

Article

# MFTransNet: A Multi-Modal Fusion with CNN-Transformer Network for Semantic Segmentation of HSR Remote Sensing Images

Shumeng He , Houqun Yang \* , Xiaoying Zhang and Xuanyu Li

College of Computer Science and Technology, Hainan University, Haikou 570228, China

\* Correspondence: yhq@hainanu.edu.cn

**Abstract:** Due to the inherent inter-class similarity and class imbalance of remote sensing images, it is difficult to obtain effective results in single-source semantic segmentation. We consider applying multi-modal data to the task of the semantic segmentation of HSR (high spatial resolution) remote sensing images, and obtain richer semantic information by data fusion to improve the accuracy and efficiency of segmentation. However, it is still a great challenge to discover how to achieve efficient and useful information complementarity based on multi-modal remote sensing image semantic segmentation, so we have to seriously examine the numerous models. Transformer has made remarkable progress in decreasing model complexity and improving scalability and training efficiency in computer vision tasks. Therefore, we introduce Transformer into multi-modal semantic segmentation. In order to cope with the issue that the Transformer model requires a large amount of computing resources, we propose a model, MFTransNet, which combines a CNN (convolutional neural network) and Transformer to realize a lightweight multi-modal semantic segmentation structure. To do this, a small convolutional network is first used for performing preliminary feature extraction. Subsequently, these features are sent to the multi-head feature fusion module to achieve adaptive feature fusion. Finally, the features of different scales are integrated together through a multi-scale decoder. The experimental results demonstrate that MFTransNet achieves the best balance among segmentation accuracy, memory-usage efficiency and inference speed.



**Citation:** He, S.; Yang, H.; Zhang, X.; Li, X. MFTransNet: A Multi-Modal Fusion with CNN-Transformer Network for Semantic Segmentation of HSR Remote Sensing Images. *Mathematics* **2023**, *11*, 722. <https://doi.org/10.3390/math11030722>

Academic Editors: Xiangtao Zheng, Jinchang Ren and Ling Wang

Received: 31 December 2022

Revised: 25 January 2023

Accepted: 28 January 2023

Published: 1 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

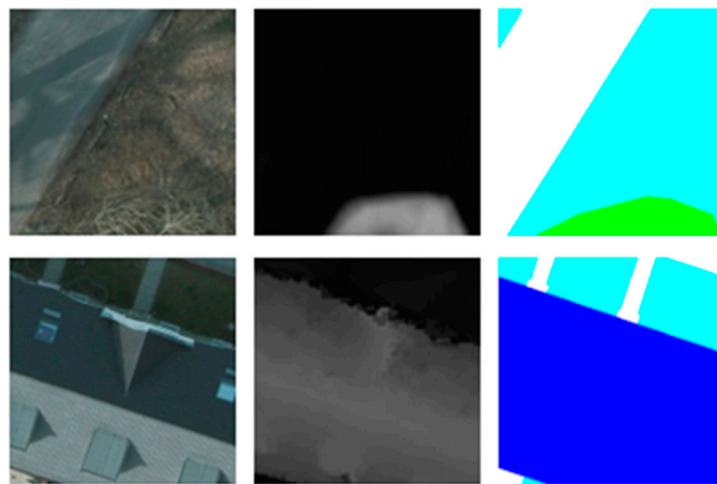
**Keywords:** semantic segmentation; high spatial resolution remote sensing images; transformer; multi-modal

**MSC:** 68T45

## 1. Introduction

HSR remote sensing images' semantic segmentation involves the application of image semantic segmentation in the field of remote sensing, which aims at extracting different types of ground objects from remote sensing images. Remote sensing image semantic segmentation technology is widely used in unmanned driving, geological detection, urban planning [1–5] and other fields [6–11], and its great significance in modern production and life is self-evident. HSR remote sensing images have rich semantic information, but the task of semantic segmentation is challenging, due to the large intra-class variance and class imbalance. For example, the difference among ground objects in remote sensing images is too small, and the sense of a boundary between different surface features is not strong. The shadow caused by illumination influence can be occluded; there is also a tremendous difference in the number of different ground objects. In Figure 1, the representation of the low vegetation (shown as cyan in the labeled image) and the tree (shown as green in the labeled image) in the first row of RGB images is so similar that it is difficult to identify the difference between the two. However it is clear in the DSM (digital surface model) image that the area where the tree is located is brighter (higher in height). In the second row,

the illumination led the house (shown as dark blue in the labeled image) to have different colors in the direction of sunrise and non-sunrise, with the non-sunrise roofs being darker. However, in the DSM image the two reversals show the same features due to being the same height. From the labeled image, we can determine that there is also a category of grass. The shading makes its representation extremely similar to that of the roofs in the non-sunrise direction, and the boundary between the two categories of surface features is very blurred.



**Figure 1.** RGB image on the left, DSM image in the middle, the label image on the right.

Since semantic segmentation is usually carried out in the feature space, the representation and learning of features is the key to realizing semantic segmentation. According to the way of expressing and learning features, the methods of semantic segmentation of remote sensing images have undergone several stages of development and fusion. The classification method based on manual feature description provides a solution for the semantic segmentation of remote sensing images in the early stage, and subsequently machine learning based on probability statistics further improves the segmentation accuracy. The emergence of CNN removes the necessity to rely on feature design completely, and makes a qualitative leap in segmentation accuracy. However, the difficulty of model visualization and the lack of datasets restricts the further improvement of segmentation accuracy. Thus, the complexity of the network has a significant impact on the segmentation effect. The deeper, wider and richer structure of the network can obtain more abstract feature representations in the data and improve the accuracy of segmentation. Nevertheless, an overly complex network model will both increase the training cost and reduce the training efficiency, and may also reduce the generalization ability of the network. How to effectively diminish the complexity of the network while ensuring the segmentation effect is one of the directions in which we are working.

In addition to the selection of methods as well as network models, we find that unimodal remote sensing image data can merely provide information attributes from a single perspective by analyzing the effects of HSR remote sensing image data forms and data structure on semantic segmentation, and it is difficult to obtain better feature extraction results. It is feasible to introduce a multi-modal remote sensing image to find semantic segmentation methods with high efficiency, accuracy and higher robustness. The fusion of multi-modal data expands the data source in regard to information dimensions and the number of samples to supplement the data requirements of the algorithm for network model change detection, and it can provide us with target image features and information from various aspects. The fusion of different features and information retains the effective discriminative information of the multiple features involved in the fusion, while avoiding the uncertainty of single data to a certain extent. It also makes the results of the semantic segmentation of remote sensing images more comprehensive and accurate.

With the development of satellite remote sensing technology and the application of artificial intelligence technology, the scale of remote sensing image datasets is now becoming larger and larger, and multi-modal remote sensing data are becoming more and more abundant. For remote sensing images, common multi-modal data include DSM images [12], NIR (near infrared ray) images [13] and SR (synthetic aperture radar) images [14]. Images of different modalities possess different characteristics and can provide richer semantic information. The DSM image realistically represents the high and low surface levels, and the areas that are closer to white represent higher heights. For the two misclassification cases generated in Figure 1, the trees and buildings show clear boundaries in the DSM images, since they are both tall objects. Using multi-modality information can ease the misclassification problem due to similar appearances and improve the segmentation efficiency of the model.

After further discussion, we determine that our goal of semantic segmentation based on multi-modal remote sensing images is to utilize complementary features from different modalities to maximize classification accuracy, while reducing the influence of the inherent noise in unimodal data and improving the generalization performance in complex application scenarios. The problem, which is the effect of noise and redundant features from different modalities, is emphasized. The direct fusion of images is likely to cause noise pollution, which will instead reduce the segmentation efficiency of the model. To cope with this problem, we creatively propose a multi-modal data fusion Transformer method drawing on the application of Vision Transformer in computer vision. The features is divided into four scales by referring to the backbone structure of Segformer [15] and modifying the encoder-decoder to apply it to the data fusion of RGB and DSM instead of pure feature extraction. Therefore, we establish the multi-modal fusion Transformer Network (MFTransNet). To evaluate the effectiveness of our model, we conducted several experiments, whose experimental results demonstrate that our model outperforms the state-of-the-art methods on the publicly available HSR image dataset Potsdam.

The research contributions of this article are as follows.

- A multi-modal semantic segmentation model MFTransNet, which combines CNN and Transformer, is proposed, and achieves a balance of accuracy and speed.
- A feature fusion module containing a multi-head attention mechanism, a feature adaptive calibration module (FACM) for adaptively calibrating features, and a complementary fusion module (CFM) for fusing multi-modal features, are proposed to achieve the efficient adaptive fusion of features.
- A multi-scale decoder is used for the multi-scale problem in remote sensing images, and DUpSampling [16] is used instead of the traditional bilinear upsampling to reduce the details lost in the upsampling process and to achieve feature aggregation at different scales.
- While ensuring task completion, a model with a lower number of parameters and more streamlined structure is obtained through channel compression and an optimized structure. Meanwhile, we skillfully use channel shuffle and PixelShuffle on feature extraction and decoder. The optimized model requires fewer computational resources and can meet a wider range of application requirements than the original model. The proposed model achieves a SOTA effect on the Potsdam dataset.

The article is organized as follows, with related work discussed and analyzed in Section 2. Section 3 describes our proposed model and discusses it. Section 4 conducts experiments on the publicly available Potsdam dataset and analyzes the results and possible improvements. Section 5 concludes and gives an outlook on our work.

