remote sensing

MDPI

Technical Note

Cross-Modal Segmentation Network for Winter Wheat Mapping in Complex Terrain Using Remote-Sensing Multi-Temporal Images and DEM Data

Nan Wang, Qingxi Wu, Yuanyuan Gui, Qiao Hu * and Wei Li

School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
wang_nan@bit.edu.cn (N.W.); viiqxxxx@bit.edu.cn (Q.W.); 3220205108@bit.edu.cn (Y.G.); liw@bit.edu.cn (W.L.)
* Correspondence: 3120210841@bit.edu.cn

Abstract: Winter wheat is a significant global food crop, and it is crucial to monitor its distribution for better agricultural management, land planning, and environmental sustainability. However, the distribution style of winter wheat planting fields is not consistent due to different terrain conditions. In mountainous areas, winter wheat planting units are smaller in size and fragmented in distribution compared to plain areas. Unfortunately, most crop-mapping research based on deep learning ignores the impact of topographic relief on crop distribution and struggles to handle hilly areas effectively. In this paper, we propose a cross-modal segmentation network for winter wheat mapping in complex terrain using remote-sensing multi-temporal images and DEM data. First, we propose a diverse receptive fusion (DRF) module, which applies a deformable receptive field to optical images during the feature fusion process, allowing it to match winter wheat plots of varying scales and a fixed receptive field to the DEM to extract evaluation features at a consistent scale. Second, we developed a distributed weight attention (DWA) module, which can enhance the feature intensity of winter wheat, thereby reducing the omission rate of planting areas, especially for the small-sized regions in hilly terrain. Furthermore, to demonstrate the performance of our model, we conducted extensive experiments and ablation studies on a large-scale dataset in Lanling county, Shandong province, China. Our results show that our proposed CM-Net is effective in mapping winter wheat in complex terrain.

Keywords: deep learning; winter wheat; semantic segmentation; remote sensing image; attention block

check for updates

Citation: Wang, N.; Wu, Q.; Gui, Y.; Hu, Q.; Li, W. Cross-Modal Segmentation Network for Winter Wheat Mapping in Complex Terrain Using Remote-Sensing Multi-Temporal Images and DEM Data. *Remote Sens.* **2024**, *16*, 1775. https://doi.org/10.3390/rs16101775

Academic Editor: Pablo Rodríguez-González

Received: 8 February 2024
Revised: 12 May 2024
Accepted: 13 May 2024
Published: 16 May 2024

1. Introduction

Winter wheat is one of the world's major food crops, and it plays an important role in ensuring food security and stabilizing national economic development. However, the production of winter wheat is facing various challenges due to climate change, land use change and crop rotation in some regions. Therefore, it is important to monitor the distribution of winter wheat cultivation for better agricultural management, land planning and environmental sustainability [1,2].

With the rapid development of remote sensing (RS) technology, RS imagery has been widely adopted in agriculture as an effective means of acquiring recurrent, large-area data [3]. This technology enables the monitoring of changes in agricultural production and resource utilization, providing timely information to support macro-level decision making [4]. Temporal information is a widely used feature for crop mapping due to the growth cycle of crops. Different crops exhibit varying growth stages and unique spectral scattering characteristics in different seasons. Numerous research studies have demonstrated that offering a model interpretation associated with crop growth features is essential in assessing the dependability of crop-mapping techniques. Therefore, most of researches distinguish winter wheat from other crops using temporal features with machine learning classifiers, such as support vector machine (SVM) [5], random forest (RF) [6,7], and decision trees (DT) [8,9], which achieves accurate winter wheat mapping.

In recent years, deep learning has developed rapidly, and has proven a superior learning ability and a strong feature presenting ability. Compared to conventional machine learning classifiers, deep learning (DL) techniques show outstanding performance in various remote sensing-related tasks, such as change detection [10] and land cover classification [11,12]. Deep learning architectures for crop mapping employ two main approaches: pixel-based classification and semantic segmentation. Pixel-based classification tasks typically start by training a CNN classifier on small image patches, then using a sliding window method to predict the category of the central pixel [13–15], of which the drawback is that the trained network only predicts the central pixel of the input image, leading to low classification efficiency. Semantic segmentation, which aims to assign a specific class label to each pixel in an image with high process efficiency, is gradually gaining attention in the crop-mapping field [16]. For example, Zhang et al. [17] combined a pyramid scene parsing network (PSPNet) [18] and GaoFen satellite images for cropland mapping. Wei et al. [19] formulated rice mapping as a task of semantic segmentation and used Unet to generate a map of rice distribution by exploiting the correlation among multi-temporal data. Ma et al. [20] proposed a rice-planting area identification attention U-Net (RIAU-Net) model to map rice with Sentinel-1 images obtained in specific months. Sai et al. [21] proposed a deep learning model that employs an over-complete representation, integrated with a backbone transfer learning-based encoder–decoder architecture to solve weed detection.

Recently, some researchers have explored multi-modal networks to improve the accuracy of semantic segmentation tasks; for instance, Garnot et al. [22] assessed the advantages of multi-modality in various tasks, and their findings indicate that by utilizing optical and radar time series data, temporal attention-based models that incorporate multiple modalities can outperform models using a single modality in terms of performance and the ability to withstand cloud cover. Li et al. [23] proposed a semantic segmentation model of multisource data fusion (MCANet). The backbone network of feature extraction in the network consists of two independent branches for the feature extraction of optical and SAR images. Optical and SAR data are complementary to each other in land-use classification, and better extraction results can be obtained by combining their advantages. Hazirbas et al. [24] proposed the FuseNet algorithm, which is based on SegNet [25] to directly add the features from both RGB and depth images to segment different objects in natural images, and they achieved an acceptable accuracy. Zhang et al. [26] proposed a feature-level fusion network named the hybrid attention-aware fusion network (HAFNet). Primarily, it enhances information fusion from multiple modalities through an attention-aware mechanism, leading to more accurate and robust classification results, which are particularly useful in complex environments. However, it can be computationally complex to implement, which may pose challenges in scaling to large study areas.

