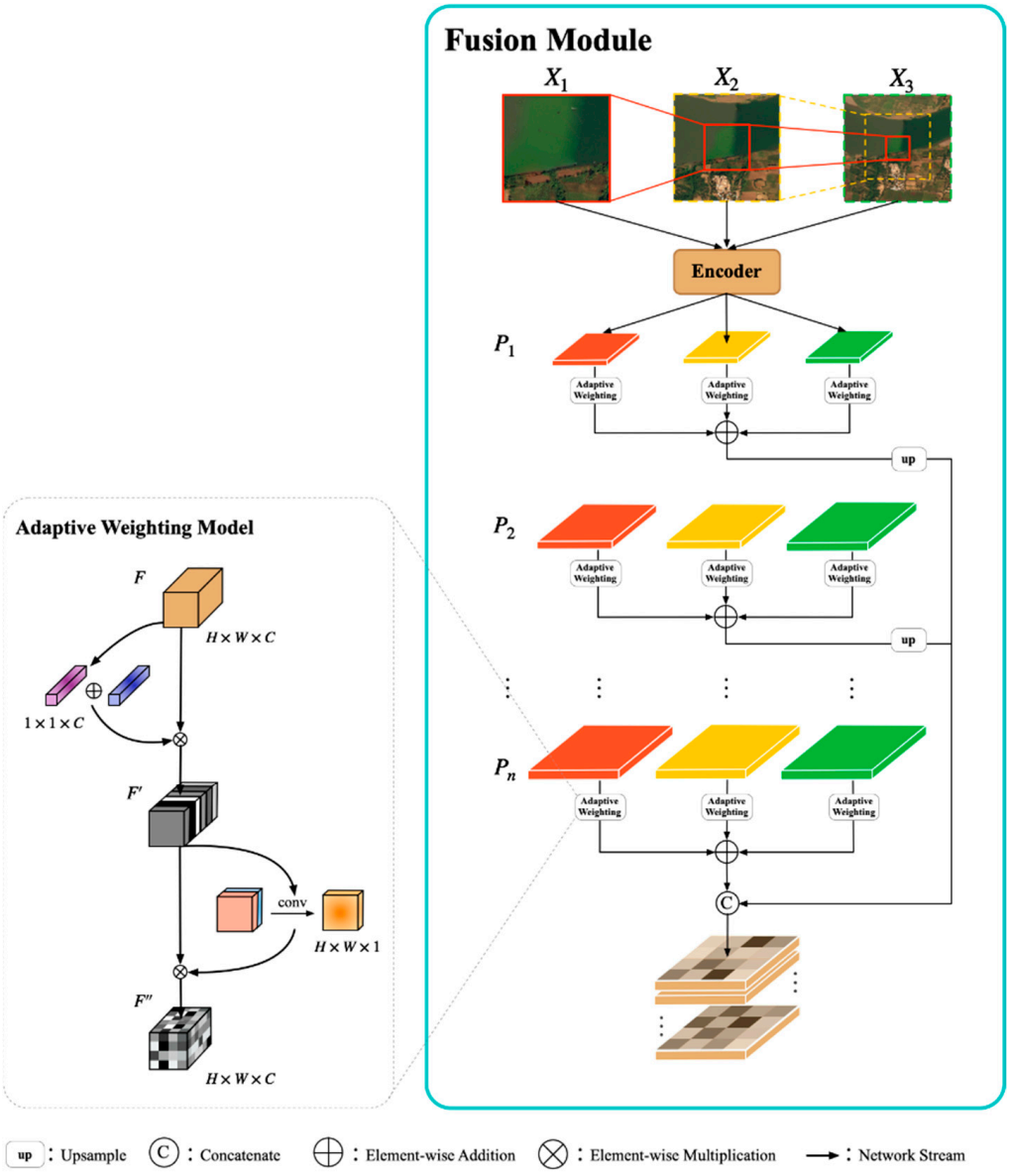

Figure 4. The multi-field context fusion module. FM adaptive fusion of multi-field semantic for local patch selected by PSM.

In regard to the channel dimension, each layer of channels in the feature map represents a distinct type of feature information that contributes differently to the task. We use the abstract property of the pooling operation to reconstruct the feature F as a $1 \times 1 \times C$ channel description. The reconstructed F is mapped through the fully connected network to obtain the channel weight matrix $M_c \in \mathbb{R}^{1\times1\times c}$; this process can be summarized as Equations (7) and (8)

$$\begin{aligned} M_c(F) &= \sigma(\boldsymbol{MLP}(\boldsymbol{Avg\ Pool}(F)) + \boldsymbol{MLP}(\boldsymbol{Max\ Pool}(F))) \\ &= \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(F^c_{avg})) + \boldsymbol{W}_1(\boldsymbol{W}_0(F^c_{max}))) \end{aligned} \tag{7}$$

$$F' = M_c(F) \odot F, \tag{8}$$

where σ denotes the sigmoid function and $\odot$ denotes element-wise multiplication. W_0 and W_1 are the shared network MLP weights.

