remote sensing

MDPI

Article

Dynamic Convolution Self-Attention Network for Land-Cover Classification in VHR Remote-Sensing Images

Xuan Wang [1], Yue Zhang [2,3], Tao Lei [2,3,*], Yingbo Wang [2,3], Yujie Zhai [2,3] and Asoke K. Nandi [4,5]

[1] Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA
[2] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China
[3] School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China
[4] Department of Electronic and Electrical Engineering, Brunel University, London UB8 3PH, UK
[5] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: leitao@sust.edu.cn

Abstract: The current deep convolutional neural networks for very-high-resolution (VHR) remote-sensing image land-cover classification often suffer from two challenges. First, the feature maps extracted by network encoders based on vanilla convolution usually contain a lot of redundant information, which easily causes misclassification of land cover. Moreover, these encoders usually require a large number of parameters and high computational costs. Second, as remote-sensing images are complex and contain many objects with large-scale variances, it is difficult to use the popular feature fusion modules to improve the representation ability of networks. To address the above issues, we propose a dynamic convolution self-attention network (DCSA-Net) for VHR remote-sensing image land-cover classification. The proposed network has two advantages. On one hand, we designed a lightweight dynamic convolution module (LDCM) by using dynamic convolution and a self-attention mechanism. This module can extract more useful image features than vanilla convolution, avoiding the negative effect of useless feature maps on land-cover classification. On the other hand, we designed a context information aggregation module (CIAM) with a ladder structure to enlarge the receptive field. This module can aggregate multi-scale contexture information from feature maps with different resolutions using a dense connection. Experiment results show that the proposed DCSA-Net is superior to state-of-the-art networks due to higher accuracy of land-cover classification, fewer parameters, and lower computational cost. The source code is made public available.

Keywords: land-cover classification; feature fusion; self-attention; lightweight

check for updates

Citation: Wang, X.; Zhang, Y.; Lei, T.; Wang, Y.; Zhai, Y.; Nandi, A.K. Dynamic Convolution Self-Attention Network for Land-Cover Classification in VHR Remote-Sensing Images. *Remote Sens.* **2022**, *14*, 4941. https://doi.org/10.3390/rs14194941

Academic Editor: Dino Ienco

Received: 24 July 2022
Accepted: 26 September 2022
Published: 3 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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/).

1. Introduction

Very-high-resolution (VHR) remote-sensing image land-cover classification refers to the process of identifying land objects according to spectral, texture, shape, and other characteristics of the objects in remote-sensing images. Nowadays, VHR remote-sensing images can clearly express the spatial structure and surface texture characteristics of ground objects and provide the conditions and basis for geoscience interpretation and analysis; they are widely used in tasks such as land-cover classification in complex scenes [1–3].

In the early VHR remote-sensing image land-cover classification methods, most were based on traditional threshold segmentation algorithm and clustering algorithms. For example, Andres et al. [4] used a threshold segmentation method along with the classifier-extraction method to analyze remote-sensing images. Zanottta et al. [5] proposed a segmentation algorithm based on regional growth to improve the classification accuracy of remote-sensing images. In addition, other methods such as fuzzy c-means clustering [6], support vector machine [7], and random forest [8] are widely used in VHR remote-sensing image land-cover classification tasks. Although the above methods realized the classification of remote-sensing images to a certain extent, they rely on manual feature extraction, thus these algorithms are usually susceptible to noise, having poor robustness and limited practical value.

In recent years, thanks to the rapid development of deep learning, convolutional neural networks (CNNs) have shown great success in image feature representation [9], which plays an important role in computer vision fields such as target detection, image segmentation, image recognition, visual reconstruction, etc. Benefiting from the emergence of CNNs, Long et al. proposed a full convolutional neural network (FCN) for image semantic segmentation [10]. Based on this, a large number of improved semantic segmentation networks have emerged, such as SegNet [11], U-Net [12], PSPNet [13], and DeepLab [14]. Some scholars have applied these improved networks to the task of VHR remote-sensing image land-cover classification, and these networks have shown excellent performance compared with traditional algorithms [15–17].

Inspired by the human visual system, the attention mechanism is widely used in neural networks since it can suppress irrelevant features and enhance sensitive features. The mainstream attention mechanism can be divided into channel attention, spatial attention, combined attention, and non-local attention. Channel or spatial attention combines pooling operation with the fully connected network to obtain the weight coefficient on the channel dimension or spatial dimension and realize feature weightings, such as SENet [18], GatherExcite [19], and GCNet [20]. In combined attention, channel attention and spatial attention are associated in series or parallel sequences to strengthen the collaborative function between channels and space, such as DANet [21] and CBAM [22]. Self-attention realizes global feature correlation modeling by learning the long-range dependency between different positions of feature maps, such as non-local and transformer. Owning to the advantages of the attention mechanism, scholars applied the attention mechanism to VHR remote-sensing image land-cover classification and proposed SCAttNet [23], TCHD-Net [24], and SCViT [25]. These networks improve the accuracy of VHR remote-sensing image land-cover classification to varying degrees compared with traditional CNNs.

Although many deep learning models have been proposed and applied to VHR remote-sensing image land-cover classification, they are still facing some problems. Firstly, due to the complexity of the high-resolution remote sensing scene, the image features extracted by the conventional convolution encoder usually have a large amount of redundant information, which will lead to incorrect classification results. At the same time, the existing encoder models are usually large-scale and have high calculation costs. Secondly, due to the complexity of high-resolution remote sensing scenes and the large-scale difference between targets, the existing feature-fusion module is difficult to effectively improve the feature representation ability of networks. Thirdly, the existing skip connection operations often ignore the differences between the features of different layers and cannot explore enough useful information from a full-scale perspective, thus cannot accurately determine the location and boundary information of the target.

To solve the aforementioned problems, in this paper we propose a dynamic convolution self-attention network (DCSA-Net) for VHR remote-sensing image land-cover classification. DCSA-Net adopts an encoder-decoder network with the residual structure as the backbone. The encoder mainly consists of two components: the lightweight dynamic convolution module (LDCM) and the context information aggregation module (CIAM). For the LDCM, we propose two strategies (LDCM_v1 and LDCM_v2) for single-mode feature fusion and multi-mode feature fusion, respectively. In LDCM_v1, we use dynamic convolution to replace traditional convolution, reduce feature redundancy, and improve network representation ability, In LDCM_v2, we combine dynamic convolution with self-attention. Introducing self-attention can make up for the defect of poor semantic information, improve the semantic representation ability, and optimize the feature map by using long-range dependency. In the CIAM module, the context information extracted through pooling windows of different sizes and deeper features are combined in a hierarchical residual fusion method to form more abundant multi-scale context information. In addition, from the perspective of full-scale capture of fine-grained details and coarse-grained semantics, we also propose a full-scale feature interaction strategy, which can transfer more abundant encoder information to the decoder, effectively improving the fusion effect of high-resolution and low-resolution semantic features. The main contributions of this paper are summarized as follows:

- We design a lightweight dynamic convolution module, which combines dynamic convolution with self-attention to extract more useful feature information at a lower cost and avoid the negative impact of redundant features on classification results.
- We design a ladder-shaped context information aggregation module, which can effectively expand the receptive field, fully integrate the multi-scale context information of different resolution feature maps, and effectively solve the problems of fuzzy target contour and large-scale changes in the remote-sensing image scene.
- We propose a full-scale multi-modal feature fusion strategy to maximize the effective fusion of high-level and low-level features, to obtain more accurate location and boundary information.

The rest of this paper is arranged as follows. The second section mainly introduces the related work. The third section describes the proposed method in detail. In the fourth section, the experiment carried out is described, and the experimental results are analyzed and discussed in detail. The experimental results and discussion of key issues are reported in the fifth section. In the sixth section, there is a discussion of the whole paper. The code is available at https://github.com/Julia90/DCSA-Net (accessed on 25 September 2022).

2. Related Work

2.1. Lightweight Network for Land-Cover Classification

At present, most methods in the field of computer vision are based on large-scale models to achieve high accuracy. However, it is not sensible to simply pursue high accuracy while ignoring the computational cost. Therefore, a series of lightweight convolutional neural networks have been proposed. For example, SqueezeNet [26] replaced part of 3 × 3 convolution kernels with 1×1 convolution while reducing the number of input channels to reduce the number of network parameters. ERFNet [27] introduced alternative thinking that used an asymmetric convolution to decompose standard 3 × 3 convolution into 1 × 3 convolution and 3 × 1 convolution to further reduce the network parameters. To capture the spatial correlation and cross-channel correlation of the feature maps respectively, MobileNet [28] proposed the depthwise separable convolution to decompose the vanilla convolution into depthwise convolution and pointwise convolution, further reducing the consumption of vanilla convolution. Inspired by asymmetric convolution and depth separable convolution, Lv et al. [29] proposed a novel lightweight network for VHR remote-sensing image land-cover classification. It uses an asymmetric depthwise separable convolution to replace the vanilla convolution to reduce network parameters and the experimental results fully show the effectiveness of this network. Inspired by the group convolution of MobileNet, ShuffleNet [30] proposed a channel shuffle operation, which can alleviate the problem of information loss caused by the lack of information exchange between channels. Based on the idea of ShuffleNet, Qiao et al. [31] designed a lightweight network for VHR remote-sensing image land-cover classification, which is based on channel attention combined with the modified ShuffleNet unit. Han et al. [32] pointed out that there is a lot of redundancy in generating rich feature maps by neural networks, thus they proposed GhostNet. This uses a series of linear transformations instead of vanilla convolution operations to generate many 'Ghost' feature maps with a reduced number of operations. Paoletti et al. [33] designed another lightweight CNN architecture and applied the modified Ghost module to the task of VHR remote-sensing image land-cover classification to achieve high-precision feature classification at a smaller cost. Additionally, in order to improve land-cover classification accuracy, Cao et al. [34] explored multiple lightweight multimodality fusion methods to achieve high accuracy and low computational complexity. Peng et al. [35] combined the idea of dense connection and fully convolutional networks to provide fine-grained semantic segmentation. Due to its dense architecture, its model complexity is proportional to the small-valued growth rate and layer number. Thus, this method achieves lower time complexity than conventional CNN. Ferrari et al. [36] proposed replacing the VGG-like encoders with the lightweight EfficientNet in a hybrid attention-aware fusion network to ensure the prediction accuracy and computation efficiency on the task of building mapping.

