Multimodal Bilinear Fusion Network With Second-Order Attention-Based Channel Selection for Land Cover Classification

Xiao Li, Lin Lei, Yuli Sun, Ming Li, and Gangyao Kuang, *Senior Member, IEEE*

***Abstract*—As two different tools for earth observation, the optical and synthetic aperture radar (SAR) images can provide complementary information of the same land types for better land cover classification. However, because of the different imaging mechanisms of optical and SAR images, how to efficiently exploit the complementary information becomes an interesting and challenging problem. In this article, we propose a novel multimodal bilinear fusion network (MBFNet), which is used to fuse the optical and SAR features for land cover classification. The MBFNet consists of three components: the feature extractor, the second-order attention-based channel selection module (SACSM), and the bilinear fusion module. First, in order to avoid the network parameters tempting to ingratiate dominant modality, the pseudo-siamese convolutional neural network (CNN) is taken as the feature extractor to extract deep semantic feature maps of optical and SAR images, respectively. Then, the SACSM is embedded into each stream, and the fine channel-attention maps with second-order statistics are obtained by bilinear integrating the global average-pooling and global max-pooling information. The SACSM can not only automatically highlight the important channels of feature maps to improve the representation power of networks, but also uses the channel selection mechanism to reconfigure compact feature maps with better discrimination. Finally, the bilinear pooling is used as the feature-level fusion method, which establishes the second-order association between two compact feature maps of the optical and SAR streams to obtain the low-dimension bilinear fusion features for land cover classification. Experimental results on three broad coregistered optical and SAR datasets demonstrate that our method achieves more effective land cover classification performance than the state-of-the-art methods.**

***Index Terms*—Attention mechanism, bilinear pooling model, convolutional neural network (CNN), feature fusion, land cover classification, multimodal learning.**

I. Introduction

LAND cover classification plays an important role in land-use analysis, environment protection, urban planning, etc., and has been a hot research field [1]–[5]. However, most of existing land cover classification methods only use unimodal remote sensing (RS) images, *e.g.*, many methods using the optical images have faced the spectral confusion issue to lower the classification accuracy [2], [3], and others using the synthetic aperture radar (SAR) images have presented poor classification because of the quality of SAR images and noise interference [4]. With the rapid development of RS techniques, it is possible to obtain multimodal RS data from the same region, and the optical and SAR images can provide a variety of information on the land properties, such as the spectral information of the optical images [5] and the scattering information of SAR images [4], [6]. Numerous studies have shown that the optical and SAR data can provide complementary information from individual sources, which is good for land cover classification [7], [8]. As shown in Fig. 1(a), the optical patches between the farmland and water are hard to distinguish, whereas their SAR patches are clearly distinguishable. The SAR patches between the tree and unknown are similar, whereas their optical patches have

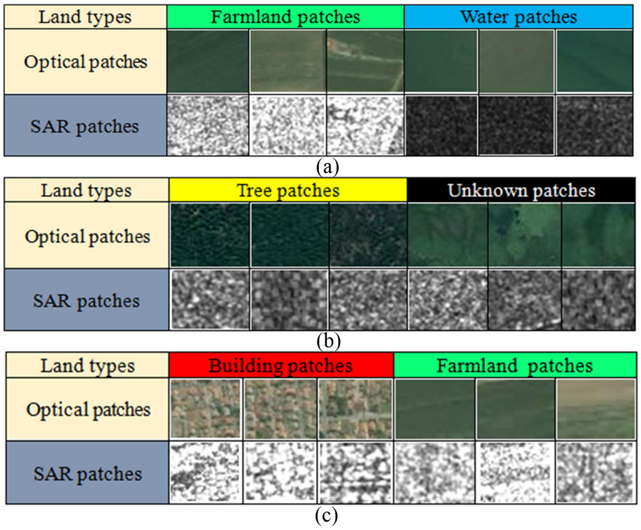


Fig. 1. Complementary of optical and corresponding SAR patches of the same land types on PoDelta. (a) Optical patches of farmland and water are similar and hard to distinguish, whereas the SAR patches are clearly distinguishable; (b) SAR patches of tree and unknown are hard to distinguish, whereas the optical patches are clearly distinguishable; (c) SAR patches of farmland and building are similar, whereas the optical patches are clearly different. Therefore, the optical and SAR patches the same land types are complementary.

