International Journal of Applied Earth Observations and Geoinformation 106 (2022) 102638

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag

# MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification

Xue Li [a], Guo Zhang [a], Hao Cui [a,*], Shasha Hou [a], Shunyao Wang [a], Xin Li [a], Yujia Chen [a], Zhijiang Li [b], Li Zhang [a]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] School of Printing and Packaging, Wuhan University, Wuhan 430079, China

ARTICLE INFO

*Keywords:*
Deep convolution neural network
Multimodal-cross attention network
Land use classification
Optical images
SAR images

ABSTRACT

Deep convolution neural network (DCNN) is among the most effective ways of performing land use classification of high-resolution remote sensing images. Land use classification by fusing optical and synthetic aperture radar (SAR) images has broad application prospects, but related research studies are few. In this study, we developed the first and largest joint optical and SAR land use classification dataset, WHU-OPT-SAR, covering an area of approximately 50,000 $km^2$, and designed a multimodal-cross attention network (MCANet). MCANet comprises three core modules: the pseudo-siamese feature extraction module, multimodal-cross attention module, and low-high level feature fusion module, which are used for independent feature extraction of optical and SAR images, second-order hidden feature mining, and multi-scale feature fusion. The land use classification accuracy of our approach on the WHU-OPT-SAR dataset was approximately 5% higher than that of optic-image-based approaches. Moreover, the accuracy of *city*, *village*, *road*, *water*, *forest*, and *farmland* classification was improved by 7%, 2%, 5%, 6%, 1%, and 0.6%, respectively, reflecting the superior performance of fusing optical and SAR images. Furthermore, the classification accuracy in Hubei Province of China, which covers an area of 190,000 $km^2$, has also increased by approximately 5%, which verifies the effectiveness of our approach.

## 1. Introduction

Deep convolution neural network (DCNN) presents a new approach toward the interpretation and mapping of large-scale remote sensing images (Zhang et al., 2016; Li et al., 2021). Land use classification based on remote sensing images is relatively more dependent on the spatial and semantic information of images (Tong et al., 2020; Ma et al., 2019), and the results can be applied to land management, environmental protection, and urban planning (Toll, 1985; Jewell, 2010; Zhang and Yang, 2020).

Human and social attributes of land use complicate the achievement of remarkable results via traditional land use classification approaches such as random forests (RF), support vector machines (SVM) (Zhang and Yang, 2020; Paneque-Gálvez et al., 2013). Large-scale land use classification is relatively dependent on the spatial and semantic features of images, and deep learning has obvious advantages. The mainstream approaches can be summarized into two categories: pixel-based and object-based classification. In pixel-based land use classification, DCNN adopts high-dimensional semantic features from the original pixel representation. It is widely used in aerial images, for instance, MCNN using multi-scale learning scheme to capture the context information of different scales (Zhao and Du, 2016), and RotEqNet using rotation invariant structure to improve the generalization ability of the model (Marcos et al., 2018). For satellite images, representative networks such as PT-GID automatically select training samples from the target domain to achieve semi-supervised training (Tong et al., 2020). Object-based land use/cover classification is used to obtain more accurate boundary information by incorporating DCNN into the object-based image analysis framework. For example, OCNN uses the object-based image analysis framework to obtain more accurate boundary information (Zhang et al., 2018). But, in practical, a single data source is often insufficient for training a land use classifier.

Imaging remote sensing mainly includes optical imaging and synthetic aperture radar (SAR) imaging. Optical images provide high spatial resolution, rich spectral, and textural information but optical sensors are susceptible to weather. SAR sensors work in all weather conditions and

* Corresponding author.
*E-mail address:* cuihaocasm@whu.edu.cn (H. Cui).

https://doi.org/10.1016/j.jag.2021.102638
Received 7 August 2021; Received in revised form 19 October 2021; Accepted 27 November 2021
Available online 30 December 2021

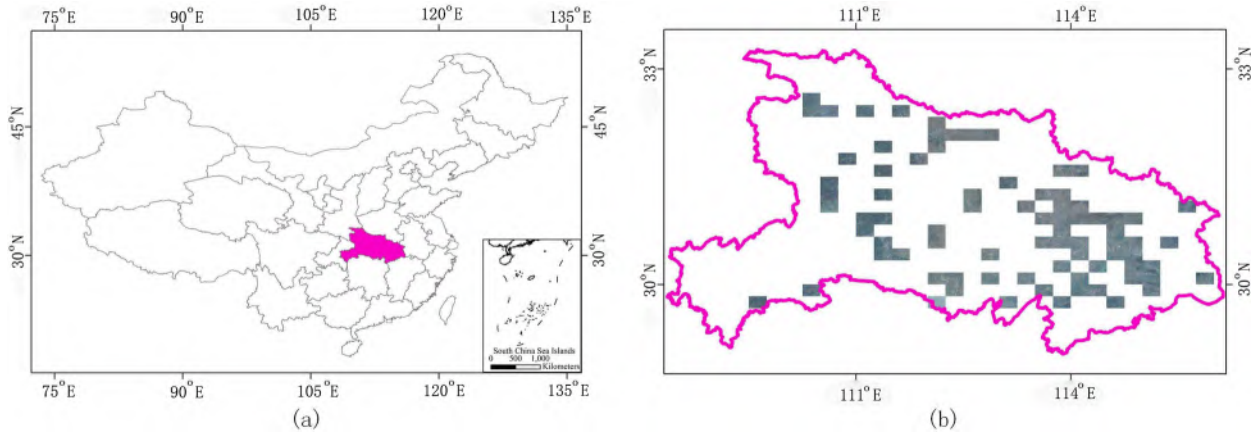


**Fig. 1.** WHU-OPT-SAR dataset. (a) Geographic location of the dataset in China. (b) Distribution of images in the dataset.

penetrate certain ground objects. And SAR images provide rich geometric information on ground objects. In certain applications, the optical and SAR images are remarkably complementary (Liu et al., 2016). Fusing optical and SAR images for remote sensing classification is regarded as a promising approach toward improving classification accuracy (Gómez-Chova et al., 2015; Schmitt and Zhu, 2016; Mou et al., 2017b).These methods mainly comprise two categories. (1) Traditional machine learning approaches, such as Markov random fields (MRFs), SVM, and other nonparametric approaches (Solberg et al., 1996; Pacifici et al., 2008). Most of these approaches rely on prior knowledge and hand-designed features, and thereby cannot represent complex high-level semantic information. (2) Deep learning approaches, which is mainly divided into decision level fusion and feature level fusion, is a current research trend. a) Decision level fusion, fusing the output of DCNN and other methods (Zhu et al., 2017). This approach can learn the semantic features of interest using various methods, but cannot extract the fused high-dimensional feature representation. b) Feature level fusion (Zhu et al., 2017). The simplest way is to stack d-dimensional features by concatenate, such as PSCNN (Mou et al., 2017a), MRSDC (Xu et al., 2017). This approach does not make full use of the characteristics of each feature and exhibits unstable performance. An alternative method involves fusing the feature representations extracted from the pseudo-siamese network. V-FesuNet explores the pros and cons of early and late fusion (Audebert et al., 2018). Although early fusion is more sensitive to noise, while late fusion recovers certain critical errors on hard pixels. MBFNet uses global average pooling and global maximum pooling to integrate multimodal features bilinearly (Li et al., 2020).

However, there are few studies on land use classification based on fusing optical and SAR images via DCNN for the following reasons. (1) Lack of organized optical and SAR image land use datasets. SEN12MS is a multi-modal land cover dataset (Schmitt et al., 2019). However, there is almost no large multi-modal land use dataset that simultaneously comprises optical and SAR images (Ma et al., 2019). (2) Lack of joint application approaches of optical and SAR images. At present, the joint application based on deep learning mainly focuses on simple feature addition, multiplication, and stacking (Hazirbas et al., 2016; Park et al., 2016; Xu et al., 2017). These approaches lack deep fusing of spatial and semantic features of optical and SAR images to yield satisfactory segmentation results.

To address the above problems, this study made the following two contributions:

-*We open-sourced a large-scale optical and SAR land use classification dataset WHU-OPT-SAR.* It comprised RGB, near infrared (NIR) optical images and corresponding SAR images, covering an area of 51448.56 $km^2$ with a resolution of 5 meters. As far as we know, WHU-OPT-SAR is the first and largest land use classification dataset that has fused high resolution optical and SAR images with sufficient annotation. It can effectively promote the development of remote sensing image interpretation from single data source to multi-sourced data.

-*We proposed a semantic segmentation framework, MCANet, which fused on optical and SAR images.* MCANet comprises three core modules: the pseudo-siamese feature extraction module, multimodal-cross attention module, and low-high level feature fusion module. The pseudo-siamese feature extraction module avoid interference. The multimodal-cross attention module enables the second-order interaction of attention maps. The low-high level feature fusion module can effectively retain the multi-scale features. The experiment was carried out on the dataset of WHU-OPT-SAR. The results show that the overall segmentation accuracy of the MCANet improved by 5% compared with the optical-image-based network. In addition, MCANet has a significant advantage over *farmland*, *city*, *village*, *water*, *road*, and *forest* classifications.

## 2. Materials and method

### 2.1. WHU-OPT-SAR dataset

The WHU-OPT-SAR dataset contains 100 optical images of 5556 × 3704 pixels and SAR images in the same area, covering an area of approximately 50000 $km^2$ in Hubei Province (30°N-33°N, 108°E-117°E), China, as shown in Fig. 1. This area has a subtropical monsoon climate, with the lowest altitude being 50 meters and the highest altitude being 3000 meters. WHU-OPT-SAR covers a wide range of remote sensing images with different terrains and different vegetation. Images in this dataset with pixel-level annotations can provide data sources for land use classification based on deep learning. WHU-OPT-SAR and its reference annotations are presented on https://github.com/AmberHen/WHU-OPT-SAR-dataset.git.

