remote sensing

MDPI

Article

Double Branch Parallel Network for Segmentation of Buildings and Waters in Remote Sensing Images

Jing Chen [1], Min Xia [1,*], Dehao Wang [1] and Haifeng Lin [2]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China
* Correspondence: xiamin@nuist.edu.cn

Abstract: The segmentation algorithm for buildings and waters is extremely important for the efficient planning and utilization of land resources. The temporal and space range of remote sensing pictures is growing. Due to the generic convolutional neural network's (CNN) insensitivity to the spatial position information in remote sensing images, certain location and edge details can be lost, leading to a low level of segmentation accuracy. This research suggests a double-branch parallel interactive network to address these issues, fully using the interactivity of global information in a Swin Transformer network, and integrating CNN to capture deeper information. Then, by building a cross-scale multi-level fusion module, the model can combine features gathered using convolutional neural networks with features derived using Swin Transformer, successfully extracting the semantic information of spatial information and context. Then, an up-sampling module for multi-scale fusion is suggested. It employs the output high-level feature information to direct the low-level feature information and recover the high-resolution pixel-level features. According to experimental results, the proposed networks maximizes the benefits of the two models and increases the precision of semantic segmentation of buildings and waters.

Keywords: double branch; CNN; semantic segmentation; buildings and waters; deep learning

Citation: Chen, J.; Xia, M.; Wang, D.; Lin, H. Double Branch Parallel Network for Segmentation of Buildings and Waters in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1536. https://doi.org/10.3390/rs15061536

Academic Editors: Sawaid Abbas, Janet E. Nichol, Faisal M. Qamer and Jianchu Xu

Received: 13 February 2023
Revised: 4 March 2023
Accepted: 9 March 2023
Published: 11 March 2023

1. Introduction

Semantic segmentation is very important in many domains such as unmanned driving, land use, ecological environment monitoring, disaster monitoring and agricultural monitoring. Identifying building and water area types from remote sensing images can provide an efficient technical approach for regional map updating, land planning, risk management [1] and regional economic development forecasting. The resolution of remote sensing images is increasing as modern computer vision and aerospace technology advance quickly, including space, spectrum and time [2]. A highly effective and affordable method for mapping a large area is to use remote sensing techniques [3]. The textural features and spatial structure properties of ground objects can be clearly expressed in high-resolution remote sensing photographs [4–6]. The development of remote sensing image technology is of great significance to the promotion of semantic segmentation tasks.

The three primary approaches used to segment traditional remote sensing images are the threshold, clustering, and maximum likelihood methods. The maximum likelihood method calculates the maximum likelihood discriminant function of each category by training set data, then substitutes the value of each pixel into the calculation, and finally evaluates the reliability of the classification outcomes. The maximum likelihood method has a high requirement for the training set, which can easily lead to very poor estimation results. The threshold method mainly defines the regional attribution of different targets in the image by threshold, but it is sensitive to the noise of the image, and in remote sensing images with a highly complicated background, the gray difference is not obvious

and the overlap of different target gray values is not obvious. The appropriate threshold is not easy to find. The clustering rule is to classify the pixels of the image, and some traditional clustering algorithms do not consider the spatial information, which can easily cause a lack of segmentation in image information and a decrease in segmentation accuracy, and the work to be performed in clustering analysis is more complicated. There are also some image segmentation methods such as the segmentation method based on the genetic algorithm, region segmentation and edge segmentation [7]. The disadvantage of region segmentation is that it is easy to cause excessive segmentation of images. The edge segmentation method cannot obtain a better regional structure, and there is a contradiction between accuracy and noise immunity [8]. Usually, the edge segmentation method is combined with the region segmentation method to obtain a better segmentation effect. In the segmentation method based on the genetic algorithm, it is difficult to determine the crossover probability and mutation probability, and the selection of fitness function is more difficult. Recently, some machine learning methods including decision tree, support vector machine, random forest and artificial shallow neural network have been determined unsuitable for massive quantities of data. Recently, fully supervised models have had great success in this area, however the lack of annotated data will result in significant performance loss [9]. In summary, on the one hand, the traditional image segmentation method is limited by the high-resolution remote sensing image spectrum. On the other hand, it has limited ability to process massive data, and there are some problems such as poor segmentation effect and poor generalization ability.

High-resolution remote sensing images have complex features and are difficult to classify. At present, some deep learning methods have performed well in natural image segmentation tasks. If they are directly applied to remote sensing image segmentation, there will be some challenges: 1. Data volume and quality: different from natural images, the data volume of remote sensing images is usually small, and it is affected by clouds, shadows, noise and other factors, resulting in unstable image quality, which may lead to overfitting or underfitting of deep learning algorithms, thereby affecting classification accuracy. 2. The multi-scale problem: remote sensing images usually involve multiple scales and resolutions. At different scales, the object in the image has different appearance and semantic information. Therefore, in the task of remote sensing image segmentation, multi-scale information needs to be processed and fused to capture global and local information, so that the model can classify the targets in the image more accurately. 3. Category imbalance: remote sensing images usually involve multiple categories, some of which may have few samples. This category imbalance may lead to bias in the model, that is, the model tends to predict common categories in training, and the prediction of rare categories is inaccurate. 4. Spatial and temporal issues: in remote sensing images, temporal and spatial information are closely related. For example, images of the same area may be taken at different times, and the location, shape, and number of targets may also change. Therefore, spatio-temporal information needs to be considered when processing remote sensing images in order to better capture change information [1,10–16]. The deep learning method can extract more and deeper feature information, which is suitable for the classification of high-resolution images [17–20]. Long et al. [21] proposed a fully-connected convolutional neural network (FCN), which migrates the feature extraction layer to the segmentation task and updates these parameters through fine-tuning [22]. At the same time, in order to achieve more precise segmentation findings, a unique structure is created to mix shallow semantic information and high-level semantic information. Chen et al. [23] proposed a semantic pixel-level segmentation network (SegNet). SegNet is a lightweight neural network that uses an upsampling layer to increase the resolution of the segmentation results to the same as the input image. In addition, SegNet uses an encoder-decoder structure with skip connections and deconvolution layers to better preserve detail and semantic information. Zhao et al. [24] proposed a pyramid network structure (PSPNet). PSPNet uses dilated convolution for the convolution of the basic ResNet [25], and in the entire encoder coding part, the features remain at the same resolution after the initial pooling layer. Auxiliary loss

is introduced in training to help ResNet learning, and the Spatial Pyrmid Pooling module is introduced to integrate semantic information in different regions to obtain better global semantic information. In the same year, a multi-path refinement network was suggested for high-resolution images segmentation (RefineNet) [26]. It is also based on making full use of shallow and deep feature information, cutting the image input, and then extracting the features, respectively. Sun et al. [27] proposed a high-resolution segmentation network (HRNet). All throughout the procedure, HRNet maintains a high-resolution representation. It achieves the purpose of strong semantic information and accurate location information by connecting high-resolution and low-resolution convolutional flow branches in parallel and continuously interacting with information between different branches. Pang et al. [8] proposed a lightweight building and water segmentation network (SGBNet), in which the channel pooling attention module extracts features through two different global pooling modules. While improving the segmentation accuracy, the network is also more efficient and lightweight. In conclusion, CNNs are usually composed of convolutional layers, pooling layers, and fully-connected layers. For semantic segmentation tasks, the encoder-decoder structure is generally used. The encoder is responsible for encoding the input image information into low-dimensional features. The decoder is responsible for mapping the low-dimensional features back to the original image size and outputting the classification results of each pixel. For each application field, the appropriate network structure and training strategy can be designed according to the specific situation to obtain the best performance.