## 2. Related Work

### 2.1. Semantic Segmentation of Remote Sensing Images

Image semantic segmentation is a type of image classification which goes with the intensive classification task targeting the pixel level. An approach based on full convolutional networks (FCN) [17] proposes an end-to-end processing method while using a

skip layer to combine the location information of the shallow layer with the deep semantic information to fill in the missing data detail. The DeepLab [18–21] series employs dilated convolution and pyramidal pooling, etc., to obtain multi-scale information while increasing the perceptual field and obtaining more robust segmentation. U-Net [22], SegNet [23], and DeconvNet [24] employ the structure of decoder-encoder, which effectively improves the boundary segmentation. Researchers have considered the inherent characteristics of HSR remote sensing images—for example, the multi-scale problem making it very difficult to locate and identify ground objects, misclassification caused by large intra-class variance, and the imbalance of foreground and background caused by too small foreground proportion. FarSeg [25] proposes a foreground-aware relational network that balances the hard samples in the foreground and background during the optimization training process. MCFNet [26] proposes a multiple context fusion network which combines both global and local information to achieve high-resolution semantic segmentation of remote sensing images. HMANet [27] proposes a category-based attention module to enhance the distinction between categories, and the non-local module is improved with sparse representation to efficiently capture regional dependencies. FactSeg [28] proposes a two-branch decoder in which the FA branch is designed to activate features of a small object and suppress large-scale context, and the semantic refinement (SR) branch aims at further differentiating small objects and enhancing the accuracy of small object semantic segmentation. DC-Swin [29] introduces the Swin Transformer as a backbone for extracting contextual information, and designed a novel densely connected feature aggregation module (DCFAM) decoder to restore resolution and generate segmentation maps.

## 2.2. Transformer

Transformer is first applied in natural semantic processing [30], which completely discards network structures such as RNN and CNN, and achieves effective results only using the attention mechanism for machine translation tasks. ViT [31] tries to apply Transformer to the field of computer vision. When a large amount of data is pre-trained and migrated to multiple small- and medium-sized image recognition benchmarks, the results show that ViT can achieve better results compared with SOTA's CNN and requires fewer training resources. Since then, Transformer has been widely used in computer vision. Setr [32] and Segformer further improve the structure of ViT to make it more usable for semantic segmentation tasks. From a theoretical point of view, Transformer can achieve better model performance compared to CNN, but the global attention mechanism imposes a tremendous computational cost, especially in shallow networks. Therefore, some of the approaches currently proposed combine Transformer with CNN to complement each other's competitiveness and achieve a model that balances accuracy and speed. BotNet [33] forms a new network structure using multi-head self-attention (MHSA) instead of the  $3 \times 3$  convolution in the ResNet bottleneck, which improves the accuracy of various classification tasks. CMT [34] designs a stage-wise transformer based on a hierarchical structure, introducing convolutional operations for fine-grained feature extraction, as well as a unique modular hierarchy for extracting local and global features. Conformer [35] proposes a dual network model structure, which consists of a CNN branch and a Transformer branch, and combines CNN-based local features with Transformer-based global representation, in order to enhance representation learning. The network structure of TransUnet [36] is modeled after Unet, with a U-shaped structure consisting of encoder and decoder. A Transformer mechanism is added to the encoder part, which gives it the advantages of both CNN and Transformer. These hybrid models improve segmentation accuracy of the model by combining various efficient structures in CNN and Transformer. Remote sensing images are generally small- or medium-sized datasets, and it is difficult to achieve good training results using Transformer as the model. There is evidence that the vision transformer needs a very large dataset to surpass CNN. Therefore, it will have a better training effect in the field of remote sensing using the CNN-Transformer hybrid structure.

### 2.3. Multi-Modal Data Fusion

In addition to model improvements, it is considered that multi-modal semantic segmentation methods can achieve more accurate segmentation by extracting information from source images of different modalities, resulting in a richer feature representation than from a single original image. SA-Gate [37] proposes a unified and efficient cross-modal bootstrap encoder that not only efficiently recalibrates RGB feature responses, but also allows for multiple stages to extract accurate depth information and alternatively aggregate the two recalibrated representations. AMFuse [38] proposes an additive-multiplicative fusion network with additive operations focusing on extracting cross-modal complementary features and multiplicative operations focusing on extracting cross-modal common features. MSDFM [39] proposes a multi-sensor data fusion model. For the first time, the single-channel DSM data is converted into a three-channel color DSM, and the color DSM is used as a supplementary input for further detailed feature extraction. At the decoder stage, data-dependent upsampling (double upsampling) method is employed instead of the common upsampling method to improve the classification accuracy of small object's pixels. TransFuser [40] proposes a novel multi-modal fusion Transformer that uses attention to integrate RGB images and LiDAR representations. C3Net [41] proposes a multi-modal semantic segmentation network that uses a cross-modal feature recalibration module to learn multi-modal features while reducing the effect of noise inherent in different modalities. A model distillation strategy is used for obtaining an accurate and compact dense prediction network. MCENet [42] proposes an end-to-end multi-source remote sensing image semantic segmentation network, and designed a co-enhanced fusion module to mine the complementary features of multi-source remote sensing images and address the intra-class variance issue. CMX [43] further improves on SA-Gate's network using Segformer as the backbone to extract features, and designed a cross modal feature correction module to perform modal calibration in both spatial and channel dimensions. HyperTransformer [44] introduces the Transformer structure into panchromatic sharpening, which is the fusion of aligned high-resolution panchromatic images (PAN) with low-resolution hyperspectral images (LR-his) to produce multi-band images with higher spatial resolution. These studies provide a novel idea for multi-modal fusion: the use of the Transformer for feature fusion modules rather than feature extraction.

We give a brief summary of the above mentioned references in Table 1. Although greater progress has been made in improving multi-modal feature fusion accuracy, the introduction of multi-modal data greatly increases the runtime for dense prediction at the pixel level. Two parallel networks are required in the backbone to perform feature extraction on images of different modalities, so using Transformer will drive the computational effort up further, which is an important factor for why we consider using Transformer for feature fusion rather than feature extraction. Therefore, achieving a balance between segmentation accuracy and segmentation speed is another focus of current research. Taking into account the situation mentioned above, we propose a multi-modal data semantic segmentation model based on the combination of CNN and Transformer: MFTransNet. The characteristics of both structures are fully utilized to achieve a balance of accuracy and speed.

**Table 1.** Summary of papers mentioned in related work.

Method	Years	Approach	Advantages	Disadvantages
FCN	2014	Use convolutional layers instead of fully connected layers	An input image of any size can be accepted	A lot of storage overhead; Coarse segmentation effect
Deeplab V3+	2018	Dilated convolution and bilinear upsampling are used	Multi-scale feature extraction is achieved	The segmentation effect of high-resolution image is poor
U-Net	2015	The encoder-decoder structure and skip connection are used	Support for a small number of trained models	Redundancy leads to slow training

Table 1. Cont.

Method	Years	Approach	Advantages	Disadvantages
SegNet	2017	Confirm the position after upsampling using pooling indices	Low memory requirements	The accuracy improvement is not high
DeconvNet	2015	Add the deconvolution layer on top of vgg	It is stronger than FCN in segmentation details	Boundary segmentation is not accurate
FarSeg	2020	Explicit foreground modeling approach	Mitigating foreground and background imbalances	-
MCFNet	2022	Confidence-based local selection criterion	Optimal balance between segmentation accuracy, storage efficiency and inference speed	-
HMANet	2020	A collection of three category-based attention modules	Improve discrimination between classes	-
FactSeg	2022	It consists of a two-branch decoder and a joint probabilistic loss	Small object features are activated and large-scale background noise is suppressed	-
DC-Swin	2022	Designed a densely connected feature aggregation module decoder	Enhanced semantic space and channel relations features	-
ViT	2021	The image patches are directly input into the Transformer for feature extraction	The effect is better than SOTA's CNN when trained with a large amount of data	A large amount of data is required for training
Setr	2021	Introducing Transformer from the perspective of semantic segmentation	The receptive field of the model is improved	Large number of parameters and computations
BotNet	2021	Replace the bottleneck in the fourth block in ResNet with the Multi-Head Self-Attention Module	Both efficiency and accuracy are improved	-
CMT	2021	The key and value computation is replaced by the deep convolution computation in the main module	The computational cost is reduced	-
Conformer	2020	It consists of a CNN branch and a Transformer branch	Enhanced global perception of local features and local details of global representations	High model complexity
TransUnet	2021	Resnet and ViT are combined using the U-Net structure	Higher performance than various methods in medical applications	High model complexity
SA-Gate	2020	Propose a unified, efficient cross-modal guidance coder	Information from different models is effectively integrated	-
MCENet	2022	Designed a co-enhanced fusion module	It is more competitive in terms of the number of parameters and inference speed	-
AMFuse	2022	Efficiently combine multiplication and addition operations	Effectiveness in fusing RGB and thermal information	-
MSDFM	2022	Color digital surface model data is used as additional input	The segmentation accuracy of small objects is improved	The segmentation accuracy is not high for low vegetation and trees
TransFuser	2021	Using attention to integrate image and LiDAR representation	Achieving state-of-the-art performance in complex driving scenarios	-
C3Net	2021	Using a cross-modal feature recombination module	A balance between accuracy and speed is achieved	-
CMX	2022	Two backbones are used to extract RGB and other modes, respectively	The generalization performance of outdoor scenes is excellent	Large number of parameters

### 3. Proposed Method

Based on the consideration of achieving a balance between the accuracy and speed of the model, we designed MFTransNet, a multi-modal semantic segmentation model combining CNN and Transformer. In Section 3.1, we first introduce the backbone network of the model. From Section 3.2 to Section 3.4, we further introduce the components of the

network, including the feature extraction module (FEM), the multi-head feature fusion module (MHFFM) and the multi-scale decoder (MSD).