Manuscript received September 7, 2019; revised December 26, 2019 and January 31, 2020; accepted February 15, 2020. Date of publication March 2, 2020; date of current version March 19, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61971426. *(Corresponding author: Lin Lei.)*

The authors are with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha 410073, China (e-mail: lxcherishm@126.com; alaleilin@163.com; sunyuli@mail.ustc.edu.cn; liming17@nudt.edu.cn; kuangyeats@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.2975252

different color characteristics, as shown in Fig. 1(b). Therefore, it is important to develop the efficient fusion models between the optical and SAR features, which can provide synergistic information on land properties.

In multimodal feature fusion models, the extraction of discriminative unimodal features plays an important role, whose quality directly affects the performance of the fusion features. Recently, the deep learning algorithms, especially the deep convolutional neural network (DCNN), as the powerful feature extractor, has been successfully applied in RS image classification [9], object detection [10], [11], and semantic image segmentation [12]. As an extension of the deep learning, multimodal deep learning models have been introduced by integrating deep learning techniques and multimodal learning methods [13]. Multimodal deep learning models can not only extract high-level unimodal features, but also capture the great association between different modal data [14]. Most of the existing multimodal deep learning models are multibranch network structure. And according to whether the weights of different branches are shared, the existing multimodal deep learning models can be roughly divided into two categories: the multimodal deep learning models with weight-sharing [14]–[16] and the multimodal deep learning models without weight-sharing [17]–[19]. For the processing of multimodal heterogeneous data, numerous research works have proven that the second category methods are more suitable [17], [20]. The reason is that the shared weights of different branches are easily temped to ingratiate dominant modal data, which will lead to the diversity loss of different modal features in the first category model. However, the unshared weights of different branches do not interfere with each other in second category model, which can avoid the drawback of the first category model. Therefore, considering the strongly different geometric and radiometric properties of optical and SAR images, we propose a novel multimodal deep learning network called MBFNet, in which two identical AlexNet [21] without weight-sharing, pretrained on ImageNet [22], are taken as the backbone network for extracting the deep semantic feature maps of optical and SAR images, respectively.

Recently, the attention mechanism of artificial intelligence has mimicked the attention mechanism of human biological systems, which has been widely applied in the natural language processing [37], object detection [38], and semantic image segmentation [39]. As a category of attention mechanism, the channel attention mechanism has been concerned to explore the interchannel relationship of the feature maps based on the channel attention maps. Therefore, it is important for channel attention module to get efficient channel attention maps. Squeeze-and-excitation block [40] suggested that the global spatial information extracted by global average pooling was entered into multilayer perceptron (MLP) [41] to generate channel attention maps. Following [40], convolutional block attention module (CBAM) [42] proved that the channel attention map computed by global average-pooling features was suboptimal, which combined global average-pooling with global max-pooling features to obtain better channel attention maps. In this article, we propose a novel channel attention module, namely the second-order attention-based channel selection module (SACSM), which is shown in Fig. 4. The SACSM adopts a more efficient bilinear approach to integrate global average-pooling and max-pooling features for obtaining fine channel attention maps with second-order statistics. Based on the channel attention maps, the SACSM performs feature recalibration to refine interchannel relationship for improving the expressive power of the network. Meanwhile, inspired by the view that the contributions of many layers are quite a little in network [43], [44], the SACSM introduces a channel selection mechanism to reconfigure compact feature maps with stronger discrimination. To provide a reasonable explanation for the channel selection mechanism in SACSM, we shows the histograms of the elements of channel attention maps P [shown in Eq. (6)] in Fig. 3, and the elements in the P determine the contributions of the corresponding channel features. From Fig. 3, we know that some elements in the P are distributed around zero, which means the corresponding channels are unimportant and redundant, and some elements in the P are high, which means the corresponding channels are important in SACSM. Therefore, based on the channel selection mechanism, the SACSM can automatically find the important channels of feature maps to reconfigure the compact features with great discrimination. Especially, it is worth mentioning that SACSM can deal with the huge dimensionality issues of bilinear fusion features [31], which will be introduced in detail later.

