remote sensing

MDPI

Article

MCAFNet: A Multiscale Channel Attention Fusion Network for Semantic Segmentation of Remote Sensing Images

Min Yuan [1,*], Dingbang Ren [1], Qisheng Feng [2], Zhaobin Wang [1], Yongkang Dong [1], Fuxiang Lu [1] and Xiaolin Wu [1]

[1] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
[2] College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou 730000, China
* Correspondence: yuanm@lzu.edu.cn

Abstract: Semantic segmentation for urban remote sensing images is one of the most-crucial tasks in the field of remote sensing. Remote sensing images contain rich information on ground objects, such as shape, location, and boundary and can be found in high-resolution remote sensing images. It is exceedingly challenging to identify remote sensing images because of the large intraclass variance and low interclass variance caused by these objects. In this article, we propose a multiscale hierarchical channel attention fusion network model based on a transformer and CNN, which we name the multiscale channel attention fusion network (MCAFNet). MCAFNet uses ResNet-50 and Vit-B/16 to learn the global–local context, and this strengthens the semantic feature representation. Specifically, a global–local transformer block (GLTB) is deployed in the encoder stage. This design handles image details at low resolution and extracts global image features better than previous methods. In the decoder module, a channel attention optimization module and a fusion module are added to better integrate high- and low-dimensional feature maps, which enhances the network's ability to obtain small-scale semantic information. The proposed method is conducted on the ISPRS Vaihingen and Potsdam datasets. Both quantitative and qualitative evaluations show the competitive performance of MCAFNet in comparison to the performance of the mainstream methods. In addition, we performed extensive ablation experiments on the Vaihingen dataset in order to test the effectiveness of multiple network components.

Keywords: semantic segmentation; transformer; channel attention module; hybrid structure

Citation: Yuan, M.; Ren, D.; Feng, Q.; Wang, Z.; Dong, Y.; Lu, F.; Wu, X. MCAFNet: A Multiscale Channel Attention Fusion Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 361. https://doi.org/10.3390/rs15020361

Academic Editors: Jiaojiao Li, Qian Du, Wei Li, Bobo Xi, Jocelyn Chanussot, Rui Song and Yunsong Li

Received: 23 October 2022
Revised: 21 December 2022
Accepted: 4 January 2023
Published: 6 January 2023

1. Introduction

Semantic segmentation assigns semantic labels to each pixel of the image [1]. In the field of remote sensing, high-resolution remote sensing images provide corresponding data and information support for the construction of smart cities. However, high-resolution urban remote sensing imagery contains rich information about ground objects, which leads to the common phenomenon of large intraclass variance and small interclass variance. Figure 1 shows the local remote sensing image of Potsdam; the orange boxes display the importance of capturing multiscale semantic information, and the black boxes illustrate the difference between small-scale objects. Therefore, extracting useful relevant information from remote sensing images has become a key issue.

In recent years, remote sensing images have developed great potential for application in the field of smart city construction. However, traditional image semantic segmentation of remote sensing images is typically performed by extracting low-level features from the image. When establishing the corresponding semantic segmentation model, there is a gap between the artificially designed features and the high-level semantic features, so the generalizability of the established semantic segmentation model is poor. The interpreted results of deep-learning-based semantic segmentation algorithms in remote sensing city images often present lump-like fuzzy boundaries, which do not sufficiently preserve the

feature information of objects. This leads to the confusion of semantic classifiers and brings great challenges to the task of semantic segmentation. Effectively segmenting small objects and improving the interpretation accuracy are still extremely challenging tasks.

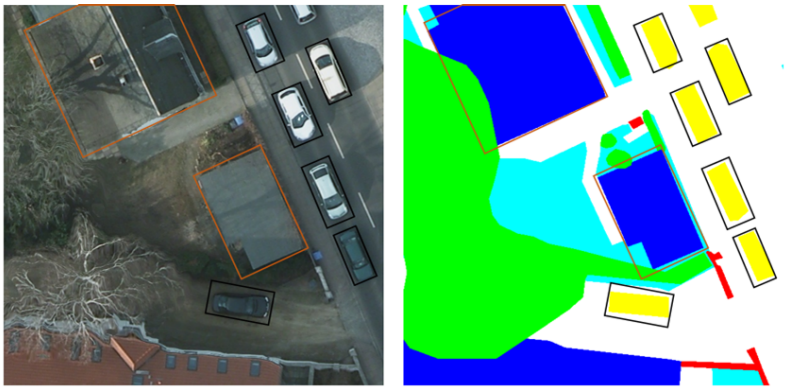

Figure 1. The challenge of urban remote sensing image interpretation.

The transformer is a concept proposed by Google in the literature [2]. Since its birth in 2017, the transformer has made rapid breakthroughs in the NLP field. With the advancement of research, it also shows great potential in the field of computer vision, providing novel solutions and achieving good results. Image Transformer [3], released in 2018, was the first to migrate the transformer architecture to the field of computer vision. Since 2019, transformer-based visual models have developed rapidly, and many new attention achievements have appeared. For example, the segmentation transformer (SETR) [4] uses a transformer encoder to completely replace the CNN backbones, discards the convolution and downsampling processes, and uses split tasks as sequence-to-sequence prediction tasks. The detection transformer (DETR) [5] applies the advantages of the transformer in the field of target detection. In July 2020, Chen et al. [6] proposed the iGPT model in order to explore the performance of the approach on images, as well as the performance of unsupervised accuracy. In October 2020, Dosovitskiy et al. [7] proposed the Vision Transformer model, an image classification scheme that is based entirely on the mechanism of self-attention, which was the first work using a transformer to substitute a standard convolution. In January 2021, Esser et al. [8] constructed the vector quantized generative adversarial network (VQGAN), which combines a transformer and CNN, and it is the first transformer architecture to generate megapixel images by semantic guidance. It is worth noting that researchers from Facebook and Berkeley [9] rechecked the design space and tested the limits that pure ConvNet can reach, indicating that the performance of a convolutional neural network is no less than that of a visual transformer, while maintaining the simplicity and effectiveness of the standard ConvNet.

Therefore, given the difficulty of identifying different scales in remote sensing images, we propose a hybrid network structure based on a transformer and CNN, which makes full use of semantic feature information at different scales. In the decoder module, a variety of effective blocks is applied to study an urban scene with complex surface features based on a CNN. The following are this paper's main contributions:

(1) We combine the ResNet-50 and a transformer hybrid model to improve the current mainstream semantic segmentation network structure, and the proposed global–local transformer block models the spatial distance correlation in the image while maintaining the hierarchical characteristics.
(2) We propose a channel attention module decoder (CAMD). In the module, a pooling fusion module is designed to enrich the feature expression of the network. We evaluated the efficiency of each part of the decoder module through ablation research.
(3) We added a fusion module to optimize the structure of the hybrid model, merge feature maps from different scales, and improve semantic representation of the underlying features.

2. Related work

2.1. Methods for Semantic Segmentation Based on Deep Learning

Deep learning [10] is a new research direction in machine learning in recent years. The field of remote sensing image interpretation has gradually implemented deep learning algorithms to deal with problems that were difficult to solve by traditional machine learning methods. At present, the mainstream deep semantic segmentation networks include three forms: a network based on a spatial pyramid structure, a multibranch network, and an encoder–decoder network. These three networks can handle problems in multiscale semantic information extraction and output resolution degradation.

The network based on the spatial pyramid structure uses the pyramid structure to capture the scale semantic information. This kind of network introduces numerous branches to reference the network's end, and each branch corresponds to a fixed scale. For example, PSPNet [11] generates input with different resolutions through an adaptive average pooling operation. PANet [12] changes the input feature map's resolution through a convolutional operation with various core sizes and step sizes. DeepLab [13,14] proposes the atrous spatial pyramid structure (ASPP), which fixes the input resolution of each branch of the pyramid structure and introduces convolutional layers with different expansion rates to expand the network's receptive field. DensAspp [15] improves the receptive field of the pyramid structure in DeepLab by introducing dense connections, making the structure suitable for large-resolution pictures.

The multibranch network sends the input image into multiple branches, and each branch has a different output resolution. For example, icnet [16] uses two spatial branches to capture small-scale targets. Reference [17] uses the branches with higher output resolution to generate a proportional fraction map, which optimized the spatial information of low-resolution branches. Bisenet [18] proposes a lightweight branch with high output resolution and introduces the attention mechanism into the fusion process of different branches, which greatly improves the network speed while maintaining the network accuracy. By combining the shallow characteristics of many branches, Fast SCNN [19] creates a multibranch network, which considerably reduces the amount of calculation consumed on the high-resolution output branches.

The encoder–decoder network gradually integrates the high-dimensional feature map into the low-dimensional feature map to improve the resolution of the output. At the same time, different levels of feature maps have different resolutions. Integrating them can enable the network to capture semantic information of different scales. Therefore, the encoder–decoder network is an effective method to address resolution degradation and multiscale complications. The first deep semantic segmentation network, FCN [20] is a famous encoder–decoder network. It generates a layer-hopping connection structure to integrate high-dimensional and low-dimensional feature maps. On this basis, U-Net [21] proposes a more efficient layer-hopping connection structure, which realizes the fusion of different feature maps with higher accuracy. SegNet [22] records the pooled index in the encoding process and uses the pooled index to supervise the decoding process, making the decoding process more standardized. Refinenet [23] introduces a large number of optimization modules to optimize the feature map fusion results, which could increase the information capture ability of the fused feature map.