### 3.1. Backbone Network

There is not a large quantity of data for remote sensing images generally, only a few hundred or a few thousand images, and this is not suitable for using a pure Transformer structure. Thus, we designed a hybrid CNN-Transformer model. As shown in Figure 2, the model is divided into four main parts, including feature extraction, feature fusion and feature re-extraction in the encoder stage, and a multi-scale decoder in the decoder stage. The CNN is used as the feature extraction network in the encoder. The multi-headed attention mechanism in the Transformer is added to the feature fusion module to achieve adaptive feature fusion, and the fused module continues to be fed into the CNN to achieve feature re-extraction. After the encoder, the four scales of features are output, fused and upsampled in a multi-scale decoder to recover the details.

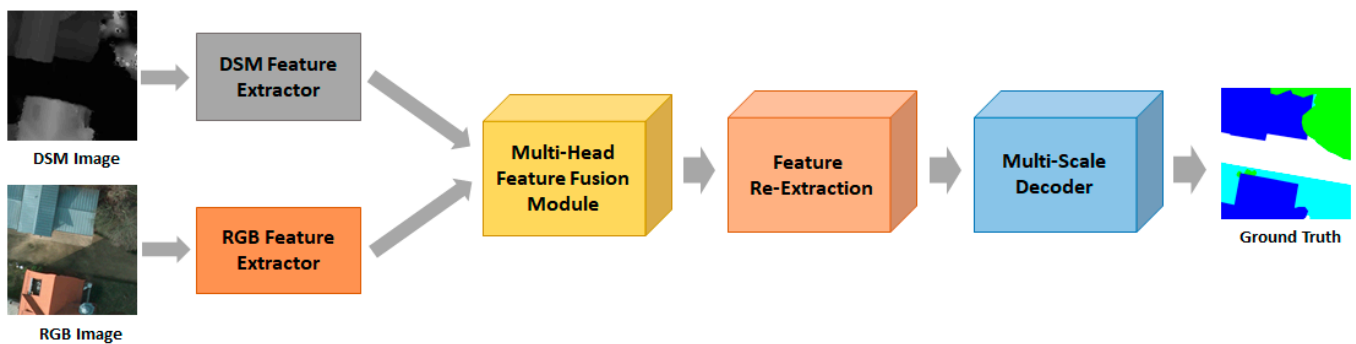


Figure 2. A flowchart of the MFTransNet.

As there can be data inconsistencies and noise between images of different modalities, the images are fed directly into the attention module for fusion, there may be situations where the information between the different modalities is not matched correctly. Therefore we first input an image into the feature extraction network to obtain the features from shallow layer, afterward send it into the feature fusion module for adaptive fusion, and finally send the fused features into the feature re-extraction module to deepen the layers of the network in order to extract richer feature information. At this point, four different scales of features are output from the encoder (as shown in Figure 3); this layering is designed to extract shallow layer features of high resolution and fine features of low resolution. In the multi-scale decoder we use DUpsampling for upsampling to obtain a finer segmentation effect.

Specifically, for the input image  $x \in \mathbb{R}^{H \times W \times C}$ , we first feed  $x$  into the feature extraction module to extract the four scales of features  $F^1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$ ,  $F^2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$ ,  $F^3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$ ,  $F^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$ , They are then fed separately into the multi-head feature fusion module for feature fusion, the fused features go into the residual module for further feature extraction, and finally all the features are incorporated into the coder for multi-scale feature fusion and upsampling for detail recovery.

$$F = Conv_{3 \times 3} \left( Concat \left( Dup \left( F^i \right) \right) \right), \forall i \tag{1}$$

where  $DUp$  is data-dependent upsampling,  $Conv_{3 \times 3}$  is the  $3 \times 3$  convolution, and  $F^i$  is the feature at layer  $i$ .

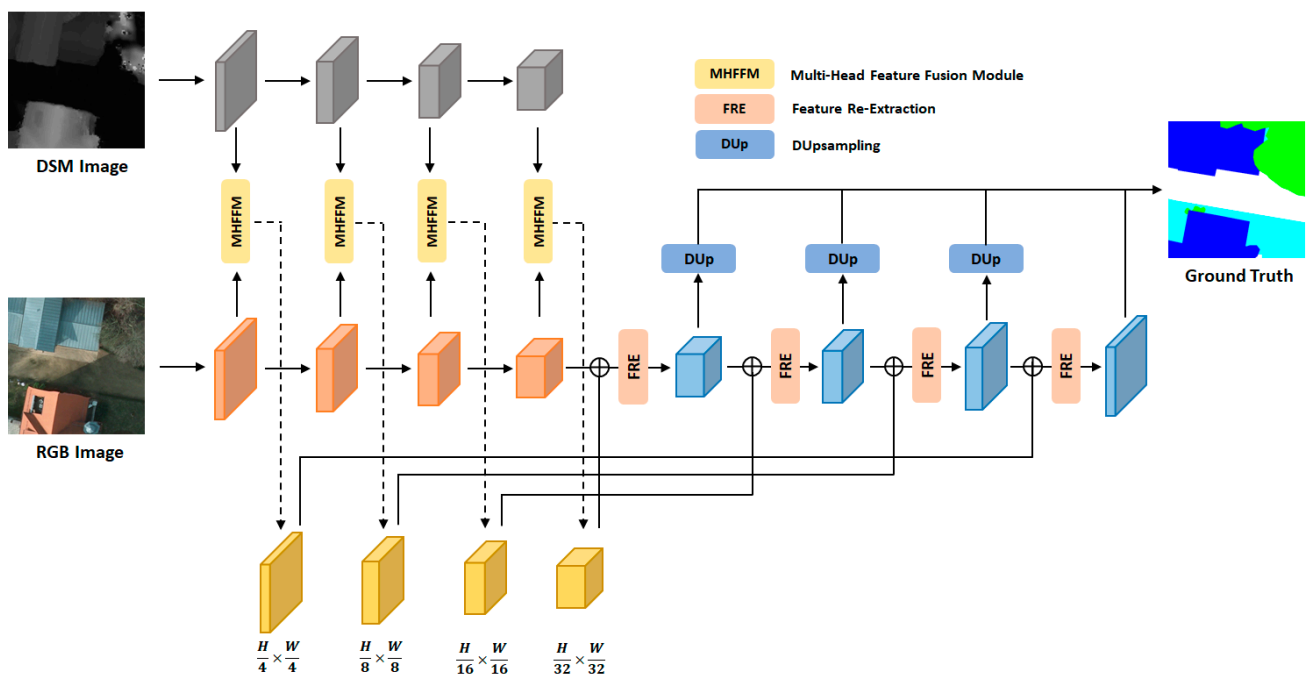


Figure 3. MFTransNet Overall Architecture.

### 3.2. Feature Extraction Module

As shown in Figure 4, we use a combination of lightweight modules in the feature extraction module. This mainly includes the depth-separable convolution module and the C4 module used in mobilenet v2 [45]. As DSM images contain only height features, whereas RGB images contain a variety of features such as color and texture, different feature extraction modules need to be designed for these two types of images. For convolutional neural networks, the more layers there are, the richer the features acquired. Therefore, we choose a deeper network when designing the RGB feature extraction module. At the same time, remote sensing images have the problem of being multi-scale, with tremendous variations in the orientation, shape, and scale of instances in the images [39]. The C4 module decreases the impact of the multi-scale problem by combining convolutions with different expansion rates, similar to a lightweight ASPP.

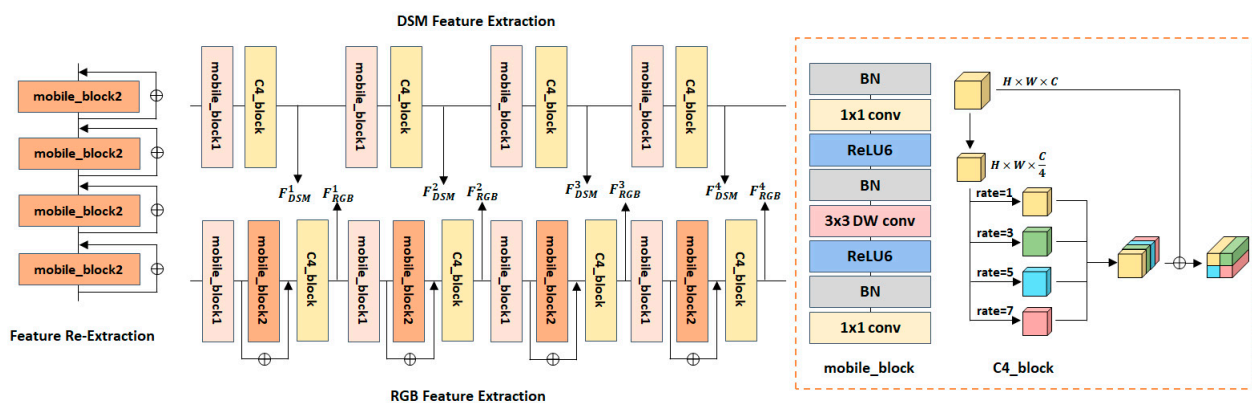


Figure 4. Schematic diagram of the feature extraction module.