For multimodal feature fusion, existing multimodal deep learning methods mainly adapt the addition fusion rule [23], [24], multiplication fusion rule [25], [26], and connection fusion rule [27], [28] to obtain fusion features. However, the ability that these fusion features explore the complex associations between optical and SAR images may be weak [29]. Bilinear pooling models [30] have been successfully embedded into DCNN for fine-grained classification [31] and visual question answering [29], [32], which could capture second-order association between the two different modalities and obtained excellent bilinear fusion features [29], [33], [36]. However, it faced a huge problem that the fusion features were high dimensional, and the problem limited the widespread use of the bilinear pooling. Therefore, the bilinear model embedded into the kernelized framework [29], [34] has been proposed, which used the polynomial kernel to map the high-dimensional fusion features and suggested the compact fusion features. Other methods could reduce the computational complexity by decomposing the parameter matrixes of the classifier, for *e.g.*, low-rank decomposition [35]. In this article, we find that the dimension of bilinear fusion features is determined by the channel number of convolutional feature maps, which is introduced in Section III-B. Therefore, the refined feature maps of optical and SAR streams can be extracted by SACSM, which have a smaller channel number and compact structure and are used to structure low-dimensional fusion features by the bilinear pooling model. The bilinear fusion features can not only capture second-order association between optical and SAR features, but also will not face the huge dimensionality issues.

In summary, we propose a novel multimodal deep learning model (MBFNet) (shown in Fig. 2) to fuse the deep features of optical and SAR images for the land cover classification. MBFNet consists of three components: feature extractor, SACSM, and the bilinear fusion module. First, the proposed

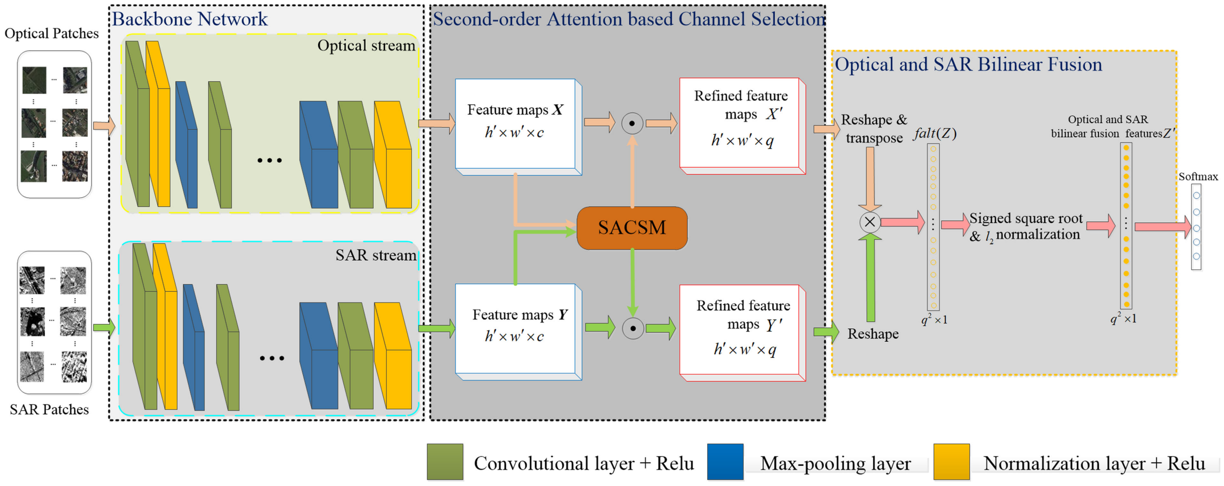


Fig. 2. Flowchart of MBFNet. ⊙ denotes the element-wise multiplication. ⊗ denotes the matrix multiplication.

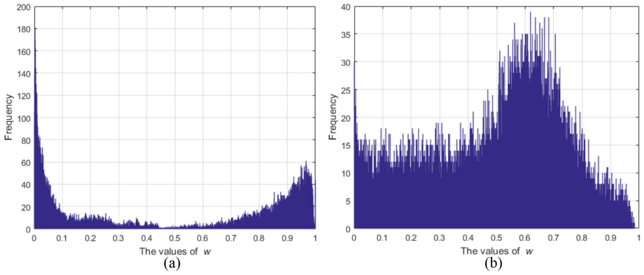


Fig. 3. Statistical histograms of the values of channel attention maps P on two image datasets. (a) Distribution of values in P on PoDelta. (b) Distribution of values in P on ChongMing. We find that some values are distributed around zero, which means the corresponding channel features are redundant and can be ignored, and some values are high, which means the corresponding channel features are important.