Although convolution operation can fuse global information by stacking multiple convolution layers, the receptive field of the convolution kernel is still limited and only focuses on the local area, which may lead to weak feature representation ability. However, the attention mechanism can effectively obtain the global relationship of the feature map, and improve the representation ability of the network from the global perspective, effectively improving the land-cover classification accuracy. For example, Li et al. [37] proposed a multi-level attention network with linear complexity to reduce a large number of computing requirements in attention operations. This design makes the combination of the attention mechanism and neural network more flexible and universal. Zhang et al. [38] improved the self-attention mechanism and proposed a lightweight double branch network with an attention module and a feature fusion module. The network can effectively reduce the interference of noise and redundant information in the feature maps while reducing the network complexity.

2.2. Feature Fusion for Land-Cover Classification

In convolutional neural networks, the input features usually go through multiple downsampling layers to gradually expand the receptive field to obtain high-level semantic features. High-level features do usually have stronger semantic information, but their resolution is usually low and poor at perceiving details. Contrary to high-level features, low-level features have high resolution and thus contain more location and detailed information, but these features go through fewer convolution layers. Thus, the semantic information is not rich and usually has serious noise interference. Because of the above phenomena, the question of how to carry out efficient image feature fusion has become a hot issue for scholars in recent years. The U-Net network proposed by Ronneberger et al. [12] combines shallow features with deep features through the idea of skip connection, which effectively improves the accuracy of the segmentation task. Liu [39] and others proposed a progressive fusion network based on the idea of skip connection of the U-Net network and applied it to the task of VHR remote-sensing image land-cover classification. Unlike U-Net, it progressively fuses features of different scales from coarse to fine, which can generate more refined fusion features, to improve the accuracy of VHR remote-sensing image land-cover classification. U-Net++ [40] redesigns the structure of skip connection based on U-Net so that the features extracted by the encoder can be more flexibly integrated into the decoder to improve the performance of the network.

Although the networks such as U-Net and U-Net++ can effectively integrate high-level and low-level features by using skip connection, thus enhancing the network's feature representation ability, they do not consider the multi-scale characteristics of the target. Therefore, when applying these networks to the VHR remote-sensing image land-cover classification tasks with large-scale differences in the target, the accuracy is often limited. The pyramid pooling module proposed by PSPNet [13] can effectively solve the problem of limited feature expression caused by large-scale changes in the target. Likewise, Wang et al. [41] proposed a multi-scale feature pyramid module to obtain the fine-grained feature maps of multi-scale global features. Though it has improved the result of land-cover classification at multiple object scales, it is a simple fusion method that has high computational complexity, resulting in low accuracy. Xu et al. [42] established different forms of feature fusion adaptive dilated space pyramid pooling module, which can not only extract and fuse more extensive multi-scale features but also reduce the degradation of model performance caused by dilated convolution. DeepLab v3+ [43] incorporates the convolutions of various dilation rates and a global average pooling into the network to obtain a multi-scale context feature. Although this method can effectively improve the accuracy, it only considers the local features and lacks the consideration of the global features. Tian et al. [44] combined the self-attention mechanism, dilated convolution, and pyramid pooling module to extract features. Afterward, the refined feature pyramid structure is used to fuse the multi-scale features. This method not only assigns adaptive weights to the local features but also considers the global. Similarly, Liu et al. [45] used a dense connectivity pattern and parallel multikernel convolution to implement various-sized receptive fields. Additionally, this uses spatial and channel relation-enhanced blocks to learn global contextual relations. Lei et al. [46] applied atrous spatial pyramid pooling with different atrous rates and a non-local block, which not only extracts and integrates multi-scale feature maps, but also enhances intra-class features by using the remote dependency of the spatial context. Shang et al. [47] employed two layers of atrous convolutions with different dilation rates and global pooling to extract multi-scale context information. To effectively fuse the semantic features, an adaptive weighted channel attention mechanism is used. Nie et al. [48] proposed a novel multiscale image generation network. It introduced a multi-attention mechanism, an edge supervised module, and a multiscale image fusion algorithm based on the Bayes model to ensure segmentation accuracy.

3. Proposed Method

3.1. Overview

DCSA-Net adopts an encoder-decoder network with the residual structure as the backbone network as shown in Figure 1. The inputs of the network are IRRG and nDSM respectively. IRRG is an image format composed of three bands of near-infrared (IR), red (R), and green (G). Compared with the RGB image format, it can provide better classification performance due to the use of near-infrared light. The digital surface model (DSM) refers to the ground elevation model including the heights of surface buildings, bridges, and trees. nDSM is a picture format after normalizing DSM. The encoder mainly consists of two components: the lightweight dynamic convolution module (LDCM) and the context information aggregation module (CIAM). For the LDCM, we propose two strategies (LDCM_v1 and LDCM_v2) for single-mode feature fusion and multi-mode feature fusion, respectively. Different from the traditional multi-modal fusion network structure, LDCM generates better feature maps at a lower cost than vanilla convolution while processing different modal data. The CIAM not only incorporates improved multi-scale contextual information but also retains the original multi-modal features. The decoder still uses a residual structure, consisting of continuous bilinear interpolation. In addition, DCSA-Net uses a full-scale feature interaction strategy to effectively improve the semantic feature fusion effect of high- and low-resolution.

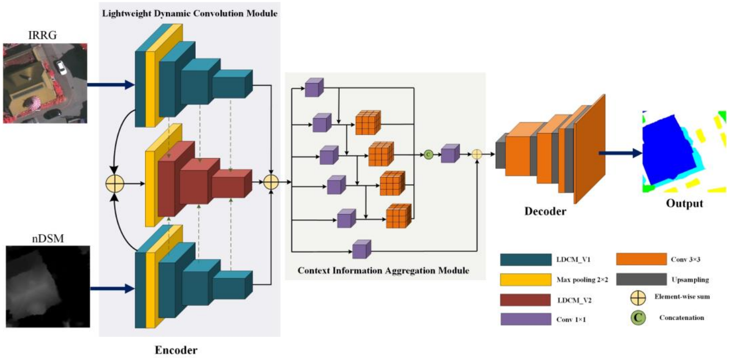


Figure 1. The overall structure of dynamic convolution self-attention network (DCSA-Net). The input comprises near-infrared, red, and green (IRRG) data and normalized digital surface model (nDSM) data. The two formats of data are encoded respectively and fused afterward by the lightweight dynamic convolution module (LDCM) to extract the useful feature. Then the feature is processed by multisize pooling kernels and then hierarchical residual fusion by the context information aggregation module (CIAM) to obtain a multiscale context feature. Finally, the feature is decoded using a full-scale feature interaction strategy to obtain the predicted output.

3.2. Lightweight Dynamic Convolution Module

To mine the image feature information deeply, we select an encoder with a three-branch structure in the feature extraction stage, including a spectral branch, a depth branch, and a fusion branch. Because the high-resolution remote sensing scene is relatively complex, the image features extracted by the encoder based on vanilla convolution are often accompanied by a large amount of redundant information, which easily causes wrong classifications, and these encoders usually need a huge number of parameters. To solve this problem, as shown in Figure 2a, we combine dynamic convolution with self-attention and design a lightweight dynamic convolution module, which can extract more useful image features at a lower cost and avoid the impact of redundant features on the final classification results.

Most of the CNNs often employ the vanilla convolution or depthwise separable convolution in the coding stage of networks. A vanilla convolution usually suffers from the problems of a large number of model parameters and a high computational complexity. Although depthwise separable convolution can effectively reduce the number of network parameters by performing depthwise convolution and pointwise convolution separately, it usually leads to the degradation of the performance of the model. No matter it's a vanilla convolution or a depthwise separable convolution, once the network training is completed, the convolution kernel parameters in the network will be fixed, and the network will process all input data equally, so adaptive feature-coding cannot be realized. In this regard, dynamic convolution realizes the adaptive coding of the network for different input data by dynamically generating a convolution kernel, which not only effectively improves the feature representation ability of the network, but also reduces the parameters of the model. Since the VHR remote-sensing image land-cover classification network is faced with the problems of complex feature types and huge network models, we use dynamic convolution to replace vanilla convolution to realize a lightweight dynamic convolution network for VHR remote-sensing image land-cover classification.