The feature extraction network we design consists of three basic modules, where mobile\_block1 upscales the number of channels, while mobile\_block2 keeps the channels unchanged and adds the residual structure C4\_block to it for the extraction of multi-scale features. Specifically, the number of feature input channels is first reduced to one-quarter of the original number, then it is fed into a convolution with different expansion rates, and



finally they are combined. A channel shuffle is added after the contacting to mix pixels between channels to enhance the correlation between channels. Compared to DSM feature extraction, the RGB feature extraction network adds the mobile\_block2 module to deepen the network and learn more features. The feature extraction module designs four scales of output  $F^1 \in \mathbb{R}^{C_1 \times \frac{H}{4} \times \frac{W}{4}}$ ,  $F^2 \in \mathbb{R}^{C_2 \times \frac{H}{8} \times \frac{W}{8}}$ ,  $F^3 \in \mathbb{R}^{C_3 \times \frac{H}{16} \times \frac{W}{16}}$ ,  $F^4 \in \mathbb{R}^{C_4 \times \frac{H}{32} \times \frac{W}{32}}$ .

$$F_{RGB}^i = C4(mo\_2(mo\_1(F_{RGB}^{i-1})) + mo\_1(F_{RGB}^{i-1})) \tag{2}$$

$$F_{DSM}^i = C4(mo\_1(F_{DSM}^{i-1})) \tag{3}$$

where  $F_{RGB}^i$  and  $F_{DSM}^i$  denote the features of the RGB image and DSM image output at each layer, respectively;  $mo\_1$  denotes mobile\_block1,  $mo\_2$  denotes mobile\_block2, and C4 denotes C4\_block.

The feature re-extraction module consists of four mobile\_block2 residual structures. Using the residual module avoids the problem of gradient disappearance and increases the depth of the network.

### 3.3. Multi-Head Feature Fusion Module

As shown in Figure 5, the multi-head feature fusion module is divided into two parts: the feature adaptive calibration module and the complementary fusion module. For complementary and redundant information in different modalities, our aim is to remove redundant information while retaining complementary information, and to enhance the representation of complementary information and attenuate redundant information by adaptively calibrating both features through a multi-headed attention mechanism. After feature calibration, we use a complementary fusion module to calculate the fusion weights of the two modality (the sum of the two weights is 1) for complementary feature fusion.

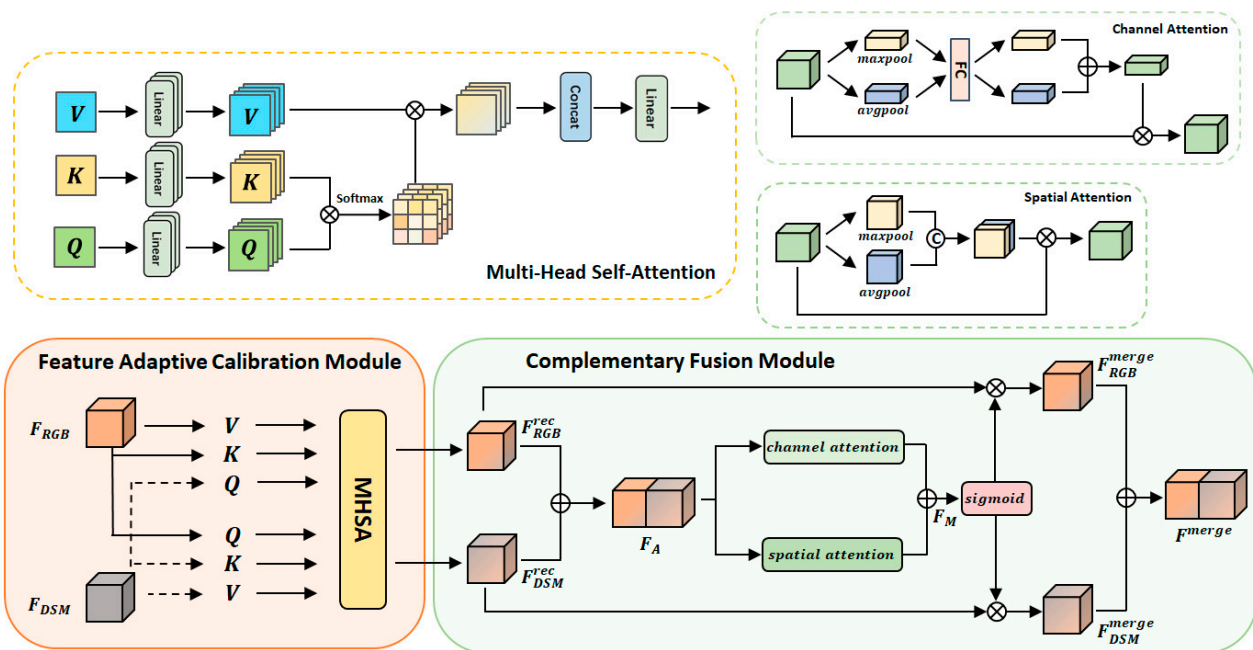


Figure 5. Diagram of the multi-head feature fusion module.

First, for the features' ( $F_{RGB}^i$  and  $F_{DSM}^i$ ) output from the feature network, they are fed into the multi-headed attention mechanism as V, K, and Q for feature calibration. Singly the calibration process of  $F_{RGB}^i$  is introduced here, and the calibration process of  $F_{DSM}^i$  is the same as that of  $F_{RGB}^i$ . Specifically,  $F_{RGB}^i$  as key and value keys and  $F_{DSM}^i$  as query

keys are mapped onto  $n$  subspaces by nonlinear transformation, and then richer feature information is captured on these  $n$  subspaces as follows:

$$v(j, i, :) = k(j, i, :) = f_{linear}^i(F_{RGB'}^i(j, :)) \tag{4}$$

$$q(j, i, :) = f_{linear}^i(F_{DSM'}^i(j, :)) \tag{5}$$

where  $f_{linear}^i$  represents the parameter representation of the  $i$ th linear layer,  $F_{RGB'}^i(j, :)$ ,  $F_{DSM'}^i(j, :)$  represents the one-dimensional representation of the  $j$ th feature map, and  $v(j, i, :)$ ,  $k(j, i, :)$ ,  $q(j, i, :)$  represents the  $i$ -th global descriptor of the  $j$ th feature map of  $v, k, q$ . In the following, to simplify notation, we will use  $v, k$ , and  $q$  to describe the operation of each layer of feature mapping. A similarity weight matrix is first obtained by multiplying, scaling, and normalizing  $k, q$ , and then multiplying it with  $v$  to obtain the calibrated features.

$$W = softmax\left(\frac{k^T \times q}{\sqrt{d_k}}\right) \tag{6}$$

$$f^{rec} = W \times v \tag{7}$$

where  $f^{rec} \in \mathbb{R}^{N \times C \times HW}$  is the calibrated feature and  $\sqrt{d_k}$  is the scaling factor.

Next, the dimensions of  $F^{rec}$  are rearranged to become  $\mathbb{R}^{C \times N \times HW}$  and a linear layer is used to reshape to obtain the final calibration feature  $F^{rec} \in \mathbb{R}^{C \times H \times W}$ :

$$F^{rec} = Linear(Reshape(f^{rec})) \tag{8}$$

After the feature calibration to obtain  $F_{DSM}^{rec}$  and  $F_{RGB}^{rec}$ , we calculate their fusion weights by an attention module that includes attention in both channel and spatial directions, and finally use an activation function that keeps the output weight values in the range  $[0, 1]$  to fuse the features of the two modalities according to different weights.

$$F_M = F_{DSM}^{rec} + F_{RGB}^{rec} \tag{9}$$

$$F_A = CA(F_M) + SA(F_M) \tag{10}$$

$$F_{RGB}^{merge} = \sigma(F_A)F_{RGB}^{rec} \tag{11}$$

$$F_{DSM}^{merge} = (1 - \sigma(F_A))F_{DSM}^{rec} \tag{12}$$

$$F^{merge} = F_{RGB}^{merge} + F_{DSM}^{merge} \tag{13}$$

where  $\sigma$  is the sigmoid function, CA is channel attention, and SA is spatial attention.

### 3.4. Multi-Scale Decoder

The multi-scale issue in remote sensing images makes it difficult to locate and identify objects; at the same time, it is difficult to recover the detailed effect accurately using the common decoder. Accordingly, we chose to construct a multi-scale decoder. The feature semantic information in the lower layer of the image is relatively small, but the shallower layer features are more detailed and the resolution of feature maps is higher. The location information is sufficient and the target location is accurate. The feature semantic information in the higher layers is richer, but the target location is coarse. The features of different layers are connected by skip-connection, and the location information of the lower layers and the semantic information of the higher layers are fused to achieve the accurate recovery of object location and details.

For  $F_{merge}^i$  and  $F_{RGB}^i$  generated from MFF and generated by upsampling  $F_{up}^i$ , we have:

$$F_{up}^1 = F_{RGB}^4 \tag{14}$$

$$F_{up}^i = PS(\sigma(Concat(F_{up}^{i-1}, F_{merge}^{i-1})) + F_{up}^{i-1}) \tag{15}$$

where  $F_{up}^i$  is  $F^i$  mentioned in Section 3.1, and PS refers to the PixelShuffle operation by which the scale of the feature is transformed to match the scale of the next layer. PixelShuffle can be seen as a special reshape operation, which moves pixels from the channel dimension to width and length dimension to achieve upsampling. In this process, the values in the tensor are not changed, and the correlation among the pixels is ensured using the pixels of other channels for filling.  $\sigma$  is a  $1 \times 1$  convolution with dimensionality reduction on the number of channels.