MBFNet model uses two identical AlexNet [21] without weight-sharing as feature extractor to extract deep semantic feature maps of optical and SAR images. Then, the SACSM is embedded into the optical and SAR streams, which integrates the global average-pooling and global max-pooling information by the bilinear operation to structure two fine channel attention maps. According to the top q values in channel attention maps, we select the corresponding channels of convolutional features maps to reconfigure compact feature maps of optical and SAR streams, respectively. Finally, we use the bilinear pooling operation to effectively fuse the abovementioned compact feature maps and obtain the discriminative fusion features for land cover classification, instead of the multimodal feature connection, element-wise summation, and element-wise production.

The main contributions of this article are the following.

First, a novel multimodal bilinear fusion network (MBFNet) with second-order attention-based channel selection is proposed to fuse the optical and SAR features for land cover classification. To the best of our knowledge, it is the first attempt to employ bilinear pooling to fuse the optical and SAR features in this paper.

Second, as a component of the MBFNet, the novel SACSM can obtain fine channel-attention maps with second-order statistics using bilinear operation. And based on the channel attention maps, the SACSM can not only explore the interchannel relationship to highlight important channels of feature maps, but can also select the important channels to obtain compact optical and SAR features for dealing with the high-dimensional problem of bilinear fusion features.

Third, we thoroughly evaluate our model on three broad coregistered optical and SAR datasets. Extensive experiments show that the proposed model obtains competitive improvements over state-of-the-art methods in land cover classification.

II. Related Work

Multimodal deep learning: Because multimodal data can be easily acquired, multimodal deep learning techniques have been widely used, which are able to build the high-level correlations between multimodal data. Ngiam *et al.* [13] were the first to propose the multimodal deep learning model by combining deep learning and multimodal learning techniques, which was used to fuse the audio and video features for phoneme classification. The experiment demonstrated that fusion features were better than the unimodal representations. Thereafter, multimodal deep learning has been widely used [45]–[47]. In the field of RS information processing, multimodal RS data have been more and more easily accessed, which provide the basis for the application of multimodal deep learning in RS images. Audebert *et al.* [24] proposed the deep two-stream fully convolution network to fuse multimodal RS images for land cover classification. Benedetti *et at.* [28] proposed the M^{3}Fusion model, in which the recurrent neural networks dealt with the high spatial resolution time-series images obtained from Sentinel-2 to extract time information, and the convolutional neural network (CNN) dealt with the very high spatial resolution time-series images obtained from Spot 6/7 to extract spatial information; two types of information were fused to improve the land cover classification. In this article, we also adopt the multimodal deep learning with bilinear model for obtaining better fusion features between optical and SAR data.

Attention mechanism: Recently, the concept of attention mechanism has become popular, which was incorporated into deep learning for diverse application domains [49], [50]. Wang *et al.* [49] introduced bottom-up, top-down feedforward attention mechanism into the residual network and proposed residual attention network. Chen *et al.* [50] learnt a soft weight map based on attention mechanism to recalibrate multiscale features for enhancing semantic segmentation. The attention mechanism in abovementioned methods is used on pixel-level. The difference is that SENet [40] produced the channel attention maps to explore the interchannel relationship of convolutional feature maps and highlighted the channels that were better for the power representation of the network. Other research works have integrated the pixel-level and channel attention mechanisms into the unitary network framework, which showed better performance in semantic segmentation tasks. PAN [51] and DANet [39] are representative examples. The proposed SACSM is classified as channel attention mechanism.