Although existing multi-modal semantic segmentation methods have achieved good segmentation results in fields such as urban areas or natural images, they are not suitable for crop mapping, especially in hilly areas. Due to different terrain conditions, the soil size, distribution, and texture of winter wheat planting fields are not consistent. Compared with plain areas, the planting units of winter wheat in mountainous areas are smaller in size and fragmented in distribution, and the growth cycle of winter wheat in mountainous areas differs from that in plain areas due to low temperatures [27,28]. Therefore, most crop-mapping research based on deep learning only focuses on temporal or spectral features' representation from optical remote sensing images and ignores the impact of topographic relief in crop distribution, so that the existing models of semantic segmentation cannot deal with the hilly area very well. To sum up, though the results of these previous studies show that cross-modal deep learning can improve the accuracy of semantic segmentation, there is still some room for improvement for crop mapping with complex topographic conditions. The following issues can be discussed.

(1) In complex terrains, both plains and hilly areas are suitable for the cultivation of winter wheat. Current cross-modal algorithms typically employ a fixed receptive field across two modal branches to extract features. These approaches overlook the fact that different modal data types offer varying perspectives on wheat features. For instance, optical remote sensing images provide insights into the growth status and require a flexible receptive field due to the varying scales of planting sizes. On the other hand, Digital Elevation Models (DEM) present terrain information, including slope details and so on, necessitating a stable receptive field for accurate computation. Consequently, the unified receptive field utilized in the present model fails to accommodate the characteristics of the dual-modal data, thereby posing challenges to its effective application in extracting the distribution of winter wheat in areas with complex terrain.

(2) In hilly regions, the small-scale planting areas, which occupy fewer pixels in the images and carry limited information, pose a significant challenge. During the downsampling and upsampling processes of the encoder–decoder convolutional neural network, the resolution of these small targets is further diminished and their feature information progressively weakens. This makes it difficult to effectively recover the information of these small targets, leading to a higher rate of omission in the mapping of winter wheat in hilly regions.

To address the problems above, we study a common network for winter wheat mapping in complex terrain. The primary contributions of this article are as follows:

- We propose a novel network named a Cross-Modal Segmentation Network (CM-Net), which has the capability to integrate temporal, spatial, and terrain features for enhanced image segmentation.
- A Diverse Receptive Fusion (DRF) module is proposed. This module applies a deformable receptive field to optical images during the feature-fusion process, allowing it to match winter wheat plots of varying scales and a fixed receptive field to the DEM to extract evaluation features at a consistent scale.
- We developed a novel spatial attention module, the Distributed Weight Attention (DWA) module. This module is specifically designed to enhance the feature intensity of our objects, thereby reducing the omission rate of planting areas, especially for the small-sized regions.

The remainder of this article is structured as follows: Section 2 details the proposed CM-Net and its constituent components. Section 3 provides an overview of the experimental data and setup. In Section 4, we present comparative experiments and ablation studies, as well as a detailed analysis of the CM-Net. Finally, we make concluding remarks in Section 5.

2. Methodology

This section delves into our proposed Cross-Modal Segmentation Network (CM-Net) as shown in Figure 1, a quintessential encoder–decoder structure. The CM-Net accepts digital elevation models (DEM) and multi-temporal remote sensing images as the inputs to extract and fuse the multimodal features. The core structure of the network is designed to extract features from both data types, culminating in the segmentation image of winter wheat.

The encoder with a dual-branch structure is comprised of a Diverse Receptive Fusion (DRF) module and a downsampling module to extract the temporal–spatial–terrain features from the multi-modal data. The decoder segment is made up of upsampling modules, and a Distributed Weight Attention (DWA) modular, which are based on attention mechanisms. The output module is composed of a 1 × 1 convolutional layer and a sigmoid activation layer.

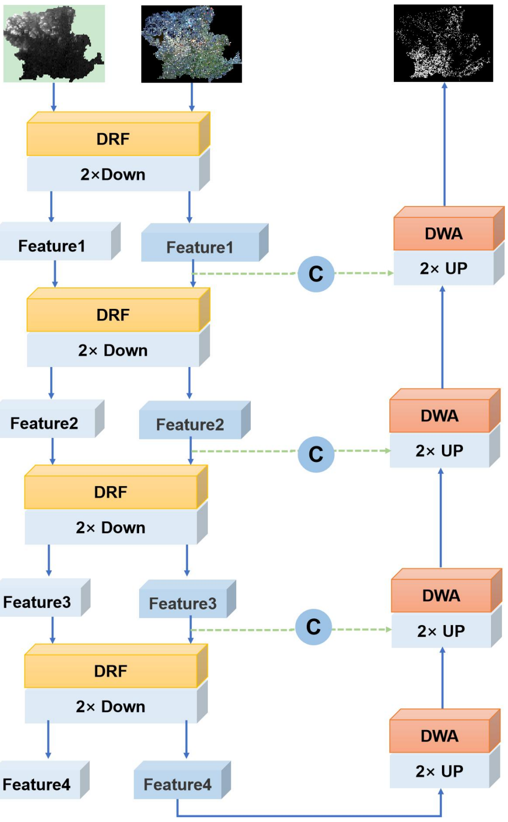

Figure 1. Structure of the CM-Net model.

2.1. Diverse Receptive Fusion Module

This paper proposes a Diverse Receptive Fusion (DRF) module as shown in Figure 2. During the feature-fusion process, a deformable receptive field is utilized on optical images, enabling the matching of winter wheat plots of various scales. Concurrently, a fixed receptive field is applied to the DEM to consistently extract evaluation features. The diverse receptive field employed in our model is designed to cater to the attributes of dual-modal data, thereby facilitating the effective extraction of the winter wheat distribution in regions with complex terrain.