In deep convolution neural networks, the attention mechanism is a strategy widely used to improve the representation ability of models. Due to the long imaging distance of remote-sensing images, the detailed information of objects is not obvious, and the long-range correlation of pixels is particularly important in VHR remote-sensing image land-cover classification. To further improve the feature representation ability of the network, we introduce a self-attention mechanism into the network model to capture the global information of feature maps and use long-range dependency to improve the feature representation ability of the network. Since multimodal data requires a three-branch encoder, considering that if the self-attention mechanism is introduced into the three-branch encoder, the computational cost is quite high, the coupling of feature representation is strong and there are too many redundant calculations. Therefore, to balance the contradiction between the computational cost and the feature representation ability, we do not introduce the self-attention mechanism to the single-mode branch and the deep branch, but introduce the self-attention mechanism to the fusion branch, because the fusion branch can take into account all the features of the single-mode branch and the deep branch, so it contains more abundant feature information.

To sum up, we have designed two structures in the LDCM module, namely LDCM_v1 and LDCM_v2, using LDCM_v1 for single-mode branches and deep branches, use LDCM_v2 for fused branches. The structures of LDCM_v1 and LDCM_v2 are shown in Figure 2b,c.

In LDCM_v1, we have introduced dynamic convolution to replace the vanilla convolution, reduce feature redundancy information, improve the feature representation ability of the network, and reduce the parameters of the model. The dynamic convolution calculation process is mainly divided into two parts. First, the dynamic convolution kernel is generated. All channel pixels at a certain spatial position of the input image are selected, which are expanded by a linear transformation to obtain the dynamic convolution kernel $K \times K \times 1$. Then the convolution is calculated. The generated dynamic convolution kernel is first duplicated and stacked into $K \times K \times C$. Then, it is multiplied by the original feature map at the original pixel position to complete the dynamic convolution calculation. In Figure 3, we show the structure of LDCM_v2. The input image is first subjected to the dynamic convolution. Then, to capture the long-range dependence of the feature map, the feature map generated from dynamic convolution is respectively passed through the self-attention channel module and the self-attention space module. Finally, the outputs of the two branches are added together to obtain the final result.

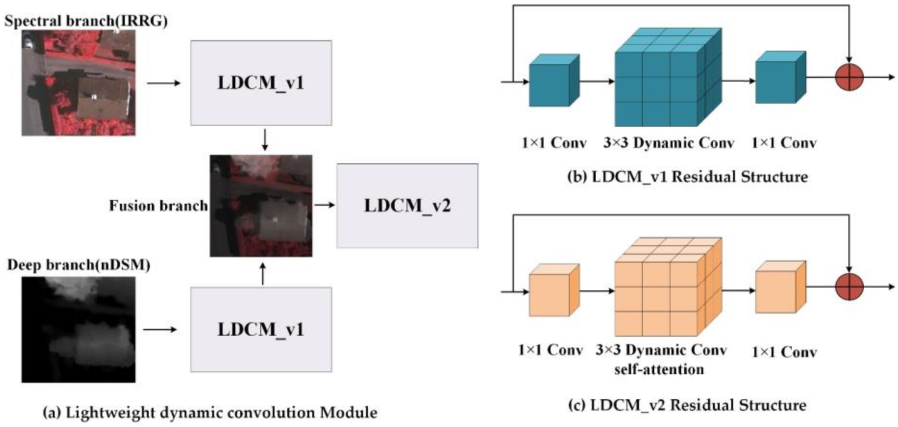

Spectral branch(IRRG)
LDCM_v1
Fusion branch
LDCM_v2
Deep branch(nDSM)
LDCM_v1
(a) Lightweight dynamic convolution Module
1×1 Conv
3×3 Dynamic Conv
1×1 Conv
(b) LDCM_v1 Residual Structure
1×1 Conv
3×3 Dynamic Conv self-attention
1×1 Conv
(c) LDCM_v2 Residual Structure

Figure 2. Lightweight dynamic convolution module.

The principle of self-attention channel module is denoted by $A^{ch}(X) \in R^{C\times1\times1}$, as shown in Equation (1).

$$A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X))))\right] \tag{1}$$

where W_z, W_v and W_q are 1×1 convolution, σ_1 and σ_2 are two-dimensional transformation operations, F_{SM} and F_{SG} represent $Softmax$ function and $Sigmoid$ function, respectively, and $\times$ is matrix dot product operation. To reduce the computation cost, the number of the internal channels W_z, W_v and W_q is set to $C/2$. The output of the self-attention pure channel filtering module is $Z^{ch} = A^{ch}(X) \otimes X \in R^{C\times H\times W}$ and $\otimes$ is the multiplication of the channel dimension.

The principle of self-attention space module is denoted by $A^{sp}(X) \in R^{1\times H\times W}$, which is defined as:

$$A^{sp}(X) = F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{AP}(W_q(X))))) \times \sigma_2(W_v(X)))] \tag{2}$$

where σ_3 is the dimension transformation operation and F_P is the average pooling operation. The output of the self-attention space filtering module is $Z^{sp} = A^{sp}(X) \otimes X \in R^{C\times H\times W}$, and $\otimes$ is the multiplication operation on the spatial dimension. The detailed module is shown in Figure 3.

The above two modules are combined to form a self-attention module (SFM), as shown in Equation (3), where + refers to pixel-level addition operations.

$$SFM(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \otimes X + A^{sp}(X) \otimes X \tag{3}$$

H × W, the dynamic convolution is calculated for each spatial location: the pixel value of all the channels at one spatial location is linearly transformed and reshaped into a dynamic convolution kernel of dimension $K \times K \times 1$, and then the feature map of each channel at this location is multiplied by this kernel, generating a new feature with dimension $K \times K \times C$. This new feature generated from dynamic convolution is then processed by the channel module and the spatial module using self-attention respectively. W_z, W_v and W_q are all 1×1 convolutions to construct corresponding attentions. The outputs from the channel and the spatial modules are added together to generate the result.

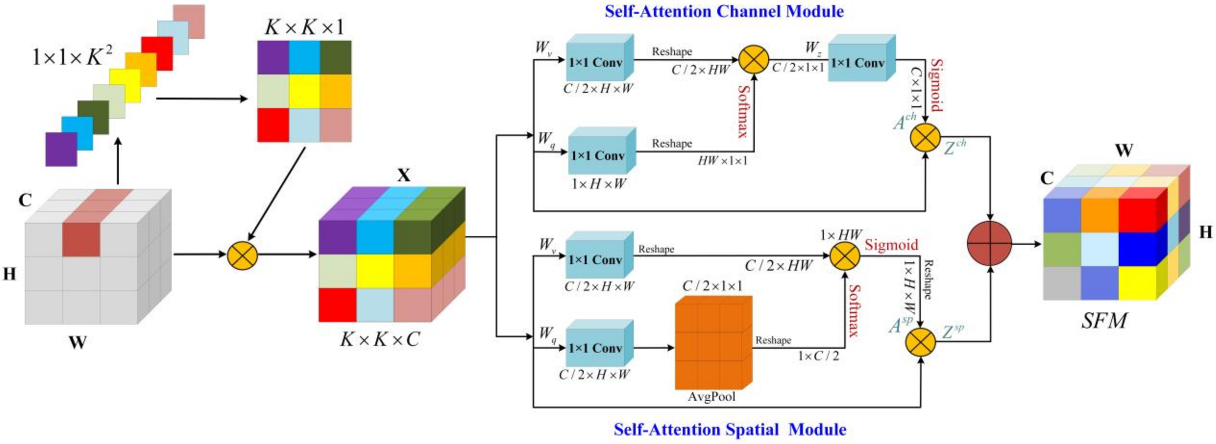


Figure 3. The structure of LDCM_v2. For an input data with dimension C × H × W, the dynamic convolution is calculated for each spatial location: the pixel value of all the channels at one spatial location is linearly transformed and reshaped into a dynamic convolution kernel of dimension $K \times K \times 1$, and then the feature map of each channel at this location is multiplied by this kernel, generating a new feature with dimension $K \times K \times C$. This new feature generated from dynamic convolution is then processed by the channel module and the spatial module using self-attention respectively. W_z , W_v and W_q are all 1×1 convolutions to construct corresponding attentions. The outputs from the channel and the spatial modules are added together to generate the result.

It can be seen from the LDCM module that we have used dynamic convolution instead of the vanilla convolution, which effectively reduces the parameters of the network. Although dynamic convolution can enhance the representation ability of spatial features, because dynamic convolution is still a local operation, it is difficult to obtain the global features of the image, which is not conducive to the classification of semantic information. The introduction of self-attention mechanism can just make up for the defect of poor semantic information and use long-range dependence to improve the ability of semantic expression and optimize the feature map.