In the spatial dimension, the use of spatial relationship features enhances the ability of the model to distinguish the image content. To direct the network to focus on local feature information of different field of view images, we assign feature weights at the spatial level to determine the spatial locations where key information is aggregated. Unlike the operations of the previous, F' performs feature compression along the channel axis to obtain a feature description of $\boldsymbol{h}\times\boldsymbol{w}\times\boldsymbol{1}$. We apply a convolution layer to generate a spatial attention map $M_s \in \mathbb{R}^{h\times w\times c}$ which encodes where to emphasize or suppress and this process can be expressed as Equations (9) and (10).

$$\begin{aligned} M_s(F') &= \sigma(f_{conv}([\boldsymbol{Avg\ Pool}(F'); \boldsymbol{Max\ Pool}(F')])) \\ &= \sigma(f_{conv}([F'^s_{avg}; F'^s_{max}])) \end{aligned} \tag{9}$$

$$F'' = \boldsymbol{M}_s(F') \odot F', \tag{10}$$

where σ denotes the sigmoid function and f_{conv} represents a convolution operation with the filter. $\odot$ denotes element-wise multiplication. The above calculation process of AWM can be simplified as Equation (11).

$$F'' = AWM(F) \tag{11}$$

At each stage of decoding, we use AWM to assign weights to the image features of different fields of view. Equation (12) presents the calculation formulas, where P_i represents the decoding of stage i. $\boldsymbol{F}''_1$, $\boldsymbol{F}''_2$ and $\boldsymbol{F}''_3$ represent the weighted feature maps of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$, respectively. In this way, the model can adaptively aggregate complementary contextual feature information. Figure 5 depicts the feature fusion process in more detail.

$$P_i = AWM\left((F'')^i_1\right) + AWM\left((F'')^i_2\right) + AWM\left((F'')^i_3\right) \tag{12}$$

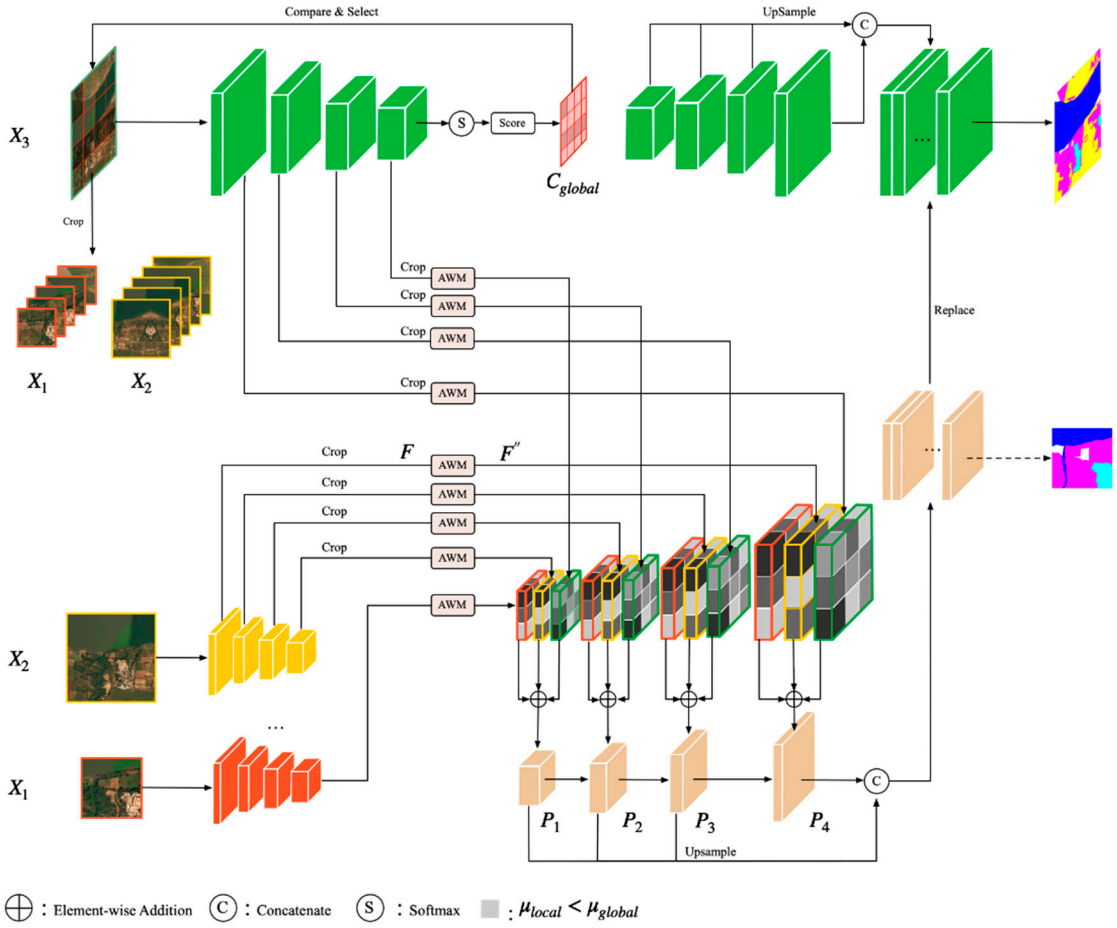

Figure 5. Interaction of PSM and FM.

4. Experiment

In this section, we describe in detail the datasets used in our experiment, introduce the evaluation indicators, and present the experimental results.

4.1. Datasets

We evaluated the performance of MCFNet on two high resolution datasets: DeepGlobe [56] and Potsdam. Table 1 contains some information about these datasets. These two datasets have great challenges for the semantic segmentation task. DeepGlobe contains the confusing categories of Agriculture and Rangeland. In addition, there are slender rivers that it are difficult for the model to segment. Potsdam contains the confusing categories of Low Vegetation and Tree, as well as smaller objects such as cars. Figure 6 shows the two datasets.

1. DeepGlobe: DeepGlobe [56] is a dataset of high-spatial-resolution satellite images. The dataset contains 803 images with a resolution of 2448 × 2448 pixels that have been annotated with seven landscape classes, including one that is an unknown class. Following the evaluation protocol of [10], the unknown class is ignored when calculating mIoU, so there are only six classes to consider. We used the same train/validation/test split as reported in [10], with 455, 207, and 142 images for training, validation, and testing, respectively.
2. Potsdam: Potsdam consists of 38 ultra high-spatial-resolution images, each with 6000 × 6000 pixels and it is representative of urban remote sensing data with its large buildings, narrow streets, and dense settlement structures. Tiles are composed of red-green-blue-infrared (RGB-IR) four-channel images. The dataset also includes a Digital Surface Model (DSM) and a normalized DSM (nDSM). In this study, we only

used RGB data. We randomly divide images into training, validation and testing sets with 26, 6 and 6 images respectively.