#### 2.1.1. Remote sensing images

The optical and SAR images in the WHU-OPT-SAR dataset are all from Hubei GaoFen Center. The optical images (four channels in RGB and near infrared and ground resolution of 2 meters.) are collected by GF-1, and the SAR images (sampling resolution of 5 meters) are collected by GF-3 fine strip II (FS II). The two are in the WGS-84 coordinate system and have a geometric correspondence of sub-pixels. In order to be consistent with the SAR image, we use bilinear interpolation to resample the optical image to a ground resolution of 5 meters to achieve the pixel-by-pixel correspondence between the two.

#### 2.1.2. Annotated datas

The annotated data were obtained from the 2017 Annual National Land Use Change Survey of China. The vector data were converted to raster data and resampled to the resolution of 5 meters as annotations of the above optical images and SAR images. Then, based on the Chinese

**Table 1**
Mapping table of the 2017 Annual National Land Use Change Survey of China and the adopted classification system.

| Category | Second level category of 2017 Annual National Land Use Change Survey of China | Proportion of pixels (%) |
|---|---|---|
| farmland | Cultivated land, irrigated land, dry land, ridges, and facility agricultural land | 35 |
| forset | Woodland, shrubland, orchard, tea garden, other garden land, other woodland | 5 |
| city | City, organizational town, port and terminal land, airport land | 6 |
| village | village | 14 |
| water | Rivers, lakes, reservoirs, ponds, inland beaches, ditches | 38 |
| road | Railroad land, highway land, rural roads | 1 |
| others | Saline land, marsh land, sandy land, bare land, mining land | 1 |

Land Use Classification criteria (GB/T21010-2017), we combined the 30 secondary categories in the selected study area into seven primary categories: *farmland*, *city*, *village*, *water*, *forest*, *road*, and *others*, as shown in Table 1. Fig. 2 shows examples of the main categories.

### 2.2. Methodology

#### 2.2.1. Architecture of MCANet

We proposed a joint semantic segmentation framework, MCANet, which uses optical and SAR images for land-use classification. As shown in Fig. 3, MCANet draws on the classic Deeplabv3 + framework (Chen et al., 2018). The input of the network is independent optical and SAR images and their annotations, and images and annotations correspond pixel by pixel.

The encoder is used for feature extraction and fusion. It mainly comprises three steps:

(1) Pseudo-siamese feature extraction. The optical and SAR images are input into pseudo-siamese networks, and the low-level and high-level feature maps of optical and SAR images are extracted independently. Low-level feature maps of optical and SAR images are expressed as $Cov_{OPT}^{low}$ and $Cov_{SAR}^{low}$. High-level feature maps of optical and SAR images are expressed as $Cov_{OPT}^{high}$ and $Cov_{SAR}^{high}$.

(2) Multimodal-cross attention computation. $Cov_{OPT}^{low}$ and $Cov_{SAR}^{low}/Cov_{OPT}^{high}$ and $Cov_{SAR}^{high}$ are fused. Low-level and high-level joint attention maps $Att_{OPT-SAR}^{low}$ and $Att_{OPT-SAR}^{high}$ are obtained using the feature cross method, which guides MCANet to focus more on the region of interest, and enhance the ability of MCANet to model high-order cross features between multimodal data.

(3) Low-high level feature fusion. For more optimized utilization, $Cov_{OPT}^{high}$ and $Cov_{SAR}^{high}$ are reduced and stacked with the $Att_{OPT-SAR}^{high}$ to obtain the high-level joint maps $A_{OPT-SAR}^{high}$. Subsequently, atrous spatial pyramid pooling (ASPP) is imputed to extract multi-scale feature representation and up sampled to obtain a multiscale high-level joint map $A_{OPT-SAR}^{high-ms}$. At the same time, the $Cov_{OPT}^{low}$, $Cov_{SAR}^{low}$ and $Att_{OPT-SAR}^{low}$ are concatenated to obtain the reduced low-level joint maps $A_{OPT-SAR}^{low}$. Then the $A_{OPT-SAR}^{low}$ is down-sampled to obtain a down-sampled low-level joint map $A_{OPT-SAR}^{low-dw}$. Finally, $A_{OPT-SAR}^{low-dw}$ and $A_{OPT-SAR}^{high-ms}$ are stacked, that is regarded as low-high level joint map $A_{OPT-SAR}^{low-high}$ and input into the decoder.

The decoder is used to decode the vector output from the encoder into the final land classification result. This decoder receives the output

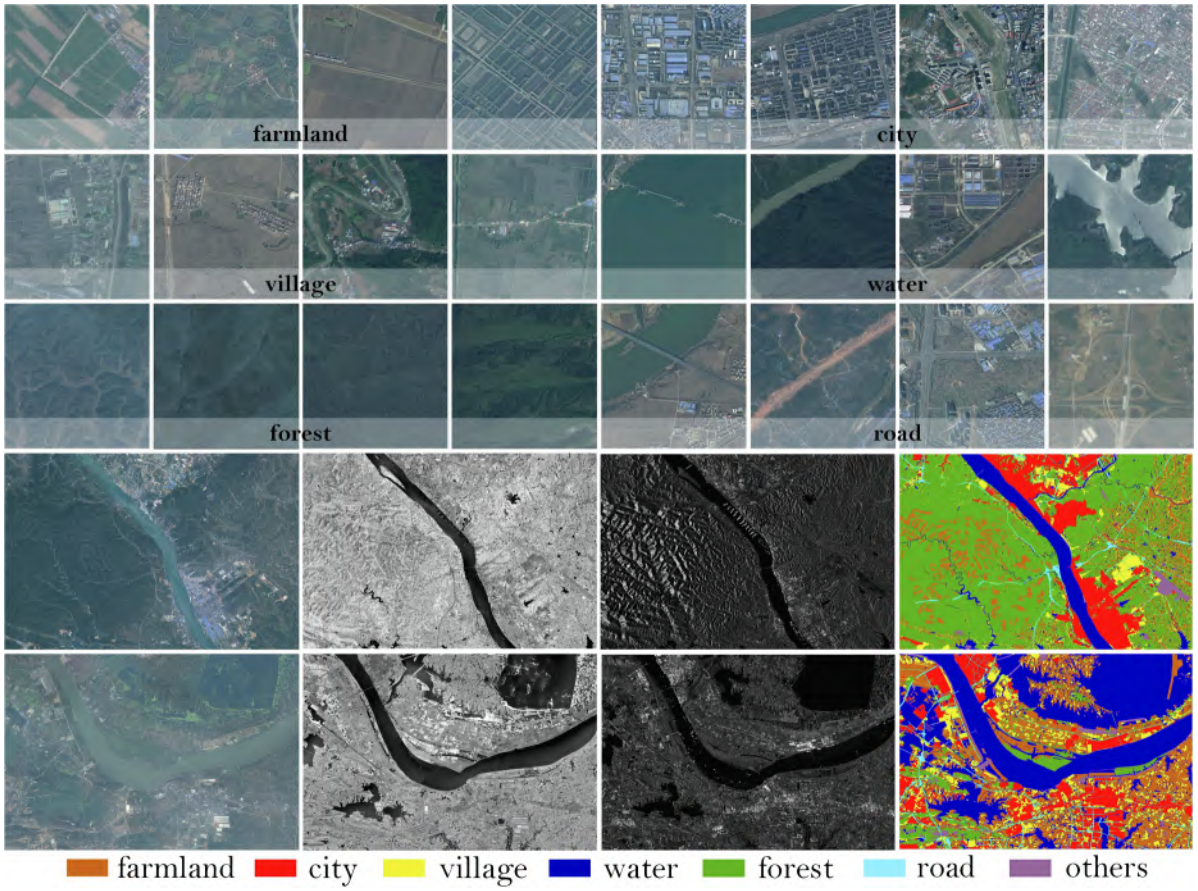


**Fig. 2.** Example of the main categories of the WHU-OPT-SAR dataset. The upper part shows the typical scene of this dataset, and the lower part is the optical images, NIR images, SAR images, and annotations from left to right.