3.3. Context Information Aggregation Module

In VHR remote-sensing images, targets usually have large-scale differences. Large-scale targets may exceed the receptive field of the full convolution neural network, resulting in discontinuous prediction. Many mispredicted areas are related to the context and the global information of different receptive fields. Therefore, it is very important to integrate effectively the context information into the VHR remote-sensing image land-cover classification. However, PSP and ASPP, which use large-scale pooling or convolution operations, will lose some detailed information. At the same time, the dilated convolution expansion rate is large, which may cause the checkerboard effect and high computational complexity. To solve the above problems, we propose a novel context information aggregation module, which can obtain very rich context information, enhance the integrity between target edges and internal tightness, and improve the accuracy of VHR remote-sensing image land-cover classification.

The specific structure of the CIAM module is shown in Figure 4. In this module, the input first passes through the average pooling operation and adaptive average pooling operation with steps of 1, 2, and 4, respectively, in parallel to generate feature maps with different resolutions. Next, the generated feature maps are inputted into a 1 × 1 convolution to change the number of channels. The 1 × 1 convolution fuses channel information, making the whole feature maps combine local and global information, but the above operation is insufficient for the aggregation of multi-scale context information. Inspired by Res2Net [49], we first upsample the feature maps outputted by 1 × 1 convolution layers, and then a larger receptive field is obtained through 3 × 3 convolution. Finally, the receptive field regions of different scales are fused by the fusion of hierarchical residuals. Due to the combined effect, more characteristic scales are generated. Finally, all feature maps are concatenated and compressed using 1 × 1 convolution. In addition, the module also adds a 1 × 1 convolution to optimize the feature-mapping. If the input is denoted by X, y_i at different scales is defined as:

$$y_i = \begin{cases} C_{1\times1}(X) & i = 1; \\ C_{3\times3}(U(C_{1\times1}(P_{AP}(X))) + y_{i-1}) & 1 < i < 5; \\ C_{3\times3}(U(C_{1\times1}(P_{AAP}(X))) + y_{i-1}) & i = 5. \end{cases} \tag{4}$$

where $C_{1\times1}$ and $C_{3\times3}$ represent 1 × 1 convolution and 3 × 3 convolution, and U represents upsampling. P_{AP} represents average pooling, and P_{AAP} represents adaptive average pooling.

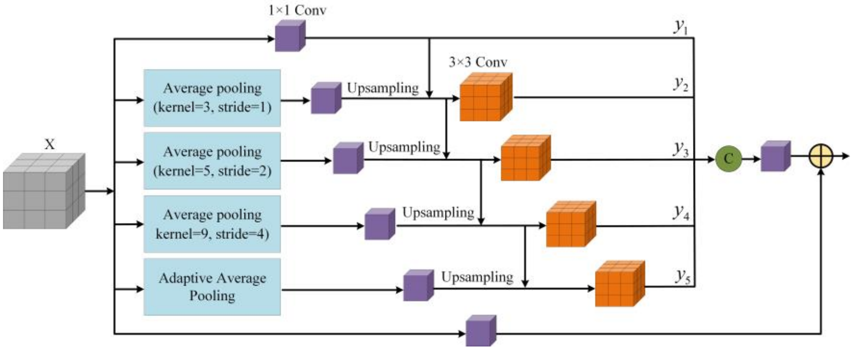

1×1 Conv
y_1
3×3 Conv
X
Average pooling (kernel=3, stride=1)
Upsampling
y_2
Average pooling (kernel=5, stride=2)
Upsampling
y_3
C
Average pooling kernel=9, stride=4)
Upsampling
y_4
Adaptive Average Pooling
Upsampling
y_5

Figure 4. Context information aggregation module.

The CIAM module combines the context information extracted from pooling windows of different sizes with deeper features to form richer multi-scale context information. Different from the previous fusion methods, it fuses the multi-scale information in the way of hierarchical residuals, which greatly improves the accuracy of ground object classification.

3.4. Full-Scale Feature Interaction Strategy

As mentioned above, we first use dynamic convolution and self-attention mechanisms to realize adaptive feature extraction and improve the effect of spatial feature extraction. Secondly, the context information aggregation module is proposed to realize the multi-scale feature fusion of context information. However, in CNNs, low-level feature maps usually have rich edge and texture details as well as spatial information. This is conducive to highlighting the boundary of segmentation targets and obtaining more refined results. In contrast, high-level feature maps pay more attention to the semantic location information of images, which is more conducive to the semantic classification of images. Therefore, the combination of low-level and high-level features is also the key to improving network performance. However, popular networks such as U-Net and U-Net++ often ignore the differences between the characteristics of different layers and cannot explore enough useful information from a full-scale perspective, so they cannot accurately determine the location and boundary information of ground object targets. Therefore, to make full use of multi-scale features, and from the perspective of capturing fine-grained details and coarse-grained semantics at full scale, we propose a full-scale feature interaction strategy (FF) as shown in Figure 5. Each decoder in the network integrates the full-scale feature maps in the encoder. Figure 5 constructs the full-scale fusion features by taking only one stage as an example (stage 2), it is fused with the features of the corresponding scale of the decoder.

Different from the skip connection in U-Net, this strategy downsamples the features from stage 1 in the encoder to the resolution of stage 2 feature map, upsamples the features from stage 3 and stage 4 to the resolution of stage 2, and finally fuses the feature map of the original resolution with the changed feature map. This not only retains the original feature information of this stage, but also the shallow edge contour information is effectively fused with the deep semantic information. It is worth noting that, in the full-scale feature interaction strategy, because our purpose is to obtain the edge texture information of shallow features, we choose the maximum pool operation that can learn the image edge texture information, and use bilinear interpolation for upsampling. The full-scale feature fusion strategy will transfer more abundant encoder information to the decoder, providing more accurate position perception and boundary perception for VHR remote-sensing image land-cover classification.

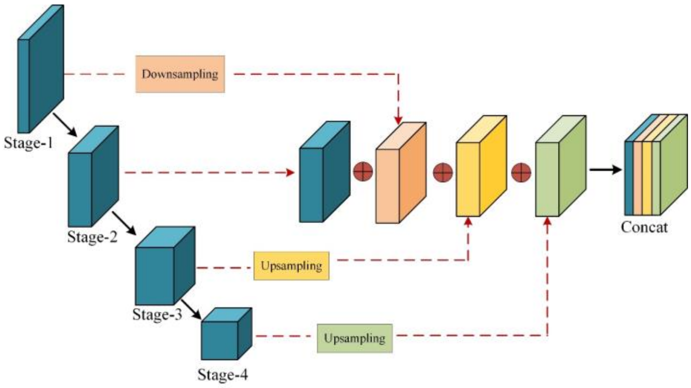


Figure 5. Full-scale feature interaction strategy taking stage 2 as an example.

4. Experimental Results and Analysis

To verify the effectiveness of DCSA-Net, we conducted experiments on the Vaihingen and Potsdam datasets of ISPRS [50]. Our experiment adopts the quantitative accuracy evaluation methods of overall accuracy (OA) and F-score to evaluate the classification results. In addition, the performance of our model is compared with state-of-the-art methods for VHR remote-sensing image land-cover classification. Finally, ablation studies are performed to verify further the superiority of DCSA-Net.

4.1. Datasets Description

The dataset is provided by the Third Committee of the International Society for Photogrammetry and Remote Sensing, containing high-resolution true orthophoto (top) slices, DSMS, and corresponding ground truth (GT) labels of two German urban areas.

All the images in the Potsdam dataset [51] are multi-model data that include four bands: near-infrared (IR), red (R), and green (G), and the corresponding normalized digital surface model (nDSM). The spatial resolution of these images is 5 cm, and image size is 6000 × 6000 pixels. Impervious surfaces, buildings, low vegetation, trees, vehicles, and debris are marked on each pixel of the 24 images. Compared with the RGB image format, the IRRG image can provide better classification performance due to the utilization of IR. Therefore, we used IRRG and nDSM image formats in our experiments. We select image sequence numbers 5_12, 6_7, and 7_9 for validation, image number 5_10 and 6_8 for prediction, and train the remaining images.

The Vaihingen dataset [52] contains 33 images with an average size of 2500 × 2100 pixels and a spatial resolution of 9 cm. In this dataset, only 16 images have real labels, and each image has the same band and label as the Potsdam dataset. We also used IRRG and nDSM image formats to improve the classification performance. In this experiment, five images are selected as the prediction set to analyze the network, and the serial numbers were 11, 15, 28, 30, and 34, respectively. Three images are used as the verification set, with sequence numbers 7, 23, and 37, and other images are used as the training set. Figure 6 shows two groups of images from these two datasets.

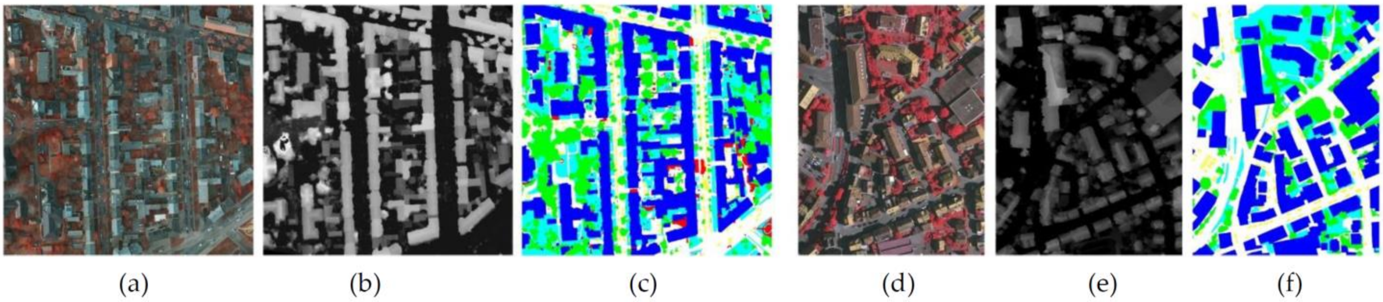


Figure 6. Sample images of Potsdam and Vaihingen datasets, digital surface models, and their corresponding labels. (**a**) Potsdam TOP, (**b**) Potsdam DSM, (**c**) Potsdam GT, (**d**) Vaihingen TOP, (**e**) Vaihingen DSM, and (**f**) Vaihingen GT.