Table 1. Details of HSR Datasets used to evaluate of framework.

Dataset	Context	Resolution	Spatial Resolution	No. Classes
DeeplGlobe	aerial scene	2448 × 2448	0.5 m	7
Potsdam	urban scene	6000 × 6000	0.05 m	6

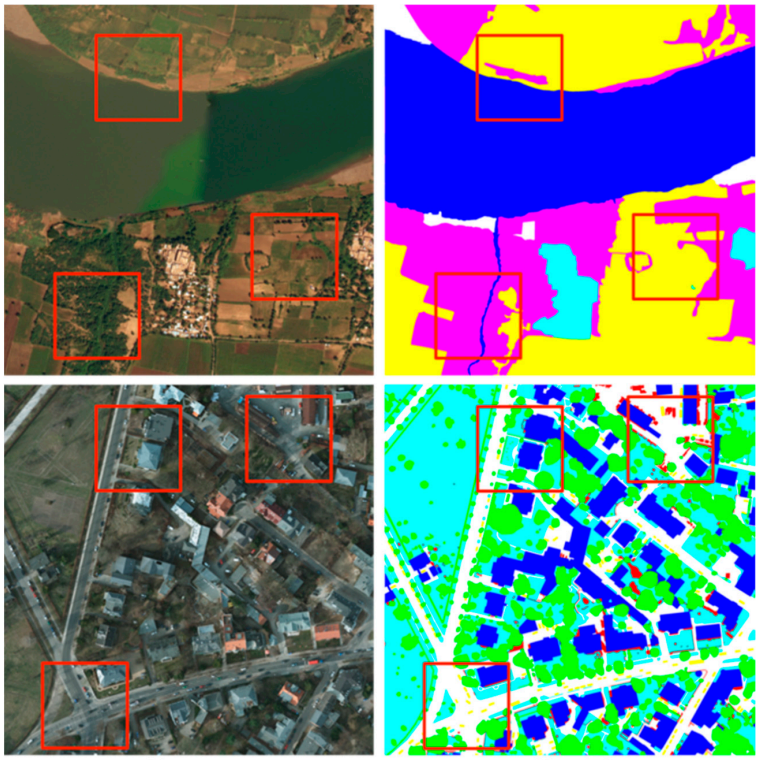

Figure 6. Dataset of DeepGlobe and Potsdam. The red area framed from the picture is the area that is difficult to segment.

4.2. Experimental Setup

The Feature Pyramid Network (FPN) [37] with a Resnet50 backbone was used as the segmentation network as in the previous work GLNet [10]. Additionally, the input sizes 512×512 px. While training our module, we randomly cropped image patches and applied the following data augmentations: rotation, and horizontal and vertical flipping. We used an Adam optimizer with decayed weight of 5×10^{-4}, and an initial learning rate of 1×10^{-3}. We trained the module for 50 epochs, and the learning rate was decreased tenfold at epoch 10, 20, 30 and 40. We used Focal Loss [57] with $fl = 2$ as the loss function for training modules. Equations (13) present the calculation formulas. We implemented MCFNet using PyTorch to start from the public implementation of FPN with ResNet50, and used a batch size of 8 for training on a workstation with a signal NVIDIA RTX 3090 GPU.

$$L_{\mathrm{FL}} = -(1 - p_t)^{\gamma} \log(p_t) \tag{13}$$

4.3. Evaluation Metrics

Semantic segmentation task often suffers from severe class imbalance problems, e.g., the category "agriculture" occupies much more area (i.e., pixels) than "water" in DeepGlobe.

The pixel-wise metric F1 and accuracy can hardly reflect how the models handle this problem. Instead, mIoU measures the average segmentation quality of each category. Thus, three metric are used in evaluating the segmentation performance of the proposed network. They are intersection over union (IoU), mean intersection over union (mIoU), and Memory Usage (MB). IoU is used to evaluate the segmentation performance of each class and mIoU is used to evaluate the average segmentation performance of all classes as well as Memory Usage is used to measure the GPU memory usage of a model. Suppose there are N classes in total. Denote P_{ii} $(i = 1, 2, \ldots\ldots, N)$ as the number of pixels of class i predicted to belong to class i, and denote P_{ij} $(i = 1, 2, \ldots\ldots, N)$ as the number of pixels predicted to belong to class j. Then mathematical formulas of IoU and mIoU can be written by

$$\mathrm{IoU} = \frac{p_{ii}}{\sum_{j=0}^{N} p_{ij} + \sum_{j=0}^{N} p_{ji} - p_{ii}},\ \mathrm{i} \in (\boldsymbol{1}, \boldsymbol{N}) \tag{14}$$

$$\mathrm{mIoU} = \frac{1}{N+1} \sum_{\mathrm{i}=0}^{\mathrm{N}} \frac{p_{ii}}{\sum_{j=0}^{N} p_{ij} + \sum_{j=0}^{N} p_{ji} - p_{ii}} \tag{15}$$

4.4. Result

4.4.1. The Result on DeepGlobe Dataset

The results of the comparison between our method and other state-of-the-art methods are presented in Table 2. They demonstrate that all models achieve higher mIoU under global inference, but consume a large amount of GPU memory. With patch-based inference, memory consumption decreases, but so does the accuracy. GLNet is a segmentation network for high-resolution images, which fuses local and global features throughout the segmentation process. It is reasonably effective, but it is still time consuming due to the fusion of global features with each local patch. In comparison, the PSM module only selects local patches with poor segmentation effect for feature fusion, thereby reducing memory usage and enhancing accuracy. MCFNet outperforms other state-of-the-art method with the segmentation accuracy of 72.6% in mIoU and only uses 1538 MB GPU memory. Figure 7 depicts the segmentation results of MCFNet.