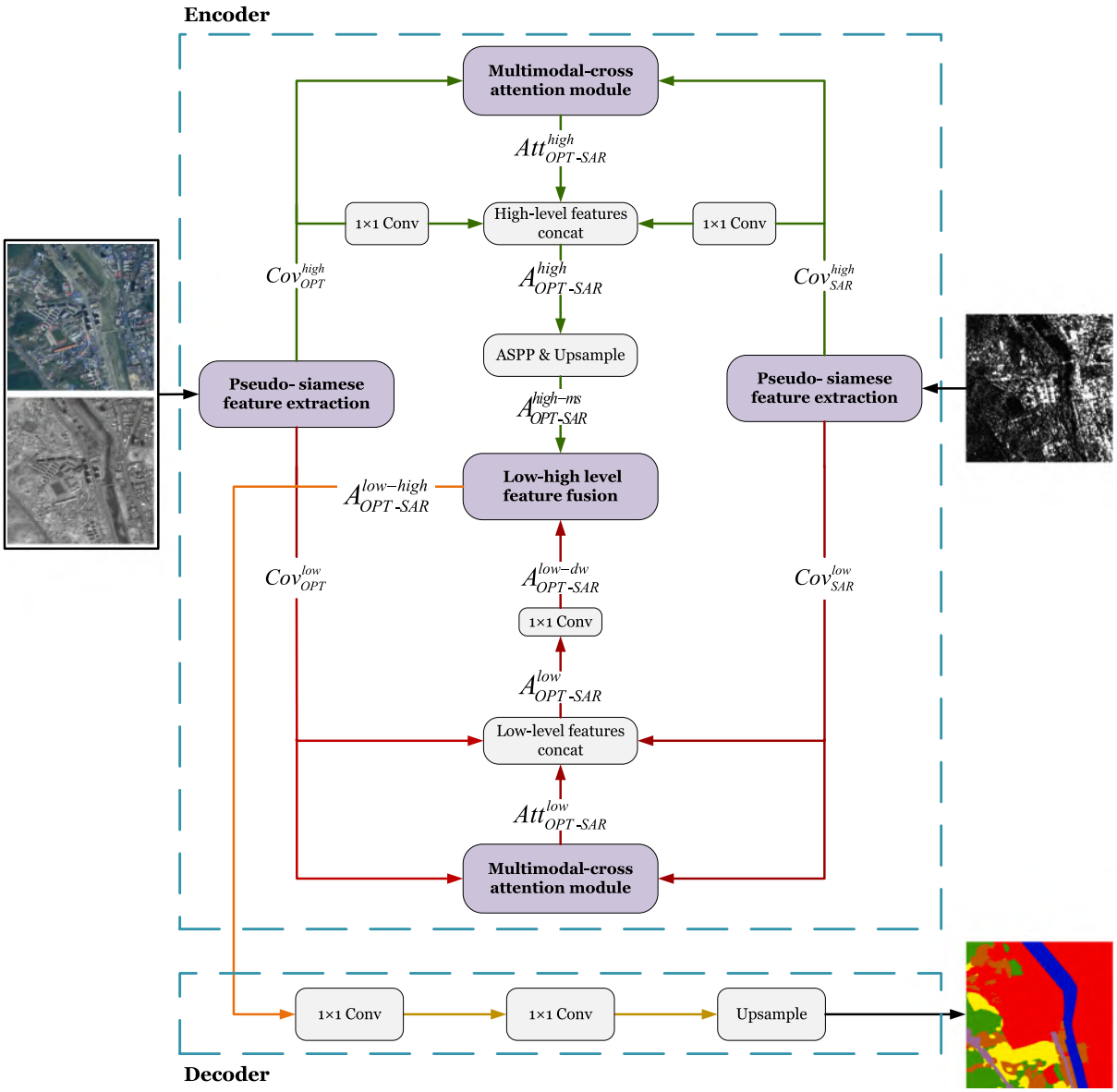


**Fig. 3.** Architecture of MCANet.

from the encoder and uses deconvolution and up sampling to obtain the joint semantic segmentation results of optical and SAR images.

### 2.2.2. *Pseudo-siamese feature extraction module*

Owing to the huge difference between the imaging characteristics of optical and SAR images, a pseudo-siamese network was designed. SAR and optical patches were input into independent convolutional streams with the same structure in parallel for feature extraction to obtain meaningful representation, thus hindering the interference between optical and SAR image due to the huge difference in ground object expression.

Considering the high-level feature stream as an example, if the optical patch $x_{OPT}$ and SAR patch $x_{SAR}$ are input into the two convolution streams of the pseudo-siamese network, through recursion, the high-level features of any deep unit, $L$, of the optical patch can be expressed as follows:

$$Cov^{high}_{OPT} = x^{l}_{OPT} + \sum_{i=l}^{L-1} F_{OPT}(x^{i}_{OPT}, W^{i}_{OPT}) \tag{1}$$

where $i$ is each convolutional layer unit, $l$ is a shallow unit, $W^{i}_{OPT}$ is a linear projection, and $F_{OPT}(x^{i}_{OPT}, W^{i}_{OPT})$ is a feature map to be learned by the optical patch. The high-level features of any deep unit, $L^{'}$ in the SAR patch are as follows:

$$Cov^{high}_{SAR} = x^{l'}_{SAR} + \sum_{j=l'}^{L'-1} F_{SAR}(x^{j}_{SAR}, W^{j}_{SAR}) \tag{2}$$

Where $j$ is each convolutional layer unit, $l^{'}$ is a shallow unit, $W^{j}_{SAR}$ is a linear projection, and $F_{SAR}(x^{j}_{SAR}, W^{j}_{SAR})$ is a feature map to be learned by the SAR patch.

### 2.2.3. *Multimodal-cross attention module*

Most conventional attention mechanisms does not have the ability to generate high-dimensional jointed features. In this section, a novel multimodal-cross attention module (MCAM) (Fig. 4), which effectively captures the position relationship of single data source feature maps, and interacts with the SAR and optical image feature maps in a two-dimensional space is proposed.

First, three 1 × 1 convolution kernels with the same parameters are

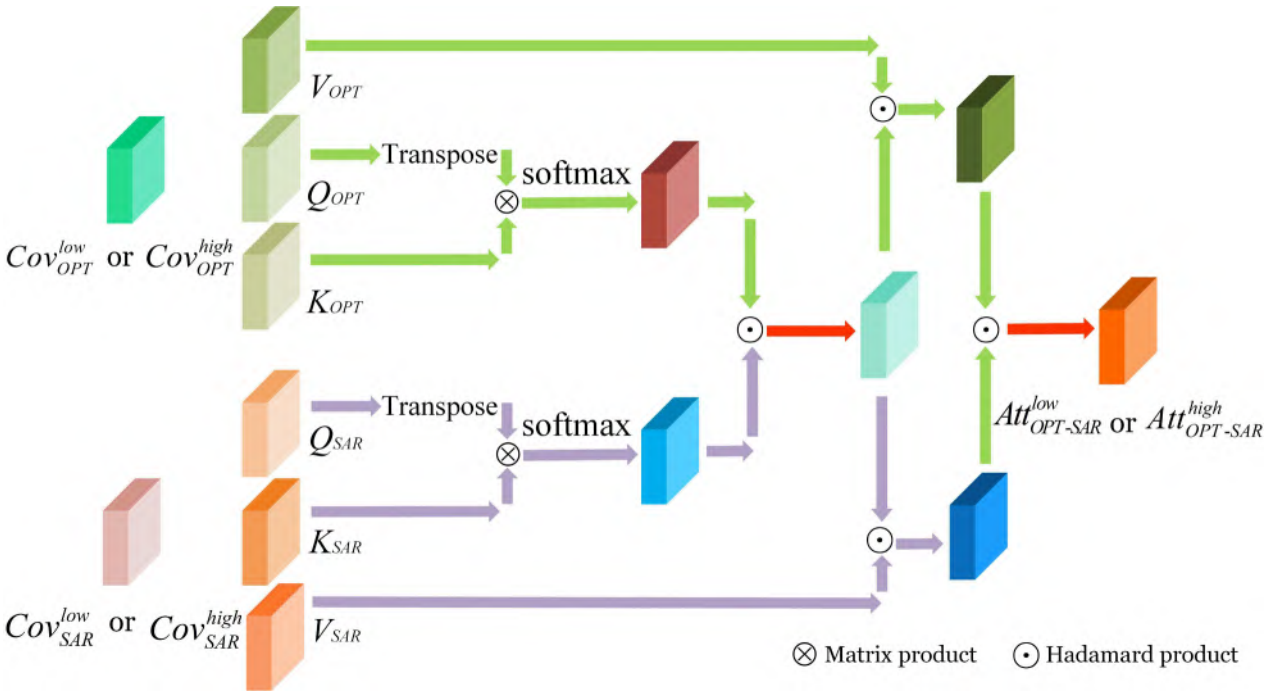


**Fig. 4.** Flowchart of the proposed MCAM.

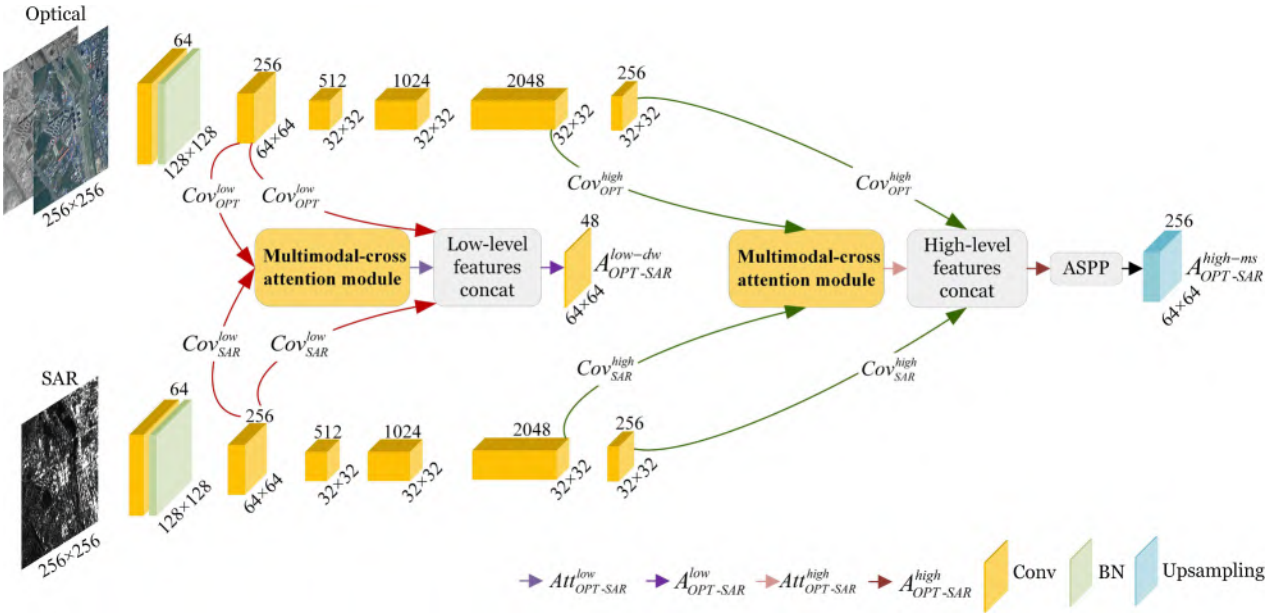


**Fig. 5.** Architecture of the encoder.

used to extract three matrices: $V_{OPT}, Q_{OPT}$, and $K_{OPT}$ from $Cov^{low}_{OPT}/Cov^{high}_{OPT}$, and $V_{SAR}, Q_{SAR}$, and $K_{SAR}$ from $Cov^{low}_{SAR}/Cov^{high}_{SAR}$, respectively. Second, the transpose of $Q_{OPT}$ is multiplied by $K_{OPT}$. The feature maps of SAR image do the same. This mechanism bears the advantage of considering the entire image as the observation range and strengthening the context information between long-distance pixels. Third, we use the Hadamard product to cross the nonlinear features of activated self-attention maps of SAR and optical images to obtain the joint attention maps. This method offers the advantage of allowing for the learning of the interactive hiding relationship between features and capturing the second-order correlation between the two kinds of features. Finally, the optical and SAR joint attention maps are multiplied by $V_{OPT}$ and $V_{SAR}$ respectively, and then features are crossed to obtain the final joint attention maps $Att^{low}_{OPT-SAR}$ / $Att^{high}_{OPT-SAR}$.