Recently, Transformer [28] has shown great value in some fields. Transformer is the first transduction model that calculates its input and output representations purely using self-attention, and does not require convolutional network or a sequence-aligned recurrent neural network. According to the conventional model, ByteNet grows logarithmically, ConvS2S increases linearly, and the number of operations needed to correlate signals at two random input or output sites increases with distance. Learning the reliance between far-off places is made more difficult as a result. In Transformer, this is reduced to a fixed number of operations. ViT is proposed as a transformer for large-scale image recognition [29]. The core process of ViT includes four parts: image block processing, image block embedding and location coding, Transformer encoder and MLP classification processing. In the task of migrating to small-scale data sets, it can achieve better performance than CNN, and successfully converts visual problems into natural language processing problems. Liu et al. [30] proposed a Transformer structure with moving windows (Swin Transformer). Swin Transformer designs a shift window and performs self-attention calculation in the shift window. The global information is fully interactive, which brings greater efficiency and performance. At the same time, the operation of the moving window can make the adjacent two windows interact with each other, thus achieving the ability of global modeling in disguise and achieving good results in the segmentation task. Based on Swin Transformer and UNet [31], swin-UNet [32] is proposed. It is based on the Swin Transformer module and constructs a symmetric encoder-decoder structure with skip connections to perform corresponding pixel-level segmentation prediction. Lu et al. [22] proposed a bilateral branch model based on the traditional Transformer, using a strip convolution module in the encoder. The information gathered by the two branches helps to guide each other and produce more accurate segmentation renderings. In conclusion, the Transformer model is a neural network structure based on a self-attention mechanism, which is usually composed of an encoder and decoder. For the semantic segmentation task, the encoder part of the Transformer model can be used as a feature extractor to extract the high-dimensional feature representation of the input image and input it into the decoder for pixel-level classification. CNN can only model local features, while Transformer can model global information. Combining them, a Transformer module can be introduced into CNN to capture longer-range contextual information. This method can improve the edge and detail information of the object in the image segmentation task.

High-resolution remote sensing photos' intricate spectral and spatial texture information not only enhances the table's finely detailed features, but also makes semantic segmentation tasks more challenging. Due to the large difference in the size of various types of features, neural networks need to effectively extract features of ground objects from different angles. For example, the shooting angle and distance of some images, the light intensity, the complexity of the terrain, including the diversity of landforms, the difference between the size of urban waters and natural lakes, the density and diversity of urban buildings and the sparseness of rural houses. More thorough criteria for the model of the semantic segmentation task are put forth by these issues. Due to the lack of global information interaction and the single calculation method, the traditional algorithm will lead to a lot of noise when predicting, and the detection and recognition ability of some areas similar to the pixels of non-target areas is insufficient. For some edge details, the loss of information is large, and on the whole, it is prone to misjudgment. Convolutional neural network-based approaches have difficulty learning explicit global and long-term semantic information relationships because of the intrinsic constraints of convolution processes. This research introduces a novel double-branch parallel image segmentation network in response to the shortcomings of the existing approaches. The network is based on Swin Transformer and CNN. In the stage of feature coding, the designed cross-scale multi-level fusion module is used to connect the two branches, and the comprehensive semantic information and spatial semantic information are extracted using CNN and Swin Transformer. The multi-scale fusion module designed by us guides feature information extracted by double branches to each other, giving full play to the characteristics of Swin Transformer's global information interaction, and making up for the errors in judgment brought on by a lack of global information and long-term semantic information interaction of CNN. During the feature decoding stage, the designed multi-scale fusion module is utilized to fuse the high-level feature information in the coding stage and the low-level feature information extracted by CNN, and the high-level feature information is used to direct low-level feature information and upsample step-by-step. Through the joint action of several modules, our network has significantly increased the segmentation precision.

2. Methods

At present, convolutional neural networks have a constrained receptive field and it is difficult to capture global information [33]. The Swin Transformer network adds a mobile sliding window to better capture global feature information and perform global information interaction. The convolutional neural network has translation invariance and global correlation, while these characteristics in the Transformer network structure are insufficient. Taking into account the aforementioned factors, we propose a parallel combination structure of both Swin Transformer and CNN. The segmentation accuracy and generalization ability of the model in segmentation tasks are greatly improved. In addition, it can effectively identify houses, waters and backgrounds in building and water tasks, and the segmented boundary details are more delicate. Figure 1 depicts the parallel network's general design, which is mostly made up of encoders and decoders. Figures 2 depict the detailed layout of each module. For a given image, first enter the encoder, enter the Swin Transformer and CNN, extract the features information, and effectively fuse the features extracted by feature fusion module designed in this paper, and pass the fusion parameters into the Swin Transformer for further feature extraction. We use Swin Transformer as a branch in the research. Compared with the traditional Transformer module, Swin Transformer designs a moving window and performs self-attention calculation in the moving window. The global information is fully interactive, which brings greater efficiency and performance. At the same time, through the operation of the moving window, the feature information interaction between the two adjacent windows can be realized, thus achieving the ability of global modeling in disguise. Through the coding stage, highly detailed information is obtained and the global information is fully understood. We propose a step-by-step fusion upsampling module in the decoding stage. The feature information obtained through

the convolutional network and the feature information obtained through the encoder are upsampled step-by-step through the multi-scale fusion module. Through the sufficient interaction of global information and the guidance of low-level feature information, some disadvantages are mitigated. The model's performance is significantly enhanced by the four fusion upsampling modules, which gradually combine feature information from high level to low level.

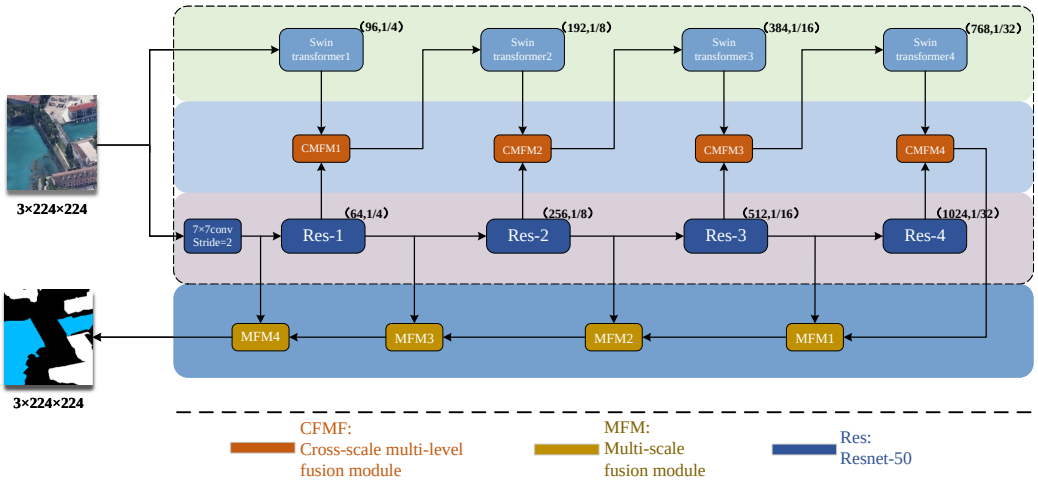

Figure 1. The overall structure of double branch parallel network.

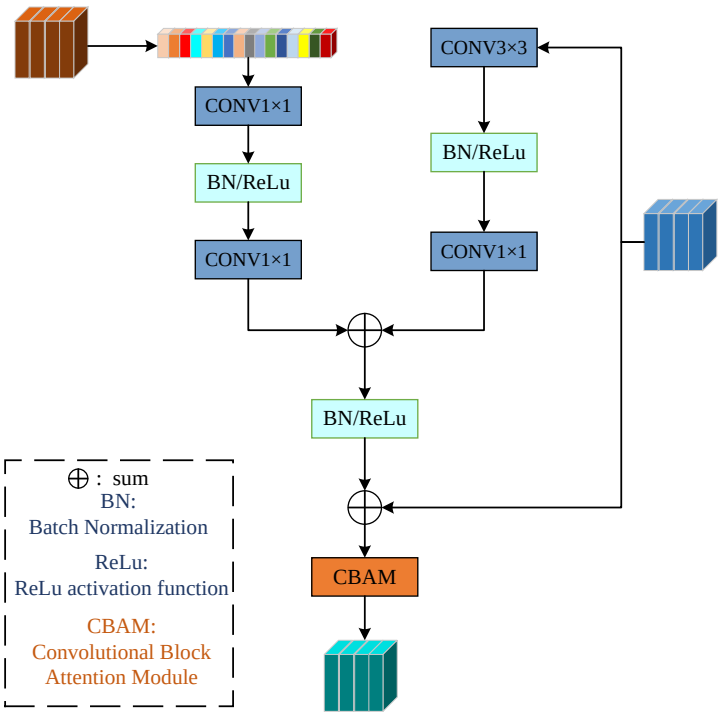

Figure 2. CMFM module structure diagram.