Compared with the above two networks, the encoder–decoder structure does not need to change the reference framework and draw additional branches to obtain small-scale semantic information. It needs only to properly optimize the semantic feature maps at different levels on the basis of the reference network. Therefore, this approach is best suited to the domain of semantic segmentation in remote sensing.

Although the encoder–decoder network has great advantages in the field of semantic segmentation, the existing studies have not found an accurate optimization and fusion method for high- and low-dimensional feature maps. Therefore, improving the optimal fusion efficiency of high- and low-dimensional feature maps is a bottleneck in the application of encoder–decoder networks.

2.2. Methods for Semantic Segmentation Based on Transformers

For image problems, convolution has a natural inherent bias translation of equivalence and locality. The transformer obviously does not have these advantages, but its core self-attention operation can obtain a large range of global information, which has obvious advantages for the information extraction range of images. The reasons for the rapid development of the transformer can be attributed to its strong ability to learn long-distance dependencies, multimodal fusion ability, and more interpretable models.

Therefore, many segmentation algorithms take ViT as the backbone network, with Segmenter [24], Segformer [7], and Swin Transformer [25] as typical representatives. Strudel et al. [24] proposed a converter model for semantic segmentation based on the research results from ViT. Segment adopts the ViT model structure in the coding stage, divides the image into blocks, performs a linear mapping, and outputs the embedded sequence after being processed by the encoder. In the decoding stage, learnable category embedding is introduced, and the output of the encoder and category embedding are sent to the decoder, which obtains the class label. Xie et al. [7] proposed SegFormer, a simple, effective, yet powerful semantic segmentation framework. SegFormer uses a hierarchical feature representation method that combines a transformer with a light multilayer perceptron (MLP). Swin Transformer [25] uses a multi-stage design similar to the convolutional neural network, and each stage has a different resolution of the feature map. This mechanism of using a local window attention fully proves that convolution, a method of extracting local feature information, can play its role.

In summary, the transformer has proven to be more powerful than a CNN in feature extraction in semantic segmentation. However, during the semantic segmentation test, the resolution of the picture is not fixed. Its requirements for pixel classification and contour details are meticulous. Transformer-based semantic segmentation methods have poor effects in processing image details. Therefore, more research is needed to build an effective transformer structure and combine it with the current CNN model.

3. The Proposed Method

In this section, we elaborate the method for the semantic segmentation of high-resolution images that combines the hybrid transformer and CNN encoder model. In Section 3.1, we describe the principle and network structures of the hybrid model based on the transformer and ResNet-50. In Section 3.2, we describe the structure of the CAMD module in the CNN-based decoder. Finally, the overall design of our network is described in Section 3.3.

We propose a multiscale channel attention fusion network (MCAFNet) in the context of the semantic segmentation of remotely sensed images. The framework of the MCAFNet is shown in Figure 2. The overall structure of the MCAFNet follows an encoder–decoder structure. A remotely sensed image of an urban area has rich spectral information and texture structure and irregular ground object boundaries, and these characteristics require a higher feature extractor. Therefore, in the encoder part, we used the hybrid model of a CNN and a transformer to extract the multilayer features of the image and optimize the structure of the transformer block. When facing the decoder part of the MCAFNet, the channel attention decoder module is introduced to learn the complex relationship between high- and low-dimensional semantic features. The fusion module is used to improve the fusion efficiency.

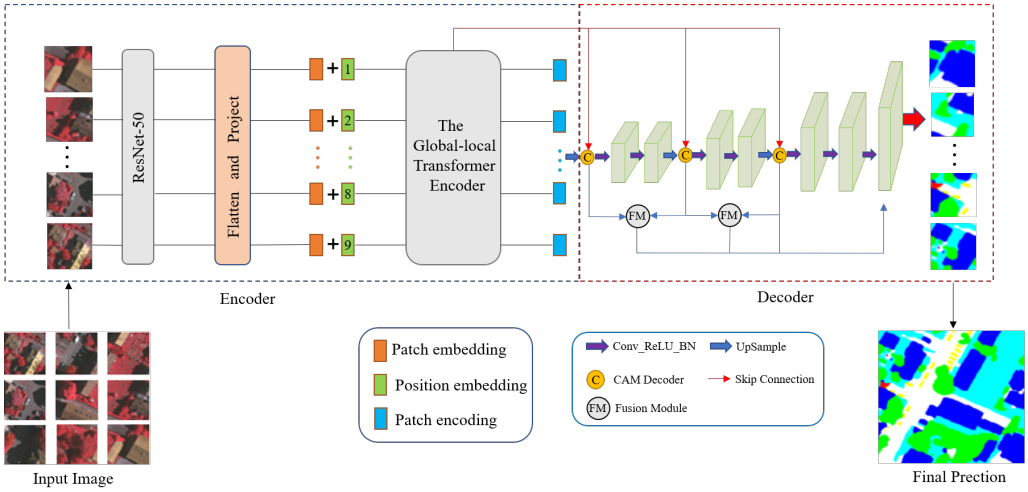

Figure 2. Overall structure diagram of the MCAFNet.

3.1. CNN-Transformer Hybrid as Encoder

The encoder module, which is presented in Figure 3, is designed as a hybrid network model of ResNet-50 and Vit-B/16 [7]. The ResNet-50 convolutional layer is used to enhance the expression of the local context information. The linear multihead self-attention of the transformer module is used to capture the global context information of urban remote sensing images.

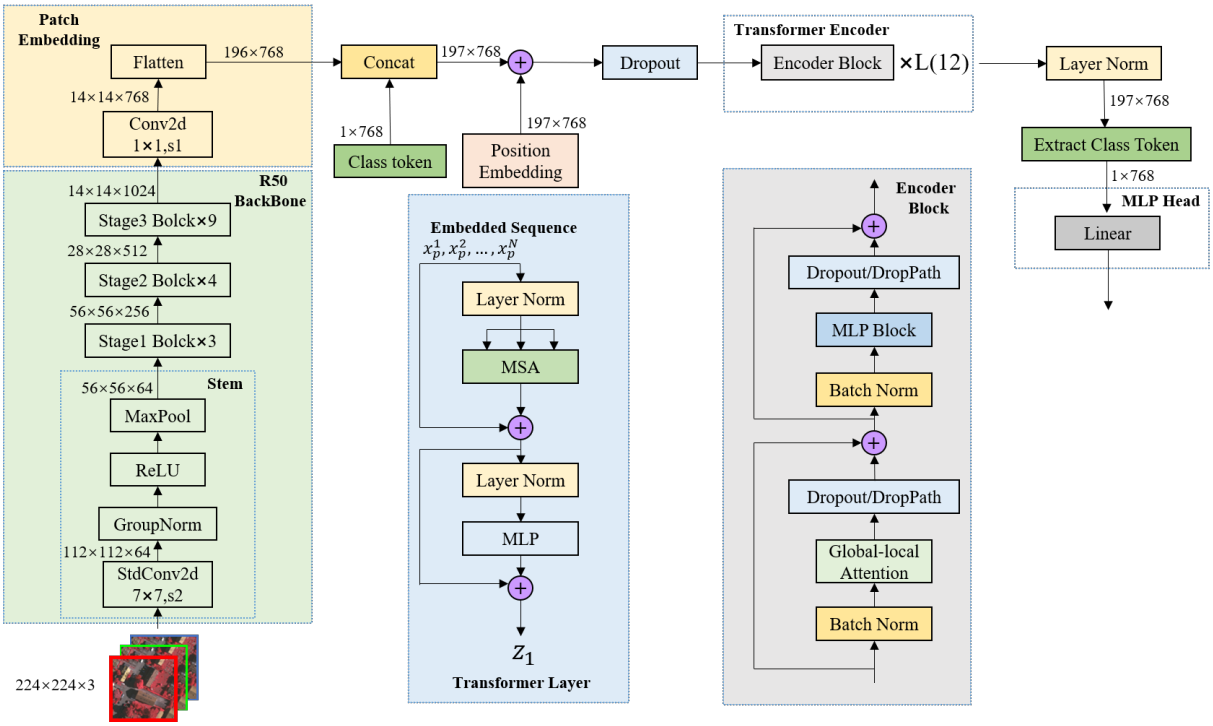

Figure 3. The encoding part structure of the MCAFNet.

We first cut the input remote sensing image x into fixed size patches $\{x = [x_1, \cdots, x_N] \in \mathbb{R}^{(N \times P^2 \times C)}\}$ for feature extraction, where $N = H \times W / P^2$ is the number of image patches and C is the number of slice channels. Then, we used ResNet-50 to perform preliminary semantic feature extraction on the patch. We flattened each patch into a one-dimensional vector $\{X_0 = [E_{X_1}, \cdots, E_{X_N}] \in \mathbb{R}^{N \times D}, E \in \mathbb{R}^{P^2 C}\}$ and, then, performed a linear projection to produce a series of patch embeddings to retain low-dimensional semantic feature in-

formation. Finally, we modeled the global image context information based on position embedding in the transformer, which perfectly removes the dependence on convolution.

Specifically, inspired by Wang et al. [26], we introduced convolutional groups to extend the multihead attention module. The specific structure is shown in Figure 4b; it adds a convolutional group branch to extract the local features of the image, while retaining the transformer's self-attention mechanism as the global feature extraction branch. In addition, we used batch normalization to solve the variable shift problem in transformer training, accelerate the convergence rate of the model, and resolve the overfitting problem.