Considering the high-level feature stream as an example, we describe the detailed operation of the proposed MCAM as follows:

(1) $1 \times 1$ Convolution is used to transform the features of the optical patch into three identical feature maps: $V_{OPT}, Q_{OPT}$, and $K_{OPT}$ with the same number of channels. In the same way, the SAR patch used to obtain three identical feature maps: $V_{SAR}, Q_{SAR}$, and $K_{SAR}$.

(2) To obtain the self-attention of a single data source, the transposes of $Q_{OPT}$ and $K_{OPT}$ are multiplied, and then activated using the softmax function:

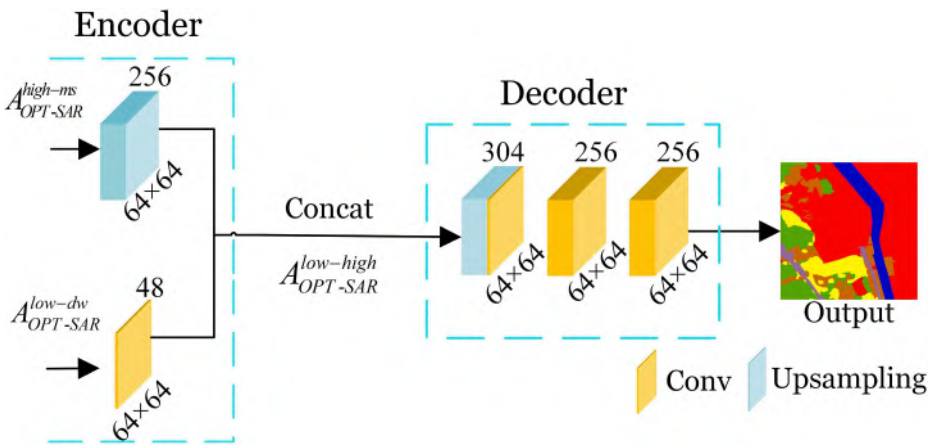


**Fig. 6.** Architecture of the decoder.

**Table 2**
Number of images in WHU-OPT-SAR dataset (256 × 256 pixels).

| Dataset | Train | Validation | Test |
|---|---|---|---|
| Number | 17640 | 5880 | 5880 |

$$S^{high}_{OPT} = softmax(Q^{T}_{OPT} \otimes K_{OPT}) \tag{3}$$

where $S^{high}_{OPT}$ is the self-attention map of the optical patch.

The self-attention map $S^{high}_{SAR}$ of the SAR patch can be expressed as follows:

$$S^{high}_{SAR} = softmax(Q^{T}_{SAR} \otimes K_{SAR}) \tag{4}$$

Through matrix multiplication, the internal correlation of features is captured, and the remote dependency between features is obtained, which can effectively model the context.

(3) The information contained in SAR and optical patches is different and mutually independent and complementary. The attention map of the high-level feature cross fusion in the $L^{high}$ layer of deep unit is expressed as follows:

$$S^{high}_{cro} = S^{high}_{OPT} \odot S^{high}_{SAR} \tag{5}$$

(4) The joint attention map, $S^{high}_{cro}$, is used to weigh the feature map of the optical patch $V_{OPT}$, and the weighted feature map $Att^{high}_{OPT}$ is expressed as follows:

$$Att^{high}_{OPT} = S^{high}_{cro} \odot V_{OPT} \tag{6}$$

Similarly, the weighted feature map of SAR patch $Att^{high}_{SAR}$ is shown as follows:

$$Att^{high}_{SAR} = S^{high}_{cro} \odot V_{SAR} \tag{7}$$

(5) The weighted feature map of SAR and optical patch is crossed to get the final joint attention map$Att^{high}_{OPT-SAR}$:

$$Att^{high}_{OPT-SAR} = Att^{high}_{OPT} \odot Att^{high}_{SAR} \tag{8}$$

### 2.3. Network details

#### 2.3.1. Encoder

As shown in Fig. 5, the optical and SAR images are input into ResNet101, and the convoluted $Cov^{low}_{OPT}$ and $Cov^{low}_{SAR}$ are input into MCAM to obtain the $Att^{low}_{OPT-SAR}$ with the size of 64 × 64 × 256, the $Cov^{low}_{OPT}$, $Cov^{low}_{SAR}$, and $Att^{low}_{OPT-SAR}$ are concatenated to obtained a 64 × 64 × 768 $A^{low}_{OPT-SAR}$. Then, the channel of the feature map is reduced to 48 to obtain $A^{low-dw}_{OPT-SAR}$. The $Cov^{low}_{OPT}$ / $Cov^{low}_{SAR}$ are convoluted to obtain the $Cov^{high}_{OPT}$ / $Cov^{high}_{SAR}$ of optical and SAR images. Subsequently, MCAM is used to get the $Att^{high}_{OPT-SAR}$ with a size of 32 × 32 × 2048. 1 × 1 convolution reduces the channels of $Cov^{high}_{OPT}$ / $Cov^{high}_{SAR}$ to 256 separately, which contributes to the enlargement of the function of $Att^{high}_{OPT-SAR}$. Then, the reduced $Cov^{high}_{OPT}$, $Cov^{high}_{SAR}$, and $Att^{high}_{OPT-SAR}$ are concatenated to obtain 32 × 32 × 2560-sized $A^{high}_{OPT-SAR}$. Finally, the feature map is input into ASPP to obtain a 32 × 32 × 256-sized feature map and up sample it to 64 × 64 × 256 that is $A^{high-ms}_{OPT-SAR}$.

#### 2.3.2. Decoder

As shown in Fig. 6, $A^{low-dw}_{OPT-SAR}$ and $A^{high-ms}_{OPT-SAR}$ are concatenated to obtain $A^{low-high}_{OPT-SAR}$, which is then input into the decoder, through two convolution layers, and the final semantic segmentation result is obtained by up sampling.

## 3. Experiments and results

### 3.1. Experimental setups

**Dataset:** The image in the WHU-OPT-SAR dataset is cropped to 256 × 256 pixel patches without overlapping. The number of images allocated to training set, validation set, and test set is shown in Table 2.

**Model Training:** Parameters of ResNet-101 are initialized using ImageNet (Deng et al., 2009). When the input is not a 3-channel image, Gaussian distribution is implemented to initialize the new conv1 layer. Specifically, we used the Adam optimizer on the Ubuntu 16.04 platform to train the model with a GTX1080Ti (memory 11 GB). The hyper parameters are set as follows: the batch size is 8, epoch number is 50, momentum value is 0.5, and initial learning rate is 0.001. When the error rate stopped decreasing, we divided the learning rate by 10 and updated the parameters with the new values. The loss function uses cross entropy loss.

**Evaluation Metrics:** To evaluate our approach, indexes such as overall accuracy (OA), Kappa coefficient (Kappa), and user's accuracy are introduced to evaluate the results (Olofsson et al., 2014).

### 3.2. Experiments on WHU-OPT-SAR dataset

In this section, we compared the accuracy of MCANet with other deep learning approaches: Deeplabv3+ (Chen et al., 2018), PSCNN (Mou et al., 2017a), MRSDC (Xu et al., 2017), V-FesuNet (Audebert et al., 2018), and MBFNet (Li et al., 2020). We conduct two sets of experiments with Deeplabv3+ (Chen et al., 2018), using optical images