2.1. Overall Structure

The article uses the parallel structure of Swin Transformer and Resnet50 convolution network to draw different levels of information of images. Swin Transformer not only has dynamic attention to focus areas, adding a moving window, but also has a global receptive field and better generalization performance. CNN with Resnet50 as the backbone has two characteristics: local perception and parameter sharing. Local perception refers to the CNN's proposal that each neuron just needs to sense the local pixels in the image rather than all of them, and that this local information may subsequently be combined at a higher

level to access all of the image's characterization information. To enhance the performance of the model, we design related modules to fully exploit the advantages of both.

2.2. Cross-Scale Multi-Level Fusion Module

To better improve the accuracy and predictive performance of models in buildings and waters segmentation tasks, we propose a CNN structure Resnet50 and Swin Transformer parallel network structure, but if we simply combine the two structures, we find that the effect of the model is not obvious, which does not meet our task requirements. Considering that two branches output different levels of characteristic information, in order to make full use of the advantages of double branch parallel network, we design a cross-scale multi-level fusion module (CMFM).

As shown in Figure 2, the fusion module we designed here has two branches. It is assumed that the size of the feature map f_1 output by the Resnet50 branch is $C_1 \times H \times W$, and the size of the feature map f_2 output by the Swin Transformer branch is $C_2 \times H \times W$, where C_1 and C_2 represent the number of channels. The f_1 is first passed through a global average pooling layer, which reduces the number of parameters. Global average pooling can better reflect the global information and avoid overfitting. On the other hand, it combines global spatial information and has stronger spatial conversion ability for input images. The next step is to go through a 1×1 convolutional layer, then through the BN and ReLu functions. Finally, the number of channels is changed from C_1 to C_2 through a 1×1 convolutional layer to obtain the other side. We will first process a 3×3 convolutional layer, then the BN and ReLu functions, and finally through a 1×1 convolutional layer. After that, f_{out1} is obtained by adding the $C_2 \times 1 \times 1$ and $C_2 \times H \times W$ size after two 1×1 convolutions, and then f_{out2} is obtained by a BN and ReLu activation function again. After that, we add the $C_2 \times H \times W$ size f_2 to it by a residual operation to obtain f_{out3}. In this paper, we obtain the output f_{out3} and then process the convolutional block attention module (CBAM) to finally obtain the output Y [34]. The above-mentioned process's calculation formula may be represented as:

$$f_1' = f^{1\times1}\Big(\delta\Big(f^{1\times1}g_a(f_1)\Big)\Big), \tag{1}$$

$$f_2' = f^{1\times1}\Big(\delta\Big(f^{3\times3}(f_2)\Big)\Big), \tag{2}$$

$$Y = CBAM\big(\delta\big(f_1' + f_2'\big) + f_2\big), \tag{3}$$

$f^{1\times1}(\cdot)$ denotes a convolution operation with a convolution kernel of 1×1, $f^{3\times3}(\cdot)$ denotes a convolution operation with a convolution kernel of 3×3, $\delta(\cdot)$ denotes ReLu activation. The ReLu function is sparse, making the sparse model to more effectively extract pertinent features and match training data. The specific calculation formula is shown in Formula (4).

$$\delta(x) = \max(0, x), \tag{4}$$

where x denotes the input of the ReLu function. $g_a(\cdot)$ in Formula (1) denotes one-dimensional average pooling, it calculates an average of all pixels of the feature map of each output channel, and can well suppress overfitting. Here, it changes the feature map of $C_1 \times H \times W$ to the size of $C_1 \times 1 \times 1$. It compresses spatial feature information into channel dimension, and integrates the global spatial information so that the global feature information can be fully utilized. The specific calculation formula is shown in Formula (5).

$$g_a = \frac{1}{H \times W} \sum_{i=1}^{H\times W} p(i), \tag{5}$$

where $p(i)$ represents the pixel value at the i-th position of the feature map. The calculation formula of $CBAM(\cdot)$ in Equation (3) is as follows (assuming the input is F):

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\Big(W_1\Big(W_0\Big(F^c_{avg}\Big)\Big) + W_1(W_0(F^c_{max}))\Big), \end{aligned} \tag{6}$$

$$\begin{aligned} M_s(F) &= \sigma\big(f^{7\times7}([AvgPool(F); MaxPool(F)])\big) \\ &= \sigma\Big(f^{7\times7}\Big(\Big[F^s_{avg}; F^s_{max}\Big]\Big)\Big), \end{aligned} \tag{7}$$

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

$$CBAM = F'' = M_s(F') \otimes F'. \tag{9}$$

In the above formula, the weights W_0 and W_1 of MLP are shared. $AvgPool(\cdot)$ and $MaxPool(\cdot)$ denote average pooling and maximum pooling operations. $M_c(\cdot)$ represents channel attention operation. $M_s(\cdot)$ denotes spatial attention operation. $f^{7\times7}(\cdot)$ denotes a convolution operation with a convolution kernel of 7×7. $\otimes$ represents tensor matrix multiplication. $\sigma(\cdot)$ represents the sigmoid activation function.The sigmoid function can play the role of normalization. Its calculation formula is as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{10}$$

where x represents the input of the *sigmoid* function.

Figure 3 shows the actual effect of our CMFM module. Among them, (a) is the original image, (b) is its label, (c) and (d) are the feature heat maps of the backbone model without CMFM module and with CMFM module, respectively. The feature heat map demonstrates that some places that were originally concerned or had relatively low attention. Following the addition of the CMFM module to the backbone network, the pixels in these regions—the red portion of the feature heat map—were given more weight by the network. This proves that the designed module is effective.

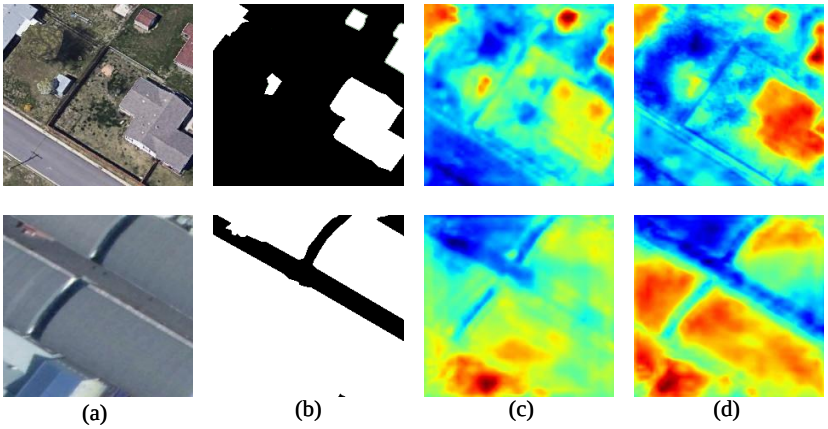

Figure 3. Characteristic heat map comparison of CMFM modules, (**a**) is the original image, (**b**) is its label, (**c**,**d**) are the effect diagrams without module and with module, respectively.

2.3. Multi-Scale Fusion Module

In the decoding stage, if only a simple and crude upsampling recovery is output, it is bound to lose more information, resulting in poor performance of the model; there will be a misjudgment of the situation. To take full advantage of the global interactivity of the Swin Transformer in our backbone network branch, and for high-level semantic features to be used to direct low-level semantic features, we designed a fusion module similar to CMFM.

The feature representation of different scales can be captured by guiding at different scales. Figure 4 is a multi-scale fusion module (MFM).