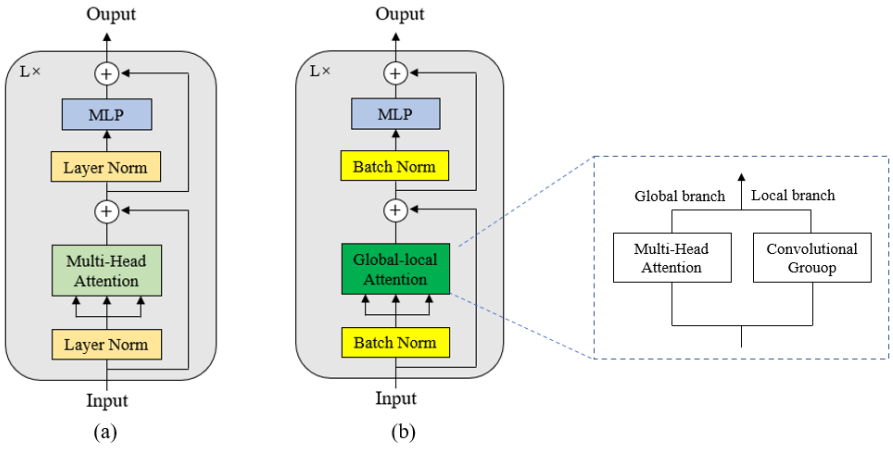

Figure 4. Optimization example of the transformer block. (**a**) is the basic block, and (**b**) is the optimized block.

Figure 5 shows the combination design idea of the module's convolution and self-attention. In order to obtain high- and low-level semantic information at the same time, we processed one part of the input image as a one-dimensional sequence based on the QKV mechanism, and the other part recovers from the sequence the two-dimensional feature map for convolution processing and, finally, splices according to the channel dimensions to output feature vectors with rich semantic information.

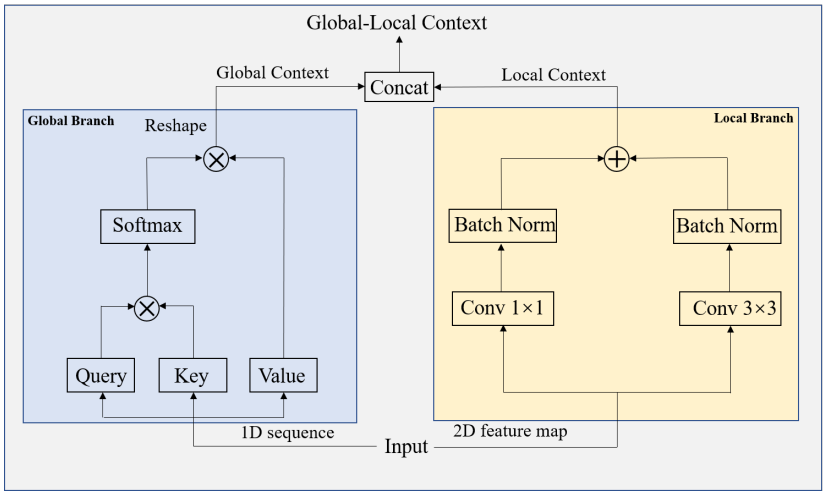

Figure 5. Specific structure design of the global–local attention module.

The global branch deploys the multihead self-attention to capture the global context, and the local branch uses two parallel convolutional layers with core sizes of 3 and 1 to extract the local context. In the transformer, the self-attention mechanism is represented

by a linear layer of three points mapped to an intermediate layer. The QKV mechanism is calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \quad (1)$$

Based on the QKV mechanism of self-attention, in this paper, we combined the deep and shallow semantic features of the input semantic features through weighted fusion, which calculates the correlation between the deep semantic features and the other shallow features in the transformer encoder.

We used the encoded feature vector corresponding to the deep feature map as the query and the value of the multihead attention mechanism and used the encoded feature vector corresponding to the shallow feature map as the key to perform attention fusion. Then, we multiplied the fusion attention map by the encoded feature vector corresponding to the deep feature map, which obtains $Attention(Q, K, V)$ through residual connections and layer normalization. Finally, more precise semantic features are output through the feed-forward network.

3.2. CNN-Based Decoder

As is shown in Figure 6, the process of the MCAFNet interpretation of urban remote sensing images can be summarized as follows: information encoding, information optimization, and information fusion. In addition, a channel attention module is added to adjust the weight of semantic features. Inspired by Ma et al. [27], we added a pooling fusion module (PFM) to enrich the feature expression of remote sensing image semantic information. Its specific structure is shown in Figure 7.

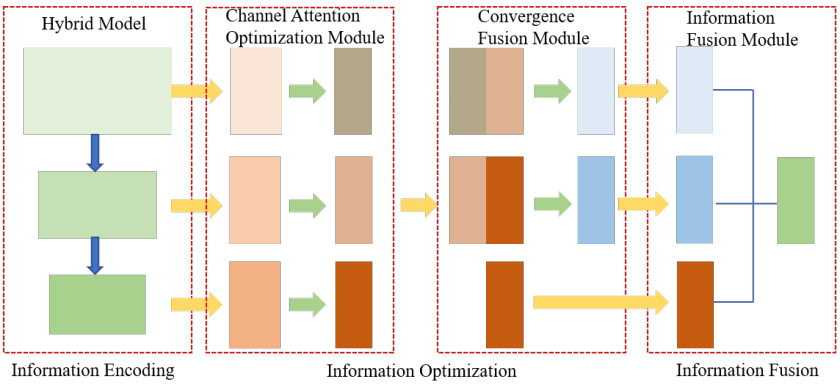

Figure 6. Three steps to decode multiscale semantic information.

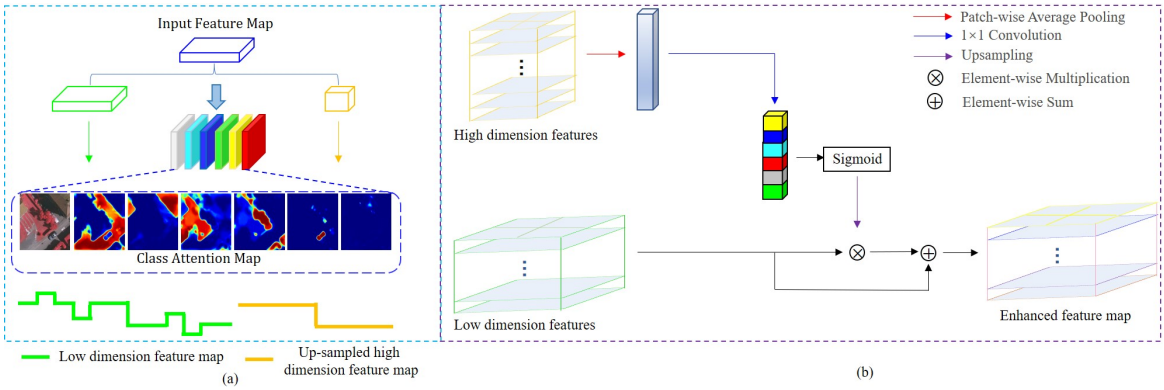

Figure 7. Design motivation and detailed structure of the PFM. (**a**) is the design motivation of PFM, and (**b**) is the specific structure of PFM.

Figure 7a shows the design motivation for the PFM. We materialized the semantic feature maps of six categories of objects and found that the requirements for high-resolution and high-level semantic segmentation contradict the design of convolutional networks. The broken lines with different fluctuations are used to represent the low-dimensional feature map and the high-dimensional feature map after upsampling. Intraclass differences can be shown by discounting small fluctuations, and the differences between classes can be represented by large discounted fluctuations. When they are merged along the spliceosome of the channel dimension, the high- and low-dimensional feature maps of each location are not equally effective, and there are uneven spatial dimensions. Simple upsampling cannot solve the semantic gap problem.

Figure 7b shows the specific structure of the PFM. The operation core of the pooling fusion module is to embed the local attention information in the high-dimensional semantic features of remote sensing images into the low-dimensional semantic features. In this manner, low-dimensional features can be fused to sense the field context information, while this original spatial information will not be lost. First, the average pooling layer is used to optimize the high-dimensional feature map, retain the background information, and obtain the channel attention vector. Then, we reused a 1×1 convolutional layer encoder of each channel weight vector that unifies the number of high- and low-dimensional feature map channels. Finally, we extracted the local attention feature map Z_c from high-dimensional semantic features. The calculation formula is as follows:

$$Z_c = \frac{1}{h_p w_p} \sum_{i=1}^{h_p} \sum_{j=1}^{w_p} x_c(i,j) \tag{2}$$

where $h_p w_p$ denotes the split window size of the average pooling operation and x_c represents a pixel from the c channel.

On this basis, we set the extracted high-dimensional feature map as $Z_H \in \mathbb{R}^{C_h \times H_h \times W_h}$, and set the original low-dimensional feature map as $X_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. Based on the move flip bottleneck convolutional operation, we generated attention maps for the low-dimensional features M_l [28] by transformation. The calculation formula is as follows:

$$M_l = F_u\{\sigma[H_l \delta(H_r Z_h)]\} \tag{3}$$

where σ and δ stand for the sigmoid and ReLU functions. A dimension-reduction convolution of 1×1 with the reduction ratio r is represented by H_r [28]; H_l adjusts the number of channels to match X_l; F_u is the upsampling operation. In addition, we added a residual design to emphasize the importance of low-dimensional features. The augmented features are computed as follows:

$$B_l = X_l + X_l M_l \tag{4}$$

Finally, the PFM outputs feature maps with both precise semantic and spatial information.