The basic structure of the DRF consists of two branches, which extract features from different modal data. The left branch takes the original image of the DEM, called the elevation branch, while the right branch takes the original image of the multi-temporal optical data, called the optical branch. The feature maps are downsized to half of their original size in both the width and height directions after passing through the DRF module, while the channel direction is doubled.

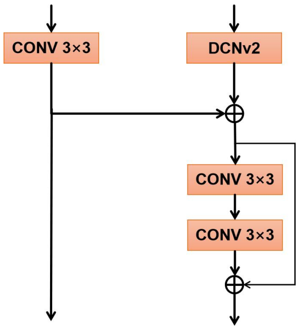

Figure 2. The Diverse Receptive Fusion (DRF) module.

The elevation branch comprises a fixed receptive field convolution block. When extracting features from the DEM data, a fixed perception can accurately capture the terrain elevation information related to the cultivation style of the winter wheat. That is, it can determine what type of terrain is suitable for wheat cultivation, what scale of cultivation is appropriate for the terrain type, and so on. This information serves as supplementary data for semantic segmentation.

Suppose that the input of the elevation branch in the multimodal fusion module of stage i is x^i_{dem}, and the output is expressed as y^i_{dem}. The calculation process of the CNN block is as follows:

$$y^i_{dem} = ReLU(BN(Conv2d(x^i_{dem}))) \tag{1}$$

The optical branch consists of a deformable convolution block with flexible receptive field, a concatenate operation, and a residual module. The input of the optical branch is multi-temporal optical images, which are fused with the features extracted from the elevation branch after going through the DCN (Deformable Convolutional Networks) block and then input into the residual module. The distribution characteristics of wheat in mountainous and plain areas are different. In mountainous and hilly regions, wheat distribution is fragmented, while in plain areas, wheat is usually distributed in large blocks, with varying sizes of wheat-planting areas. However, the receptive field of the conventional convolutional kernels is fixed and cannot adapt to geometric transformations in the spatial domain. By contrast, deformable convolution incorporates trainable offsets in the convolution module, which enables the convolution sampling points to shift, resulting in a receptive field that automatically adapts to changing sampling positions. In each layer of the CM-Net network, deformable convolution modules are added to introduce variability and automatically adjust the receptive field, thereby enhancing the perception of winter wheat at both large and small scales and achieving the accurate extraction of winter wheat with varying scales.

Specifically, assuming that the input of the optical branch in the i-th stage of the multimodal fusion module is represented as x^i_{opt}:

$$\begin{aligned} y^i_{dcn} &= ReLU(BN(DCNv2(x^i_{opt}))) \\ y^i_{cat} &= concatenate(y^i_{dem}, y^i_{dcn}) \\ y^i_{res} &= ResBlock(y^i_{cat}) \end{aligned} \tag{2}$$

y^i_{dcn} is the output of DCN Block. After that, y^i_{dcn} is spliced with the feature matrix y^i_{dem} output from the elevation branch along the channel direction to obtain y^i_{cat}, and finally y^i_{res} is obtained by ResBolck.

2.2. Distributed Weight Attention Module

A distributed weighted attention module (DWA) was designed to enhance the feature intensity in the decoding process of winter wheat. By using global pooling and expanding the range of the activation function, the intensity of target features can be enhanced while the intensity of non-target features is suppressed, so as to effectively recover the information of small planting areas.

As shown in Figure 3, the DWA does not alter the shape of the feature map. Assuming that the size of the feature map input into the DWA is C (channel) × H (height) × W (width), the feature map undergoes channel-based global max pooling and global mean pooling, resulting in two 1 × H × W feature maps. These feature maps are then concatenated along the channel dimension and passed through a convolutional layer with a kernel size of 3 × 3 and a sigmoid activation layer, resulting in a 1 × H × W spatial-attention weight matrix W_{sam}.

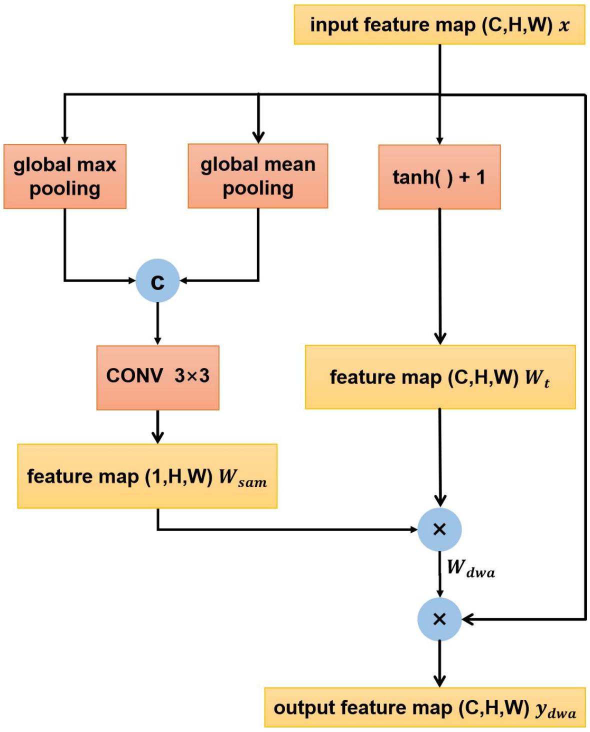
input feature map (C,H,W) x
global max pooling
global mean pooling
tanh( ) + 1
c
CONV 3×3
feature map (C,H,W) W_t
feature map (1,H,W) W_{sam}
×
W_{dwa}
×
output feature map (C,H,W) y_{dwa}

Figure 3. The distributed weighted attention (DWA) module.

$$\begin{aligned} y &= concatenate(AvgPool(x), MaxPool(x)) \\ W_{sam} &= Sigmoid(Conv2d(y)) \end{aligned} \tag{3}$$

x is the input feature map to the pooling layers. After that, y is the concatenated output resulting from the application of both average pooling and max pooling operations to x. The aforementioned operations are the computation process of spatial attention, where the feature map is subjected to average pooling and max pooling along the channel dimension. The global max pooling operation extracts the maximum value for each pixel of the feature map C × H × W in the channel direction. The global mean pooling calculates the average value for each pixel of the feature map C × H × W in the channel direction.