Due to the limitation of GPU memory, it is necessary to crop the original images to adapt to the network input size. Therefore, each image was cropped to 256 × 256 pixels, with an overlap of 128 pixels, and the final prediction was spliced. To avoid overfitting the network, we applied the random flip and the random rotation on images to achieve data enhancement. Therefore, the network could effectively avoid the overfitting during the training process, and thus shows strong robustness.

4.2. Training Details

We used ResNet50 [53] pretrained on ImageNet [54] as the backbone network. We implemented the proposed method with the Pytorch framework and trained it on an NVIDIA GeForce GTX 3090 GPU with 25.4GB VRAM, and an Adam optimizer with a momentum of 0.9, weight decay of 0.004, and initial learning rate of 0.001 was applied to optimize the network. We used cross-entropy as the loss function of the network. The total number of training rounds and the total batch size of training were set to 200 and 16, respectively.

4.3. Metrics

We adopted the most common evaluation metrics in the field of land-cover classification, F-score (*F*1), Overall Accuracy (*OA*), and mean intersection-over-union (*mIoU*), to evaluate the performance of the different methods. The metrics are defined as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{TP + FP + FN} \tag{7}$$

where

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

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. These results usually convert the segmentation image into a single channel category index and conduct a pixel-by-pixel statistical comparison through the confusion matrix [55] and the real label to evaluate the segmentation accuracy.

4.4. Results and Analysis

In this section, we present the performance analysis between DCSA-Net and comparative methods including DeepLab v3+ [43] (2018), MANet [47] (2020), DSMFNet [34] (2019), REMSNet [45] (2020), DP-DCN [35] (2019), and MMAFNet [46] (2021). These networks can be mainly divided into two groups: the networks with spatial relations and the networks without spatial relations, in which both REMSNet and MMAFNet introduce spatial relationships. It is worth noting that different from other networks, DeepLab v3+ does not use digital surface models. We evaluated DCSA-Net and comparative methods on Potsdam and Vaihingen datasets. The experimental results are shown in Tables 1 and 2.

Table 1. Experimental results of F1, OA, and mIoU on the Potsdam dataset (%). The best results are shown in bold.

Methods	Imp. Surf.	Building	Low veg.	Tree	Car	Mean F1	OA	mIoU
Deeplab v3+ [43]	89.88	93.78	83.23	81.66	93.50	88.41	87.72	79.35
MANet [47]	91.33	95.91	85.88	87.01	91.46	90.32	89.19	81.42
DSMFNet [34]	93.03	95.75	86.33	86.46	94.88	91.29	90.36	82.47
DP-DCN [35]	92.53	95.36	87.21	86.32	95.42	91.37	90.45	82.55
REMSNet [45]	93.48	96.17	87.52	87.97	95.03	92.03	90.79	83.56
MMAFNet [46]	93.61	96.26	87.87	88.65	95.32	92.34	91.04	84.03
DCSA-Net	**93.69**	**96.34**	**88.05**	**88.87**	**95.63**	**92.52**	**91.25**	**84.24**

The experimental results on the Potsdam dataset are shown in Table 1. We compared DCSA-Net with DeepLab v3+, MANet, DSMFNet, DP-DCN, REMSNet, and MMAFNet. The experimental data show that the average F1, OA, and mIoU of our DCSA-Net are higher than the comparative methods. Compared with DeepLab v3+ and DP-DCN, the increment of DCSA-Net on the average F1 is 4.11% and 1.15%, respectively, which proves that the LDCM makes full use of DSM features and has better feature fusion ability. Compared with REMSNet, DCSA-Net improves the classification of low vegetation and tree categories by 0.53% and 0.90%.

To compare the classification performance of different methods intuitively, we visualized the experimental results on some images from the Potsdam dataset in Figures 7 and 8. From the comparison of the marked areas in the dotted box, it can be seen that all the comparative methods perform badly on segmenting the vehicles covered by trees. Additioanally, the trees next to the vehicles are mistakenly classified as vehicles. It is especially hard to distinguish between low vegetation and trees, as well as to recognize the vehicles under the tree shelter for comparative methods. Moreover, both REMSNet and MMAFNet generate too much redundant information, thus the low vegetation is almost completely classified as trees. Contrary to the above methods, the proposed DCSA-Net can clearly judge the junction of the two types of targets in the area and thus provides better land-cover classification results.

Figure 8 shows the comparison of classification results of a complete image in the Potsdam prediction set. As the low vegetation contains complex context information and the debris recognition also has challenging structure and texture, the comparative methods are prone to misjudge these categories. However, DCSA-Net uses the CIAM to obtain richer context information and repeatedly model the relationship between the global spatial dimension and channel dimension many times. The proposed method obtains better classification results and alleviates some misclassification. Therefore, our model has stronger robustness.

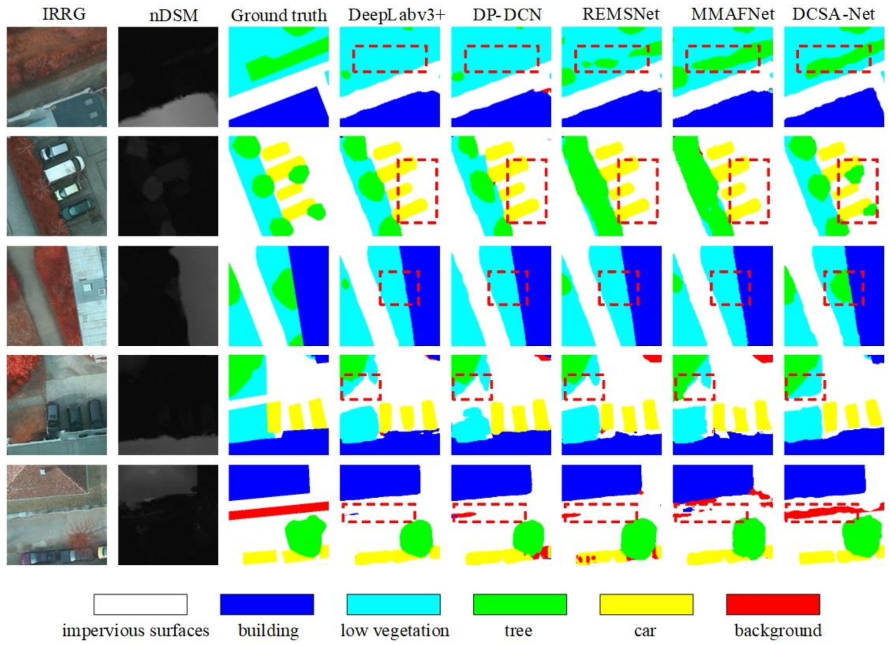
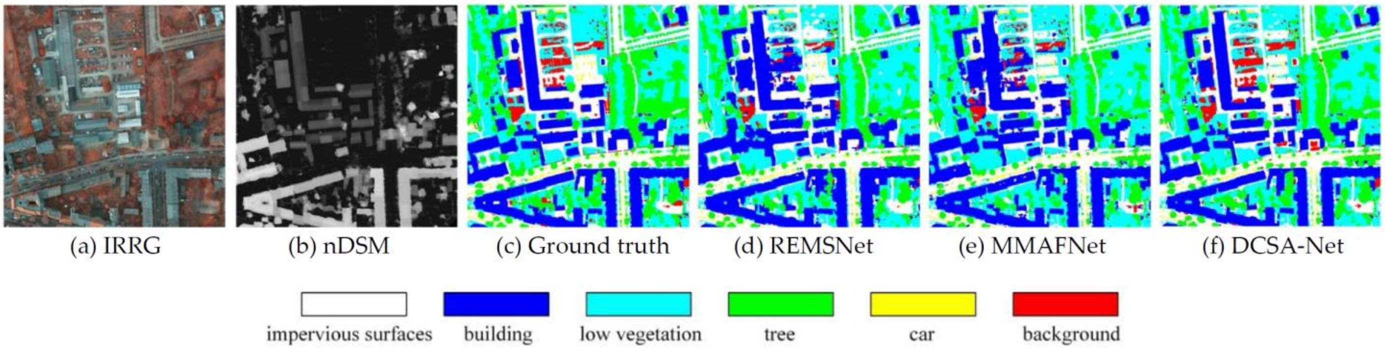


Figure 7. Comparison of visual results of different methods on cropped images from the Potsdam dataset.