The attention mechanism can greatly improve the information capture ability of feature maps. Through the observation of urban remote sensing images, it was found that there are not too many irregular boundaries between adjacent objects in the image, and there was less detail information in the image, making the spatial information of high- and low-dimensional feature maps more accurate. Therefore, this paper used the channel attention mechanism in the soft attention mechanism and did not introduce the spatial attention branch to optimize the spatial information of low-dimensional feature maps. Therefore, we designed the channel attention decoder by combining the channel attention module and the PFM. Figure 8 shows the specific structure of the channel attention module decoder (CAMD).

First, we improved the semantic and spatial information acquisition ability of the MCAFNet based on the channel attention module (CAM). Then, we used upsampling and a 3×3 convolutional operation the further optimize and unify the resolution and channel number of the redefined feature map. Finally, the PFM is used to emphasize

the underlying feature information in the key high-dimensional features and filter the background information, restore the pixel position of the target category, and output the enhanced feature information.

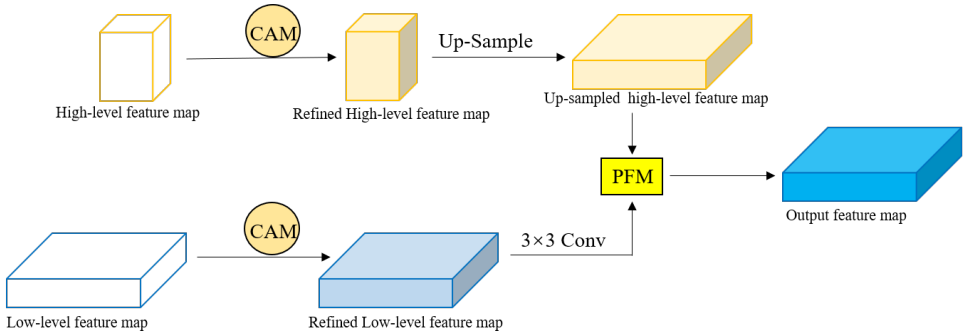

Figure 8. Specific structure of channel attention module decoder.

3.3. Network Architecture

Inspired by Peng et al. [29], we took the output feature maps of the hybrid model as the optimization targets, in order for the network to be able to capture semantic information for three different scales. The characteristic images of Output 1 and Output 2 have a large resolution, which makes them more sensitive to small-scale targets in urban remote sensing images. Therefore, they are the most-important optimization objectives of the network. The overall network structure of the hierarchical encoding and decoding network is shown in Figure 9.

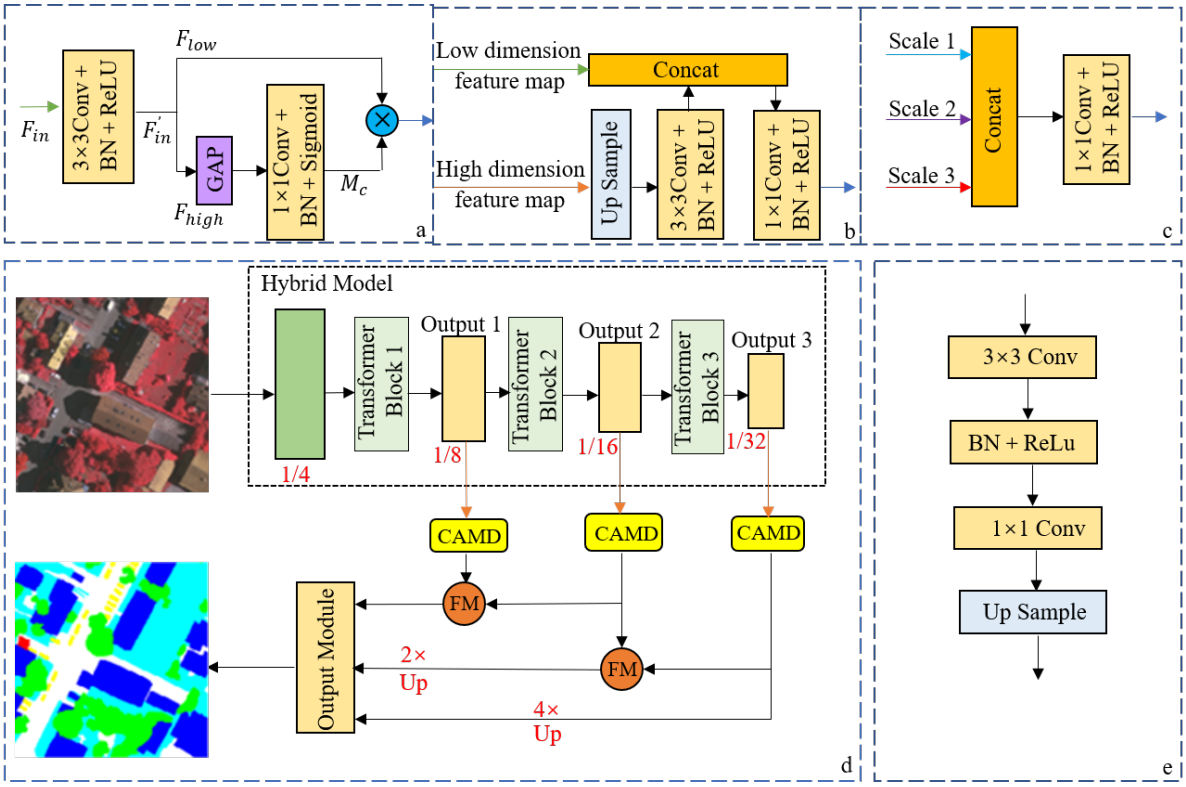

Figure 9. Each module of the network and its overall network structure. (**a**) Channel attention optimization module network structure, (**b**) convergence module network structure, (**c**) information fusion module network structure, (**d**) overall network structure, and (**e**) output module network structure.

A transformer is more suitable for extracting global image information and has a limited ability to capture local semantic information. Directly fusing its intermediate feature map to obtain multiscale semantic information leads to poor segmentation results. Therefore, before fusing the multilevel feature maps, we added a 3×3 convolutional layer combined with BN and ReLU to the front-end of the channel attention branch so that it has the function of unifying the number of output channels, which optimizes the feature maps of each dimension, especially low-dimensional feature maps. We assumed that the feature map $F_{in} \in R^{C \times h \times w}$ is the input of the channel attention mechanism and $M_c \in R^{C \times 1 \times 1}$ is the channel attention mask. The calculation process of the output feature map F_{cout} after the channel attention mechanism optimization is as follows. The improved channel attention branch is shown in Figure 9a.

$$F_{cout} = F_{low} \bigotimes M_C \tag{5}$$

where $\otimes$ denotes multiplication by elements. However, the network is still unable to extract accurate small-scale semantic information based on the improved channel attention branch. Therefore, after using the channel attention branch to optimize the feature maps of each dimension, we continued to use the pooling fusion module of the CAMD to optimize the low-dimensional feature maps. The specific fusion module structure used in this article is illustrated in Figure 9b.

The fusion module that we adopted has two input feature maps F_{cout} and F_{high}. The resolution of the high-dimensional feature map is half of the resolution of the low-dimensional feature map. This setting limits the information spread between the two feature maps, which is convenient for information fusion. The fusion module first uses the bilinear interpolation operation UP(·) so that the resolution of the high-dimensional feature map is consistent with the low-dimensional feature map. Thereafter, we adopted a 3×3 convolutional layer $CV_{3\times3}$ (·) combining BN and ReLU to optimize the high-dimensional feature maps after upsampling. Finally, the optimized high- and low-dimensional feature maps are aggregated through the concat operation C (· , ·), and the aggregation is optimized using a 1×1 convolutional layer $CV_{1\times1}$ (·) that combines BN and ReLU to generate the output. The process of realizing the entire fusion module can be expressed as follows:

$$F_{cfout} = CV_{1\times1}(C(F_{low}, CV_{3\times3}(UP(F_{high})))) \tag{6}$$

Through the fusion module, the low-dimensional feature map obtains more abstract information, and its ability to capture semantic information is also significantly improved. In addition, the fusion module has a simple structure, which quickly improves the ability of low-dimensional feature maps.

We introduce the channel attention module and fusion module to optimize the feature maps of each dimension, the feature maps of each dimension are simultaneously sent to the information fusion module shown in Figure 9c. This module uses cascade operations to fuse the characteristics of different scales, optimizes the fusion result with a 1×1 convolutional layer combining BN and ReLU, and finally, outputs a feature map that captures multiscale semantic information.

The specific MCAFNet architecture is illustrated in Figure 9d. First, we used a hybrid network model to extract global and local features from remotely sensed imagery. Then, the channel attention branches were used to optimize the feature maps from different levels within the network to improve their ability to capture semantic information. Next, the feature maps of Output 1 and Output 2 are sent to the same fusion module, and the feature maps of Output 2 and Output 3 are sent to another fusion module to greatly improve the ability of the low-dimensional feature map to capture small-scale semantic information. Unlike encoder–decoder networks, our network does not introduce the information of Output 3 to optimize Output 1 in the decoding process, which avoids the impact of the information gap on the optimization efficiency of Output 1. After optimizing the feature maps of each dimension, they are upsampled at the same time to make them have the same resolution as Output 1, and they are input into the information fusion module, which can

improve the ability of capturing semantic information by output feature maps. Finally, the output result of the information fusion module is processed by the output module to generate the final network output feature map, as shown in Figure 9e.