Spatial attention learns a weight for each pixel on the feature map, enhancing the representation of important regions while suppressing unimportant regions, resulting in weighted features.

Considering that the input feature map x of the DWA contains information for extracting winter wheat from the DEM and multispectral images, each channel of the feature map has a corresponding role and cannot be simply suppressed or activated. Therefore, the 1 × H × W spatial attention weight matrix cannot be directly multiplied with the C × H × W input feature map x. Thus, DAM remaps all information through an activation function to obtain the overall distribution of the discrete feature map.

The DAM uses the activation function $tanh() + 1$ to map the value range of feature maps to 0–2, resulting in a matrix W_t of the shape C × H × W. This matrix, W_t, is then multiplied with the weight matrix W_{sam}, where values smaller than 1 decrease the weights of the pixel corresponding to W_{sam}, and values larger than 1 increase the weights of the pixel corresponding to W_{sam}. This process generates a final matrix W_{dwa} of the shape C × H × W. By "loosening" the features before multiplying with the weight map, important information can be highlighted and redundant information can be reduced. Additionally, while preserving the information from each channel of the input features, W is adjusted to obtain attention weights for each channel. The calculation process is as follows:

$$
\begin{aligned}
&W_t = tanh(x) + 1 \\
&W_{dwa} = W_t * W_{sam} \\
&y_{dwa} = W_{dwa} \times x
\end{aligned}
\tag{4}
$$

3. Experimental Data and Setup

3.1. Introduction of the Wheat Dataset

The experimental study area is located in Lanling County, Linyi City, Shandong Province, China, between 117°41″∼118°18″ east longitude and 34°37″∼35°06″ north latitude. It belongs to the warm temperate monsoon continental climate zone, with long and dry winters and hot and humid summers. The main variety of wheat planted is winter wheat. The terrain of Lanling County gradually decreases from northwest to southeast, including low mountains, hills, plains, and depressions. The terrain is relatively complex. Wheat in mountainous areas is distributed in a fragmented manner, while wheat in plain areas is distributed in large blocks and is relatively flat.

This experiment used two modalities of data, remote sensing data acquired by the Sentinel-2 satellite and corresponding digital elevation model (DEM) data.

Sentinel-2 is a high-resolution multispectral imaging satellite developed by the European Space Agency. It carries a multispectral imaging instrument (MSI) and consists of two satellites, Sentinel-2A and Sentinel-2B, which are placed at opposite sides of the Earth with an orbital difference of 180°. The revisit period of each satellite is 10 days, and 5 days for both. The Sentinel-2 satellite covers 13 working bands, with ground resolutions of 10 m, 20 m, and 60 m for different bands. In this study, 10 bands with a resolution of 10 m were used. The Google Earth Engine (GEE) platform provides global-scale geospatial analysis services, which support the extraction of agricultural information and provide convenient data and feature collection functions for Sentinel-series remote sensing images.

To enhance computational efficiency, we selected images from two key growth periods of winter wheat as the optical branch inputs for the network, each with a size of 5282 × 6767. The 10-channel remote sensing images acquired on 18 December 2017 and 15 April 2018 were stacked along the channel direction to obtain a dual-temporal remote sensing image with 20 bands. This complete remote sensing image was then divided into patches of size 256 × 256, with patches located outside the Lanling County area being discarded, resulting in a total of 629 image patches. The proportion of wheat in each remaining image patch was calculated, and data-augmentation methods such as flipping and scaling were used to enhance the patches according to the proportion. Finally, these processed patches were utilized to compile the dataset, yielding a total of 2274 image patches. The dataset was divided into training, testing, and validation sets at a ratio of 6:2:2.

Digital elevation model (DEM) data represent ground elevation in the form of an ordered array of numerical values. This enables the digital simulation of ground terrain based on limited terrain elevation data. The DEM data used in this experiment are from the second version of the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM), with a resolution of 30 m. The DEM was then upsampled with ENVI to a resolution of 10 m.

The ground-truth map of winter wheat in Lanling County was provided by Qingdao Geosea Sky Information Co., Ltd., Qingdao, China which was plotted based on actual investigation in the spring of 2018.

3.2. Settings of Experiments

The experiments were conducted on a Linux server equipped with two Nvidia GeForce RTX 3080 graphics processing units (GPUs). The Adam optimizer was used in the experiments with a learning rate of 0.0001, and the training was performed using cosine annealing. The batch size was set to 10, and a total of 200 epochs were trained. The model with the best performance on the validation set was selected for testing on the test set. The input image size for the network was set to 256 × 256.

The network employs a cross-entropy loss function. In the experiments, the number of classification categories is represented by M, and the number of samples is denoted by N. The formula for the cross-entropy loss function is provided below:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c}\log(p_{i,c}) \tag{5}$$

$y_{i,c}$ is the label of the i-th sample category c, whose value is 0 or 1; $p_{i,c}$ is the probability of the i-th sample category c.