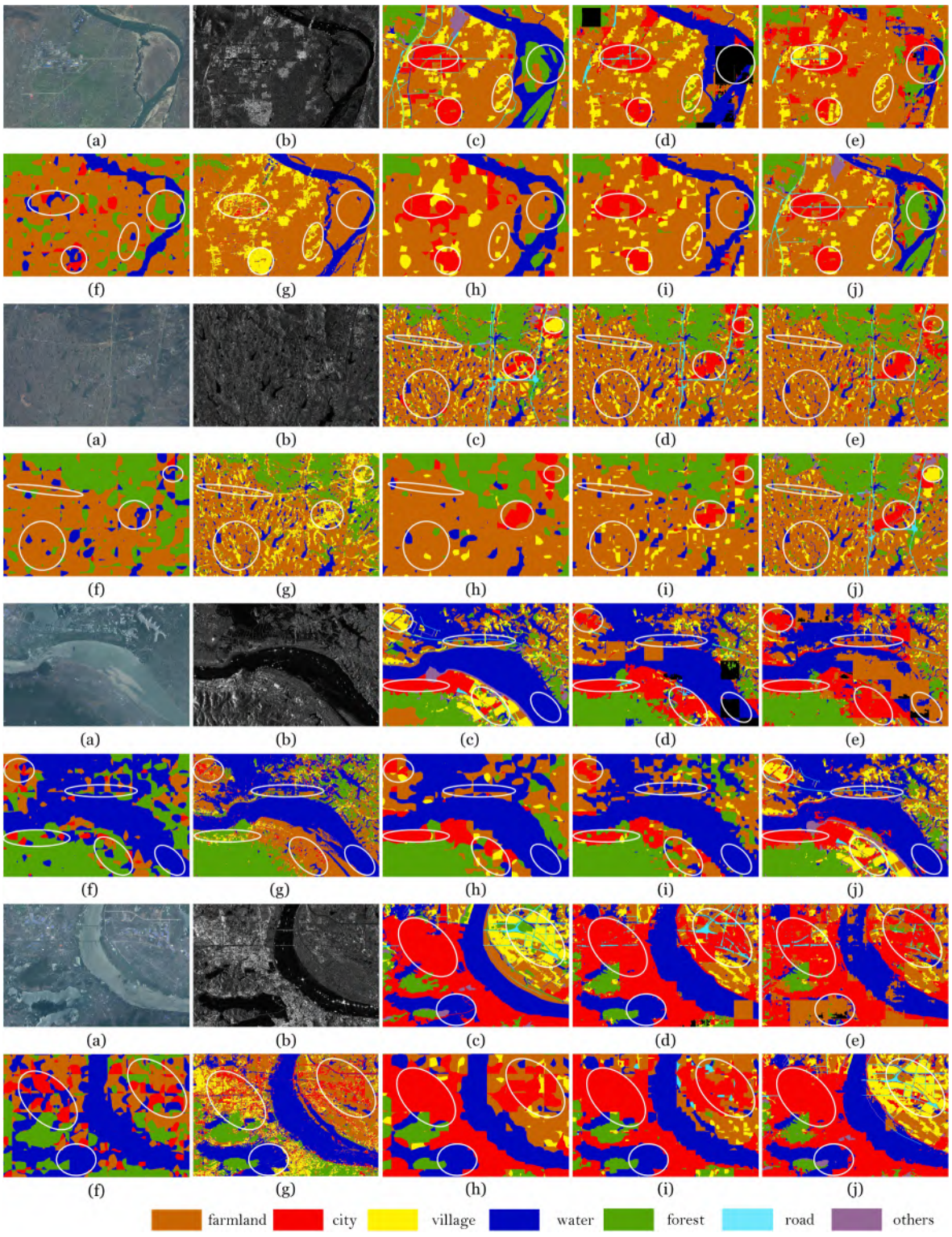


**Fig. 7.** Land use classification results of WHU-OPT-SAR dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) Deeplabv3+(RGBN) classification results. (e) Deeplabv3+(RGBN-SAR) classification results. (f) PSCNN classification results. (g) MRSDC classification results. (h) V-FesuNet classification results. (i) MBFNet classification results. (j) MCANet classification results.

(concatenated RGB and NIR channels) as the dataset (named Deeplabv3+(RGBN)), and optical-SAR images (concatenated RGB, NIR and corresponding SAR channels) as the dataset (named Deeplabv3+(RGBN-SAR)). PSCNN (Mou et al., 2017a) introduced pseudo-siamese network into the processing of multi-source remote sensing image for the first time, so we borrowed its network structure, yet its loss function was changed into cross entropy loss of semantic segmentation. MRSDC (Xu et al., 2017), and V-FesuNet (Audebert et al., 2018) are multi-modal semantic segmentation networks with concatenated features. MBFNet (Li et al., 2020) uses spatial attention for fusion.

Fig. 7 shows the results of land classification in four groups from top to bottom. The overall classification effect of each approach is similar,

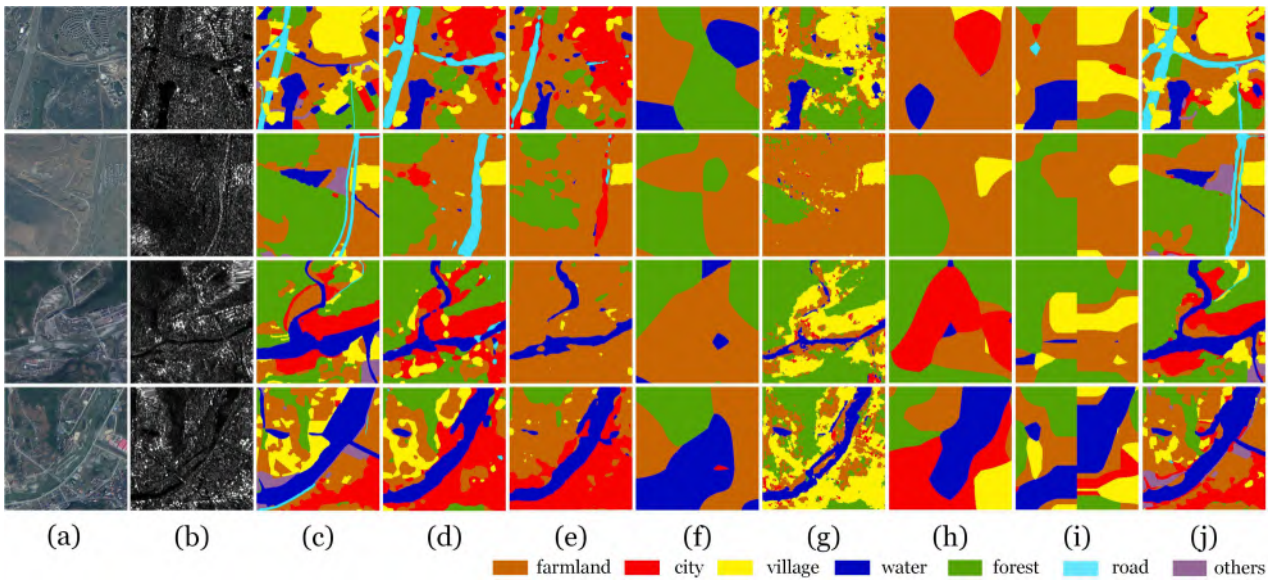


**Fig. 8.** Land use classification results of 256 × 256 pixel images in the WHU-OPT-SAR dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d-j) Deeplabv3+(RGBN) classification results, Deeplabv3+(RGBN-SAR) classification results, PSCNN classification results, MRSDC classification results, V-FesuNet classification results, MBFNet classification results, and MCANet classification results.

**Table 3**
Land use classification accuracy of different approaches on WHU-OPT-SAR dataset.

| Methods | OA | Kappa | User's Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | farmland | city | village | water | forest | road | others |
| Deeplabv3+(RGBN) | 0.798 | 0.690 | 0.737 | 0.556 | 0.509 | 0.644 | 0.895 | 0.257 | 0.074 |
| Deeplabv3+(RGBN-SAR) | 0.786 | 0.673 | 0.644 | 0.587 | 0.414 | 0.534 | 0.942 | 0.240 | 0.061 |
| PSCNN | 0.703 | 0.520 | 0.588 | 0.095 | 0.002 | 0.413 | 0.943 | 0.081 | 0.010 |
| MRSDC | 0.776 | 0.648 | 0.665 | 0.131 | 0.500 | 0.530 | 0.925 | 0.058 | 0.024 |
| V-FesuNet | 0.783 | 0.655 | 0.668 | 0.493 | 0.115 | 0.498 | 0.953 | 0.073 | 0.033 |
| MBFNet | 0.783 | 0.658 | 0.702 | 0.429 | 0.262 | 0.478 | 0.940 | 0.064 | 0.035 |
| **MCANet (Our's)** | **0.829** | **0.737** | **0.743** | **0.622** | **0.531** | **0.657** | **0.955** | **0.310** | **0.098** |

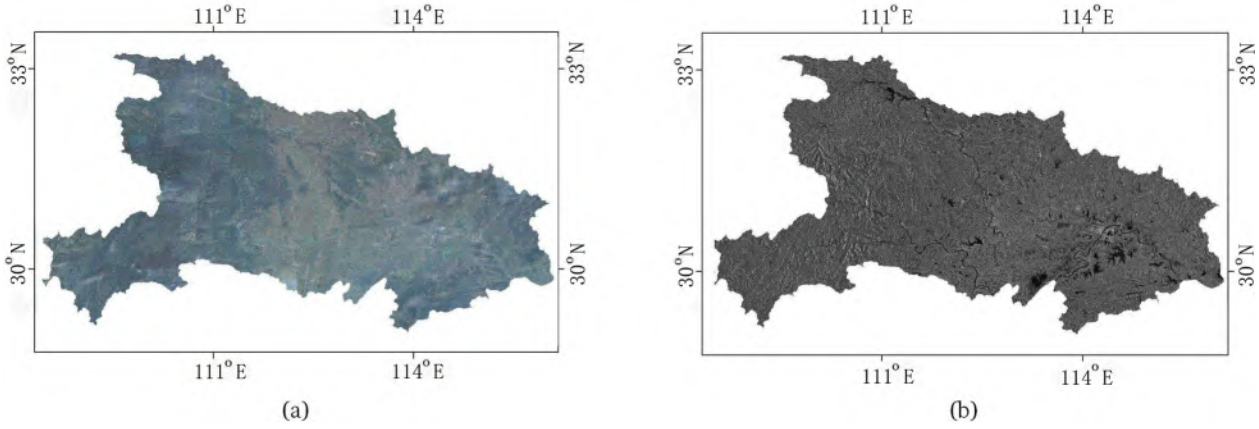


**Fig. 9.** Optical and SAR images of Hubei Province. (a) Optical image of GF-1 in Hubei Province. (b) SAR image of GF-3 in Hubei Province.