(a) IRRG (b) nDSM (c) Ground truth (d) REMSNet (e) MMAFNet (f) DCSA-Net

Figure 8. Comparison of visual results of different methods on original images from the Potsdam dataset.

The evaluation results of the experiment on the Vaihingen dataset are shown in Table 2. The proposed DCSA-Net achieves average F1 of 88.81%, OA of 90.58%, and mIoU of 78.93%. Compared with DeepLab v3+, the average F1, OA, and mIoU increased by 3.33%, 3.36%, and 3.49%, respectively. Compared with MMAFNet, the average F1, OA, and mIoU increased by 0.20%, 0.31%, and 0.28%, respectively. Among all categories, the clear improvement is the classification of trees, which proves that DCSA-Net makes better use of DSM features to assist classification. We also visualized the experimental results of this data as shown in Figures 9 and 10.

Table 2. Experimental results of F1, OA, and mIoU on the Vaihingen dataset (%). The best results are shown in bold.

Methods	**Imp. Surf.**	**Building**	**Low veg.**	**Tree**	**Car**	**Mean F1**	**OA**	**mIoU**
Deeplab v3+ [43]	87.67	93.95	79.17	86.26	80.34	85.48	87.22	75.44
MANet [47]	90.12	94.08	81.01	87.21	81.16	86.72	88.17	76.79
DP-DCN [35]	91.47	94.55	80.13	88.02	80.25	86.89	89.32	77.09
DSMFNet [34]	91.47	95.08	82.11	88.61	81.01	87.66	89.80	77.76
REMSNet [45]	92.01	95.67	82.35	89.73	81.26	88.20	90.08	78.16
MMAFNet [46]	92.06	96.12	82.71	90.01	82.13	88.61	90.27	78.65
DCSA-Net	**92.11**	**96.19**	**83.04**	**90.31**	**82.39**	**88.81**	**90.58**	**78.93**

Figure 9 shows the classification diagram obtained by five different models on the Vaihingen dataset. Compared with other models, DCSA-Net obtains a more coherent and accurate classification map, especially for the distinction between trees and low vegetation in the dotted box marked area in the first and third lines. In addition, it can be observed that our method can obtain more fine-grained classification results for vehicle recognition in the marked area as shown in the second line. Combined with the F1 of vehicles in Table 2, it is obvious that our method has effectively improved the classification results of small targets.

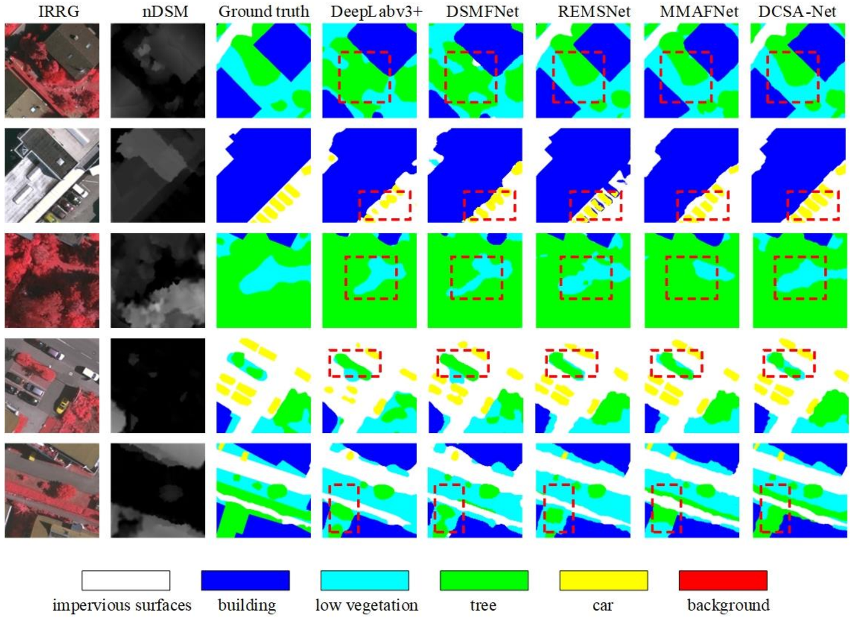


Figure 9. Comparison of visual results of different methods on cropped images from the Vaihingen dataset.

As shown in Figure 10, the comparative methods cannot distinguish between regular targets and irregular targets at the same time, especially for the buildings and roads, leading to inaccurate vehicle positioning. The main reason is that these methods always focus on exploring spatial context information using features extracted by convolution, which is not enough to provide useful clues to distinguish different objects, in particular the isolated objects. Innovatively, DCSA-Net introduces the dynamic convolution and a more complex context information aggregation module to convert the relationship between different categories of objects into adaptive, which can obtain clearer boundaries in the segmentation of all categories and has higher accuracy in complex high-resolution remote sensing scenes.

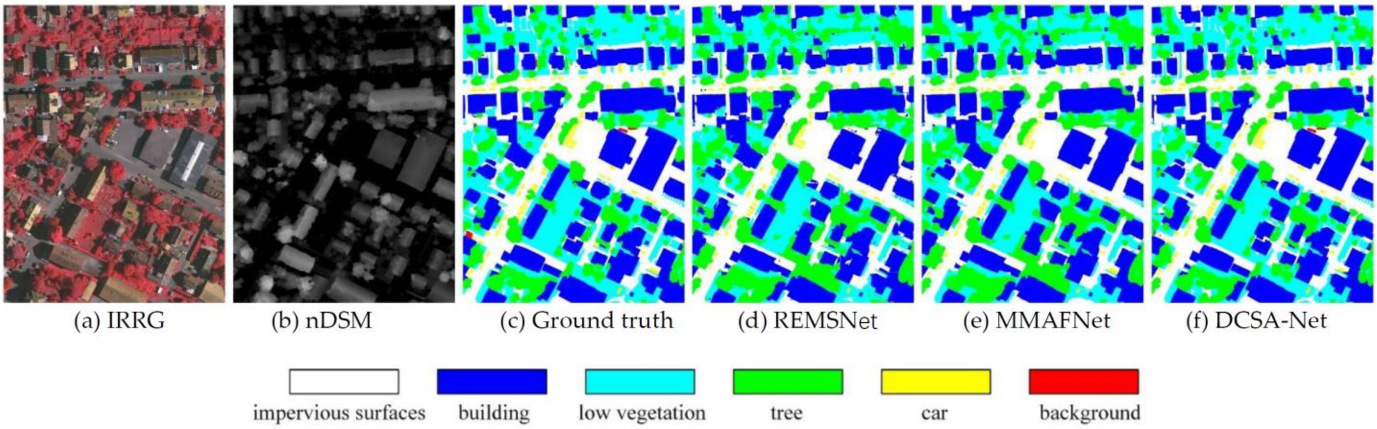


Figure 10. Comparison of visual results of different methods on original images from the Vaihingen dataset.

4.5. Model Complexity

To further evaluate the model performance, the experiment mainly uses the floating-point operation rate per second (GFLOPs) and the model parameters (Params) to evaluate the model complexity. The specific results are given in Table 3. We can see that the proposed DCSA-Net has a lower model complexity and higher classification accuracy than the comparative methods. Different from vanilla convolution, the proposed dynamic convolution kernels are different in the spatial space but shared along the channel dimension. As the number of feature channels decides the dimension of dynamic convolution kernels, the model complexity is proportional to it, which has been greatly reduced. Compared with MMAFNet, the number of parameters was reduced by 71%, yet OA was improved by 0.31%. The results demonstrate that our proposed DCSA-Net not only reduces the number of parameters and model size but also achieves higher accuracy for land-cover classification.

Table 3. Performance comparison of different methods on Vaihingen dataset.

Methods	GFLOPs (GB)	Params (M)	Mean F1	OA
DeepLab v3+ [43]	89	47	85.48	87.22
MANet [47]	63	85	86.72	88.17
DP-DCN [35]	25	28.5	86.89	89.32
DSMFNet [34]	53	52	87.66	89.80
MMAFNet [46]	69	93	88.61	90.27
DCSA-Net	**21**	**27**	**88.81**	**90.58**

4.6. Ablation Studies

We applied the ablation experiments on the Vaihingen dataset to verify the effectiveness of different modules proposed by DCSA-Net by comparing them with the baseline model. The evaluation indexes were F1 and OA. The experiment decomposed and combined the proposed modules, including Res50, Res50+LDCM, Res50+CIAM, Res50+FF, Res50+LDCM+CIAM, Res50+LDCM+FF, Res50+CIAM+FF, and All Modules (DCSA-Net). The specific contents are shown in Table 4.

Table 4. Quantitative analysis of Ablation Experiment in Vaihingen dataset (%). The best results are shown in bold.

Method	Imp. Surf.	Building	Low veg.	Tree	Car	Mean F1	OA
Res50	86.94	89.67	75.83	84.42	77.40	82.85	84.98
Res50+LDCM	88.23	93.81	78.36	86.99	80.31	85.54	87.42
Res50+CIAM	88.17	92.22	77.80	85.88	79.06	84.63	86.58
Res50+FF	89.81	94.04	79.15	87.36	81.49	86.37	87.98
Res50+LDCM+CIAM	89.03	93.90	78.76	87.21	80.95	85.97	87.86
Res50+LDCM+FF	91.57	95.62	81.94	89.22	81.53	87.98	89.85
Res50+CIAM+FF	90.68	94.97	81.18	88.48	81.46	87.35	89.19
DCSA-Net	**92.11**	**96.19**	**83.04**	**90.31**	**82.39**	**88.81**	**90.58**

Res50 refers to a baseline consisting of two pre-trained ResNet50 that extract features from different modalities and fuse these features after each res block. We chose this model as the baseline for ablation experiments.