Table 2. Performance of MCFNet and other segmentation models on DeepGlobe Dataset.

Model	Patch Inference		Global Inference	
	mIoU(%)	**Memory(MB)**	**mIoU(%)**	**Memory(MB)**
UNet	37.3	949	38.4	5507
ICNet	35.5	1195	40.2	2557
PSPNet	53.3	1513	56.6	6289
SegNet	60.8	1139	61.2	10,339
Deeplabv3+	63.1	1279	63.5	3199
FCN-8s	64.3	1963	70.1	5227
	mIoU(%)		**Memory(MB)**	
GLNet	71.6		1865	
MCFNet	72.6		1538	
MCFNet-All Fusion	73.0		1538	

We present the segmentation performance of baseline, GLNet, MCFNet and MCFNet -all fusion for all DeepGlobe classes in Table 2. Among all categories, the "Agriculture" class gains the highest classification accuracy of 78.9%, because it has a large proportion in the dataset. In contrast, category "Rangeland" and "Barren" has the lower prediction accuracy, because class "Agriculture", "Rangeland" and "Barren" are three classes of objects that are similar in appearance but different in class. Compared with baseline, MCFNet improves the segmentation accuracy by 2.5% and 4.7% on Rangeland and Barren respectively. Our experimental results demonstrate that MCFNet has good discriminative ability for ambiguous categories, such as Agriculture and Rangeland. It benefits from

our integration of multi-field contextual semantic information. The visualization results are shown in Figure 8. In Table 3, all referred to all classes average, and the rest single class codes are as follows: U.-Urban, A.-Agriculture, R.-Rangeland, F.-Forest, W.-Water, B.-Barren.

Table 3. Segmentation performance measured in IoU/mIoU(%) on DeepGlobe.

Class	U.	A.	R.	F.	W.	B.	All
Baseline	78.3	87.2	39.0	78.9	82.4	59.4	70.8
GLNet	78.1	86.8	38.6	79.8	82.6	63.6	71.6
MCFNet	78.9	87.3	41.5	80.6	83.1	64.1	72.6
MCFNet-All Fusion	79.3	87.3	43.2	80.7	83.7	63.7	73.0

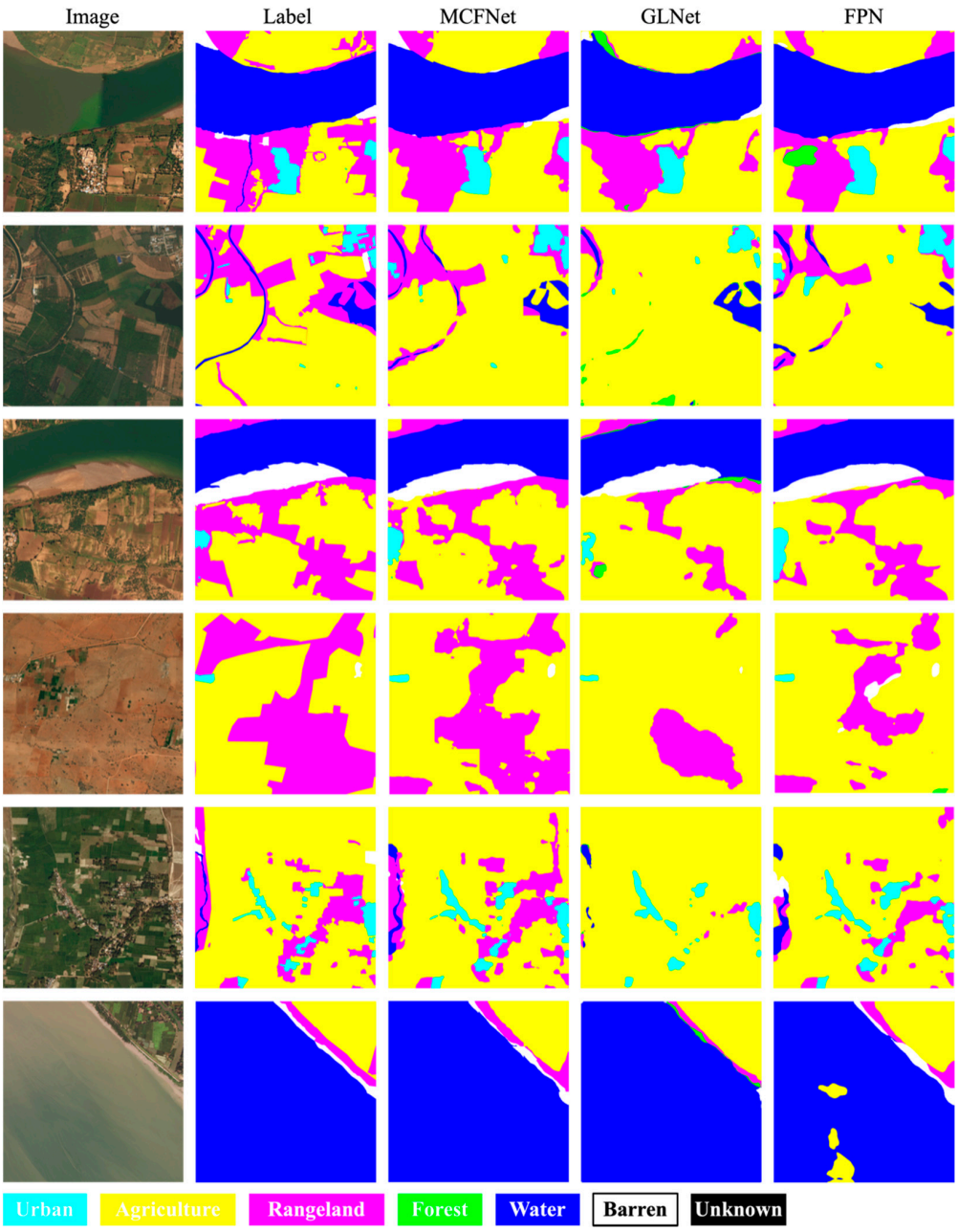

Figure 7. Visual comparison of semantic segmentation results on the DeepGlobe.