4. Experiment Setup

4.1. Datasets and Evaluation Metrics

All the data needed for repeating the experiments described in the paper are available at https://www.isprs.org (accessed on 10 April 2022). The dataset includes two sub-datasets: the ISPRS Vaihingen and Potsdam 2D semantic segmentation datasets, which correspond to two high-resolution urban remote sensing images of Vaihingen and Potsdam in Germany. The ground objects in the image are marked and distinguished with bright colors, rich ground structures, and representative categories and are suitable for verifying the generalization and robustness of semantic segmentation models. According to the experimental results officially given by ISPRS, generally, only the classification accuracy of the first five categories is evaluated. Therefore, we conducted ablation and interpretation experiments based on them.

Since the image resolution of the experimental dataset is very high, it cannot be directly sent to the GPU for training on the network. We cut the tif image into nonoverlapping blocks to produce 10,000 images, each with a size of 256 × 256 as the training set and test set. The image as rotated, flipped, tilted, translated, elastically transformed, perspective, cropped, and zoomed to complete the expansion of the dataset. Figure 10 shows some examples of the data augmentation.

We calculated the confusion matrix of these datasets and extracted the overall accuracy (OA), the mean F_1-score, and the mean intersection over union (MIoU) [30] of each class in order to assess the semantic segmentation results. The confusion matrix was obtained by comparing the segmentation result of the predicted output with the labeled image. The formula for the calculation is:

$$CM = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

The definitions of the relevant evaluation indicators are shown in the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$OA = \frac{\sum_{i=1}^{n} c_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \tag{9}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

$$MIoU = \frac{1}{n} \sum_{i=1}^{n} \frac{c_{ii}}{\sum_{i=1}^{n} c_{ij} + \sum_{j=1}^{n} c_{ji} + c_{ii}} \tag{11}$$

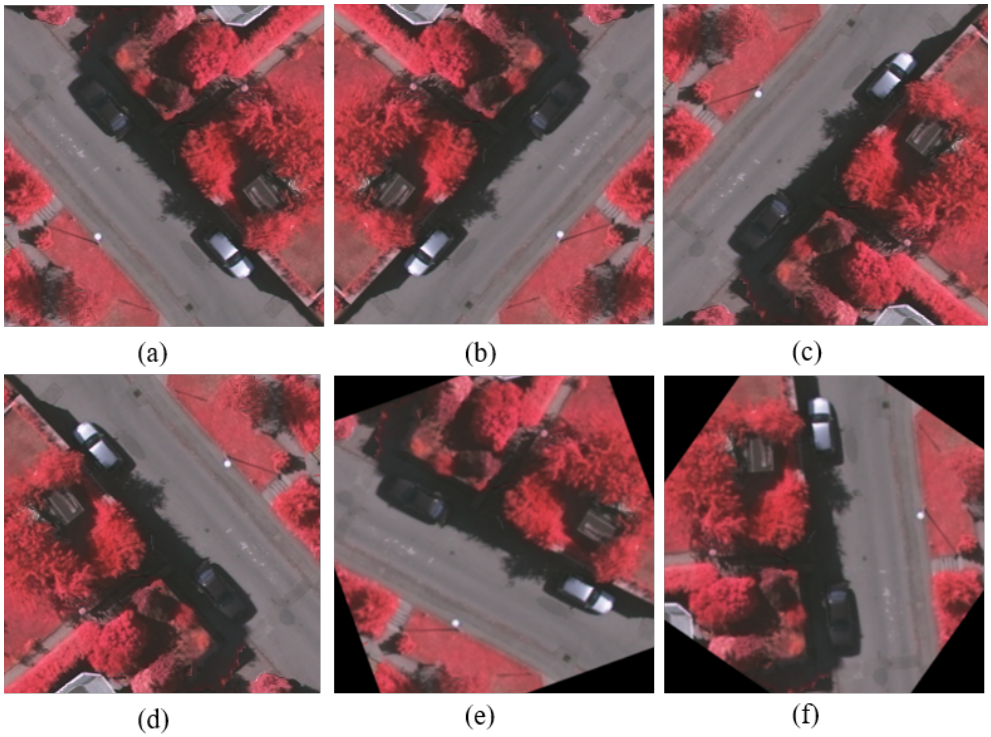

Figure 10. Examples of experimental dataset enhancements. (**a**) is the reference image, and (**b**–**f**) is examples of multiple enhanced operations based on opencv.

4.2. Implementation Details

We used Python 3.6 and the open-source deep learning framework PyTorch for the experiments in this paper. The optimization algorithm we adopted is the random gradient descent algorithm with momentum equal to 0.9 and weight attenuation equal to 0.0005. In order to train the proposed model on both datasets, the number of iterations was fixed at 15,000. Some experimental parameter settings were as follows: the batch size was set to six; the initial learning rate was set to 0.01; the initial learning rate of the encoder in the network was 0.005. The image was randomly scaled during the training process to be between 0.5- and 2-times higher than the original resolution. At the same time, the image was randomly flipped horizontally to increase the robustness of the model. On this basis, the input image was cropped and padded to a 224 × 224 resolution rate to unify the resolution in each batch of training data.

The skewed distribution of ground objects in remote sensing image sample sets leads to class imbalance. Inspired by D. Eigen et al. [31], we introduced a focal loss function to make the model focus on complex and difficult samples. The definition of the loss function is:

$$MFB\_Focal_{loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c \cdot l_c^{(n)}(1-p_c^{(n)})^2 \cdot log(p_c^{(n)}) \quad (12)$$

where N represents the number of samples of a minibatch, C represents the number of categories, w_c represents the weight corresponding to category c, $l_c^{(n)}$ represents the true label corresponding to sample n, and $p_c^{(n)}$ represents the probability of sample n for category c.

5. Experiments And Results

5.1. Result Display

To confirm the efficiency of our approach, we visualize the results of the model segmentation on the ISPRS dataset, as shown in Figures 11 and 12. The segmentation

results of our proposed network model are almost close to the label values, and the interpretation effect is outstanding in the high-resolution urban remote sensing scene with a dense distribution of ground objects. The boundary of the segmentation results of different types of ground objects is smooth and accurate, and the confusion of classification rarely occurred.

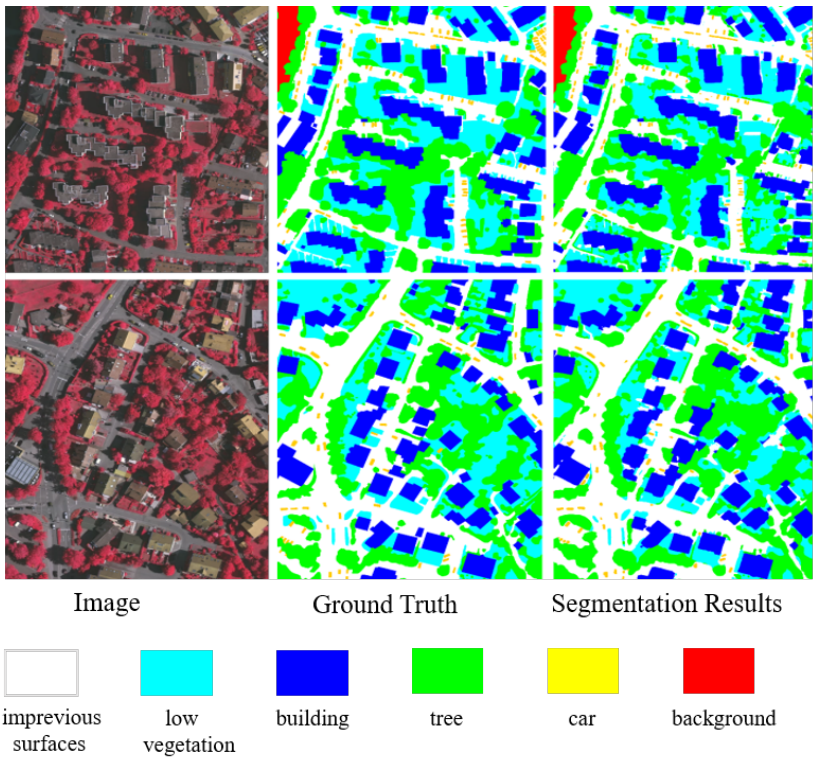

Figure 11. The segmentation results of the MCAFNet on the Vaihingen dataset.