but MCANet distinguishes the details of each ground object more accurately. As shown in (d), the classification results of Deeplabv3+(RGBN) show that many places such as *water* and *forest* have not been successfully detected. As shown in (e) in the first, third and fourth groups, there is a large area of confusion between *water* and *farmland* in the results of the Deeplabv3+(RGBN-SAR), because the stacking of optical and SAR image at the image level causes them to interfere with each other in the feature extraction process. As shown in (j), the above situation has been improved in the results of MCANet. Because of the low backscattering coefficient of *water* in SAR images, the *water* appears black with clear outlines, which are clearly distinguished from *farmland*. As shown in the (f) in the first to fourth group, in the results of PSCNN (Mou et al., 2017a), large areas of the *city* and *village* have not been detected. The (g) in the first, third and fourth groups shows that MRSDC (Xu et al., 2017) did not detect narrow *city* areas between *forest* and *water*. The (h) and (i) in the first, second and third groups show that V-FesuNet (Audebert et al., 2018) and MBFNet (Li et al., 2020) missed a large number of small and scattered *village*. This may be because those

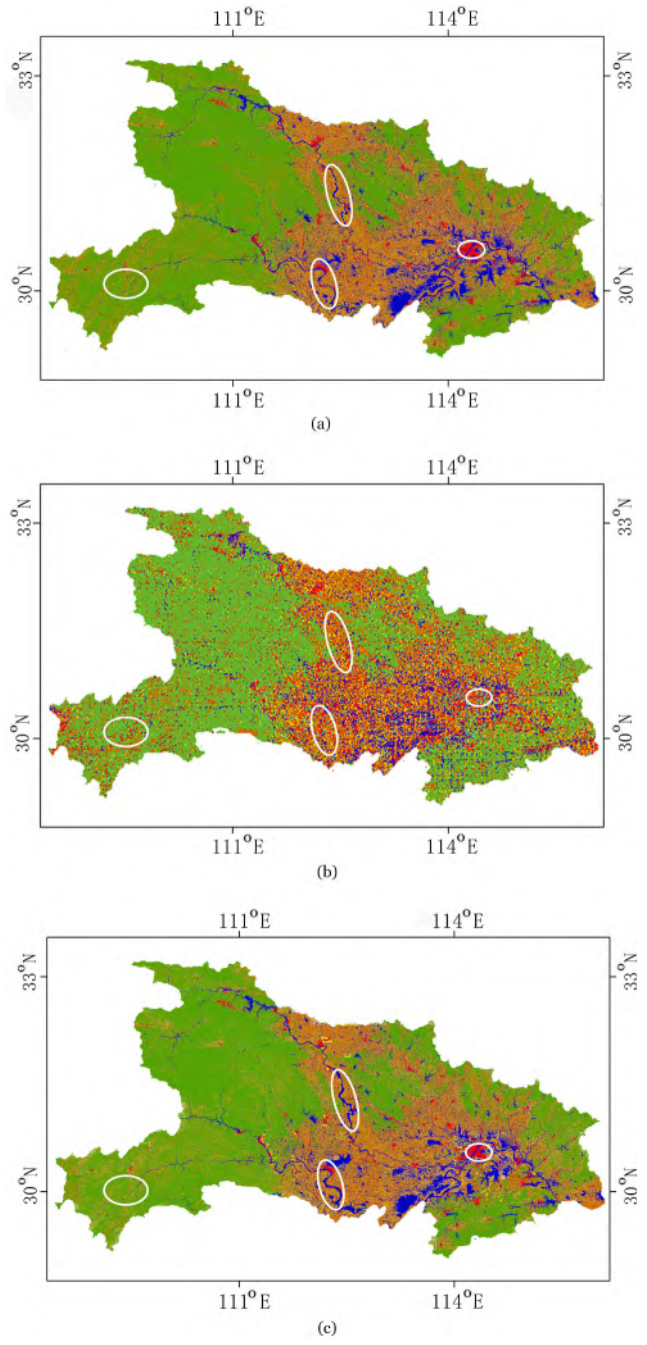


**Fig. 10.** Land use classification results of Hubei Province. (a) Land use classification annotation in Hubei Province. (b) Deeplabv3+(RGBN) classification result of land use in Hubei Province. (c) MCANet classification result of land use in Hubei Province.

networks cannot fuse multi-scale features. However, both low-level spatial features and high-level semantic features are fused in our approach, thus reducing the loss of effective features. Therefore, as shown in (j), our approach rarely misses *city* and *village* segmentations. The (g), the (h) and (i) in the fourth group show that MRSDC (Xu et al., 2017), MBFNet (Li et al., 2020) and V-FesuNet (Audebert et al., 2018) identify a large *village* as *city*. But our approach can use the joint attention maps comprising optical and SAR images to guide the entire MCANet to better identify the difference between *city* and *village* at the high-dimensional level, so that the two categories are more distinguishable. In addition, as shown in (c)-(j) in the second to fourth groups, our approach can detect a *road* with higher continuity compared to other approaches. These results show that our approach can establish long dependence between pixels and effectively model the contextual semantic information on the *road* to improve the continuity of the *road*. Furthermore, as shown in Fig. 8, the edge noise of each category extracted by our approach is smaller than those of Deeplabv3+(RGBN) and Deeplabv3+(RGBN-SAR), and the edge details are more accurate than those of the MBFNet (Li et al., 2020), V-FesuNet (Audebert et al., 2018), MRSDC (Xu et al., 2017), and PSCNN (Mou et al., 2017a).

Quantitative evaluation was also conducted to compare the effectiveness of these approaches. As shown in Table 3, MCANet has the highest OA (0.829) and Kappa (0.737), which exceed those yielded by Deeplabv3+(RGBN) (OA of 0.798 and Kappa of 0.690) and Deeplabv3+(RGBN-SAR) (OA of 0.786, and Kappa of 0.673). Compared with other approaches, the improvements of all indicators were consistent. Compared with MBFNet (Li et al., 2020), the OA and Kappa of MCANet increased by 4.6% and 7.9%, respectively, because MCANet crosses the high-level semantic features of optical and SAR images in second order, so that the extracted categories are more accurate. Among them, the accuracy of *city* and *village* segmentation was nearly five times higher than that of MRSDC (Xu et al., 2017) (from 0.131 to 0.622) and V-FesuNet (Audebert et al., 2018) (from 0.115 to 0.531).The accuracy of *road* segmentation was three to four times higher than that of the four comparison approaches.

### 3.3. Experiments on Hubei Province, China, remote sensing images

To test the performance of our approach on large-scale remote sensing images, we conducted land use classification experiments in Hubei Province, China, covering an area of approximately 190000 $km^2$. Fig. 9 shows all the optical and SAR images of Hubei Province.

Fig. 10(a) shows the annotation of land use classification in Hubei Province. The imbalance of each category was obvious. Fig. 10(b) and (c) show the land use classification results of Hubei Province using Deeplabv3+(RGBN) and MCANet, respectively, in the pre-training model of the WHU-OPT-SAR dataset. The overall performances of Deeplabv3+(RGBN) and MCANet were similar. However, Deeplabv3+(RGBN) results show the poor continuity of the edges of the ground objects led to more severe pixel confusion. The overall extraction performance of our approach was more optimized; particularly, the boundary detection of *water* and *city* was more accurate, and the confusion between *forest* and *farmland* relatively less.

As shown in Fig. 11, details 15 use partial images with a small range and single category to show the classification results of each representative ground object. Detail 1 mainly comprised *farmland* with scattered *water*. Deeplabv3+(RGBN) mistakenly segmented a large number of *farmland* into *forest*, and our method avoided this error. This may be due to the greater discrimination between *forest* and *farmland* in optical and SAR images. Detail 2 comprised *village* with extensive *water*. Detail 3 is a slender *water* passing through a *city*. Because artificial buildings are similar in spectrum and texture, Deeplabv3+(RGBN) misclassify *city* as *road* and *village*. However, our approach can fully mine the key semantic information that distinguishes *village* from *city* on multi-source images, without relying on superficial visual features and fuse this information at a higher level to obtain more accurate classification results. The continuity of the long and narrow *water* in details 3 and 4, and the long and narrow *road* in details 5 are better classified via our method than via Deeplabv3+(RGBN).

The prediction accuracy of Hubei Province are shown in Table 4. The OA and Kappa of our approach was 5% and 7% higher than that of Deeplabv3+(RGBN). In addition, our approach improved the prediction accuracy of each category, such as the accuracy of *farmland*, *city*, *village*, *water*, *forest*, and *road* classification by approximately 4.7%, 4.9%, 4.8%, 2.3%, 7.4% and 3%. The proportions of the various categories were very unbalanced. However, our approach learned the multi-modal relationship of low-level space and high-level space and can better handle the situation. Therefore, the model trained by our approach is more robust