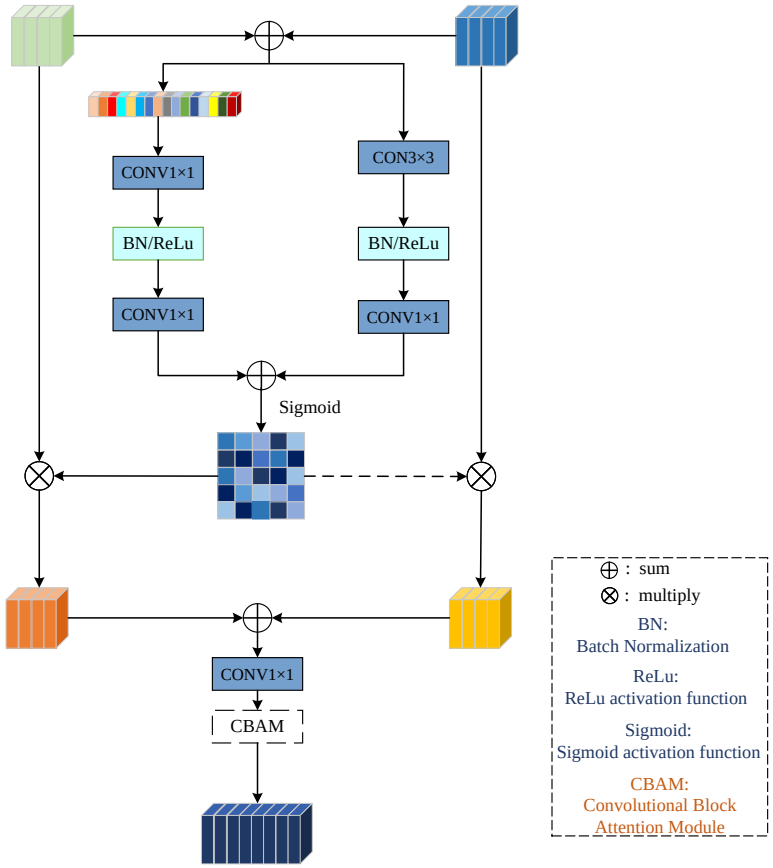

Figure 4. MFM module structure diagram.

Two feature maps with the same scale size make up the module's input, whose structure is shown in Figure 4. We assume that the two inputs are X_1 and X_2, respectively, and the size is $C \times H \times W$, where C denotes the number of channels of the feature map, H and W denote the height and width of the feature map. First, we add the two inputs to obtain X_3, and then parallel out of the two branches X_{31}, X_{32}; X_{31} through the global average pooling, and then through a convolution kernel of 1×1 two-dimensional convolution, then through BN and ReLu function, and finally through a convolution kernel of 1×1 convolution operation to obtain the output X'_{31}. On the other side we direct X_{32} through a convolution kernel for 3×3 two-dimensional convolution operation, the activation functions for BN and ReLu also follow. At last, the output X'_{32} is obtained by convolution operation with convolution kernel of 1×1. Then we add the output of the two to obtain the output X_4, which is activated by sigmoid to obtain the weight s. Since the weight s obtained after sigmoid activation is distributed between (0,1), here we use s and $(1 - s)$ as weight coefficients to weight X_1 and X_2, respectively to obtain X_{1out} and X_{2out}, and then add them. To obtain the number of channels matching the next stage, the final output Y is obtained by changing the channel through a two-dimensional convolution with a convolution kernel of 1×1 (the dotted boxed CBAM module is only used in the last MFM module, so Formula (15) is only used in the last MFM. The calculation formula has been described above, so it is not repeated here). The calculation formula of the above-mentioned process can be represented as:

$$X'_{31} = f^{1\times1}\left(\delta\left(f^{1\times1}(g_a(X_1 + X_2))\right)\right), \tag{11}$$

$$X'_{32} = f^{1\times1}\Big(\delta\Big(f^{3\times3}(X_1 + X_2)\Big)\Big), \tag{12}$$

$$X_4 = X'_{31} + X'_{32}, \tag{13}$$

$$Y = f^{1\times1}((\sigma(X_4) \otimes X_1 + (1 - \sigma(X_4)) \otimes X_2)), \tag{14}$$

$$Y = CBAM(Y), \tag{15}$$

where $g(\cdot)$ denotes global average pooling, $f^{3\times3}(\cdot)$ denotes a 2D convolution with a 3×3 convolution kernel, $f^{1\times1}(\cdot)$ represents a 2D convolution with a 1×1 convolution kernel, $\delta(\cdot)$ includes Batch Normalization and ReLU function activation, $\otimes$ represents tensor matrix multiplication, $\sigma(\cdot)$ represents the sigmoid activation function.

The above modules together constitute our algorithm network. The network adopts a two-branch parallel method. The convolution branch extracts detailed information to obtain low-level semantic information, and the Swin Transfomer branch extracts contextual information to obtain high-level semantic information. In the coding stage, the CMFM module completely interacts with the contextual information by utilizing the benefits of the two branch networks. In the decoding stage, the process gives full play to the advantages of MFM module, step-by-step fusion recovery, and makes the segmentation boundary more delicate, where in the case of complex backgrounds it can better identify waters and buildings.

3. Results

3.1. Building and Water Dataset

In order to test its effectiveness in the semantic segmentation job of buildings and waters, this paper created a buidings and waters dataset to train and validate the model. In comparison to some other datasets, the dataset created in this experiment has a large spatial span, more angles, and a complex background due to the low angle of view, which necessitates more complex algorithms. The dataset comprises of 10,000 pairs of Google Earth photos divided into the following categories: a riverfront residential complex in China, a private villa in North America, and so on. After that, the photos were divided into 224×224 images, and the data was enhanced on these images. There are three different kinds of strategies: the 50% horizontal flip, the 50% vertical flip, and the 10% random spin, for example. The enlarged dataset can be improved, but it can also raise model training process interference and improve the model's generalizability. Architecture, water, and background are the three object categories that have been manually assigned to these pictures. Figure 5 displays the sliced image together with its label. A single type of image was eliminated, and the remaining photos were then randomly split into an 8:2 training and validation set. The dataset created in this experiment contains a rich variety of backgrounds, which meet the experimental requirements and ensure that the experimental segmentation accuracy will not be biased due to the single dataset type.

The following characteristics are present in the dataset: (1) of the objects in the dataset selected in this paper, such as vehicles, some containers and some similar buildings, some of them have large differences in color, as shown in Figure 5d, and some of the buildings have a great similarity to the surrounding background color, as shown in Figure 5e, so this puts forward higher requirements for the proposed model detection ability. (2) In this dataset, we selected more coverage scenarios to better and more comprehensively test the performance of our model. (3) Because the remote sensing satellite is different in angle and spatial distance when shooting, the difficulty of segmentation is increased to a certain extent. (4) In some dense buildings, some high-rise buildings' shadows cast on nearby low-rise structures causes a certain degree of interference to the segmentation of the model.

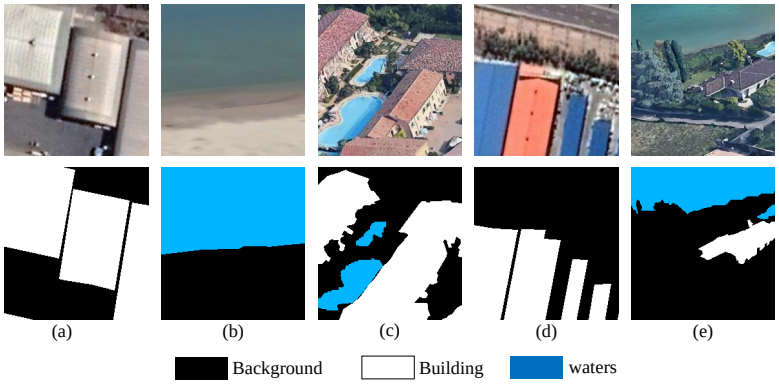

Figure 5. Partial display of the land cover dataset. (**a**–**e**) are the display of different remote sensing images and corresponding labels, respectively.

3.2. Waters Dataset

This paper selects the waters dataset for verification in order to more thoroughly reflect the network's ability to cope with edge characteristics, e.g., water area, to better validate the generalizability of the network model it proposes. In this experiment, we select China HJ-1A (HJ-1B) multi-spectral environmental remote sensing satellite images as the required dataset. Additionally, the model's capacity for generalization is examined.

Three visible and near-infrared spectral bands from the HJ-1A (HJ-1B) satellite charge coupled device camera are used in the water dataset. In order to effectively utilize the waters information, we selected a mix of bands 1, 2, and 4 to produce a three-channel waters image. To prevent over-fitting and ensure the accuracy of experimental segmentation accuracy, we cut the original image into 256 × 256 images and randomly flip, rotate, and scale the experimental image. Here, we created 8000 waters datasets and divided them into training sets and verification sets according to the 8:2 ratio column. The water dataset is a binary dataset, that is, the model identifies two semantic categories of waters and background. Figure 6 displays the cropped image and its labels.