Owing to the introduction of LDCM, the extracted features of each layer contain richer semantic information, thus Res50+LDCM increases the average F1 and OA by 2.69% and 2.44%, respectively, compared with the baseline. Among all categories, the clearest improvement was for buildings, with an increase of 4.14%. Through repeatedly modeling the self-attention feature map many times, LDCM can readjust the filtered weights to the multimodal features, which enhances the most useful information for classification and reduces the misclassification of similar categories. It demonstrates that a reasonable and efficient fusion of multimodal features can improve the classification performance of our proposed network.

Res50+CIAM integrates multi-scale context information, aggregates and optimizes deeper semantic information, and expands the global relationship at the image level. Mitigating some misclassification through richer contextual information, compared with the baseline, the average F1 and OA are increased by 1.78% and 1.60%, respectively, which strengthens the performance of the network and makes the model more robust.

Res50+FF adopts the full-scale feature fusion strategy in the feature fusion stage and integrates the feature-mapping of each layer based on retaining the original resolution features. The benefit of interaction between high-resolution features and low-resolution features is huge, thus the improvement of the model compared with the baseline is also the highest, with an average F1 and OA increased by 3.52% and 3.00%, respectively.

To further verify the effectiveness of our proposed module, we superimposed the modules in pairs, and the results are shown in Table 4. Res50+LDCM+CIAM can not only extract more useful feature information, but also obtain more abundant context information, compared with res50+LDCM and res50+CIAM, F1 and OA were improved by 0.43%, 0.44% and 1.34%, 1.28%, respectively. Res50+LDCM+FF can extract more useful feature information while fully interacting with high and low-resolution features, compared with res50+LDCM and res50+FF, F1 and OA are improved by 2.44%, 2.43% and 1.61%, 1.87%, respectively. Res50+CIAM+FF not only obtains richer context information, but also fully interacts with high and low-resolution features, compared with res50+CIAM and res50+FF, F1 and OA are improved by 2.72%, 2.61% and 0.98%, 1.21%, respectively.

Finally, all modules are integrated into the baseline to form the proposed DCSA-Net. Compared with the baseline, the proposed DCSA-Net improves the average $F1$ by 5.96% and OA index by 5.60%. All the above ablation experiments clearly show that the DCSA-Net proposed in this paper can significantly improve the land-cover classification accuracy for VHR remote-sensing images.

5. Discussion

5.1. Discussion on the Effectiveness of the LDCM

LDCM can extract more useful image features at a lower cost, avoiding the impact of redundant features on the final classification results. The visualization results are shown in Figure 11, which further verifies its effectiveness. Specifically, on the Vaihingen verification dataset, the outputs of the encoder with and without the introduction of the LDCM are shown in Figure 11, with red indicating the high attention area and blue indicating the low attention area. In Figure 11d,e show that the LDCM is not introduced and that the LDCM is introduced, respectively. We can see that after the LDCM is introduced, the boundaries between categories are more obvious, the network pays more attention to the target object, and reduces the influence of irrelevant features, which is conducive to improving the classification accuracy.

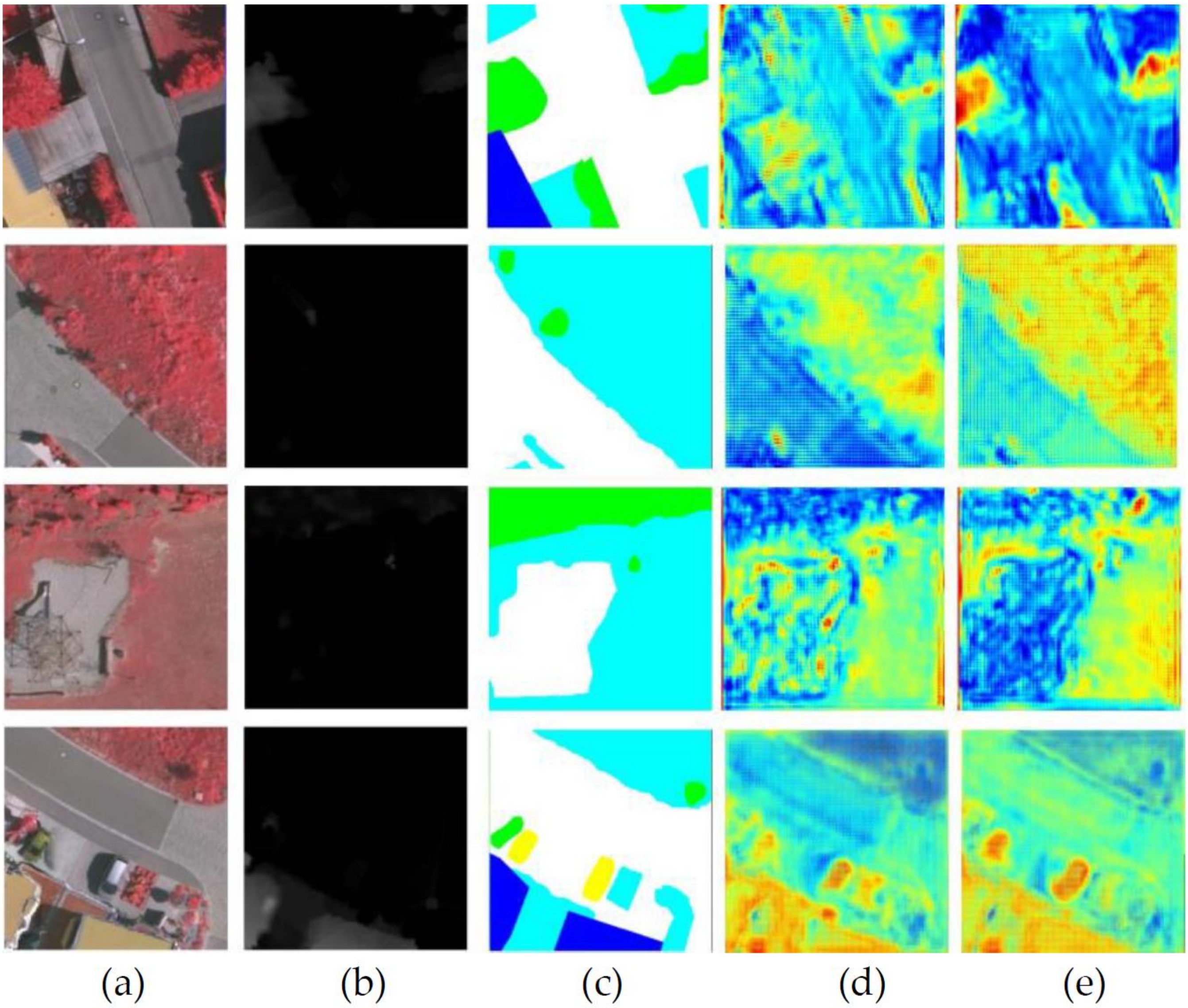


Figure 11. Visualization with and without LDCM. (**a**) IRRG, (**b**) nDSM, (**c**) ground truth, (**d**) output of LDCM is not introduced, (**e**) output after the introduction of LDCM.

5.2. Discussion on the Optimal Selection of Key Parameters in the CIAM

CIAM is to obtain rich context information, enhance the integrity between target edges and internal compactness, and improve the accuracy of remote-sensing image feature classification. In this module, the context information extracted from different-size pooling windows is combined with deeper features to form richer multi-scale context information. However, it is a problem to determine an appropriate window size. In view of this problem, we conducted experiments on different scales of windows and their combinations, and the experimental results are shown in Table 5. It can be found that when the scale value is too large or too small, CIAM cannot achieve the best performance. Finally, the scale of (3,5,9) produces the best performance in our experiment.

Table 5. Comparison of different pool size combinations in CIAM on the Vaihingen dataset (%). The best results are shown in bold.

Method	Imp. Surf.	Building	Low veg.	Tree	Car	Mean F1	OA
CIAM(3,5,7)	92.01	96.11	82.64	90.21	82.28	88.65	90.47
CIAM(3,7,11)	92.08	96.16	82.98	90.26	82.33	88.76	90.55
CIAM(3,5,11)	92.10	96.16	83.00	90.29	82.37	88.78	90.56
CIAM(3,5,9) DCSA-Net	**92.11**	**96.19**	**83.04**	**90.31**	**82.39**	**88.81**	**90.58**

6. Conclusions

In this paper, we analyzed the problems existing in the current VHR remote-sensing image land-cover classification task based on deep learning and proposed a dynamic convolution self-attention network, named DCSA-Net. The main innovation of this work is to ensure high-quality multimodal internal features with low computational costs. In terms of network structure, the LDCM is introduced into the encoder to improve the multimodal characteristics. At the same time, the CIAM is introduced at the end of the encoder to expand the deeper semantic information of multi-scale context. Finally, richer full-resolution mixed features are obtained by a full-scale feature fusion. The experiments show that the proposed network is superior to the mainstream networks on F1 and OA, and has lower model complexity. In addition, in the ablation experiments on the Vaihingen dataset, we have verified the performance of the proposed modules, and further verified the effectiveness of DCSA-Net in VHR remote-sensing image land-cover classification.