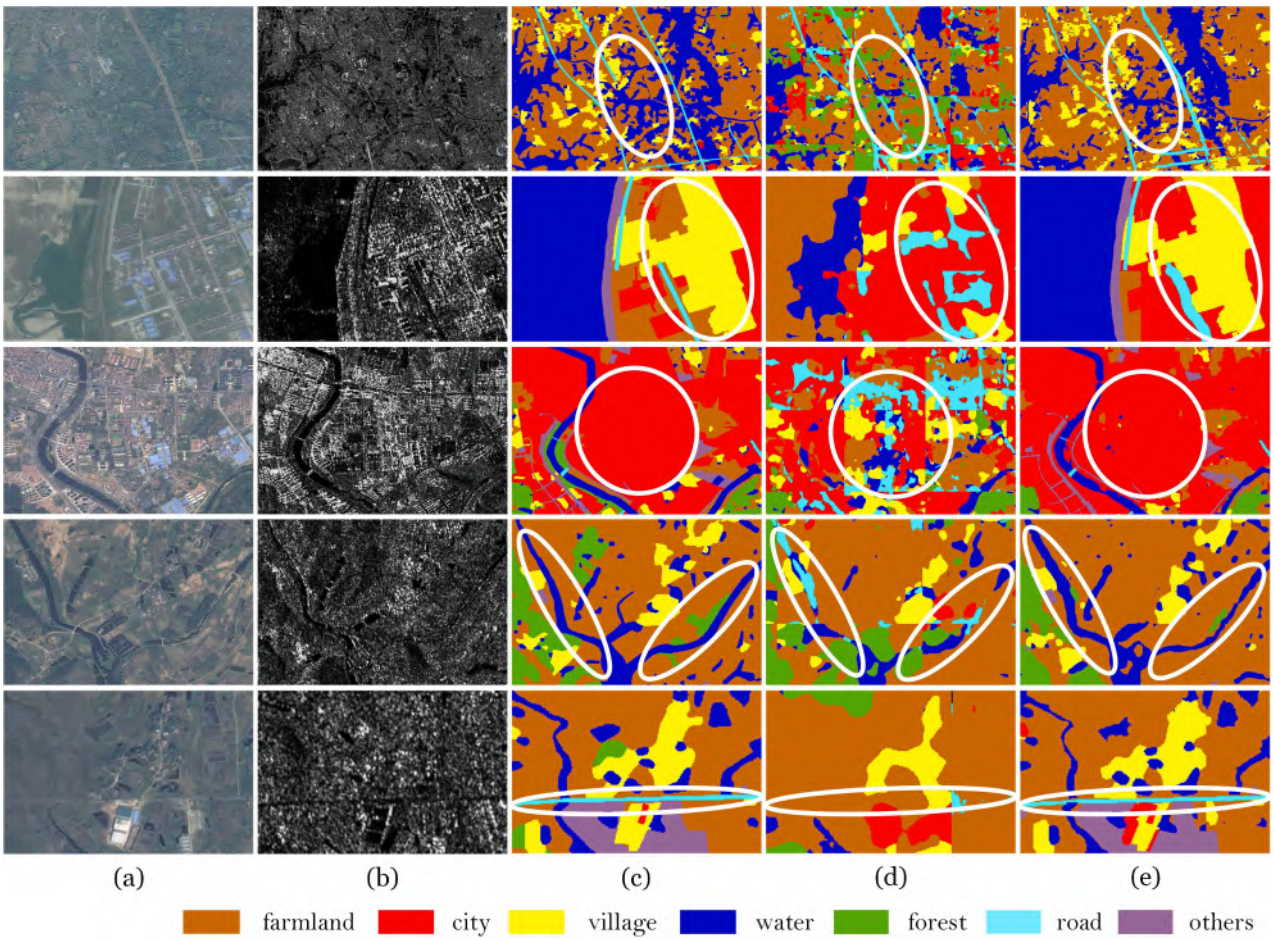


**Fig. 11.** Land use classification results of Hubei Province dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) Deeplabv3+(RGBN) classification results. (e) MCANet classification results. From top to bottom, the first row to the seventh row is details 17.

**Table 4**
Land use classification accuracy of different approaches on Hubei Province dataset.

| Methods | OA | Kappa | User's Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | farmland | city | village | water | forest | road | others |
| Deeplabv3+(RGBN) | 0.759 | 0.638 | 0.693 | 0.487 | 0.471 | 0.625 | 0.848 | 0.254 | 0.053 |
| **MCANet (Our's)** | **0.809** | **0.706** | **0.740** | **0.495** | **0.519** | **0.648** | **0.922** | **0.284** | **0.075** |

and has a stronger generalization ability than that trained by Deeplabv3+(RGBN).

## 4. Discussion

### 4.1. Advantages of MCANet

Considering the different characteristics of optical and SAR images, we propose MCANet whose advantages are as follows. (1) It inputs the optical and SAR images into pseudo-siamese networks, avoiding the loss of feature caused by the mutual influence of the optical and SAR images in the feature extraction part owing to the huge difference. (2) MCAM was designed in the fusion stage of optical and SAR image features. The feature maps of the optical and SAR images are transposed and then multiplied separately, which can establish the middle and long dependence between the pixels in the remote sensing image and capture the global relationship of the image. (3) After obtaining the respective attention maps of optical and SAR, they are cross-fused using the Hadamard product. Compared with the general DCNN, this approach has the advantage of fitting nonlinear features adequately by establishing the second-order representation of the attention map of optical and SAR images, and improve the expression ability of the entire semantic segmentation model. (4) Our approach can perform second-order attention intersection on the low-level and high-level features.

### 4.2. Significant role of SAR images in ground object extraction

Fig. 12 shows an overall consistent segmentation result of all approaches, yet the details are quite different. Fig. 12(d) indicates that Deeplabv3+(RGBN) confuses *city* and *village* due to the lack of supplementary information provided by SAR images, while Deeplabv3+(RGBN-SAR) cannot avoid the mutual interference of SAR and optical images, which results in poor *road* continuity and mix-up of *city* and *village*. The fusion mechanisms of optical and SAR images by PSCNN (Mou et al., 2017a), MRSDC (Xu et al., 2017), V-FesuNet (Audebert et al., 2018) and MBFNet (Li et al., 2020) are relatively straightforward, which easily leads to information loss. However, MCANet have a lower miss and false rate of these kind of objects. This is because a global association of pixels in the image is established by MCAM to obtain more accurate context space and semantic information, which improves the

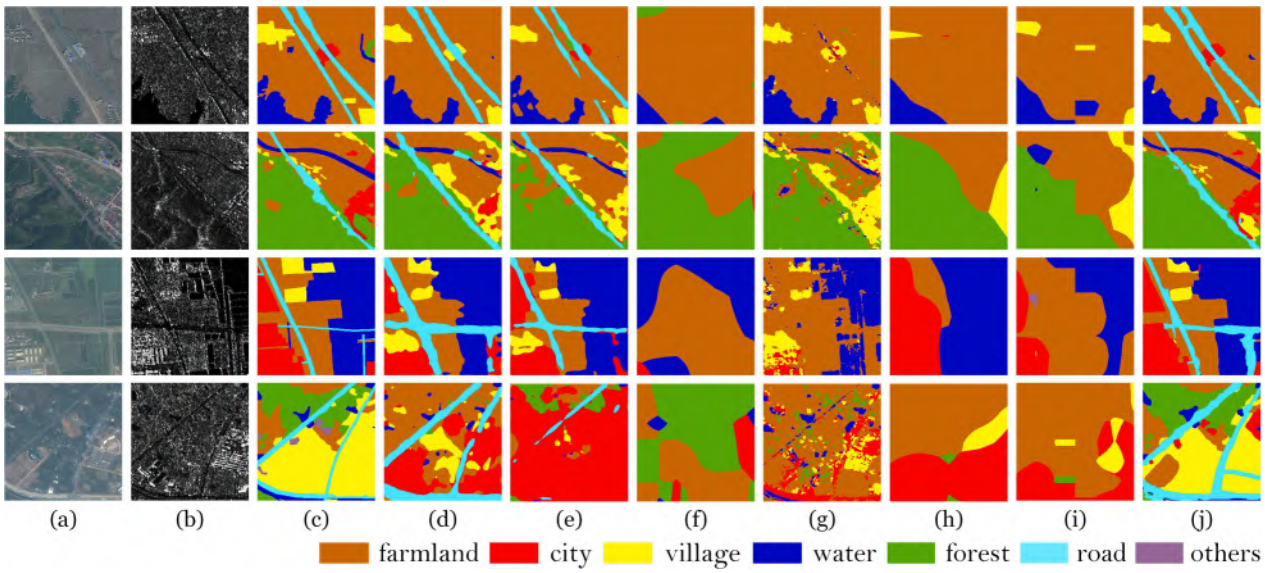


**Fig. 12.** Land use classification results of the main categories of the WHU-OPT-SAR dataset. (a) Optical images. (b) SAR images. (b) Annotations. (d) Deeplabv3+(RGBN) classification results. (e) Deeplabv3+(RGBN-SAR) classification results. (f) PSCNN classification results. (g) MRSDC classification results. (h) V-FesuNet classification results. (i) MBFNet classification results. (j) MCANet classification results.

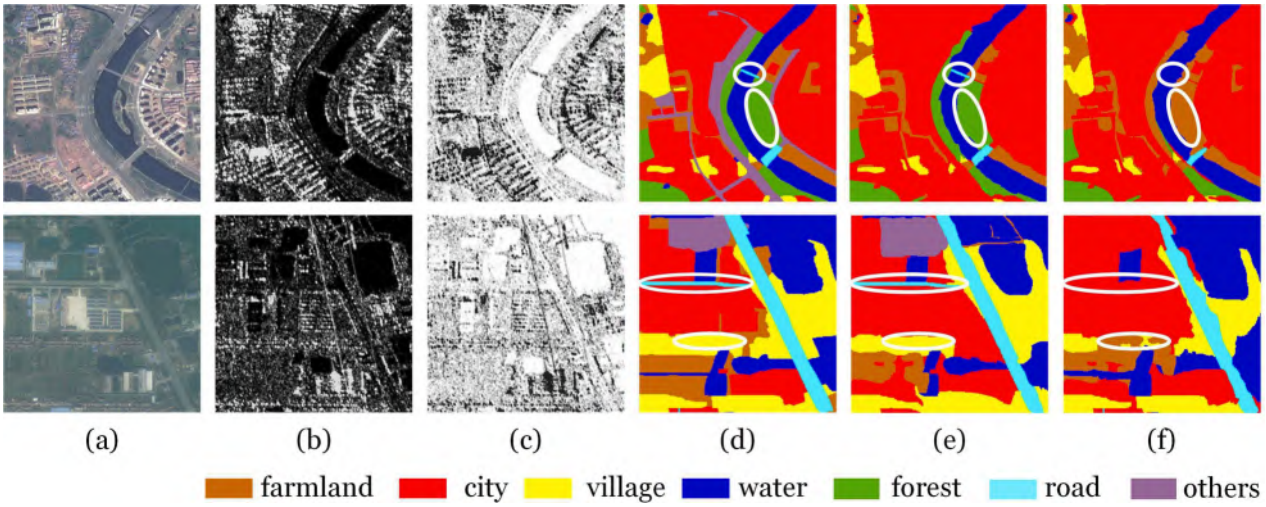


**Fig. 13.** Land use classification results of SAR images and 255-SAR. (a) Optical images. (b) SAR images. (c) 255-SAR images. (d) Annotations. (e) SAR segmentation results. (f) 255-SAR segmentation results.