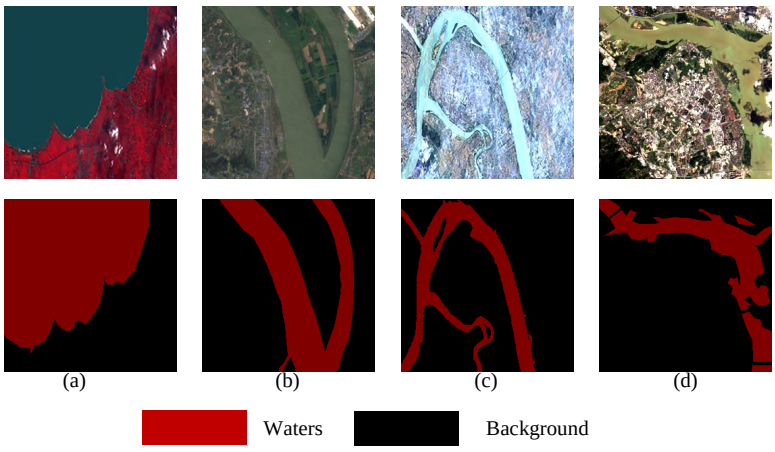

Figure 6. Partial representation of the waters dataset. (**a**–**d**) are the display of different remote sensing images and corresponding labels, respectively.

3.3. Inria Dataset

In order to further verify the generalization ability of the proposed model, we selected the Inria Aerial Image Labeling Dataset. It is a public data set for computer vision and machine learning research developed by the French National Institute of Digital and Automation (Inria). The dataset contains a set of aerial images taken from high altitude, covering some cities in southern France, and the ground coverage types include buildings, trees and roads. Its resolution is 5000 × 5000 pixels. Here, we cut it into 256 × 256

images and divided them into a training set and validation set according to the ratio of 8:2. As shown in Figure 7, part of the data set is displayed.

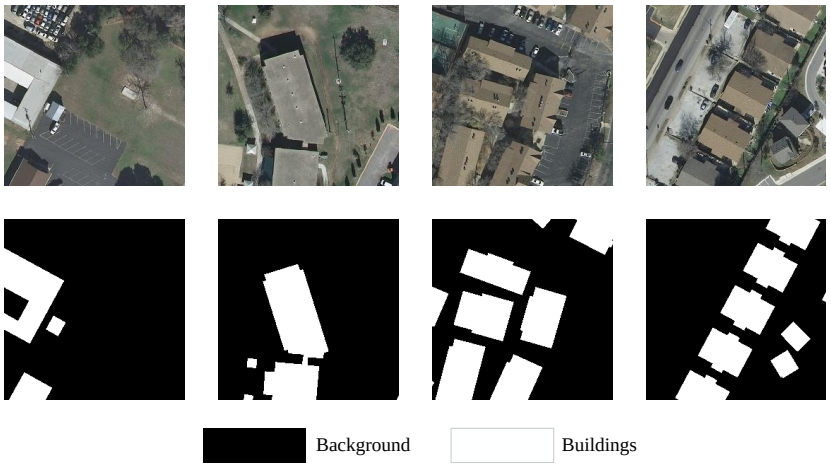

Figure 7. Partial representation of the Inria dataset.

A machine with an NVIDIA RTX3080 graphics card was used to carry out all the experiments in this study. The operating system adopted in this experiment is Windows 10. The construction of the experimental model in this paper is based on the deep learning framework of pytorch (2017). In terms of optimizer, this paper uses an adaptive moment estimation optimizer, it combines the advantages of both momentum and RMSProp optimization algorithms, the first-order moment estimation of the gradient and the second-order moment estimation, comprehensive consideration, and then calculates the update step. The number of iterations is set to 300 in all of the experiments in this work because, according to experimental observation, most experiments tend to converge after 200 iterations. The loss function used in the experiment is BCEWITH-LogitsLoss. Due to the physical memory limitations of the computer's graphics GPU, the experiment's batch size was set to 8. The experimental index is an important reference for evaluating the effect of the model. Here, we use the mean pixel accuracy (MPA), pixel accuracy (PA) and mean intersection over union (MIOU) on the union set as evaluation indicators. The following are the MPA, PA, and MIOU formulae.

$$MPA = \frac{1}{k}\sum_{i=0}^{k}\frac{p_{i,j}}{\sum_{j=0}^{k}p_{i,j}}, \tag{16}$$

$$PA = \frac{\sum_{i=0}^{k}p_{i,j}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{i,j}}, \tag{17}$$

$$MIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{i,j}}{\sum_{j=0}^{k}p_{i,j}+\sum_{j=0}^{k}p_{j,i}-p_{i,i}}, \tag{18}$$

where k denotes the class of object segmentation (excluding background), $P_{i,i}$ shows the real number, $P_{i,j}$ represents the number of pixels that belong to category i but are predicted to be j.

3.4. Ablation Experiment

We initially use Resnet50 as the backbone network, and then upsample each layer and connect them for output. Then we add the Swin Transformer branch in parallel based on Resnet50, and each layer of the two branches is added and up-sampled step by step for addition and output. Then, to verify the efficiency of the models and modules created in this research, we gradually add each module to the model. Here, the model's primary evaluation metric is MIOU. The ablation experiments are shown in Table 1

Table 1. Results of ablation experiments in land cover datasets.

Method	MIOU (%)
Resnet	82.59
Resnet + Swin Transformer	84.40
Resnet + Swin Transformer + CMFM	85.53
Resnet + Swin Transformer + CMFM + MFM	87.86

(1) To more effectively extract the image's feature information, extract more scale spatial location information, and fully carry out global information interaction, we first simply add the Swin Transformer branch in parallel to the Resnet50 branch. Through the experimental results, we find that the MIOU value can be increased to 84.40% when the two branches extract features in parallel. (2) Two branches are added to the connection module in parallel. For two branches in parallel, if each layer is simply added, there will be some lost spatial and semantic position information. To fully extract the feature information and improve how the high-level feature information directs the underlying feature information, the overall recognition ability of the model and the processing of some detailed features are improved. Based on this, this paper designs a cross-scale multi-level fusion module. It is found that the MOIU value of the model was increased to 85.53% after adding the CMFM module. (3) Ablation (MFM) for high-low information mutual guidance fusion module: since the shape and size of some rivers are not constant in the building and water segmentation task, some of the previous methods for river boundary treatment are not delicate enough. To restore the characteristic information of the edge of the river including some buildings, in this paper, we integrate low-level feature information extracted through convolutional network and high-level feature information through the Swin Transformer, utilizing the high-level feature information to direct the low-level feature information. We processed the global feature pooling in the module, once again fully incorporating global feature information interaction, improving the accuracy of the category area identification and model performance. Finally, the MIOU value of our model reached 87.86%.

3.5. Contrast Experiment

We contrast our model with some excellent models for building and water segmentation in this part, such as DABNet [35], FCN8s, PSPNet, DeeplabV3plus [36], Shufflenetv2 [37], BisenetV2 [38], Dual-branch [22], and so on. The Deeplab series mainly uses dilated convolution and pyramid pooling ASPP [25]. By using different dilated convolutions on a given feature layer, it can effectively resample and construct convolution kernels of different receptive fields to obtain information of multi-scale objects. The Bisenet series fuses the extracted deep feature information and spatial information through a spatial branch and a semantic branch, and supervises model training through an auxiliary loss function. FCN solves the issue of semantic-level picture segmentation, as the fully-connected layer of a traditional CNN is changed to a convolutional layer, which classifies images at the pixel level [39]. The pyramid pooling module, which is the primary component of PSPNet, may gather contextual information from several regions to boost global information acquisition capacity [23]. The Shufflenet series uses grouping convolution to group different features of the input layer, and then uses different convolution kernels to convolve each group, thereby reducing the amount of convolution calculation, mainly playing a lightweight effect. DABNet proposes a deep non-decomposable bottleneck module, which effectively uses asymmetric convolution kernel dilated convolution to construct the bottleneck layer, generates sufficient acceptance domain, intensively uses context information, and greatly reduces parameters.