3.3. Evaluation Metrics

In this paper's semantic segmentation experiments, the samples were divided into two categories: winter wheat and background excluding winter wheat. In order to evaluate the semantic segmentation performance, the IOU, accuracy, and F1-score were used as evaluation indicators for the winter wheat category in the experimental results, and overall accuracy, average accuracy (AA), and mean intersection over union (MIoU) were used as evaluation indicators for the overall segmentation performance. The confusion matrix was used to calculate the above evaluation indicators. The confusion matrix is a contingency table that summarizes the prediction results of a classification model in machine learning. True Positive (TP) represents the number of positive samples predicted as positive, False Negative (FN) represents the number of positive samples predicted as negative, False Positive (FP) represents the number of negative samples predicted as positive, and True Negative (TN) represents the number of negative samples predicted as negative. Positive samples refer to the winter wheat category, and negative samples refer to the background category. Below are the formulas for each indicator:

$$F1 - score = \tfrac{2TP}{2TP+FP+FN} \tag{6}$$

$$IOU = \tfrac{TP}{FP+FN+TP} \tag{7}$$

$$mIOU = \tfrac{1}{2}(\tfrac{TP}{FP+FN+TP} + \tfrac{TN}{FP+FN+TN}) \tag{8}$$

$$OA = \tfrac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

$$AA = \tfrac{1}{2}(\tfrac{TP}{TP+FP} + \tfrac{TN}{TN+FN}) \tag{10}$$

4. Experiment and Analysis

4.1. Comparative Tests

In this paper, we compare our proposed model with classic semantic segmentation models, such as the UnetFormer [29], DeepLabv3+ [30], Segnet [25], Pspnet [18], Upernet [31], and (MS)2-Net [32] models. Among them, DeepLabv3+, Segnet, Pspnet, and Upernet are single-modality networks that take the dual-temporal remote sensing data and DEM data concatenated along the channel dimension as input, with a total of 21 input channels. (MS)2-Net is a dual-modality network specifically designed for remote-sensing data segmentation, where the 20-channel dual-temporal data and 1-channel DEM data are fed into two separate branches of the network.

The (MS)2-Net is a new network designed for segmenting multi-modal remote sensing data. It uses a multi-stage fusion module to combine different types of information and a multi-source attention module to enhance the discriminability of features from different modalities. As shown in Table 1, (MS)2-Net achieved an F1-score of 92.34% for the winter wheat category. DeepLabv3+, a classic semantic segmentation network introduced in 2017, addresses the challenge of segmenting objects at multiple scales. It utilizes an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale contextual information and incorporates a decoder module, inspired by the U-Net architecture, for up-sampling to refine the edge precision. When applied to the semantic segmentation of winter wheat in remote sensing images, DeepLabv3+ achieves an F1-score of 89.80% for the winter wheat category. The Pyramid Scene Parsing Network (PSPNet) aggregates contextual information from different regions, enabling the model to understand global contextual information. For winter wheat segmentation, we selected DenseNet as the backbone for PSPNet and achieved an F1-score of 86.54% for the winter wheat category. SegNet, proposed in 2015, consists of an encoder and a decoder. In this study, we used ResNet as the backbone for winter wheat segmentation and achieved an F1-score of 86.45% for the winter wheat category. unetFormer introduced a Global–Local Transformer Block (GLTB) based on transformers, which efficiently captures global and local contextual information for the real-time semantic segmentation of urban scenes. When applied to winter wheat segmentation, unetFormer achieved an F1-score of 90.39% for the winter wheat category. UPerNet is a unified perceptual resolution network for scene understanding based on the Feature Pyramid Network (FPN). When used for winter wheat segmentation, UPerNet achieved an F1-score of 90.30% for the winter wheat category.

However, our proposed CM-Net network, specifically designed for winter wheat semantic segmentation, outperformed the other models with a segmentation F1-score of 93.55%. In addition to the F1-score, CM-Net also performed the best in IOU and F1-score for the winter wheat category, and its overall segmentation evaluation metrics such as OA and MIOU were superior to the other networks.

The segmented results of the CM-Net and other comparison networks on the test set are shown in Figure 4, primarily depicting the winter wheat predictions in mountainous and hilly regions. It can be observed that, overall, the CM-Net network achieved the best prediction performance for sparsely distributed winter wheat, followed by (MS)2-Net. CM-Net effectively captured various details in wheat distribution, while (MS)2-Net also performed well. On the other hand, unetFormer, DeepLabv3+, and UPerNet could predict the general distribution of winter wheat but exhibited average performance in terms of detail prediction. SegNet and PSPNet, however, demonstrated poor performance in segmenting details, with evident instances of both false-positive and false-negative detections in the prediction results.

Table 1. Comparison of our proposed model with classic semantic segmentation models.

Model	**Overall**			**Winter Wheat**	
	OA	**AA**	**MIoU**	**IoU**	**F1**
Ours	**94.67**	**94.52**	**89.60**	**87.88**	**93.55**
(MS)2-Net [32]	93.66	93.47	87.76	85.77	92.34
unetFormer [29]	92.04	91.77	84.87	82.47	90.39
UperNet [31]	91.98	91.73	84.75	82.31	90.3
DeepLabv3p [30]	91.57	91.31	84.04	81.50	89.8
SegNet [25]	88.73	88.34	79.27	76.13	86.45
PSPNet [18]	88.67	88.21	79.22	76.27	86.54