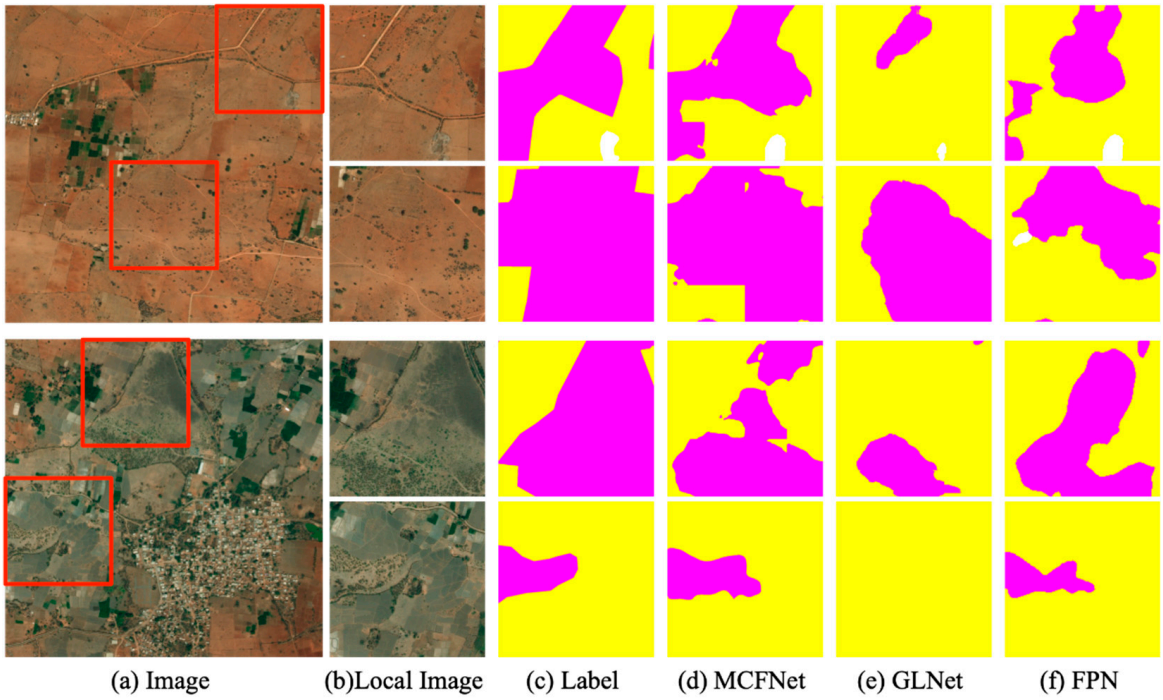

Figure 8. Visualization of semantic segmentation results for ambiguous categories Agriculture and Rangeland.

4.4.2. The result on Potsdam dataset

Compared with DeepGlobe, Potsdam has a higher image resolution and cannot be directly input to image for model training. Accordingly, we only compared the patch segmentation results for Potsdam. In those methods, the backbone used to extract high-level features and low-level features was also Resnet50. The testing results are shown in Table 4 where MCFNet yields mIoU of 70.1% and is quantitatively superior to other methods in terms of both accuracy and memory usage. The "Impervious Surface" class gains the highest classification accuracy of 79.6%. It s indicated that most of the pixels of this class are correctly classified because the contextual information of the "Impervious Surface" is not complicated, and could be easily extracted. In addition to the "background" class, the "car" class gains the lowest classification accuracy of 63.9%. It is because cars belong to the densely located small objects and that cars have intra-class variability, such as different colors and parking locations. As expected, the segmentation performance of MCFNet is superior to other models and it results in the 8% accuracy improvement compared to baseline. Figure 9 depicts the segmentation results of MCFNet.

Table 4. Performance of MCFNet and other segmentation models on DeepGlobe Dataset.

Model	mIoU(%)	Memory(MB)
UNet	60.4	3480
PSPNet	64.8	3108
FCN-8s	65.9	4496
FPN	66.2	4044
Deeplabv3+	66.8	3424
MCFNet	70.1	2594
MCFNet-All Fusion	72.2	2594

To further explore the advantages of our model in each class, we present the segmentation results of all types of Potsdam in Table 5. It is not difficult to see that our method has a huge advantage in the segmentation of small objects, which benefits from our contextual

fusion network. This is also confirmed in Figure 10, which shows the local segmentation results. In Table 5, all referred to all classes average, and the rest single class codes are as follows: B.-Building, T.-Tree, Cl.-Clutter/Background, V.-Low Vegetation, C.-Car, S.-Impervious Surface.

Table 5. Segmentation performance measured in IoU/mIoU(%) on Potsdam.

Class	B.	T.	Cl.	V.	C.	S.	All
Baseline	77.0	58.1	61.0	68.3	55.9	76.4	66.2
DeepLab v3+	76.2	57.1	56.9	66.4	67.3	76.9	66.8
MCFNet	78.2	65.9	61.4	71.6	63.9	79.6	70.1
MCFNet-All Fusion	78.9	68.4	61.5	73.0	70.6	81.2	72.2