**Table 5**
Land use classification accuracy of WHU-OPT-SAR dataset before and after SAR image inversion.

| Methods | OA | Kappa | User's Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | farmland | city | village | water | forest | road | others |
| MCANet (SAR) | **0.829** | **0.737** | **0.743** | 0.622 | **0.531** | 0.657 | **0.955** | **0.310** | **0.098** |
| MCANet (255-SAR) | 0.805 | 0.699 | 0.629 | **0.634** | 0.440 | **0.662** | 0.950 | 0.296 | 0.074 |

semantic segmentation accuracy of imbalanced datasets.

Table 3 shows that MCANet is conducive for the semantic segmentation of *farmland*, *city*, *village*, *water*, *forest*, *road* and *others*. Deeplabv3+(RGBN-SAR) exhibits lower accuracy than the Deeplabv3+(RGBN) approach. For example, the accuracy *city* detection was approximately 6.5 and 4.7 times higher than that by PSCNN (Mou et al., 2017a) and MRSDC (Xu et al., 2017), the accuracy of *village* detection was approximately 4.6 times higher than that by V-FesuNet (Mou et al., 2017a), and the accuracy of *road* detection 4.8 times higher than that by MBFNet (Li et al., 2020). This is attributed to SAR images are more sensitive to structures, which can remarkably supplement the semantic information.

### 4.3. Key feature presented by SAR images

In land use classification, the relationship between the information provided by the SAR image and the optical image is closely related to the structural design of the multi-modal fusion network. We are interested in whether the intensity distribution of SAR image pixels is potentially

correlated to the optical image spectrum, and what information SAR images provide in semantic segmentation. Based on this, we designed the following experiment.

The inverted image was obtained by subtracting the pixel value of the SAR image from the pixel value of 255, as shown in Fig. 13(c). The outlines of the *city* and *water* are relatively obvious. Compared with Fig. 13(e), Fig. 13(f) shows that the result of the SAR image after inversion is not significantly adversely affected. The segmentation results of *water* and *city* are more accurate. Although the boundary noise of *road*, *village* and *farmland* has increased and certain detections are missed, the overall classification accuracy has not decreased severely.

Table 5 shows the land use classification accuracy of WHU-OPT-SAR dataset before and after SAR inversion.MCANet(SAR) and MCANet(255-SAR) respectively represent the original and inverse SAR and optical image joint segmentation. As seen in Table 5, compared with the MCANet (SAR), the OA of the MCANet (255-SAR) is reduced by approximately 3.8%, and the Kappa is reduced by approximately 2.4%. The accuracy of *city* and *water* detection increased slightly (1.9 times and 0.8 times, respectively). This may be because these two categories still have obvious contour in the inverted image. After destroying the intensity distribution of the SAR image, the contour features of the SAR remain. The intensity distribution of SAR images is not directly related to the spectral information of optical images. Therefore, when designing MCANet, instead of simply stacking or multiplying optical and SAR images, MCAM is designed for their fusion.

### 4.4. Disadvantages of MCANet

The above experiments have demonstrated the effectiveness of our algorithm; however, our approach exhibit certain shortcomings.The side-view imaging characteristics of SAR sensors cause shadows and overlaps in its images, especially in mountain areas. Our approach is not designed specifically for these areas. Therefore, based on the above shortcomings, we will further improve this approach to adapt to different situations in future.

## 5. Conclusion

In this study, land use classification based on optical and SAR images is studied in depth as follows. (1) A WHU-OPT-SAR dataset, a large-scale optical and SAR images land use classification dataset including seven types of ground objects: *farmland*, *city*, *village*, *water*, *forest*, *road* and *others*, covering approximately 50,000 $km^2$. (2) MCANet, a multimodal-cross attention network, comprising a pseudo-siamese feature extraction, multimodal-cross attention, and low-high level feature fusion modules is developed. (3) We tested our algorithm and performed two groups of experiments on the WHU-OPT-SAR dataset and Hubei Province dataset, which proved the advantages of MCANet in land use classification joint with optical and SAR images. Furthermore, MCANet significantly improves the classification accuracy of *city*, *village*, *road*, *water*, *forest* and *farmland*. In addition, SAR images are observed to provide contour information in semantic segmentation of remote sensing images, which presents high reference value for future multimodal data fusion.

## Funding

This work was supported by the Postdoctoral Innovation Talent Support Program (BX2021222), Civil Aerospace Technology Pre Research Project(D040107), Natural Science Foundation of China projects (NSFC) (No: 41801397), National Key R&D Program of China (2018YFC0825803).

## CRediT authorship contribution statement

**Xue Li:** Conceptualization, Methodology. **Shasha Hou:** Investigation.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors thank the China Land Surveying and Planning Institute for its valuable contribution with labeling of Hubei Province dataset. We also would like to thank the Supercomputing Center of Wuhan University for providing computing power to conduct experiments.

## References

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proc. Eur. Conf. Comput. Vis. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Gómez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G., 2015. Multimodal classification of remote sensing images: A review and future directions. Proc. IEEE Inst. Electr. Electron. Eng. 103, 1560–1584. https://doi.org/10.1109/JPROC.2015.2449668.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. Comput. Vis. Image. Underst. 213–228. https://doi.org/10.1007/978-3-319-54181-5_14.

Jewell, N., 2010. An evaluation of multi-date spot data for agriculture and land use mapping in the united kingdom. Int. J. Remote Sens. 10, 939–951. https://doi.org/10.1080/01431168908903936.

Li, R., Duan, C., Zheng, S., Zhang, C., Atkinson, P.M., 2021. Macu-net for semantic segmentation of fine-resolution remotely sensed images. IEEE Geosci. Remote. Sens. Lett. 1–5 https://doi.org/10.1109/LGRS.2021.3052886.

Li, X., Lei, L., Sun, Y., Li, M., Kuang, G., 2020. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. Int. J. Appl. Earth Obs. Geoinf. 13, 1011–1026. https://doi.org/10.1109/JSTARS.2020.2975252.

Liu, J., Gong, M., Qin, K., Zhang, P., 2016. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans. Neural Netw. Learn. Syst. 29, 545–559. https://doi.org/10.1109/TNNLS.2016.2636227.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. ISPRS J. Photogramm. Remote Sens. 145, 96–107. https://doi.org/10.1016/j.isprsjprs.2018.01.021.

Mou, L., Schmitt, M., Wang, Y., Zhu, X.X., 2017a. Identifying corresponding patches in sar and optical imagery with a convolutional neural network. In: Proc. IEEE Int. Geosci. Remote Sens. Symp., pp. 5482–5485. https://doi.org/10.1109/LGRS.2018.2799232.

Mou, L., Zhu, X., Vakalopoulou, M., Karantzalos, K., Paragios, N., Le Saux, B., Moser, G., Tuia, D., 2017b. Multitemporal very high resolution from space: Outcome of the 2016 ieee grss data fusion contest. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 10, 3435–3447. https://doi.org/10.1109/JSTARS.2017.2696823.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

Pacifici, F., Del Frate, F., Emery, W.J., Gamba, P., Chanussot, J., 2008. Urban mapping using coarse sar and optical data: Outcome of the 2007 grss data fusion contest. IEEE Geosci. Remote Sens. Lett. 5, 331–335. https://doi.org/10.1109/LGRS.2008.915939.

Paneque-Gálvez, J., Mas, J.F., Moré, G., Cristóbal, J., Orta-Martínez, M., Luz, A.C., Guéze, M., Macía, M.J., Reyes-García, V., 2013. Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity. Int. J. Appl. Earth Obs. Geoinf. 23, 372–383. https://doi.org/10.1016/j.jag.2012.10.007.

Park, E., Han, X., Berg, T.L., Berg, A.C., 2016. Combining multiple sources of knowledge in deep cnns for action recognition. IEEE Win. Conf. Appl. Comput. Vis. 1–8. https://doi.org/10.1109/WACV.2016.7477589.

Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. ISPRS J. Ph RS. IV-2/W7, 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019.

Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing: An ever-growing relationship. IEEE Geosci. Remote Sens. Mag. 4, 6–23. https://doi.org/10.1109/MGRS.2016.2561021.

Solberg, A.H.S., Taxt, T., Jain, A.K., 1996. A markov random field model for classification of multisource satellite imagery. IEEE Trans. Geosci. Remote Sens. 34, 100–113. https://doi.org/10.1109/36.481897.

Toll, David L., 1985. Analysis of digital LANDSAT MSS and SEASAT SAR data for use in discriminating land cover at the urban fringe of Denver, Colorado. Int. J. Remote Sens. 6 (7), 1209–1229. https://doi.org/10.1080/01431168508948273.

Tong, X.Y., Xia, G.S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322. https://doi.org/10.1016/j.rse.2019.111322.

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., Zhang, B., 2017. Multisource remote sensing data classification based on convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56, 937–949.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (ocnn) for urban land use classification. Remote Sens. Environ. 216, 57–70. https://doi.org/10.1016/j.rse.2018.06.034.

Zhang, F., Yang, X., 2020. Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection. Remote Sens. Environ. 251, 112105. https://doi.org/10.1016/j.rse.2020.112105.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4, 22–40. https://doi.org/10.1109/MGRS.2016.2540798.

Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 113, 155–165. https://doi.org/10.1016/j.isprsjprs.2016.01.004.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307.