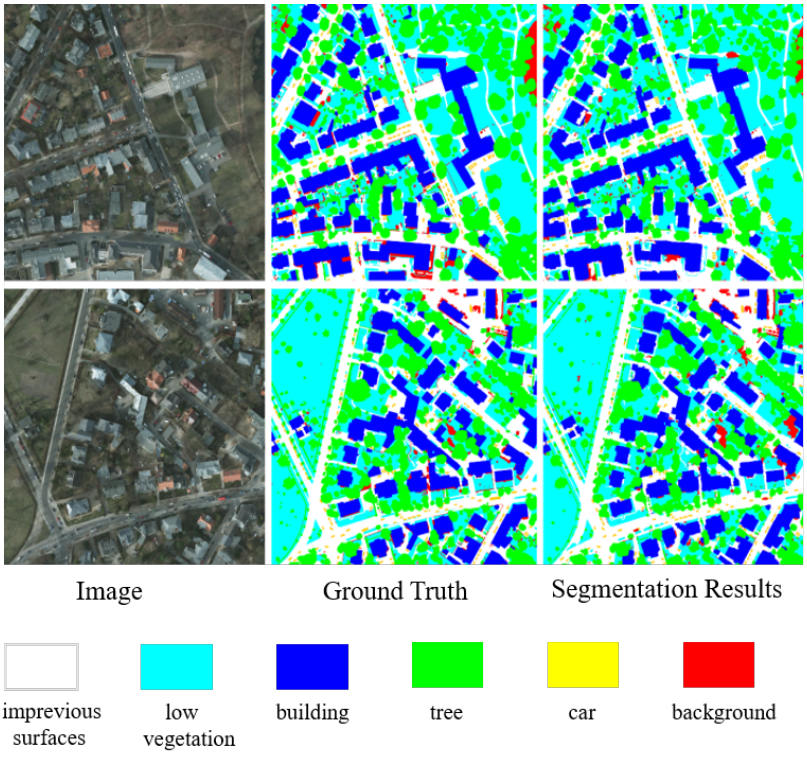

Figure 12. The segmentation results of the MCAFNet on the Potsdam dataset.

5.2. CAM Visualization Analysis

We took the output feature map of the three stages in the MCAFNet encoding part as the optimization goal. According to the change of the resolution of the output feature map, we dynamically adjusted the weight of each channel, so that the CAMD can better optimize the small-scale target. To make the effect more intuitive, we visualize the heat map of the attention mechanism of three stage, as shown in Figure 13. The weight of small target features increases with the deepening of the process, which makes the model more sensitive to small-scale targets.

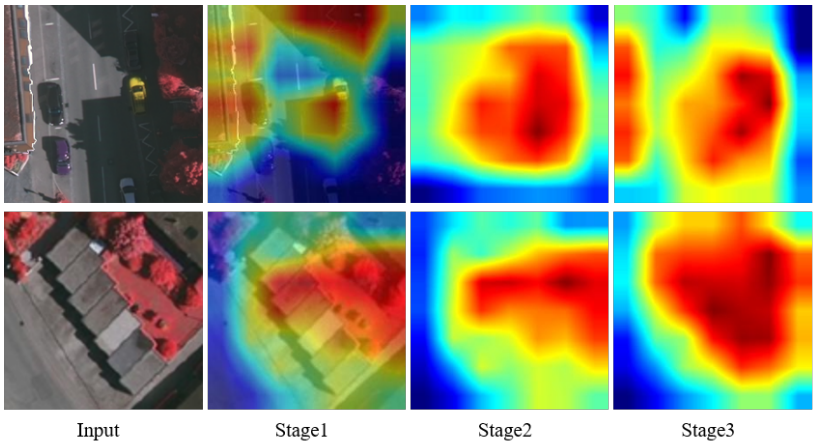

Figure 13. Heat map of small target objects in different stages.

5.3. Architecture Ablation Study

Comprehensive ablation experiments were designed to analyze the efficiency of every part in the MCAFNet. The influence of each module of the network is shown in Table 1, and the improvement effect of the overall and local remote sensing images is shown in Figure 14. It can be seen from Table 1 that different modules of the MCAFNet improved the segmentation performance. However, the performance gain of the transformer and FM is relatively marginal. This is because transformer applications in computer vision are less compatible with urban remote sensing scenes, and the improvement of semantic segmentation accuracy in multi-category scenes is limited. The effect will be obvious if it is combined with the CAMD modules. FM has a simple structure and a small amount of calculation, which can increase the network depth of low-dimensional feature maps, so as to improve the ability of low-dimensional feature maps to capture small-scale targets. It is mainly for the optimized high- and low-dimensional feature maps, which are placed after the CAMD. If a feature map does not go through the channel attention optimization branch, even after FM optimization, this limits the accuracy of the segmentation due to the sufficient depth of the baseline network.

Table 1. The influence of each module on the MCAFNet's performance.

	Transformer	CAMD	FM	Mean F_1 (%)
MCAFNet	×	×	×	81.24
	√	×	×	83.25
	√	√	×	85.46
	√	×	√	83.78
	√	√	√	88.41

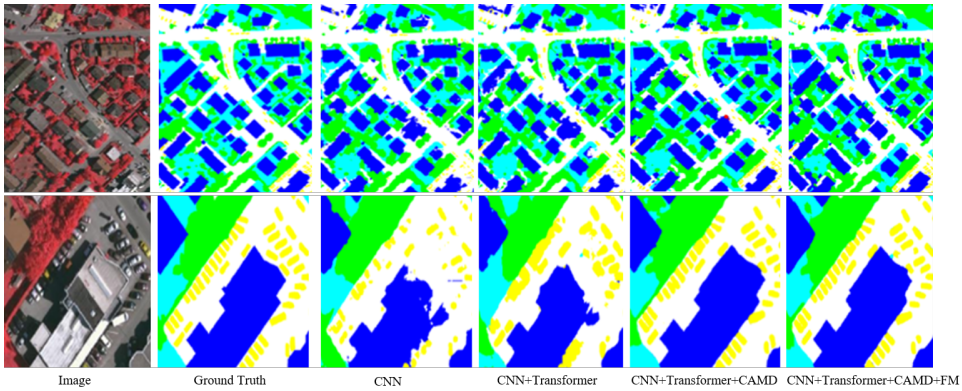

Figure 14. The improvement effect of the overall and local remote sensing images.

(1) Baseline network:

In order to evaluate the performance improvement brought by the architecture in the CNN transformer hybrid as the encoder, we carried out visual analysis on the attention map of different ground object categories for the models before and after the transformer was removed, as shown in Figure 15. The upper row represents the attention map of the figure category of the CNN transformer hybrid model, and the lower row represents the attention map of the category after the transformer is removed. Through comparison, it can be clearly seen that the CNN transformer structure plays an important role in distinguishing the semantic features of different kinds of ground objects when interpreting ground objects. In the process of feature reconstruction, more attention is paid to the pixels of the same category. After the transformer is removed, the model is relatively seriously interfered with by the features of other categories when interpreting a single category of ground objects, which effectively proves that the CNN transformer structure can extract global–local context feature information and improve the segmentation accuracy of multiple types of ground objects.

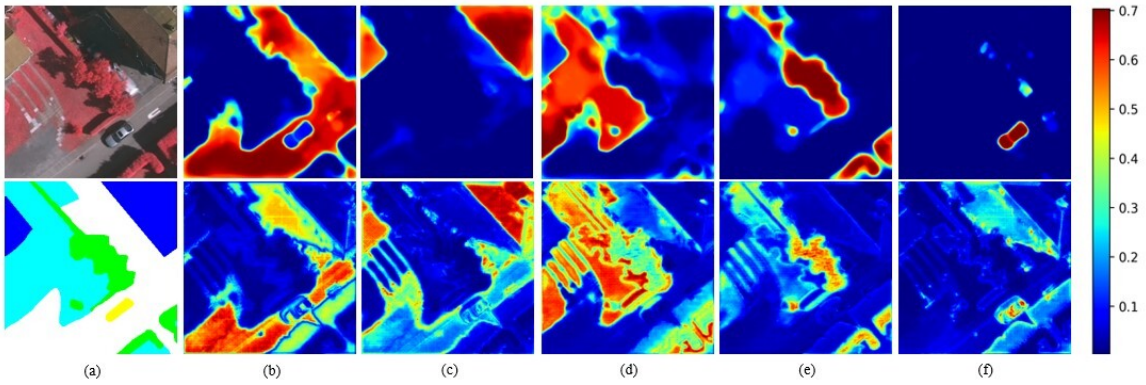

Figure 15. The effect of baseline network on the remote sensing semantic segmentation results. (**a**) is the input image and ground truth, and (**b**–**f**) are different types of surface feature attention maps.

(2) Channel attention optimization module:

In the ablation experiment, the channel attention decoder was removed, and only the front-end convolutional layers were kept. At this time, the mean F_1-score of the network dropped from 88.41 to 83.78, indicating that the channel attention decoder module plays a very large role in remote sensing scene segmentation. Figure 16 shows the effect of the CAMD on the remote sensing semantic segmentation results.

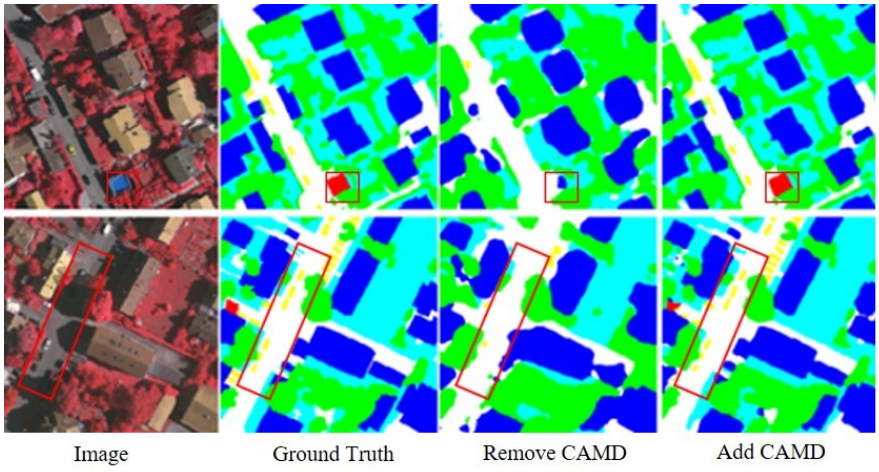

Figure 16. The effect of the CAMD on the remote sensing semantic segmentation results.