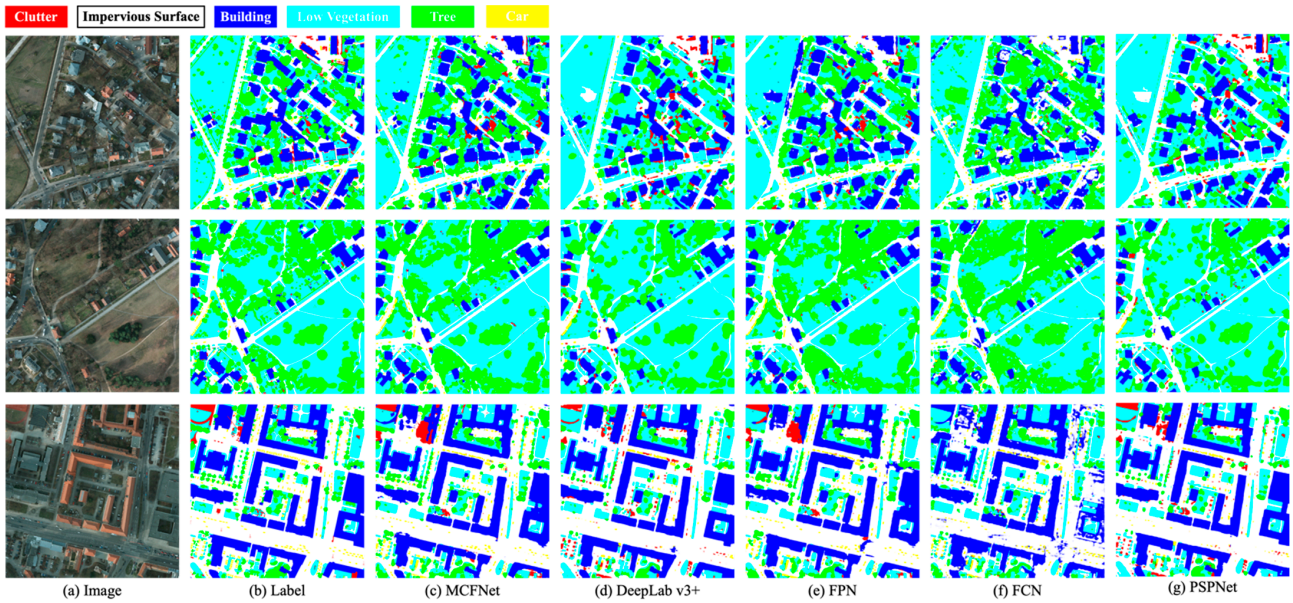

Figure 9. The display of segmentation output for Potsdam dataset.

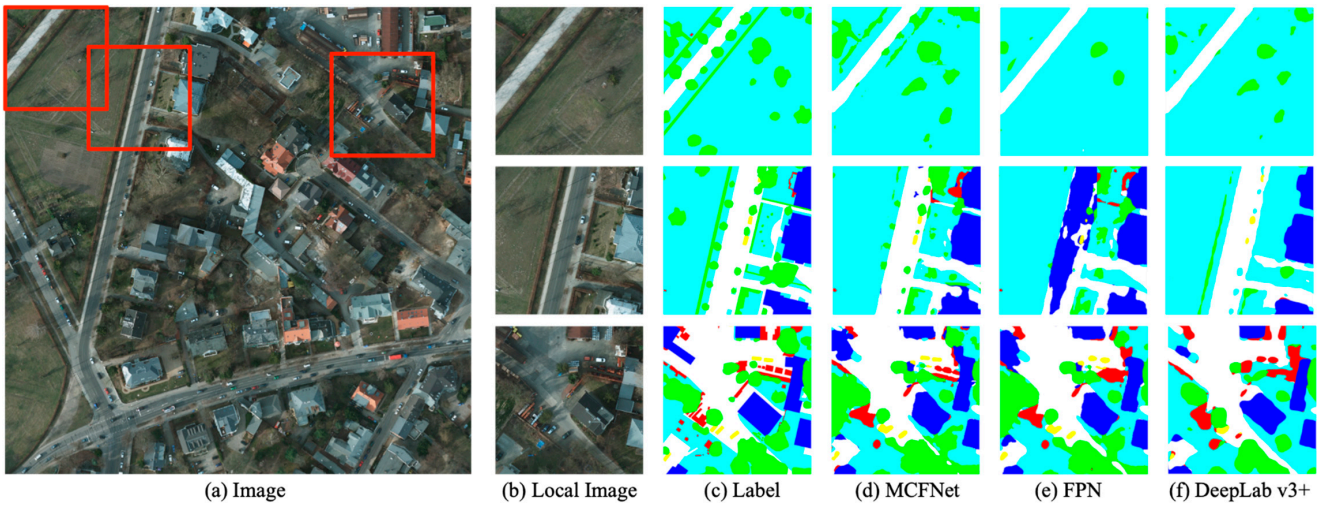

Figure 10. Local details of the Postdam image through different methods.

4.5. Ablation Study

To further test the efficiency of our model, we designed a series of ablation experiments. The ablative study is to test the importance and performance of each part of the model. For the two parts of our model, PSM and FM feature correlation module, we conducted

ablation research. Furthermore, we added a time metric to demonstrate the effectiveness of our module. Time represents the time consumed by the model to predict an entire image. In Tables 6–8 Random Select (R.S.) represents a random selection of five local images for enhancement. Since the memory usage of the GPU does not change much with or without the addition of these modules, we added inference time as a metric to evaluate the efficiency of the model.

Table 6. The effects of different components of our method on the results.

	Model	Random Select	PSM	FM	mIoU(%)	Time(s)
DeepGlobe	Baseline				70.8	1.1
	Ours	√		√	71.1	1.6
			√	√	72.6	1.8
				√	73.0	2.7
Potsdam	Baseline				66.2	10.3
	Ours	√		√	68.1	12.4
			√	√	70.1	12.4
				√	72.2	15.2

Table 7. Segmentation performance measured in IoU(%) on DeepGlobe.

Class	U.	A.	R.	F.	W.	B.
Baseline	78.3	87.2	39.0	78.9	82.4	59.4
+ R.S. & FM	77.3	86.4	36.5	78.9	81.7	60.1
+ PSM & FM	78.9	87.3	41.5	80.6	83.1	64.1
+ FM	79.3	87.3	43.2	80.7	83.7	63.7

Table 8. Segmentation performance measured in IoU(%) on Postdam.