Table 2 shows that SegNet has the worst segmentation effect, and the MIOU and MPA values are only 80.06% and 89.06%. In general convolutional neural networks, PSPNet (backbone adopts Resnet50) network has the highest segmentation accuracy, where the MIOU value and MPA value can reach 86.61% and 92.88%, respectively. At the same

time, the double-branch parallel network designed in this paper is compared with the network swin UNet that improves the swin Transformer, where MPA is 90.00% and 94.11%, and MIOU is 81.14% and 87.86%. Compared with these algorithms, our double-branch parallel network achieves the optimal value in three indicators. Comparing our approach to other models, it can be shown that it significantly improves segmentation, and our model has a strong pertinence for semantic segmentation tasks of buildings and waters. None of the semantic segmentation networks in the table pre-loaded the training weight, and the training requirement parameters were set uniformly to ensure the fairness of the comparison experiment.

Table 2. Experimental results compared to other algorithms.

Method	PA (%)	MPA (%)	MIOU (%)
SegNet [40]	89.88	89.06	80.06
UNet [31]	92.96	92.19	86.27
PAN [41]	92.78	93.22	85.78
HRNet [27]	92.37	92.68	84.78
DFNet [42]	91.74	90.31	84.42
DeepLabV3+ [36]	92.83	93.42	86.18
BiSeNetV2 [38]	91.19	91.13	82.81
DANet [43]	92.16	92.50	84.25
DABNet [35]	92.61	92.33	85.38
ShuffleNetV2 [37]	90.63	89.75	81.08
PSPNet [24]	93.40	92.88	86.61
FCN8s [21]	92.66	92.81	85.91
SwinUNet [32]	89.68	90.00	81.14
Dual-branch [22]	92.80	93.25	86.02
MFANet [12]	93.02	93.73	86.62
Ours	93.64	94.11	87.86

Figure 8 compares the prediction maps of some networks. We comprehensively verify our network by comparing the prediction effect maps of seven remote sensing images. Figure 8i is the label graph. Figure 8b–g are the experimental comparison results, Figure 8h is the proposed network prediction rendering. The comparison shows that the effect map predicted by the network model in this research is more accurate in detecting the buildings and waters, and there is no omission overall. This is due to the fact that our double-branch parallel structure fully exploits its own benefits. On the basis of the global information interaction of Swin Transformer, the two modules we designed are added to the global average pooling again. With the cooperation of these modules, the global information interaction is continuously carried out. In addition, by using high-level feature information to direct low-level feature information, some edge feature information of the segmented target are repaired. Finally, we find that the effect prediction graph of the double-branch parallel network model is the closest to the label graph, and our prediction accuracy is the best.

Figure 8a is a remote sensing image, Figure 8b–h represent DeepLabV3, FCN-8s, DABNet, PAN, PSPNet, UNet and the paper designed a double-branch parallel network, Figure 8i is the label graph. To show the superiority of our network model over other networks more directly, the red box is used to mark some areas for more intuitive comparison. It is clear that the double-branch parallel network designed in this research reduces the occurrence of misjudgment and missed judgment by grasping the global information and processing some details. The processing of some edge information is also relatively good, and the segmentation accuracy is improved.

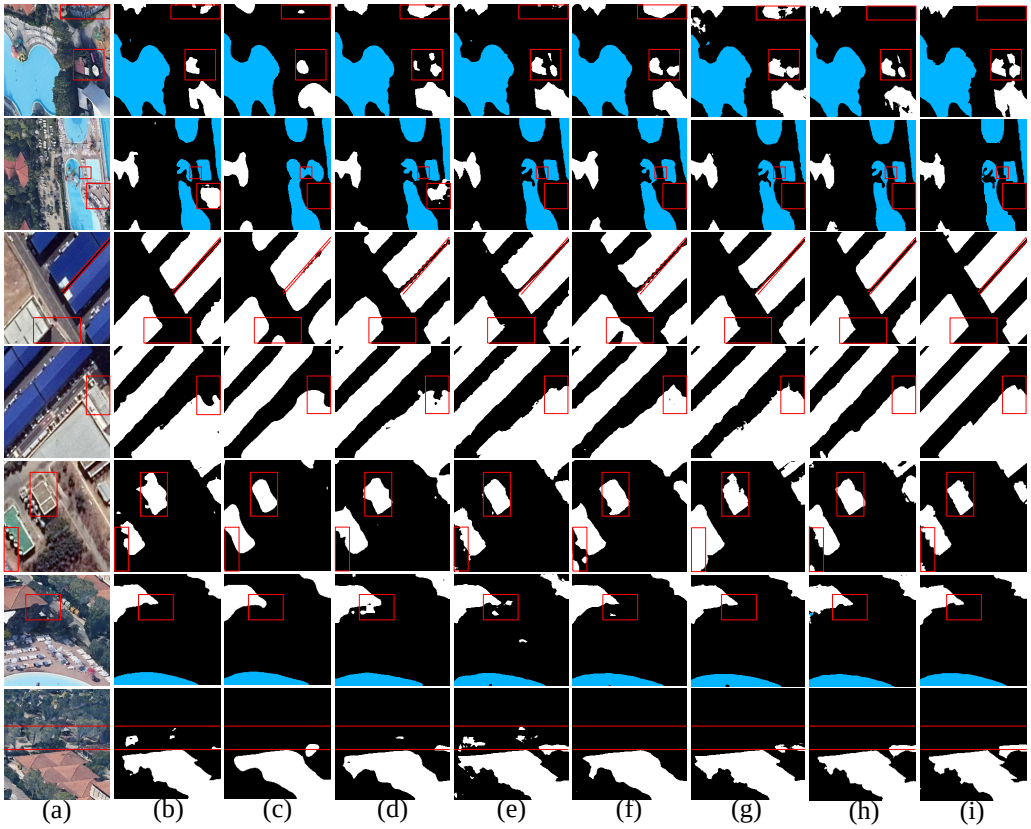

Figure 8. Prediction comparison of different algorithms on land cover dataset. (**a**) Original image; (**b**) DeepLabV3; (**c**) FCN-8s; (**d**) DABNet; (**e**) PAN; (**f**) PSPNet; (**g**) UNet; (**h**) our model; (**i**) Label.

3.6. Generalization Experiment

3.6.1. Waters Dataset

Considering the generalization ability of the double-branch parallel network model, we chose to conduct experiments on the water dataset. Compared with the building and water dataset, the water dataset can detect the segmentation ability of our model in more complex background environments. In this research, we more thoroughly demonstrate the superiority of the network using experimental comparison using several network models on the water dataset, which increases the diversity of trials.

Here, our model is contrasted with some land cover neural networks, such as ESP-NetV2, SegNet, DeepLabV3+, PSPNet, FCN8s and other traditional land segmentation networks. This article also compares some of the latest improved networks on Transformer, such as PVT (Pyramid Vision Transformer), VIT (Vision Transformer), CVT, conformer. The experiment was carried out under the same conditions. The segmentation accuracy is shown in Table 3. It is evident from the table that the segmentation accuracy of the model utilized in this investigation, which came in at 96.38%, is the highest.

Figure 9 compares the predicted representations of various models from the waters dataset. Among them, Figure 9a is the remote sensing image, Figure 9b–h represent CVT, DeepLabV3+, DFN, FCN-8s, SegNet, ShuffleV2 and the double-branch parallel network designed in this paper, Figure 9i is the label graph. Here, some areas are marked with yellow boxes for more intuitive comparison. From the figure, we can see that our network can still detect rivers well under different complex background conditions. Although other models can also detect the river, in some small tributaries there will always be some missing parts, and our model can be a good detector of small tributaries with the information. This is because the two branches of our model can fully extract spatial feature information and detail feature information, under the action of CMFM module, the global feature information is fully interacted, which can better grasp the global information and accurately

detect the location of the river. Finally, the MFM module guides the low-level feature information by using the advanced feature information, making up for the lack of some feature information.

Table 3. Compare experimental results with different models on a water dataset.