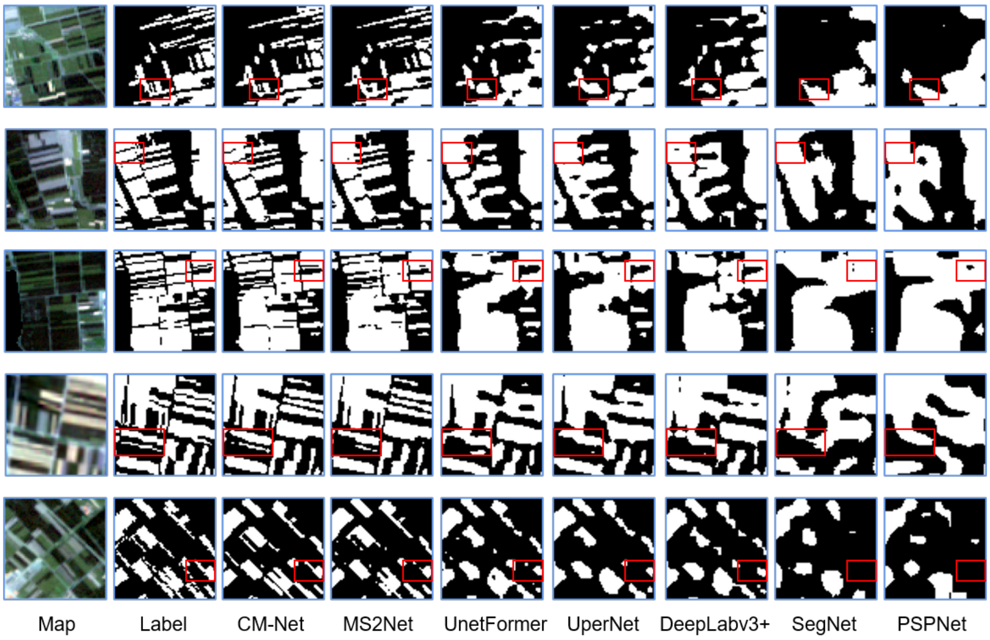

Figure 4. From left to right in the figure are the original multispectral remote sensing image, the ground truth map of wheat distribution (the white area represents winter wheat, and the black area represents other land cover types), and the results predicted by CM-Net, (MS)2Net, UnetFprmer, UperNet, DeepLabv3+, SegNet, and PSPNet. The red squares highlight the visual effects of different methods.

4.2. Ablation Experiments

To verify the effectiveness of each module, control variable ablation experiments were conducted, and the results are shown in Table 2. Experiments CM-Net, Net1, Net2, and Net3 verified the effectiveness of each module mentioned in Section 2 under the condition of inputting dual-temporal remote sensing data and DEM data. CM-Net is the proposed network, Net1 and Net2 were compared without and with DRF, and Net3 is without adding the DWA module. The experiment showed that DRF performed slightly better. The comparison between Net1 and Net3, as well as CM-Net and Net2, verified that the DWA module helps to improve the semantic segmentation effect of winter wheat.

The experimental results of CM-Net, Net1, Net2, and Net3 showed that the model using the DRF module in the encoding stage and the DWA module in the decoding stage performed the best. Experiments Net4 and Net5 continued to explore the impact of different data on the task based on CM-Net. It can be seen that the segmentation performance of dual-temporal data plus DEM data was the best.

Figure 5 shows the prediction results of winter wheat distribution in a part of Lanling County from ablation experiments CM-Net to Net5. From left to right, Figure 5 shows the remote sensing image of a certain area in Lanling County in 2018, the ground truth map of winter wheat distribution in that area, and the predicted winter wheat distribution results of experiments CM-Net to Net5 in that area. The white area in the figure represents winter wheat, and the black area represents other land cover types.

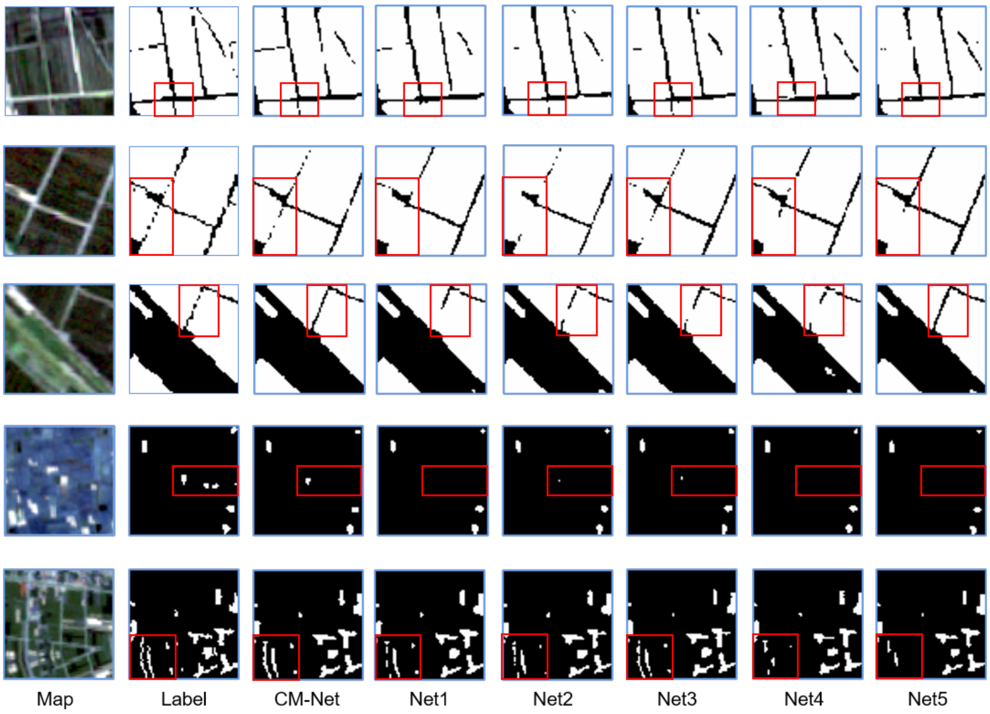

Figure 5. From left to right in the figure are the original multispectral remote sensing image, the ground truth map of wheat distribution (the white area represents winter wheat and the black area represents other land cover types), and the prediction results of experiments CM-Net, Net1, Net2, Net3, Net4, and Net5. The red squares highlight the visual effects of different methods.

Table 2. The results of various ablation experiments.The bold font indicates optimal precision. "o" indicates the modules or data contained in the current model.