Accurately recovering the resolution of the feature maps during upsampling is key to achieve the recognition of a small object. Traditional upsampling is often achieved using a combination of bilinear interpolation and convolution. However, bilinear upsampling does not take into account the correlation among the predictions of each pixel and can lose some details of a small object. We therefore use a data-dependent upsampling method here, especially for low-resolution feature mapping, to achieve better segmentation accuracy. For features with different scales, DUpsampling is used to recover the features with small resolution to the maximum resolution uniformly, and finally the final classification result is obtained through concatenation and convolution operations. Figure 6 shows the details in the decoder.

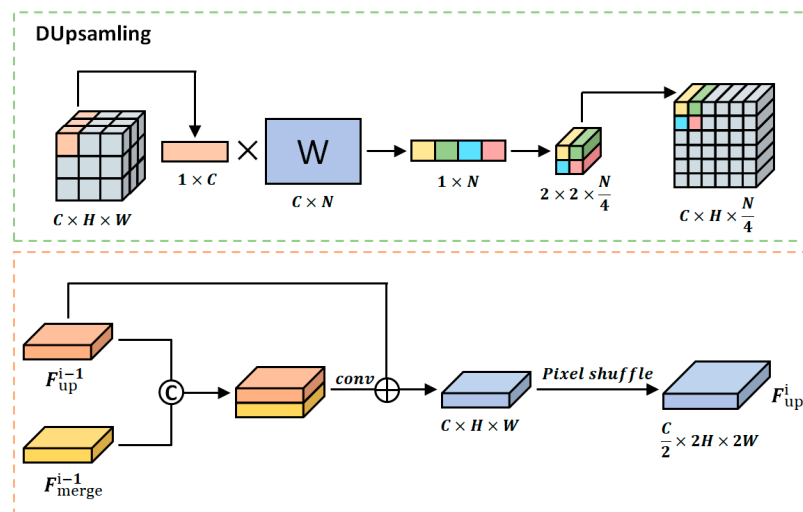


Figure 6. Decoder-related details (including composition of DUpsampling and steps after feature fusion).

## 4. Experiment







### 4.1. Dataset

The dataset we used is from the Potsdam urban classification dataset provided by ISPRS. Each dataset has six categories, namely impervious surface, building, low vegetation, tree, car, and cluster. This dataset has nDSM images (normalized digital surface model) in addition to RGB images and labeled maps.

The Potsdam dataset includes 38 images of  $6000 \times 6000$  pixels, which are divided into smaller images of  $512 \times 512$  pixels. The training set consists of {2\_10, 2\_11, 2\_12, 2\_13, 3\_10, 3\_11, 3\_12, 3\_13, 4\_10, 4\_11, 4\_12, 4\_13, 5\_10, 5\_11, 5\_12, 5\_13, 6\_7, 6\_8, 6\_9, 6\_10, 6\_11, 7\_7, 7\_8, 7\_9, 7\_10, 7\_11}, and the remainder is the test set. After cropping, there are 3744 images in the training set and 1728 images in the test set.

In Table 2, we counted the distribution of the dataset. The training dataset and test dataset have similar data distributions in general. It is worth noting that there is an unbalanced proportion of categories in the dataset. The tree, car, and cluster categories have a small proportion in the dataset, especially the car category, which only be accounted for 1.55% in the training dataset. In the subsequent experiments, we took some measures to suppress this imbalance.

**Table 2.** The detailed information of Potsdam for experiments.

Attributes		Categories				
Name	Impervious Surface	Building	Low Vegetation	Tree	Car	Clutter
Color						
Train set	29.95%	24.56%	23.43%	15.53%	1.55%	4.98%
Test set	34.29%	25.39%	18.84%	14.99%	2.04%	4.45%

#### 4.2. Implementation Details

The model is implemented on the pytorch deep learning platform. We use the SGB optimizer with an initial learning rate of 0.0005. The beam and weight attenuation coefficients are set to 0.9 and 0.001, respectively. We train for 200 epochs on the Potsdam dataset with the learning rate decreasing by cos. The size of each batch is set to eight. We employed an NVIDIA 3080 GPU to train our network.

**Loss function:** In the loss function we chose dice loss instead of the commonly used cross entropy function. Remote sensing images have the characteristic of category imbalance, and the dice loss has a better effect on the category imbalance problem. Therefore we use the loss function as follows:

$$L = L_{CE} + L_{Dice} \quad (16)$$

The Dice coefficient is an ensemble similarity measure usually used for calculating the similarity of two sample points:

$$S = \frac{2|X \cap Y|}{|X| + |Y|} \quad (17)$$

where  $X \cap Y$  is the intersection between  $X$  and  $Y$ , and  $|X|$ ,  $|Y|$  denote the number of elements of  $X$  and  $Y$ .

$$DiceLoss = 1 - S \quad (18)$$

As can be seen from the definition of the dice loss, dice loss is a region-dependent loss, meaning that the loss and gradient value of a pixel point is not only related to the label and predicted value of that point, but also to the label and predicted value of other points.

#### 4.3. Evaluation

We evaluated the segmentation accuracy of each category using IoU (intersection over union) and PA (pixel accuracy), which is the ratio of the intersection of the model's predicted and true values for a given category to the union:

$$IoU = \frac{TP}{TP + FP + FN} \quad (19)$$

PA is the number of pixels with the correct category predicted as a proportion of the total number of pixels:

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

The overall model accuracy was also assessed using mIoU, mPA, Acc, and F1-score, where mIoU and mPA are the cumulative average of IoU and PA for each category, respectively. Acc represents the number of correctly classified samples as a proportion of the total sample:

$$Acc = \frac{TP}{TP + TN + FP + FN} \quad (21)$$

F1-score is the summed average of precision and recall, where precision represents the proportion of positive samples classified into the positive sample that is correctly classified

and recall represents the proportion of positive samples classified into the positive sample that are true of all positive samples:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (22)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (23)$$

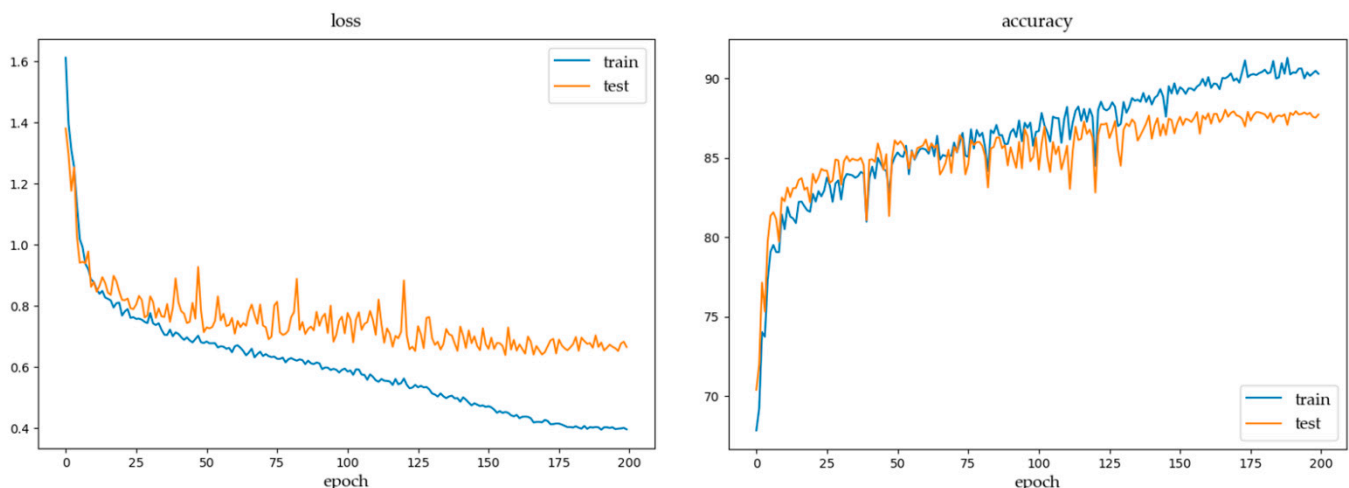
$$\text{Recall} = \frac{TP}{TP + FN} \quad (24)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

#### 4.4. Experiment

In order to verify the effectiveness of our model, a feature visualization operation, contrast experiment and ablation experiment were carried out on the Potsdam dataset. The feature visualization is mainly to verify the effectiveness of the feature fusion module and the help of multi-modal fusion for semantic segmentation. In the comparison experiments, we compared MFTransNet with several SOTA methods in regard to model accuracy and number of parameters, and verified that our model achieves a good balance between accuracy and speed. The ablation experiments were conducted for different components in the feature fusion module to verify the effectiveness of each component.