Class	B.	T.	Cl.	V.	C.	S.
Baseline	77.0	58.1	61.0	68.3	55.9	76.4
+ R.S. & FM	77.8	61.3	61.5	70.1	60.0	77.6
+ PSM & FM	78.2	65.9	61.4	71.6	63.9	79.6
+ FM	78.9	68.4	61.5	73.0	70.6	81.2

In Table 6, aggregating contexts obviously improves the segmentation performance, which the improvement from 70.8% to 73% in DeepGlobe and from 66.2% to 72.2% in Potsdam. Specially, a small field of view context provides the nonlocal self-correlation information which facilitates in the segmentation of small objects. However, for the global image, a large field of view context can improve the accuracy of model perception. Hence, better results can be achieved by integrating the context information from different field of view. For instance, the segmentation accuracy for these two datasets is the highest when FM for all local patch, reaching 73% and 72.2% respectively. We show the DeeplGlobe local details as shown in Figure 11. Furthermore, as shown in Tables 7 and 8, compared to baseline, the addition of the FM module has improved the segmentation accuracy for each category.

To demonstrate the effectiveness of the PSM module, we set up a random selection of local patches for comparison. From the above table, the higher accuracy of segmentation selected by PSM proves the effectiveness of our scoring mechanism. In terms of algorithmic efficiency, PSM selects only the local patches that are less effective in segmentation which decreases the inference time compared to semantic fusion for all local patches. In Table 7, all referred to all classes average, and the rest single class codes are as follows: U.-Urban, A.-Agriculture, R.-Rangeland, F.-Forest, W.-Water, B.-Barren. In Table 8, the rest single class codes are as follows: B.-Building, T.-Tree, Cl.-Clutter/Background, V.-Low Vegetation, C.-Car, S.-Impervious Surface.

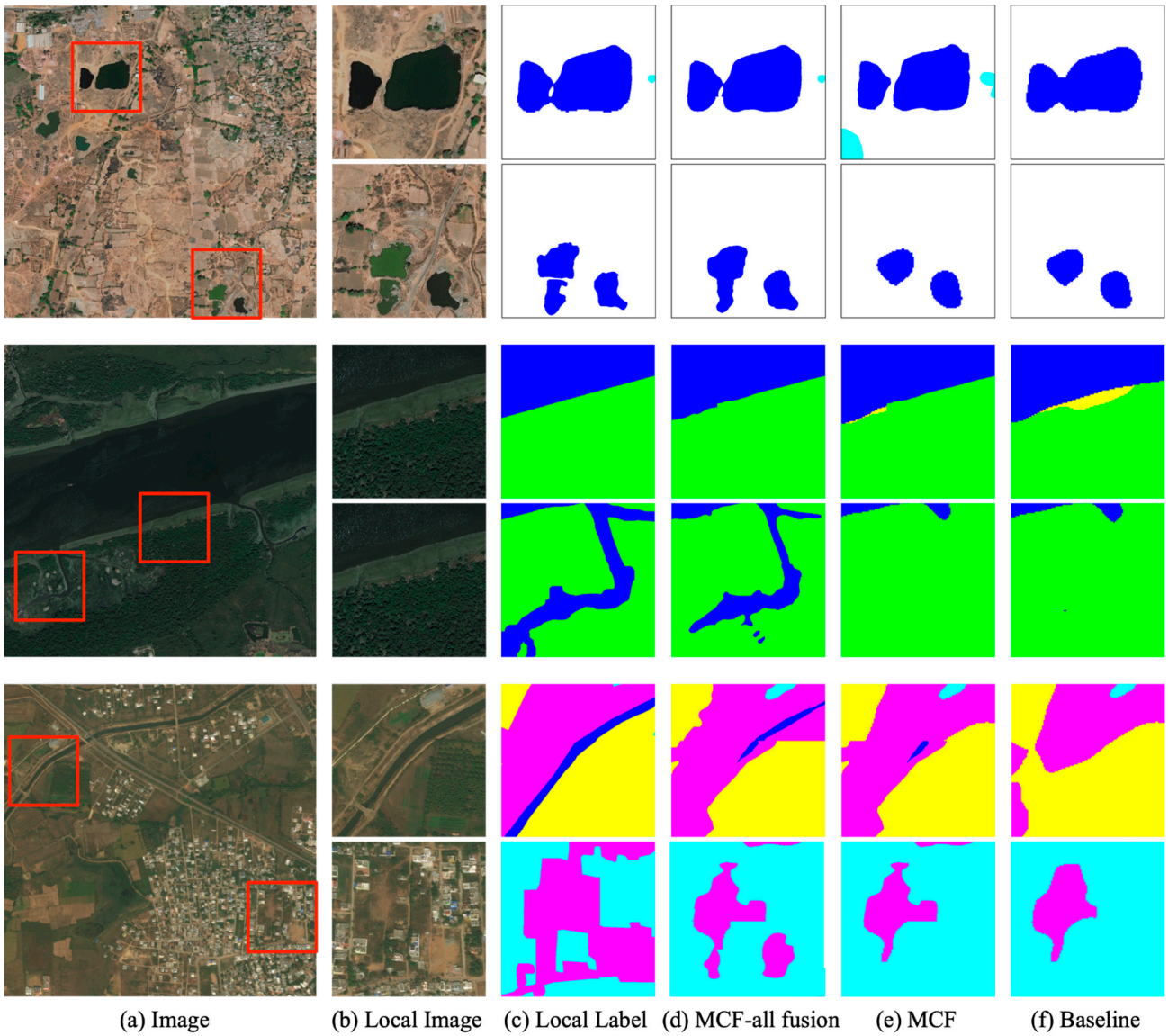

Figure 11. Local details of the DeepGlobe image through different methods.