(3) Fusion module:

To test the gains in performance brought by the fusion module to the network, we removed it from the network and tested the change in network performance. After removing it, the mean F_1-score of the network decreased from 88.41 to 85.46, which fully proved the importance of the fusion module. From the visualization results shown in the figure, the fusion module increased the network depth of the low-dimensional feature map, which improved the ability of the low-dimensional feature map to capture small-scale targets, and it can efficiently optimize the segmentation results of small-scale targets. Figure 17 shows the effect of the FM on the remote sensing semantic segmentation results.

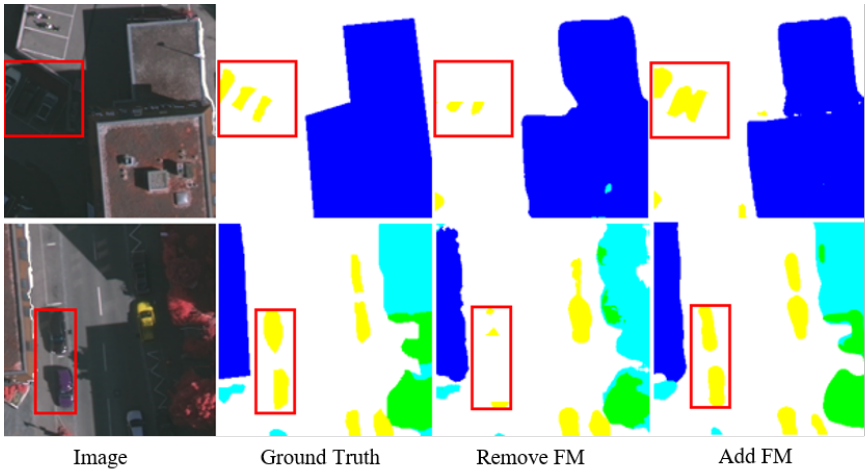

Figure 17. The effect of the FM on the remote sensing semantic segmentation results.

(4) Advanced contrast:

Compared to the performance of U-Net, SegNet, PSPNet, HRNetV2 [32], DeepLab V3+ [33], TransUnet [34], SegFormer, Inception-ResNetV2 [35], and Swin Transformer, the performance of our MCAFNet model in urban remote sensing image interpretation was significantly improved. In Table 2, we provide four test-set-based assessment indices for various models. Compared with the mainstream semantic segmentation model, our model achieved greater improvement in various indicators.

Table 2. The metrics (%) of the semantic segmentation models in the testing phase.

Method	Overall Accuracy	Recall	Mean F_1	MIoU
U-Net	87.5	83.6	82.7	81.2
SegNet	89.4	86.9	86.7	83.6
PSPNet	89.7	87.1	86.9	84.6
HRNetV2	87.2	84.1	83.2	81.4
DeepLab V3+	89.8	87.0	86.7	85.2
TransUnet	90.1	87.2	87.3	86.2
SegFormer	89.5	86.8	87.1	85.9
Inception-ResNetV2	88.1	86.5	86.4	85.5
Swin Transformer	90.2	87.3	87.9	87.3
MCAFNet	**90.8**	**87.9**	**88.4**	**88.2**

To prove the superiority of the method in remote sensing scenes, based on the same experimental conditions, comparison experiments were carried out from local small-scale object segmentation and whole remote sensing scene interpretation. Some visualization results on Vaihingen and Potsdam are shown in Figures 18 and 19. The mean F_1-score of the different models is shown in Tables 3 and 4.

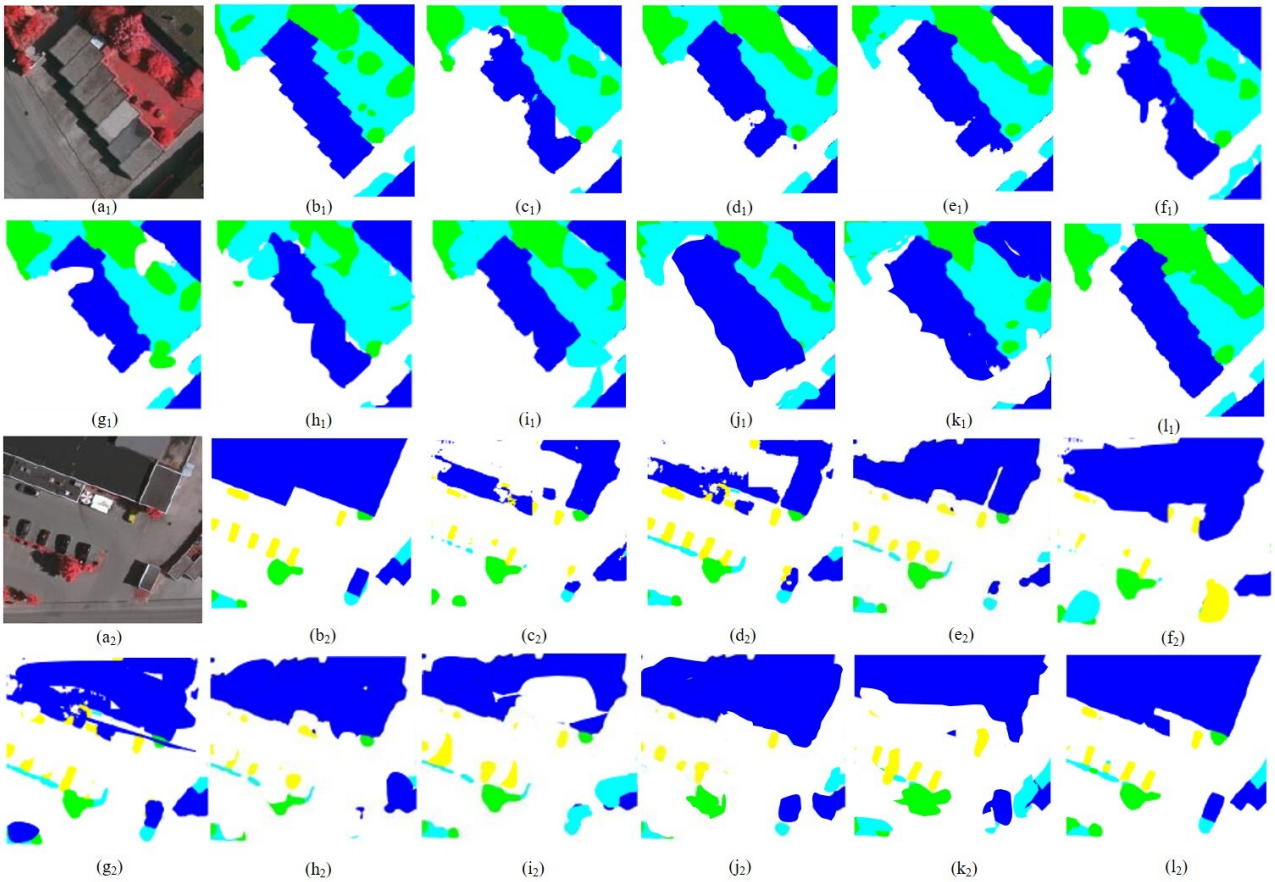

Figure 18. Semantic segmentation results on Vaihingen. (**a1**,**a2**) Original image, (**b1**,**b2**) ground truth, (**c1**,**c2**) U-Net, (**d1**,**d2**) SegNet, (**e1**,**e2**) PSPNet, (**f1**,**f2**) HRNetV2, (**g1**,**g2**) DeepLab V3+, (**h1**,**h2**) TransUnet, (**i1**,**i2**) SegFormer, (**j1**,**j2**) Inception-ResNetV2, (**k1**,**k2**) Swin Transformer, and (**l1**,**l2**) MCAFNet.