Method	PA (%)	MPA (%)	MIOU (%)
Conformer [44]	96.81	96.19	91.65
SegNet [40]	97.48	97.04	93.35
ESPNetV2 [45]	98.23	98.09	95.27
GhostNet [46]	96.80	96.32	91.61
DeeplabV3+ [36]	98.14	97.50	95.08
DFNet [47]	98.11	97.82	94.96
CVT [48]	97.34	96.61	93.03
ShuffleNetV2 [37]	98.17	98.02	95.10
FCN8s [21]	97.45	97.01	93.28
Dual-branch [22]	98.28	98.13	95.39
PVT [49]	97.88	97.75	94.34
VIT [29]	87.81	85.51	70.66
MFANet [12]	98.25	97.96	95.33
Ours	98.65	98.39	96.38

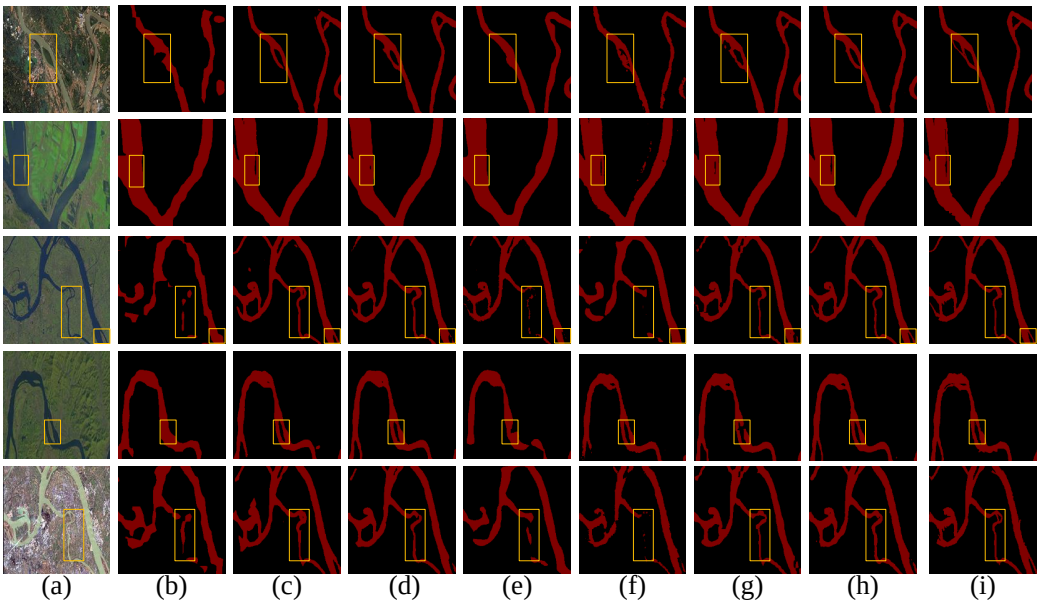

Figure 9. Comparison of prediction effects of different models in water dataset. (**a**) Original image; (**b**) CVT; (**c**) DeepLabV3+; (**d**) DFN; (**e**) FCN-8s; (**f**) SegNet; (**g**) ShuffleV2; (**h**) our model; (**i**) Label.

3.6.2. Inria Dataset

The main task of the experimental model in this paper is to segment the remote sensing images of buildings and waters. After the generalization experiment results on the waters dataset, our model has a better segmentation effect on the waters. In order to further fully reflect the ability of our model, we conducted generalization experiments on the Inria dataset and compared it with some land segmentation network models, as shown in Table 4.

Table 4. Compare experimental results with different models on Inria dataset.

Method	PA (%)	MPA (%)	MIOU (%)
DFNet [47]	93.95	91.20	82.98
CVT [48]	90.91	86.34	75.72
GhostNet [46]	91.40	87.01	76.93
MFANet [12]	94.47	92.04	84.28
SegNet [40]	94.21	91.20	83.78
HRNet [27]	93.89	90.08	82.85
DeeplabV3+ [36]	94.33	91.37	84.09
FCN8s [36]	93.21	89.31	81.49
SGBNet [8]	94.12	91.52	83.36
PVT [49]	93.95	91.26	82.48
Dual-branch [22]	94.38	92.16	83.96
ShuffleNetV2 [37]	94.22	91.34	83.75
OCRNet [50]	94.77	92.27	85.13
Ours	95.27	92.66	86.19

From the table, we can see that the model designed in this paper has achieved the best in all three indicators. The data show that the generalization ability of the model is strong and persuasive. Figure 10 shows the comparison between our model and other models. From the figure, we can see that the model designed in this paper still achieves good results on the public data set Inria. Since our two-branch model complements the advantages of CNN and swin Transformer through the designed fusion module, on the prediction effect diagram, we can see that our model has achieved good results in both the processing of edge details and the problem of misjudgment, which proves that our model has strong generalization ability.

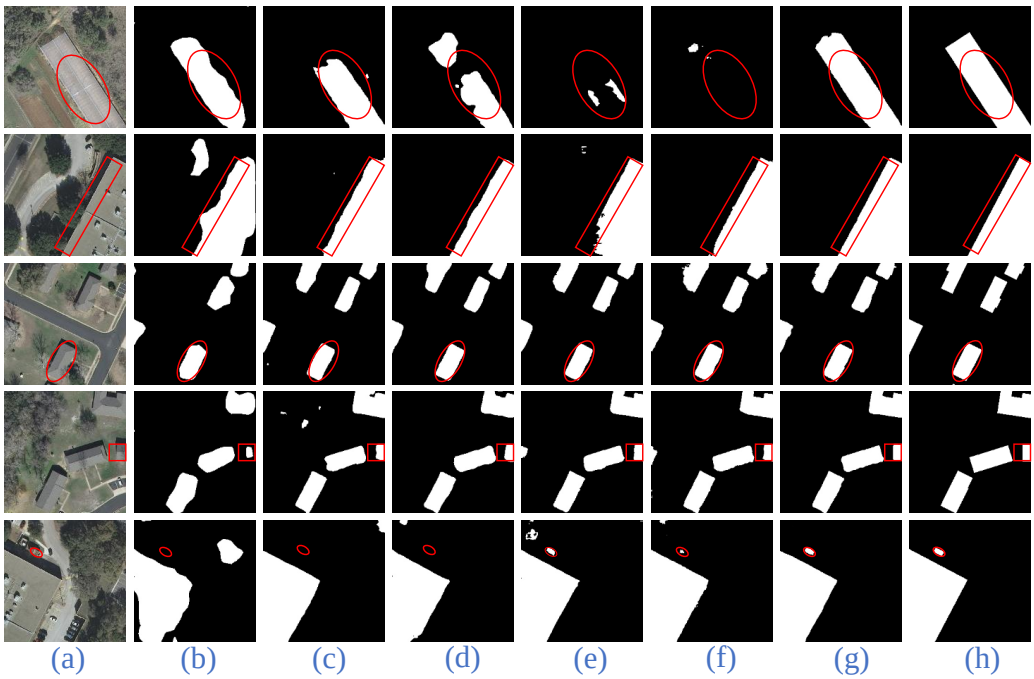

Figure 10. Comparison of prediction effects of different models in Inria dataset. (**a**) Original image; (**b**) CVT; (**c**) DeepLabV3+; (**d**) DFN; (**e**) Dual-branch; (**f**) SegNet; (**g**) our model; (**h**) Label.

4. Discussion

4.1. About the Model

The network is based on Swin Transformer and CNN. In the stage of feature coding, the designed cross-scale multi-level fusion module is used to connect the two branches,

and the comprehensive semantic information and spatial semantic information are extracted using CNN and Swin Transformer. The multi-scale fusion module designed by us guides feature information extracted by double branches to each other, giving full play to the characteristics of Swin Transformer's global information interaction, and making up for the judgment errors brought on by a lack of global information and long-term semantic information interaction of CNN. During the feature decoding stage, the designed multi-scale fusion module is utilized to fuse the high-level feature information in the coding stage and the low-level feature information extracted by CNN, and the high-level feature information is used to direct low-level feature information and upsample step-by-step. Through the joint action of several modules, our network has significantly increased the segmentation precision. The following are this paper's main contributions:

1. A double-branch parallel network of Swin Transformer and CNN is proposed. The two network structures extract feature information separately and aggregate the extracted feature information, which can better improve the accuracy and generalization of segmentation. Swin Transformer makes up for the deficiency of the limited receptive field of convolutional neural network (CNN) and can better perform global information interaction; in addition, CNN can make up for the lack of translation in the variance of Transformer.
2. Considering the difference of feature information extracted from two branches, a cross-scale bilateral feature aggregation module is proposed. This method can effectively aggregate different levels of feature information and guide each other, so that more feature information can be globally interacted. It effectively reduces the occurrence of misjudgment. In the upsampling stage, an aggregation module is also proposed, which fully utilizes high-level semantic information to direct low-level semantic information, and recovers high-resolution pixel-level feature information and edge feature information.