5. Conclusions

In this article, we propose the MCFNet framework for HSR remote sensing images semantic segmentation in order to balance accuracy and efficiency of high spatial resolution image segmentation. Our network consists primarily of PSM and FM modules. Specifically, PSM is proposed to select local patches with poor segmentation effects, while FM performs multi-field of view contextual semantic fusion on these local patches.

MCFNet can not only extract the contextual, information from the global image, but also refine the details of objects in the local image. Therefore, it can alleviate the intra-class similarity and inter-class dissimilarity of remote sensing images and improve segmentation accuracy. Furthermore, an additional benefit of PSM is that context fusion of local patches does not take up much memory.

We have demonstrated the benefits of MCFNet on two challenging high-resolution remote sensing images. In future research, we will attempt to directly use a transformer to represent information characteristics in a way that allows global contextual relationships to be obtained from the beginning, to compensate for the inherent shortcomings of CNN.

Author Contributions: This work was conducted in collaboration with all authors. H.Y. and X.D. defined the research theme. H.Y. and C.W. supervised the research work and provided experimental facilities. X.D. and S.H. designed the semantic segmentation model and conducted the experiments. X.D. checked the experimental results. This manuscript was written by X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hainan Province Science and Technology Special Fund (Grant No. ZDYF2022GXJS228; ZDYF2021GXJS006); Hainan Provincial Natural Science Foundation of China (Grant No. 620RC559).

Data Availability Statement: We are grateful to CVPR and ISPRS for providing the open benchmarks for 2D remote sensing image semantic segmentation. The data in the paper can be obtained through the following link. DeepGlobe: DEEPGLOBE-CVPR18-Home, accessed on 30 October 2021 and Potsdam: 2D Semantic Labeling Contest-Potsdam (isprs.org), accessed on 2 March 2022.

Acknowledgments: We gratefully appreciate the editor and anonymous reviewers for their efforts and constructive comments, which have greatly improved the technical quality and presentation of this study.

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

References

1. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying Urban Functional Zones by Coupling Remote Sensing Imagery and Human Sensing Data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
2. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
3. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
4. Zheng, X.; Gong, T.; Li, X.; Lu, X. Generalized Scene Classification From Small-Scale Datasets With Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
5. Zhao, C.; Lu, Z. Remote Sensing of Landslides—A Review. *Remote Sens.* **2018**, *10*, 279. [CrossRef]
6. Tomás, R.; Li, Z. Earth Observations for Geohazards: Present and Future Challenges. *Remote Sens.* **2017**, *9*, 194. [CrossRef]
7. Yao, H.; Qin, R.; Chen, X. Unmanned Aerial Vehicle for Remote Sensing Applications—A Review. *Remote Sens.* **2019**, *11*, 1443. [CrossRef]
8. Liang, J.; Zhou, J.; Qian, Y.; Wen, L.; Bai, X.; Gao, Y. On the Sampling Strategy for Evaluation of Spectral-Spatial Methods in Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 862–880. [CrossRef]
9. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Trans. Geosci. Remote Sens.* **2016**, *4*, 22–40. [CrossRef]
10. Chen, W.; Jiang, Z.; Wang, Z.; Cui, K.; Qian, X. Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-High Resolution Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8916–8925.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI), Munich, Germany, 5–9 October 2015.
13. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging.* **2020**, *39*, 1856–1867. [CrossRef] [PubMed]
14. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
16. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2015**, arXiv:1412.7062.
17. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR, Viena, Austria, 4 May 2021.
19. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3285–3294.

20. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886.
21. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
22. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4095–4104.
23. 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]
24. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 701. [CrossRef]
25. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
26. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision(ECCV), Munich, Germany, 8–14 September 2018.
27. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.X.; Wang, W.J.; Zhu, Y.K.; Pang, R.M.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October 2019–2 November 2019.
31. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
32. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Wang, J. Structured Knowledge Distillation for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
33. Chen, L.C.; Yi, Y.; Jiang, W.; Wei, X.; Yuille, A.L. Attention to Scale: Scale-Aware Semantic Image Segmentation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
34. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for Object Segmentation and Fine-grained Localization. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 447–456. [CrossRef]
35. Liu, C.; Chen, L.C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.L.; Li, F.F. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
36. Zhong, Z.; Lin, Z.Q.; Bidart, R.; Hu, X.; Wong, A. Squeeze-and-Attention Networks for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
37. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017.
38. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Computer Society, Honolulu, HI, USA, 21–26 July 2016.
39. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. Munich, Germany, 8–14 September 2018.
40. Sun, K.; Xiao, B.; Liu, D.; Wang, J.D. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
41. Yuan, Y.; Chen, X.; Wang, J. Object-Contextual Representations for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
42. Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; Sang, N. Context Prior for Scene Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
43. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Yan, Y. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
44. Liu, W.; Rabinovich, A.; Berg, A.C. ParseNet: Looking Wider to See Better. *arXiv* **2015**, arXiv:1506.04579.
45. Huynh, C.; Tran, A.T.; Luu, K.; Hoai, M. Progressive Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 21–24 June 2021.

46. Cheng, H.K.; Chung, J.; Tai, Y.W.; Tang, C.K. CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement. *arXiv* **2020**, arXiv:2005.02551.
47. Zhang, Q.; Yang, G.; Zhang, G. Collaborative Network for Super-Resolution and Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
48. Chen, L.; Dou, X.; Peng, J.; Li, W.; Sun, B.; Li, H. EFCNet: Ensemble Full Convolutional Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
49. 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), Las Vegas, NV, USA, 27–30 June 2016.
50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Kai, L.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Miami, FL, USA, 20–25 June 2009.
51. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
52. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
53. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
54. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
55. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
56. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
57. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 318–327. [CrossRef] [PubMed]