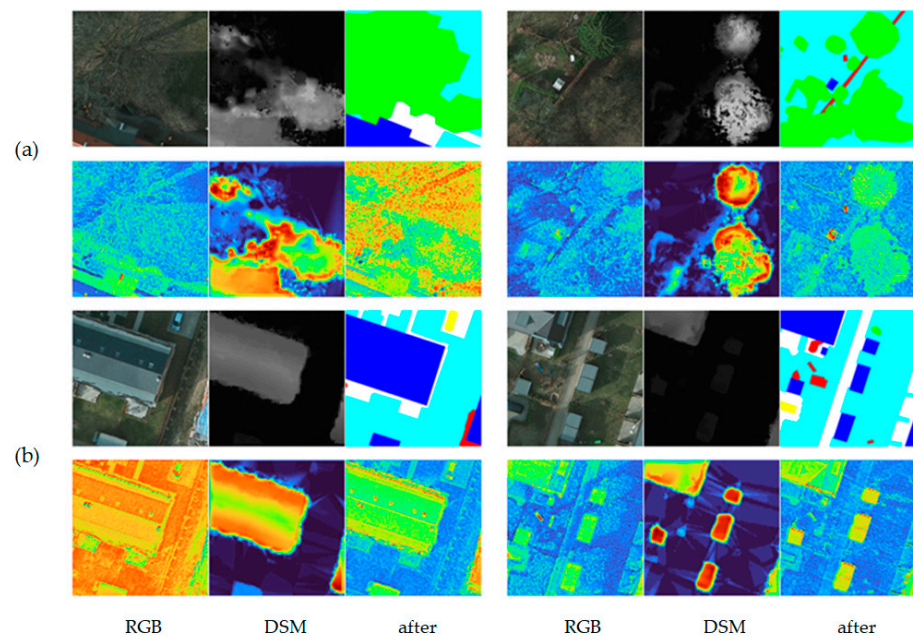
Figure 7 shows a graph of the change in loss function and accuracy during the training process. It can be seen that after 150 epochs, the training loss is still gradually decreasing, while the test loss has a slight increase. Thus, we chose to stop training after 50 epochs. Between 175–200 epochs, the training loss has converged and the test loss fluctuation value has become smaller. Finally, we chose the model with the smallest test loss between 175–200 epochs as the final evaluated model parameter. The same situation can be observed in the line graph of accuracy.



**Figure 7.** Line chart of loss and accuracy during training.

#### 4.5. Feature Visualization

The results of feature visualization are included in Figure 8. For images with different modalities, it is clear from the results of feature visualization that they have different valid information, with RGB images being more sensitive to color and texture information and DSM images being more sensitive to height information. Effectively combining the information from different modalities will be of great help in improving the semantic segmentation accuracy.



**Figure 8.** The figure contains the RGB images, the DSM images, and the feature maps after feature extraction, and the feature maps after the feature fusion module, aligned in the columns with the GT (ground truth) images. (a) The visualization focuses on the trees; (b) The visualization focuses on the building.

In group (a), we performed feature-map visualization containing low vegetation and trees. It can be found that there is almost no difference between the two categories in the feature maps of RGB images, but the tree information is more obvious in DSM images. After feature fusion, the height information is added on the basis of the original details, which makes the difference between low vegetation and trees in the feature map more obvious. In group (b), we visualized the feature map including buildings, and observed that the RGB image feature map paid more attention to the texture information, and the height information in the fused feature map enhanced the buildings, which retained the texture information and enhanced the height information at the same time. Therefore, we can conclude that the feature fusion module can fuse the complementary information between images of different modalities well.

#### 4.6. Comparative Experiments

We compared MFTransNet with several SOTA methods.

U-Net, DeepLab V3+ and TransFuse\_S [46] are unimodal semantic segmentation models. We only used RGB images as input, and comparing with these models can show the advantages of multi-modal models very well. In addition, SA-Gate, ACNet [47], RedNet [48] are RGBD models, so we use DSM information instead of depth information, and since both DSM and HHA represent height information data, we assert there will be generality among these models.

From Table 3, we can see that the multi-modal models all outperform the unimodal models in terms of accuracy and have better results in three categories: building, low vegetation, and trees, which are able to rely on height data to get better discrimination. This indicates that the use of multi-modal inputs plays an important role in improving segmentation accuracy. Our model largely outperforms most methods compared to other models, surpassing RedNet in mPA compared to it, but slightly lower in mIoU. We have bolded the two highest metrics in Table 1 for observation and, in terms of segmentation results for each category, our model is effective in improving the segmentation of low vegetation and trees, with the increased height information effectively distinguishing between these two similar-looking features. The model also performs better in the segmentation of buildings, where

the height information effectively removes the blurring between buildings and shaded boundaries. The experiments show that our model can effectively eliminate redundant information and enhance complementary information to achieve efficient multi-modal data fusion.

**Table 3.** Experiment results on Potsdam dataset.

Method	Impervious Surface		Building		Low Vegetation		Tree		Car		Clutter		mIoU	mPA
	IoU	PA	IoU	PA	IoU	PA	IoU	PA	IoU	PA	IoU	PA		
U-Net	81.78	88.32	89.47	95.14	69.28	86.4	<b>72.89</b>	<b>83.6</b>	<b>82.34</b>	87.89	32.85	<b>44.8</b>	71.44	81.03
DeepLab v3+	78.43	91.72	87.33	91.24	65.99	80.51	71.2	77.66	81.73	89.15	<b>40.06</b>	<b>55.02</b>	70.79	80.89
TransFuse_S	80.86	91.09	88.4	93.42	68.74	85.71	70.58	78.93	81.47	89.58	32.01	39.69	70.34	79.74
ACNet	<b>83.44</b>	91.15	<b>92.73</b>	<b>97.26</b>	69.05	83.57	71.57	<b>85.13</b>	81.84	87.6	30.08	35.6	71.45	80.05
RedNet	82.83	<b>92.45</b>	<b>92.4</b>	96.09	69.64	83.68	71.5	81.47	<b>83.06</b>	<b>89.89</b>	35.96	44.52	<b>72.57</b>	<b>81.35</b>
SA-Gate	82.74	<b>91.82</b>	91.33	<b>96.42</b>	<b>70.52</b>	<b>86.87</b>	69.88	77.9	79.56	85.31	<b>36.88</b>	44.08	71.82	80.4
MTransNet	<b>83.26</b>	91.09	91.92	95.54	<b>70.28</b>	<b>86.44</b>	<b>72.26</b>	82.37	82.08	<b>90.83</b>	33.87	43.96	<b>72.34</b>	<b>81.7</b>

The two largest terms of the value are bolded.

Figure 9 shows a comparison of the segmentation plots for the various models. In the first and second rows, the segmentation results for trees are highlighted in the dashed box. The results show that the multi-modal model is very effective at distinguishing between trees and low vegetation, whereas the unimodal model struggles to achieve accurate segmentation results when distinguishing between these two types of objects with similar appearances. Our model and RedNet segmentation results are the closest to the labelled map. We also observe that in the middle left part of the first row, the DSM image appears as a clear bright area and most models identify trees. However, the labeled map does not show that this is a tree, possibly due to incorrect manual labelling. In rows 3–5, the segmentation effect of the buildings is highlighted. The unimodal models show a large number of misclassifications at the boundaries, incorrectly identifying impervious surfaces as clutter. The multi-modal model works better, with our model and ACNet boundary segmentation being the best. Since clutter is the most difficult of these categories to identify, these models all show relatively poor segmentation results. Our model is still very similar to the labeled graph in general shape. From these visualizations, it can be concluded that the multi-modal model is better at addressing interclass similarity (segmentation of trees and low vegetation) and intraclass variability (segmentation of impervious surfaces at building boundaries). Our model is also in the better segmentation category of the multimodal model.

As can be seen from Table 4 and Figure 10, the multi-modal model outperforms the unimodal model in terms of accuracy metrics. However, with the same backbone, the parameters of the multi-modal model are more than twice as large as those of the unimodal model. The multi-modal model enriches the information sources due to its multiple source inputs, but likewise introduces a huge number of parameters and large amount of computational effort, and it requires two backbone networks to handle the different modal inputs. As can be seen in Table 4, our model has half the number of parameters and 1/10 the computational effort of U-Net, and our model achieves the effect of the multi-modal model in all accuracy metrics while reducing the model size as much as possible. In addition, we do not use training weights in our experiments, but the accuracy still reaches that of other models using pre-trained weights. This evidence illustrates that MTransNet achieves a good balance between model size and performance.