Author Contributions: Conceptualization, X.W. and Y.Z. (Yue Zhang); experiment design, T.L. and Y.W.; formal analysis, X.W. and Y.Z. (Yue Zhang); figures and graphs, Y.Z. (Yue Zhang) and Y.Z. (Yujie Zhai); writing—original draft preparation, T.L.; review and editing, A.K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (Program No. 61871259, 62271296, 61861024), in part by Natural Science Basic Research Program of Shaanxi (Program No. 2021JC-47), in part by Key Research and Development Program of Shaanxi (Program No. 2022GY-436, 2021ZDLGY08-07), in part by Natural Science Basic Research Program of Shaanxi (Program No. 2022JQ-634, 2022JQ-018), and in part by Shaanxi Joint Laboratory of Artificial Intelligence (No. 2020SS-03).

Data Availability Statement: The datasets used in this study have been published, and their addresses are https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/ (accessed on 30 January 2021) and https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/ (accessed on 30 January 2021).

Acknowledgments: The author would like to thank International Society for Photogrammetry and Remote Sensing (ISPRS) for providing data and all colleagues for their contribution to the land cover results of VHR remote-sensing images.

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

References

1. Lv, Z.; Liu, T.; Benediktsson, J.A.; Falco, N. Land Cover Change Detection Techniques: Very-high-resolution optical images: A review. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 44–63. [CrossRef]
2. Lei, T.; Wang, J.; Ning, H.; Wang, X.; Xue, D.; Wang, Q.; Nandi, A.K. Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
3. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
4. Troya-Galvis, A.; Gançarski, P.; Berti-Equille, L. Remote sensing image analysis by aggregation of segmentation-classification collaborative agents. *Pattern Recognit.* **2018**, *73*, 259–274. [CrossRef]
5. Zanotta, D.C.; Zortea, M.; Ferreira, M.P. A supervised approach for simultaneous segmentation and classification of remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 162–173. [CrossRef]

6. Lei, T.; Jia, X.; Zhang, Y.; Liu, S.; Meng, H.; Nandi, A.K. Superpixel-Based Fast Fuzzy C-Means Clustering for Color Image Segmentation. *IEEE Trans. Fuzzy Syst.* **2019**, *27*, 1753–1766. [CrossRef]
7. Yu, H.; Gao, L.; Li, J.; Li, S.S.; Zhang, B.; Benediktsson, J.A. Spectral-Spatial Hyperspectral Image Classification Using Subspace-Based Support Vector Machines and Adaptive Markov Random Fields. *Remote Sens.* **2016**, *8*, 355. [CrossRef]
8. Dong, L.; Du, H.; Mao, F.; Han, N.; Li, X.; Zhou, G.; Zhu, D.; Zheng, J.; Zhang, M.; Xing, L.; et al. Very High Resolution Remote Sensing Imagery Classification Using a Fusion of Random Forest and Deep Learning Technique—Subtropical Area for Example. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *13*, 113–128. [CrossRef]
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Proc. Adv. Neural Inf. Process. Syst* **2012**, *25*, 1097–1105. [CrossRef]
10. Long, J.L.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. 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]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
13. 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 (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 2881–2890.
14. 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]
15. Zhang, X.X.; Chen, T. Segmentation of High Spatial Resolution Remote Sensing Image based on U-Net Convolutional Net-works. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Wikoloa, HI, USA, 26 September 2020–2 October 2020; pp. 2571–2574.
16. Mustafa, N.; Zhao, J.P.; Liu, Z.Y.; Zhang, Z.H.; Yu, W.X. Iron ORE Region Segmentation Using High-Resolution Remote Sensing Images Based on Res-U-Net. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Wikoloa, HI, USA, 26 September 2020–2 October 2020; pp. 2563–2566.
17. Niu, Z.; Liu, W.; Zhao, J.; Jiang, G. DeepLab-Based Spatial Feature Extraction for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 251–255. [CrossRef]
18. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E.H. Squeeze-and-excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2011–2023. [CrossRef]
19. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting Feature Context in Convolutional Neural Networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 3–8 December 2018; pp. 9423–9433.
20. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
21. Fu, J.; Liu, J.; Tian, H.J.; Li, Y.; Bao, Y.J.; Fang, Z.W.; Lu, H.Q. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
22. Woo, S.; Park, J.C.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network with Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. [CrossRef]
24. Zhang, C.; Jiang, W.S.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C.J. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
25. Lv, P.; Wu, W.; Zhong, Y.; Du, F.; Zhang, L. SCViT: A Spatial-Channel Feature Preserving Vision Transformer for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
26. Iandola, F.N.; Han, S.; Matthew, W.M.; Ashraf, K.; William, J.D.; Keutzer, K. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5 MB Model Size. *arXiv* **2016**, arXiv:1602.07360. preprint.
27. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]
28. Howard, A.G.; Zhu, M.L.; Chen, B.; Kalenichenko, D.; Wang, W.J.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. preprint.
29. Lv, L.; Guo, Y.; Bao, T.; Fu, C.; Huo, H.; Fang, T. MFALNet: A Multiscale Feature Aggregation Lightweight Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 2172–2176. [CrossRef]
30. Zhang, X.Y.; Zhou, X.Y.; Lin, M.X.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.

31. Qiao, S.; Dong, X.-M.; Peng, J.; Sun, W. LiteSCANet: An Efficient Lightweight Network Based on Spectral and Channel-Wise Attention for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11655–11668. [CrossRef]
32. Han, K.; Wang, Y.H.; Tian, Q.; Guo, J.Y.; Xu, C.J.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
33. Paoletti, M.E.; Haut, J.M.; Pereira, N.S.; Plaza, J.; Plaza, A. Ghostnet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10378–10393. [CrossRef]
34. 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]
35. Peng, C.; Li, Y.; Jiao, L.; Chen, Y.; Shang, R. Densely Based Multi-Scale and Multi-Modal Fully Convolutional Networks for High-Resolution Remote-Sensing Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2612–2626. [CrossRef]
36. Ferrari, L.; Dell'Acqua, F.; Zhang, P.; Du, P. Integrating EfficientNet into an HAFNet Structure for Building Mapping in High-Resolution Optical Earth Observation Data. *Remote Sens.* **2021**, *13*, 4361. [CrossRef]
37. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
38. Zhang, Y.J.; Chen, Y.L.; Ma, Q.J.; He, C.T.; Cheng, J. Dual Lightweight Network with Attention and Feature Fusion for Semantic Segmentation of High-Resolution Remote Sensing Images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2755–2758.
39. Liu, Q.; Xiao, L.; Yang, J.; Wei, Z. Multilevel Superpixel Structured Graph U-Nets for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
40. Zhou, Z.W.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J.M. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep learning in medical image analysis and multimodal learning for clinical decision support, Granada, Spain, 20 September 2018; pp. 3–11.
41. Wang, Y.; Sun, Z.; Zhao, W. Encoder- and Decoder-Based Networks Using Multiscale Feature Fusion and Nonlocal Block for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1159–1163. [CrossRef]
42. Xu, Q.; Yuan, X.; Ouyang, C.; Zeng, Y. Attention-Based Pyramid Network for Segmentation and Classification of High-Resolution and Hyperspectral Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3501. [CrossRef]
43. Chen, L.C.; Papandreou, G.; Shroff, 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.
44. Tian, Q.; Zhao, Y.; Li, Y.; Chen, J.; Chen, X.; Qin, K. Multiscale Building Extraction with Refined Attention Pyramid Networks. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
45. Liu, C.; Zeng, D.; Wu, H.; Wang, Y.; Jia, S.; Xin, L. Urban Land Cover Classification of High-Resolution Aerial Imagery Using a Relation-Enhanced Multiscale Convolutional Network. *Remote Sens.* **2020**, *12*, 311. [CrossRef]
46. Lei, T.; Li, L.; Lv, Z.; Zhu, M.; Du, X.; Nandi, A.K. Multi-Modality and Multi-Scale Attention Fusion Network for Land Cover Classification from VHR Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3771. [CrossRef]
47. Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; Stolkin, R. Multi-scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 872. [CrossRef]
48. Nie, J.; Wang, C.; Yu, S.; Shi, J.; Lv, X.; Wei, Z. MIGN: Multiscale Image Generation Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Multimedia* **2022**, 1–14. [CrossRef]
49. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P.H. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
50. Gerke, M.; Rottensteiner, F.; Wegner, J.D.; Sohn, G. ISPRS Semantic Labeling Contest. In Proceedings of the Photogrammetric Computer Vision (PCV), Zurich, Switzerland, 5–7 September 2014.
51. ISPRS Potsdam 2D Semantic Labeling Dataset. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-labelpotsdam.html (accessed on 30 January 2021).
52. ISPRS Vaihingen 2D Semantic Labeling Dataset. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-semlabel-vaihingen.html (accessed on 30 January 2021).
53. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Miami (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
55. Huang, B.; Zhao, B.; Song, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]