	Data			Model		Overall			Winter Wheat	
	Single-Temporal + Dem	**Bi-Temporal**	**Bi-Temporal + Dem**	**DRF**	**DWA**	**OA**	**mIOU**	**AA**	**IoU**	**F1**
CM-Net			o	o	o	**94.67**	**89.60**	**94.52**	**87.88**	**93.55**
Net1			o			94.52	89.31	94.35	87.56	93.37
Net2			o	o		94.56	89.39	94.40	87.65	93.42
Net3			o		o	94.58	89.44	94.43	87.70	93.45
Net4		o		o	o	93.93	88.25	93.72	86.35	92.68
Net5	o			o	o	93.93	88.24	93.75	86.33	92.66

4.2.1. Ablation for the DRF

To verify the effectiveness of DRF modular, we replaced the deformable convolution block with a regular convolution module, resulting in the same receptive filed in multi-modal data. Experiments Net1 and Net2 compare the performance of the DRF module without the Dynamic Weighted Aggregation (DWA) module. In experiment Net2, the prediction results for the winter wheat category show improvements 0.09 and 0.05 percentage points in terms of intersection over union (IOU) and F1-score, respectively, compared to Net1.

Experiments CM-Net and Net3 involve comparing the encoder structures based on the DRF, with the addition of the DWA module at the decoder stage. In experiment CM-Net, the prediction results for the winter wheat category show improvements 0.18 and 0.1 percentage points in the IOU and F1-score, respectively, compared to Net3. The experimental results demonstrate that incorporating diverse receptive fields in the multi-modal structure yields better results than using the same receptive filed.

In Figure 5, the top two rows depict narrow roads. In experiment Net1, several road areas were not segmented properly, while in experiment Net2 with the DRF module, the segmentation performance improved, allowing for the recognition of more road information compared to Net1. The comparison of segmentation results between CM-Net and Net3 also indicates that the DRF module can identify more detailed information.

4.2.2. Ablation for the DWA

In experiments Net1 and Net3, two experiments were conducted in the decoder stage, one without adding the DWA module and one with adding the DWA module. The comparison of the experimental results of Net1 and Net3 showed the effect of adding the DWA module. The overall and winter wheat category segmentation performance of Net3 was better than that of Net1. In terms of the segmentation performance of the winter wheat category, the IOU and F1 indicators of Net3 were respectively 0.14% and 0.08% higher than those of Net1.

In experiments CM-Net and Net2, the DRF was used in the encoder part, and two experiments were conducted in the decoder stage, one without adding the DWA module and one with adding the DWA module. In experiment CM-Net, the segmentation effect was better, and the IOU and F1 indicators of the winter wheat category's segmentation results were respectively 0.23% and 0.13% higher than those of Net2.

From Figure 5, it can be observed that experiment CM-Net had the best overall segmentation effect compared to the other experiments shown in the figure. Figure 5 shows that CM-Net identified road information that was not identified by Net2 and performed better for relatively fragmented wheat distributions.

4.2.3. Ablation for the Data Source

Experiments CM-Net, Net4, and Net5 were conducted using the same network structure but with different dataset configurations. CM-Net used dual-temporal remote sensing data and DEM data, with DEM data as input to the elevation branch and dual-phase remote sensing data as input to the optical branch. Net4 only used dual-phase data, with 2017 remote sensing data as input to the elevation branch and 2018 remote sensing data as input to the optical branch. Finally, the dual-phase data of 2017 and 2018 were input to the network. Net5 used single-phase data and DEM data, with DEM data as input to the elevation branch and 2018 remote sensing data as input to the optical branch.

Experiments CM-Net and Net4 in Table 1 verified the contribution of multimodal data to the task. In terms of winter wheat category segmentation results, the IOU and F1 indicators of the CM-Net were respectively 1.53% and 0.87% higher than those of Net4. Experiments CM-Net and Net5 verified the advantages of dual-phase data compared to single-phase data. In terms of winter wheat category segmentation results, the IOU and F1 indicators of the CM-Net were respectively 1.55% and 0.89% higher than those of Net5. The experimental results show that the combination of dual-phase remote sensing data and DEM multimodal data in CM-Net had the best performance for winter wheat semantic segmentation.

From the predicted results shown in Figure 5, it can be observed that there were misclassifications or omissions in the segmentation results of Net4 using only dual-phase remote sensing data and Net5 using single-phase remote sensing data and DEM data. The performance of Net4 and Net5 was inferior to that of experiment CM-Net.

5. Conclusions

In this paper, we introduce a deep network called CM-Net that can segment objects using cross-modal multi-temporal optical images and DEM images. The network achieves excellent segmentation results by leveraging two key aspects:

(1) The Diverse Receptive Fusion (DRF) module is proposed. This module applies a deformable receptive field to optical images during the feature fusion process. It allows the network to match winter wheat plots of varying scales by adapting to

their characteristics using a fixed receptive field for DEM images. This enables the extraction of evaluation features at a consistent scale, accommodating the dual-modal data.

(2) The distributed weighted attention module (DWA) has been meticulously engineered to optimize feature intensity during the crucial decoding phase of winter wheat segmentation. By integrating sophisticated global pooling techniques with a broadened scope of activation functions, the DWA adeptly enhances the salience of essential features specific to winter wheat. Concurrently, it effectively diminishes the presence of irrelevant features. This dual capability of selective enhancement and suppression is vital for accurately extracting minute yet significant details from small planting areas.

Comparative experiments and ablation studies demonstrated that CM-Net exhibits strong competitive performance and generalization capabilities.

Author Contributions: Conceptualization, N.W. and Q.H.; methodology, Q.H.; software, Y.G.; validation, Q.W.; writing—original draft preparation, Q.H. and N.W.; writing—review and editing, Q.W.; visualization, Q.W.; supervision, W.L.; funding acquisition, N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science for a Better Development ofInner Mongolia Program, grant number 2022EEDSKJXM003-2.