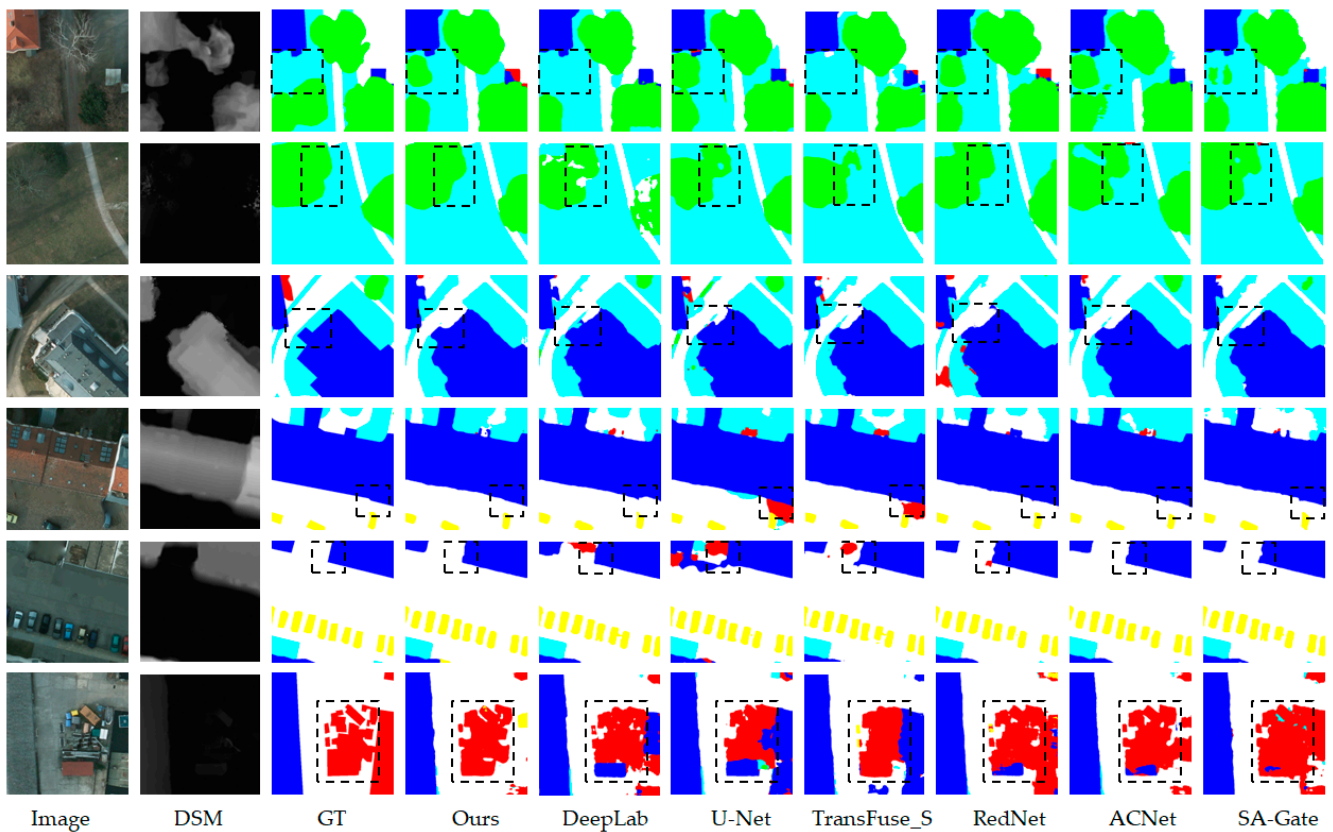


Figure 9. Visualization results of MFTransNet and SOTA methods (white: impervious surfaces; cyan: low vegetation; blue: buildings; yellow: cars; green: trees; and red: clutter/background).

Table 4. Comparison of model size and accuracy.

Method	Backbone	FLOPs (G)	Params (M)	F1-Score (%)	Acc (%)
U-Net	Resnet50	92.12	43.93	81.87	87.55
DeepLab v3+	Xception	83.44	54.71	81.94	85.69
TransFuse_S	Resnet34	58.36	37.61	81.42	86.53
RedNet	Resnet50	85.26	81.95	82.94	87.89
SA-Gate	Resnet101	165.1	110.85	82.75	87.71
ACNet	Resnet50	106.25	116.60	82.16	87.83
MFTransNet	-	9.91	23.20	82.57	87.93

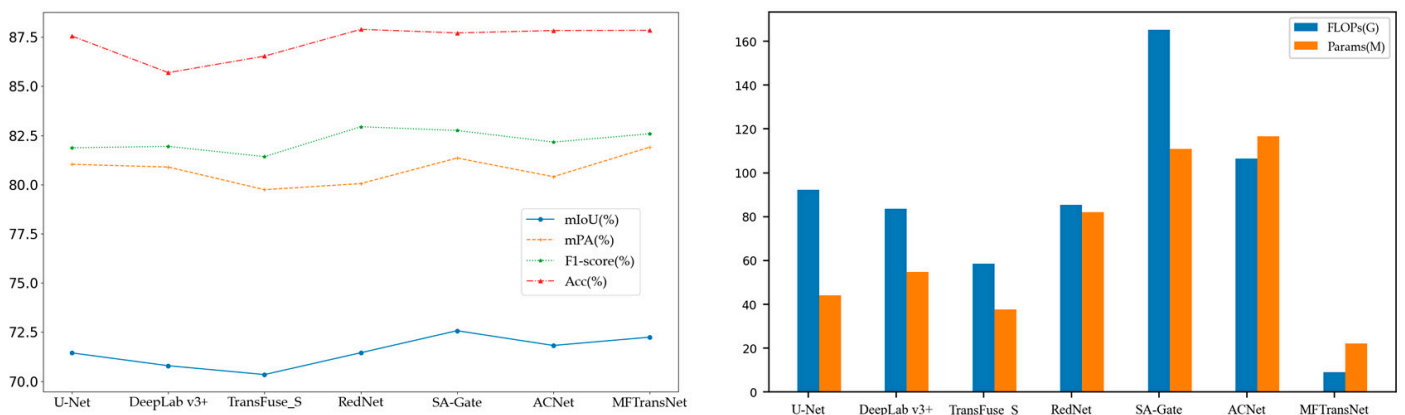


Figure 10. Comparison chart of model size and accuracy (On the left is a line graph of each indicator representing accuracy, and on the right is a bar graph of each indicator representing size).



#### 4.7. Ablation Experiments

In order to verify the effectiveness of each component in the feature fusion module, we conducted an ablation experiment on it to compare the effect of adding different components to the model. AMfuse+ is used to replace the original feature fusion module as baseline. The evaluation results are listed in Tables 5 and 6, where  $\checkmark$  indicates that the corresponding module is retained. The results show that these two components have a certain optimization effect on the accuracy of the model, and the combination effect is better.

**Table 5.** The effects of different components of our method on the results.

Method	FACM	CFM	mIoU	mPA	F1-Score (%)	Acc (%)
Baseline			70.91	81.07	81.99	86.99
Ours	$\checkmark$		71.57	80.77	81.84	87.64
Ours		$\checkmark$	71.59	80.5	82.9	87.68
Ours	$\checkmark$	$\checkmark$	72.34	81.7	82.57	87.93

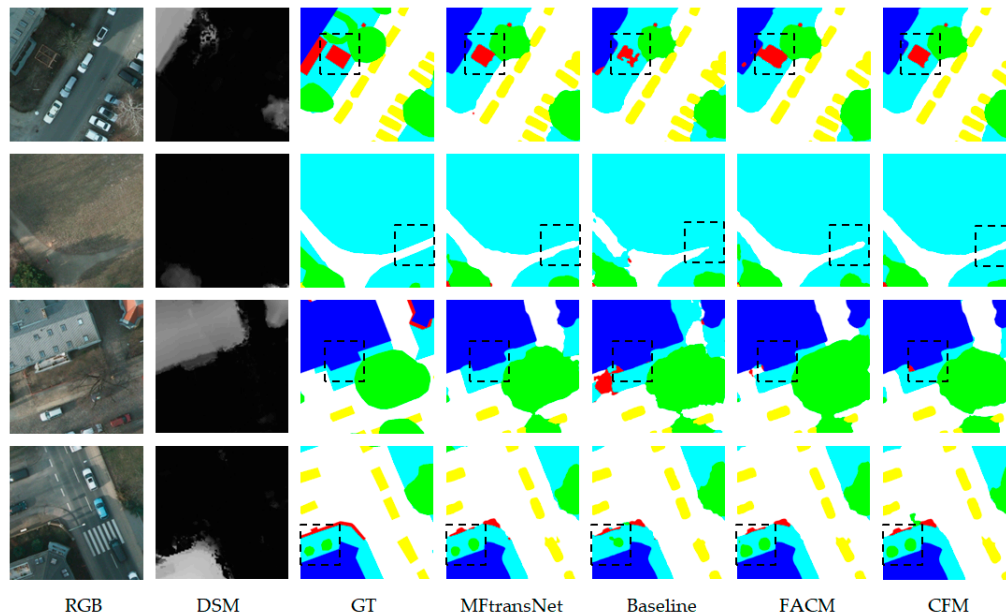
**Table 6.** Evaluation results for each category.

Method	Impervious Surface		Building		Low Vegetation		Tree		Car		Clutter	
	IoU	PA	IoU	PA	IoU	PA	IoU	PA	IoU	PA	IoU	PA
Baseline	81.34	88.09	91.38	97.2	68.13	85.91	70.13	81.09	79.57	91.47	34.94	42.64
With FACM	82.34	91.41	91.7	95.27	70.64	83.89	72.64	84.62	81.24	88.86	30.88	40.57
With CFM	82.52	91.26	91.56	95.83	69.98	86.0	71.65	81.83	80.12	87.73	33.74	40.33
MFTransNet	83.26	91.09	91.92	95.54	70.28	86.44	72.26	82.37	82.08	90.83	33.87	43.96

Figure 11 shows the visualization results of the ablation experiment. In the first row it is observed that the segmentation of the object by baseline is poor, after the addition of FACM the segmentation of the interior of the object becomes better, and after the addition of CFM the segmentation boundary of the object is even better. In the third row, there is a large number of misclassifications in the lower left corner of the dashed box of baseline, which improves with the addition of FACM. However, there are still a small number of misclassifications, which improve with the addition of CFM. The best segmentation results are obtained when both modules are added. With the addition of FACM, the information of the two modalities was calibrated. The redundant information between the modalities is not eliminated due to the fusion module not distributing the weights well, and the boundary processing is rather rough. With the addition of CFM, the segmentation is significantly improved and the weights between the different modalities are well distributed. Nevertheless the complementary information is not enhanced due to the lack of calibration of the modalities and the segmentation is still less than perfect. The best result is achieved using both modules together, with the features being calibrated and then fused to extract the complementary information between the modalities, while suppressing redundant information and achieving a more accurate segmentation.