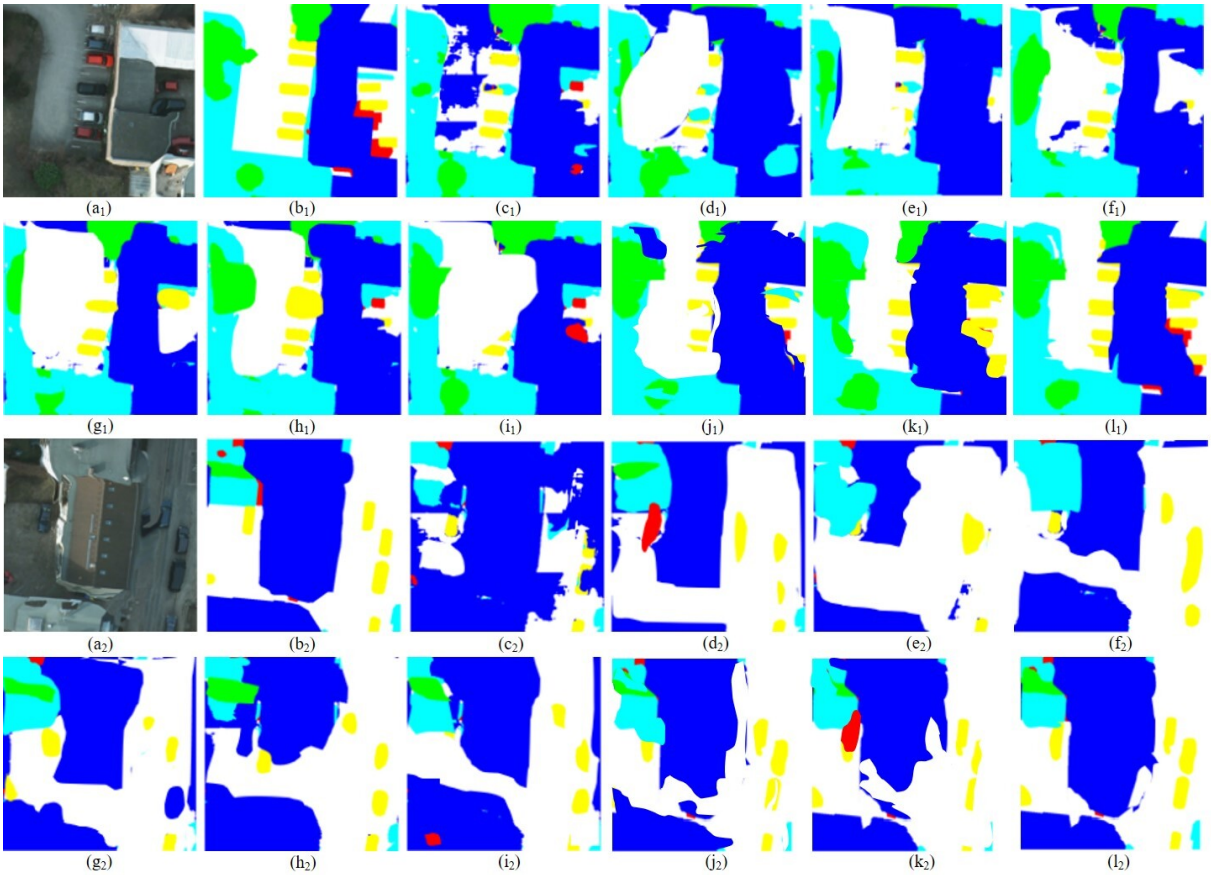

Figure 19. Semantic segmentation results on Potsdam. (**a1**,**a2**) Original image, (**b1**,**b2**) ground truth, (**c1**,**c2**) U-Net, (**d1**,**d2**) SegNet, (**e1**,**e2**) PSPNet, (**f1**,**f2**) HRNetV2, (**g1**,**g2**) DeepLab V3+, (**h1**,**h2**) TransUnet, (**i1**,**i2**) SegFormer, (**j1**,**j2**) Inception-ResNetV2, (**k1**,**k2**) Swin Transformer, and (**l1**,**l2**) MCAFNet.

The performance of these mainstream semantic segmentation networks may be related to their own structure. There are many convolutional layers and pooling layers with steps in the networks, but the convolution lacks an overall understanding of the image itself; moreover, it cannot model feature dependence and does not dynamically adapt to changes in the input. The networks performed poorly in capturing the semantic information of different scales and processing the spatial output resolution of the network. However, we applied a hybrid model network architecture that can reduce the influence of the convolutional operation. It also better integrates semantic information and spatial information through the attention module to optimize the segmentation accuracy of urban features.

Table 3. Performance of ground objects interpreted by the different models on the Vaihingen dataset.

Method	#Param	Building	Car	Low_veg	Imp	Tree	GFLOPs
U-Net	118 M	88.2	75.2	80.2	86.9	85.4	135.4
SegNet	104 M	90.1	84.2	81.7	90.5	86.8	82.9
PSPNet	121 M	91.2	85.7	82.9	91.1	86.4	20.5
HRNetV2	40 M	90.8	76.8	80.4	87.5	86.5	51.5
DeepLab V3+	223 M	91.7	81.4	82.1	88.9	87.6	72.3
TransUnet	257 M	92.3	85.1	83.2	89.7	87.1	112.4
SegFormer	246 M	91.1	81.3	81.5	86.9	86.9	88.7
Inception-ResNetV2	153 M	90.7	84.9	82.5	89.1	86.7	98.5
Swin Transformer	238 M	92.4	85.5	84.1	91.3	87.2	131.4
MCAFNet	334 M	**93.6**	**86.4**	**84.9**	**92.6**	**88.1**	**164.2**

Table 4. Performance of ground objects interpreted by the different models on the Potsdam dataset.

Method	#Param	Building	Car	Low_veg	Imp	Tree	GFLOPs
U-Net	114M	87.2	76.2	80.8	86.2	84.6	123.5
SegNet	97M	89.4	83.6	82.3	90.1	85.9	80.5
PSPNet	114M	90.3	84.7	81.9	89.6	85.4	16.1
HRNetV2	38M	91.2	77.8	80.1	86.9	85.6	43.8
DeepLab V3+	207M	91.4	80.9	81.8	88.2	87.2	62.7
TransUnet	231M	91.8	84.6	82.8	89.1	86.7	98.7
SegFormer	220M	90.7	81.1	81.1	86.4	86.3	81.2
Inception-ResNetV2	141M	90.1	83.9	80.9	87.7	85.5	87.3
Swin Transformer	217M	91.6	85.1	83.1	90.4	87.4	108.7
MCAFNet	320M	**92.4**	**86.1**	**83.9**	**91.3**	**88.3**	**153.3**

6. Discussion

The MCAFNet model we proposed effectively reduces the probability of the misclassification of ground objects in the interpretation of urban remote sensing images and improves the accuracy by integrating low-level semantic features, such as the shape and boundary of ground objects and the high-level semantic information of ground object categories. Two factors ensure the superiority of the model. First, the MCAFNet model realizes the structural innovation in the encoder part and fully combines the advantages of the CNN and transformer when processing semantic segmentation tasks. Second, the proposed network adopts a pooling fusion module in the decoder section. This elaborate design alleviates the information gap and improves the utilization of low-dimensional feature maps. However, the parameters of our model are relatively large, and how to better simplify the transformer structure and combine the advantages of CNNs requires further exploration.

7. Conclusions

We proposed the MCAFNet to realize fast and high-precision semantic segmentation of remote sensing images. The designed network structure successfully integrates the advantages of the transformer and CNNs. Furthermore, we used a channel attention decoder to emphasize the key areas, especially the small-scale target semantic information. The research of our method realizes the robustness and generalization of the model. However, in exchange for high accuracy, the proposed model relies on a large amount of computation, and the CAMD is mainly aimed at the feature extraction of small-scale objects. Its performance is not good when dealing with large-scale target objects. At the same time, there are far more than three bands available in the remote sensing task of image segmentation, and the DSM in the remote sensing image can be used for auxiliary segmentation. This information was not used in this paper. Therefore, future research needs to consider how to reduce the amount of computation of the encoding part, further improve the proposed CAMD to focus on multi-category and large-scale feature extraction, and utilize multiple band information of urban remote sensing images.

Author Contributions: Methodology and writing—original draft preparation, D.R.; writing—review and editing and funding acquisition, M.Y.; resources, Q.F.; data curation, Y.D. and X.W.; supervision, Z.W.; project administration, F.L.; funding acquisition, M.Y. All authors have read and agreed to the published version of this manuscript.

Funding: This work was funded by the Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2021-ct09, the Science and Technology support program of Gansu Province of China under Grant No. 21JR7RA457, the Science and Technology innovation Project of Forestry and Grassland Bureau of Gansu Province (kjcx2022010) and the National Natural Science Foundation of China under Grant No. 62176108.

Data Availability Statement: All the data needed for experiments described in our paper are available at https://www.isprs.org (accessed on 10 April 2022).

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

References

1. Tapasvi, B.; Udaya Kumar, N.; Gnanamanoharan, E. A Survey on Semantic Segmentation using Deep Learning Techniques. *Int. J. Eng. Res. Technol.* **2021**, *9*, 50–56.
2. 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*, 5998–6008.
3. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
4. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
5. Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.C. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5463–5474.
6. He, L.; Zhou, Q.; Li, X.; Niu, L.; Cheng, G.; Li, X.; Liu, W.; Tong, Y.; Ma, L.; Zhang, L. End-to-end video object detection with spatial-temporal transformers. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 1507–1516.
7. 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 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
8. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12873–12883.
9. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 11976–11986.
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
11. 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.
12. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
13. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
14. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 472–480.
15. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
16. Borjigin, S.; Sahoo, P.K. Color image segmentation based on multi-level Tsallis–Havrda–Charvát entropy and 2D histogram using PSO algorithms. *Pattern Recognit.* **2019**, *92*, 107–118. [CrossRef]
17. Wu, Z.; Shen, C.; Hengel, A.V.d. Real-time semantic image segmentation via spatial sparsity. *arXiv* **2017**, arXiv:1712.00213.
18. Xu, Q.; Ma, Y.; Wu, J.; Long, C. Faster BiSeNet: A Faster Bilateral Segmentation Network for Real-time Semantic Segmentation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
19. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
20. 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.
21. 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; pp. 234–241.
22. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder–decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
23. 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.
24. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.

26. Wang, L.; Fang, S.; Zhang, C.; Li, R.; Duan, C. Efficient Hybrid Transformer: Learning Global-local Context for Urban Scene Segmentation. *arXiv* **2021**, arXiv:2109.08937.
27. Peng, C.; Tian, T.; Chen, C.; Guo, X.; Ma, J. Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. *Neural Netw.* **2021**, *137*, 188–199. [CrossRef] [PubMed]
28. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
29. Peng, C.; Zhang, K.; Ma, Y.; Ma, J. Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5601313. [CrossRef]
30. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building extraction in very high resolution imagery by dense-attention networks. *Remote Sens.* **2018**, *10*, 1768. [CrossRef]
31. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multiscale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
32. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
33. 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] [PubMed]
34. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
35. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017.