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

References

1. Van Tricht, K.; Gobin, A.; Gilliams, S.; Piccard, I. Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: A case study for Belgium. *Remote. Sens.* **2018**, *10*, 1642. [CrossRef]
2. Guo, Y.; Jia, X.; Paull, D.; Benediktsson, J.A. Nomination-favoured opinion pool for optical-SAR-synergistic rice mapping in face of weakened flooding signals. *ISPRS J. Photogramm. Remote. Sens.* **2019**, *155*, 187–205. [CrossRef]
3. Wu, M.; Huang, W.; Niu, Z.; Wang, Y.; Wang, C.; Li, W.; Hao, P.; Yu, B. Fine crop mapping by combining high spectral and high spatial resolution remote sensing data in complex heterogeneous areas. *Comput. Electron. Agric.* **2017**, *139*, 1–9. [CrossRef]
4. Hao, P.; Zhan, Y.; Wang, L.; Niu, Z.; Shakir, M. Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. *Remote. Sens.* **2015**, *7*, 5347–5369. [CrossRef]
5. Zheng, B.; Myint, S.W.; Thenkabail, P.S.; Aggarwal, R.M. A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 103–112. [CrossRef]
6. Yang, G.; Li, X.; Liu, P.; Yao, X.; Zhu, Y.; Cao, W.; Cheng, T. Automated in-season mapping of winter wheat in China with training data generation and model transfer. *ISPRS J. Photogramm. Remote. Sens.* **2023**, *202*, 422–438. [CrossRef]
7. Zhong, L.; Gong, P.; Biging, G.S. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. *Remote. Sens. Environ.* **2014**, *140*, 1–13. [CrossRef]
8. Tian, H.; Zhou, B.; Chen, Y.; Wu, M.; Niu, Z. Extraction of winter wheat acreage based on GF-1 PMS remote sensing image on county scale. *J. China Agric. Univ.* **2017**. [CrossRef]
9. Tian, H.; Huang, N.; Niu, Z.; Qin, Y.; Pei, J.; Wang, J. Mapping winter crops in China with multi-source satellite imagery and phenology-based algorithm. *Remote. Sens.* **2019**, *11*, 820. [CrossRef]
10. Li, S.; Huo, L. Remote sensing image change detection based on fully convolutional network with pyramid attention. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4352–4355.
11. Guo, J.; Ren, H.; Zheng, Y.; Nie, J.; Chen, S.; Sun, Y.; Qin, Q. Identify urban area from remote sensing image using deep learning method. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7407–7410.
12. Alp, G.; Sertel, E. Deep learning based patch-wise land cover land use classification: A new small benchmark sentinel-2 image dataset. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3179–3182.
13. Garnot, V.S.F.; Landrieu, L.; Giordano, S.; Chehata, N. Satellite image time series classification with pixel-set encoders and temporal self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12325–12334.
14. Conrad, C.; Dech, S.; Dubovyk, O.; Klein, S.F.D.; Löw, F.; Schorcht, G.; Zeidler, J. Derivation of temporal windows for accurate crop discrimination in heterogeneous croplands of Uzbekistan using multitemporal RapidEye images. *Comput. Electron. Agric.* **2014**, *103*, 63–74. [CrossRef]

15. Rußwurm, M.; Korner, M. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
16. Tarasiou, M.; Güler, R.A.; Zafeiriou, S. Context-self contrastive pretraining for crop type semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–17. [CrossRef]
17. Zhang, D.; Pan, Y.; Zhang, J.; Hu, T.; Zhao, J.; Li, N.; Chen, Q. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. *Remote. Sens. Environ.* **2020**, *247*, 111912. [CrossRef]
18. 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.
19. Wei, P.; Chai, D.; Lin, T.; Tang, C.; Du, M.; Huang, J. Large-scale rice mapping under different years based on time-series Sentinel-1 images using deep semantic segmentation model. *ISPRS J. Photogramm. Remote. Sens.* **2021**, *174*, 198–214. [CrossRef]
20. Ma, X.; Huang, Z.; Zhu, S.; Fang, W.; Wu, Y. Rice Planting Area Identification Based on Multi-Temporal Sentinel-1 SAR Images and an Attention U-Net Model. *Remote. Sens.* **2022**, *14*, 4573. [CrossRef]
21. Sai, G.U.; Tejasri, N.; Kumar, A.; Rajalakshmi, P. Deep learning based overcomplete representations for paddy rice crop and weed segmentation. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6077–6080.
22. Garnot, V.S.F.; Landrieu, L.; Chehata, N. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS J. Photogramm. Remote. Sens.* **2022**, *187*, 294–305. [CrossRef]
23. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [CrossRef]
24. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Proceedings of the Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part I 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 213–228.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
26. Zhang, P.; Du, P.; Lin, C.; Wang, X.; Li, E.; Xue, Z.; Bai, X. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data. *Remote. Sens.* **2020**, *12*, 3764. [CrossRef]
27. Vos, J.; Heuvelink, E. Concepts to model growth and development of plants. In Proceedings of the 2006 Second International Symposium on Plant Growth Modeling and Applications, Beijing, China, 13–17 November 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 3–10.
28. Tian, Z.; Gao, Z.; Xu, Y.; Chen, H. Impacts of climate change on winter wheat production in China. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium, Seoul, Republic of Korea, 29 July 2005; IGARSS'05; IEEE: Piscataway, NJ, USA, 2005; Volume 1, p. 4.
29. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote. Sens.* **2022**, *190*, 196–214. [CrossRef]
30. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
31. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
32. Zhao, J.; Zhou, Y.; Shi, B.; Yang, J.; Zhang, D.; Yao, R. Multi-stage fusion and multi-source attention network for multi-modal remote sensing image segmentation. *ACM Trans. Intell. Syst. Technol. (TIST)* **2021**, *12*, 1–20. [CrossRef]

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