Overall, our model meets the design requirements, but there are still some problems. Firstly, the remote sensing dataset suffers from an insufficient volume of data and an insufficient dataset. The insufficient dataset caused the model to fall into overfitting easily during training, a problem that arose for all models during the experiments; the larger the model size, the earlier it would start to overfit. The insufficient dataset prevents us from verifying the performance of the model on other datasets. Multi-modal remote sensing semantic segmentation datasets are currently scarce, and we will contribute to the development of the field by autonomously annotating datasets in subsequent research. Furthermore, the accuracy of our model is still relatively low for the category clutter, as can be seen in Table 3. This category is different from the others in that it is very diverse in terms of size, height, color, and various attributes, and is one of the most difficult types to

recognize. It requires richer contextual information to improve the accuracy of this category. Our model is designed to fuse information from different modalities and scales, and may lack some consideration of context in order to reduce the number of parameters. Future research will investigate the integration of contextual information in more depth without considering the number of parameters.



**Figure 11.** Visualization results of the ablation experiment.

## 5. Conclusions

In this article, we propose the MFTransNet framework for the semantic segmentation of multi-modal remote sensing images as a way of balancing accuracy and efficiency in high spatial resolution image segmentation. Our network mainly consists of CNN-Transformer modules. Specifically, the feature fusion module MFF with a multi-head attention mechanism is designed, in which ACM can adaptively calibrate features, and CFF is used to fuse multi-modal features to achieve efficient feature adaptive fusion. In view of the multi-scale problem existing in remote sensing images, a multi-scale decoder structure and DUpsampling is used instead of traditional bilinear upsampling to reduce the details lost during upsampling and achieve feature aggregation at different scales.

MFTransNet not only overcomes the problem of sample imbalance and improves the confidence level of intra-class objects, but also improves the segmentation of feature edges. The model effectively mitigates the intra-class similarity and inter-class dissimilarity of remote sensing images and improves segmentation accuracy. In addition, thanks to the application of Transformer with multi-modal data fusion, the model does not require much memory. We have demonstrated the benefits of MFTransNet on challenging high-resolution remotely sensed images. However, as multi-modal data operations need to consider fusing data from different sources and dealing with different levels of noise and missing data, the model still suffers from poor complementarity between modality, alignment difficulties, and redundancy between modalities. The ability to design similarity measures between modalities to represent HSR remote sensing image data in a meaningful and valuable way is critical to our model. Considering that the joint representation approach can handle more than two modalities, while coordinated representation can currently only handle two modalities, in future work, we will continue to improve the Transformer model structure and strive to achieve joint multi-modal fusion in order to further improve the accuracy of HSR remote sensing image segmentation while controlling the model size.

**Author Contributions:** This work was conducted in collaboration with all authors. H.Y. defined the research theme, and supervised the research work and provided experimental facilities. S.H. and X.Z. designed the semantic segmentation model and conducted the experiments. S.H. and X.L. checked the experimental results. S.H. revised the paper according to the requirements of the review. This manuscript was written by S.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Hainan Province Science and Technology Special Fund (Grant No. ZDYF2022GXJS228); Hainan Provincial Natural Science Foundation of China (Grant No. 620RC559); Haikou Science and Technology Plan Project (2022-007, 2022-015, 2020-056).

**Data Availability Statement:** We are grateful to CVPR and ISPRS for providing the open benchmarks for 2D remote sensing image semantic segmentation. The data in the paper can be obtained through the following link. Potsdam: 2D Semantic Labeling Contest—Potsdam (isprs.org), accessed on 2 March 2022. Our code and trained model will be available at <https://github.com/bukaqiii/MFTransNet>, accessed on 2 March 2022.

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

## References

1. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017.
2. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [[CrossRef](#)]
3. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X.; Wei, X. Semantic Segmentation of Aerial Images with Shuffling Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 173–177. [[CrossRef](#)]
4. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [[CrossRef](#)]
5. Zheng, X.; Gong, T.; Li, X.; Lu, X. Generalized Scene Classification from Small-Scale Datasets with Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [[CrossRef](#)]
6. Le, N.Q.K. Potential of deep representative learning features to interpret the sequence information in proteomics. *Proteomics* **2022**, *22*, e2100232. [[CrossRef](#)]
7. Kha, Q.-H.; Ho, Q.-T.; Le, N.Q.K. Identifying SNARE Proteins Using an Alignment-Free Method Based on Multiscan Convolutional Neural Network and PSSM Profiles. *J. Chem. Inf. Model.* **2022**, *62*, 4820–4826. [[CrossRef](#)] [[PubMed](#)]
8. Albulayhi, K.; Smadi, A.A.; Sheldon, F.T.; Abercrombie, R.K. IoT Intrusion Detection Taxonomy, Reference Architecture, and Analyses. *Sensors* **2021**, *21*, 6432. [[CrossRef](#)]
9. Abu Al-Haija, Q.; Krichen, M. A Lightweight In-Vehicle Alcohol Detection Using Smart Sensing and Supervised Learning. *Computers* **2022**, *11*, 121. [[CrossRef](#)]
10. Alsulami, A.A.; Abu Al-Haija, Q.; Alqahtani, A.; Alsini, R. Symmetrical Simulation Scheme for Anomaly Detection in Autonomous Vehicles Based on LSTM Model. *Symmetry* **2022**, *14*, 1450. [[CrossRef](#)]
11. Kareem, S.S.; Mostafa, R.R.; Hashim, F.A.; El-Bakry, H.M. An Effective Feature Selection Model Using Hybrid Metaheuristic Algorithms for IoT Intrusion Detection. *Sensors* **2022**, *22*, 1396. [[CrossRef](#)]
12. Cao, Z.; Fu, K.; Lu, X.; Diao, W.; Sun, H.; Yan, M.; Yu, H.; Sun, X. End-to-End DSM Fusion Networks for Semantic Segmentation in High-Resolution Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1766–1770. [[CrossRef](#)]
13. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *32*, 6700–6713. [[CrossRef](#)]
14. Lang, F.; Yang, J.; Yan, S.; Qin, F. Superpixel Segmentation of Polarimetric Synthetic Aperture Radar (SAR) Images Based on Generalized Mean Shift. *Remote Sens.* **2018**, *10*, 1592. [[CrossRef](#)]
15. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
16. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3121–3130.
17. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [[CrossRef](#)]
18. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*; Springer International Publishing: Cham, Switzerland, 2018; pp. 833–851.
19. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.

20. 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](#)]
21. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
23. 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](#)]
24. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
25. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4095–4104.
26. Du, X.; He, S.; Yang, H.; Wang, C. Multi-Field Context Fusion Network for Semantic Segmentation of High-Spatial-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5830. [[CrossRef](#)]
27. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [[CrossRef](#)]
28. Ma, A.; Wang, J.; Zhong, Y.; Zheng, Z. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [[CrossRef](#)]
29. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
31. 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 6x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
32. 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. 6877–6886. [[CrossRef](#)]
33. Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. *arXiv* **2021**, arXiv:2101.11605.
34. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12165–12175. [[CrossRef](#)]
35. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020.
36. 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.
37. Chen, X. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. *arXiv* **2020**, arXiv:2007.09183.
38. Liu, H.; Chen, F.; Zeng, Z.; Tan, X. AMFuse: Add-Multiply-Based Cross-Modal Fusion Network for Multi-Spectral Semantic Segmentation. *Remote Sens.* **2022**, *14*, 3368. [[CrossRef](#)]
39. Weng, Q.; Chen, H.; Chen, H.; Guo, W.; Mao, Z. A Multisensor Data Fusion Model for Semantic Segmentation in Aerial Images. *IEEE Geoscience Remote Sens. Lett.* **2022**, *19*, 1–5. [[CrossRef](#)]
40. Prakash, A.; Chitta, K.; Geiger, A. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
41. Cao, Z.; Diao, W.; Sun, X.; Lyu, X.; Yan, M.; Fu, K. C3Net: Cross-Modal Feature Recalibrated, Cross-Scale Semantic Aggregated and Compact Network for Semantic Segmentation of Multi-Modal High-Resolution Aerial Images. *Remote Sens.* **2021**, *13*, 528. [[CrossRef](#)]
42. Zhao, J.; Zhang, D.; Shi, B.; Zhou, Y.; Chen, J.; Yao, R.; Xue, Y. Multi-source collaborative enhanced for remote sensing images semantic segmentation. *Neurocomputing* **2022**, *493*, 76–90. [[CrossRef](#)]
43. Liu, H.; Zhang, J.; Yang, K.; Hu, X.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *arXiv* **2022**, arXiv:2203.04838.
44. Wele, G.; Patel, V.M. HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening. *arXiv* **2022**, arXiv:2203.02503.

45. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [[CrossRef](#)]
46. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; pp. 14–24. [[CrossRef](#)]
47. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNET: Attention based network to exploit complementary features for rgb-d semantic segmentation. In Proceedings of the IEEE Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
48. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. RedNet: Residual encoderdecoder network for indoor RGB-D semantic segmentation. *arXiv* **2018**, arXiv:1806.01054.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.