4.2. About the Experiment

In order to verify the ability of our model, this study conducted comparative experiments and generalization experiments on the building water dataset and the water dataset and the public Inria dataset. In the comparative experiment, our model is superior to other classical network models in the three indicators. The PA value, MPA value and MIOU value reached 93.64%, 94.11% and 87.86%, respectively. In the prediction effect diagram, the model designed in this paper is compared with other networks regarding the problem of misjudgment, as well as dealing with some edge feature information, which is used to reflect the advantages of our dual branch network. CNN and Swin Transformer give full play to their respective advantages under the action of the fusion module; more feature information is globally interacted, and advanced feature information is used to guide low-level feature information, making edge detail features more delicate. We conducted generalization experiments on the water data set and the Inria data set. Similarly, from the numerical and prediction effect diagrams, our network model is still superior to other models, which proves that our model has better generalization ability.

4.3. Limitations and Future Prospects of the Model

Since both CNN and Transformer models have large computational overhead, our work in the future is to further optimize the structure of the model while ensuring the segmentation accuracy of the model, design more efficient model structure and more effective training strategies to reduce training complexity and training difficulty.

5. Conclusions

In remote sensing images, house waters are important geographical indications. They have important practical significance for land planning, water resources protection planning and geographic mapping. The segmentation task of buildings and waters is also an important part of the land cover segmentation task. Existing image segmentation network

models primarily employ CNNs to extract feature information from images. In order to make up for the deficiency of CNN in feature extraction and better fully interact with global semantic information, this paper presents a double branch parallel network structure algorithm for segmentation task. In the coding process of the algorithm, we use Resnet50 and Swin Transformer to extract features for the two branches, obtain rich context information and spatial information, utilize the benefits of the two branches' feature extraction information to the fullest extent, and fuse the feature information extracted by different branches through our fusion module, which provides rich pixel information for the upsampling information recovery. In the process of decoding, we use a fusion module designed to fuse the encoded high-level feature information with the feature information of ResNet50 branch, and use high-level feature information to direct low-level feature information. Upsampling can gradually refine and restore high-resolution images and obtain more spatial details. Compared with some of the current semantic segmentation network models, our model has greatly improved the accuracy of segmentation in buildings and waters. From the performance of different datasets, our model has good anti-interference and recognition capabilities, and can accurately determine the location of waters and houses in complex background environments, while the segmented edges are also more delicate. In the future, in order to enhance the practical applications, we will further optimize the model structure under the assumption of ensuring segmentation precision, and improve the model training speed.

Author Contributions: Conceptualization, J.C. and M.X.; methodology, M.X. and J.C.; software, J.C.; validation, D.W. and H.L.; formal analysis, M.X.; investigation, J.C.; resources, M.X.; data curation, J.C.; writing–original draft preparation, J.C.; writing–review and editing, M.X.; visualization, M.X.; supervision, M.X.; project administration, M.X.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of PR China (42075130).

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

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

References

1. Shu, Q.; Pan, J.; Zhang, Z.; Wang, M. DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102940. [CrossRef]
2. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial Cross Attention Meets CNN: Bi-Branch Fusion Network for Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 32–43. [CrossRef]
3. Yu, Y.; Huang, L.; Lu, W.; Guan, H.; Ma, L.; Jin, S.; Yu, C.; Zhang, Y.; Tang, P.; Liu, Z.; et al. WaterHRNet: A multibranch hierarchical attentive network for water body extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103103. [CrossRef]
4. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
5. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
6. Hu, K.; Li, M.; Xia, M.; Lin, H. Multi-Scale Feature Aggregation Network for Water Area Segmentation. *Remote Sens.* **2022**, *14*, 206. [CrossRef]
7. Wang, T.; Lan, J.; Han, Z.; Hu, Z.; Huang, Y.; Deng, Y.; Zhang, H.; Wang, J.; Chen, M.; Jiang, H.; et al. O-Net: A novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification. *Front. Neurosci.* **2022**, *16*, 876065. [CrossRef]
8. Pang, K.; Weng, L.; Zhang, Y.; Liu, J.; Lin, H.; Xia, M. SGBNet: An Ultra Light-weight Network for Real-time Semantic Segmentation of Land Cover. *Int. J. Remote Sens.* **2022**, *43*, 5917–5939. [CrossRef]
9. Chen, J.; Sun, B.; Wang, L.; Fang, B.; Chang, Y.; Li, Y.; Zhang, J.; Lyu, X.; Chen, G. Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102881. [CrossRef]
10. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 5940–5960. [CrossRef]

11. Wang, Z.; Xia, M.; Lu, M.; Pan, L.; Liu, J. Parameter Identification in Power Transmission Systems Based on Graph Convolution Network. *IEEE Trans. Power Deliv.* **2022**, *37*, 3155–3163. [CrossRef]
12. Chen, B.; Xia, M.; Huang, J. Mfanet: A multi-level feature aggregation network for semantic segmentation of land cover. *Remote Sens.* **2021**, *13*, 731. [CrossRef]
13. Ma, Z.; Xia, M.; Weng, L.; Lin, H. Local Feature Search Network for Building and Water Segmentation of Remote Sensing Image. *Sustainability* **2023**, *15*, 3034. [CrossRef]
14. Ding, Y.; Zhao, X.; Zhang, Z.; Cai, W.; Yang, N.; Zhan, Y. Semi-supervised locality preserving dense graph neural network with ARMA filters and context-aware learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
15. Zhang, Z.; Ding, Y.; Zhao, X.; Siye, L.; Yang, N.; Cai, Y.; Zhan, Y. Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification. *Expert Syst. Appl.* **2023**, *217*, 119508. [CrossRef]
16. Ding, Y.; Zhang, Z.; Zhao, X.; Hong, D.; Cai, W.; Yu, C.; Yang, N.; Cai, W. Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification. *Neurocomputing* **2022**, *501*, 246–257. [CrossRef]
17. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote Sens.* **2022**, *43*, 5874–5894. [CrossRef]
18. Hu, K.; Weng, C.; Zhang, Y.; Jin, J.; Xia, Q. An Overview of Underwater Vision Enhancement: From Traditional Methods to Recent Deep Learning. *J. Mar. Sci. Eng.* **2022**, *10*, 241. [CrossRef]
19. Hu, K.; Ding, Y.; Jin, J.; Weng, L.; Xia, M. Skeleton Motion Recognition Based on Multi-Scale Deep Spatio-Temporal Features. *Appl. Sci.* **2022**, *12*, 1028. [CrossRef]
20. Ding, Y.; Zhang, Z.; Zhao, X.; Cai, Y.; Li, S.; Deng, B.; Cai, W. Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-branch Network for Cloud and Cloud Shadow Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
23. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
24. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
27. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
29. 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. *arXiv* **2020**, arXiv:2010.11929.
30. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9355–9366.
31. 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, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
32. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
33. Gao, J.; Weng, L.; Xia, M.; Lin, H. MLNet: Multichannel feature fusion lozenge network for land segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 016513. [CrossRef]
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
35. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
36. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
37. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.

38. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
39. Fang, L.; Liu, J.; Liu, J.; Mao, R. Automatic segmentation and 3d reconstruction of spine based on fcn and marching cubes in ct volumes. In Proceedings of the 2018 10th International Conference on Modelling, Identification and Control (ICMIC), Guiyang, China, 2–4 July 2018; pp. 1–5.
40. 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]
41. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
42. Jiang, W.; Wu, Y.; Guan, L.; Zhao, J. Dfnet: Semantic segmentation on panoramic images with dynamic loss weights and residual fusion block. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5887–5892.
43. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
44. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 367–376.
45. 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, Long Beach, CA, USA, 16–20 June 2019; pp. 9190–9200.
46. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1580–1589.
47. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1857–1866.
48. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
49. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
50. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.