Bilinear model: As two-factor model with the mathematical property of separability, the bilinear model was first introduced by Tenenbaum and Freeman [30], which routinely separated perceptual systems into "content" and "style." Thereafter, the bilinear models have been successfully used in both classical methods and deep learning, such as Fisher vector [52], VLAD [53], and CNN [29], [31]–[36]. BCNN [31] replaced first-order pooling layer (global max or average pooling) with bilinear pooling layer to take the second-order statistics into consideration, which achieved great performance in fine-grained classification. However, there is a problem that the bilinear fusion features are high dimensional. To deal with this problem, Gao *et al.* [34] proposed a compact bilinear pooling model, and the discriminative compact bilinear pooling features were obtained based on Random Maclaurin (RM) and tensor sketch (TS) projection methods. Following [34], the multimodal compact bilinear pooling model generated the fusion features between image and text features by compact bilinear pooling, which was used for visual question answering. However, the abovementioned models have ignored matrix structure of the bilinear pooling features. Kong *et al.* [35] proposed the low-rank bilinear pooling model, which classified the bilinear feature matrix by using the bilinear classifier and maintained the structural information. Kin *et al.* [33] proposed the low-rank pooling using Hadamard product, which mapped the bilinear features to low-dimension feature space. Factorized bilinear layer was proposed [36], and it modeled the feature relationship with linear complexity. Meanwhile, to prevent the overfitting, the DropFactor was used. For the high-dimension problem of bilinear fusion features, we propose different solution, SACSM.

III. Methodology

The proposed MBFNet deals with the problem of learning fusion features between the optical and SAR modals for improving land cover classification. Given the optical and corresponding SAR patches, we design the MBFNet such that the predicted label $\hat{l}_i$ matches the correct label l_i^*. The predicted label can be formulated as

$$\hat{l}_i = \arg\max_{l_i \in \Phi} p\left(l_i \left| x_{\text{opt}}, y_{\text{sar}}; \theta_{\text{opt}}, \theta_{\text{sar}}\right.\right) \qquad (1)$$

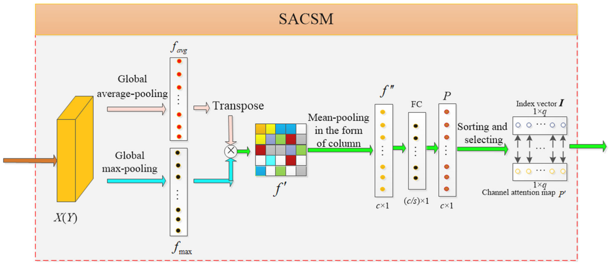


Fig. 4. Flowchart of the proposed SACSM.

where x_{opt} and y_{sar} denote the optical and SAR patch pairs, respectively, θ_{opt} and θ_{sar} are the parameters of optical and SAR streams, respectively, and Φ denotes the set of label. The architecture of MBFNet is shown in Fig. 2. Different from the previous bilinear fusion work, such as BCNN [30], we use the SACSM to explore the interchannel relationship of the feature maps $X \in \mathbb{R}^{h'\times w'\times c}$ and $Y \in \mathbb{R}^{h'\times w'\times c}$, respectively. X and Y denote the normalized feature maps of convolution layer *conv5*, and h', w', and $\boldsymbol{c}$ indicate the height, weight, and the channel number of feature maps, respectively; we select the significant channels of X and Y to structure compact and discriminative feature maps $X' \in \mathbb{R}^{h'\times w'\times q}$ and $Y' \in \mathbb{R}^{h'\times w'\times q}$, where q denotes the number of selected important channels, $q < c$. Then, we use the bilinear pooling to obtain fusion features with second-order statistics providing (through softmax function) final answer in (1).

In this section, we describe in detail the implementation process of the proposed MBFNet. First, we detail the SACSM, which captures two compact and discriminative unimodal feature maps from different modalities, respectively, in Section III-A. Then, in Section III-B, we discuss the fusion of optical and SAR features based on the bilinear pooling model. Finally, we describe the training process of the MBFNet in Section III-C.

A. *Second-Order Attention-Based Channel Selection Module*

As a novel channel attention mechanism module, the proposed SACSM (shown in Fig. 4) can effectively exploit the interchannel relationship of feature maps. In the channel attention module, the channel attention maps are keys, which can exploit better interchannel relationship to improve the power of network. SENet [40] fed the first-order statistics into MLP [41] to obtain channel attention maps, which have been proven to be suboptimal [42]. However, the proposed SACSM by bilinear integrating global average-pooling and max-pooling features introduces the second-order statistics for inferring finer channel attention maps. We confirm that exploiting global average-pooling and max-pooling features greatly improves the representation power of networks rather than using each independently. And more close to our work, CBAM [42] fed the global average-pooling and global max-pooling features into MLP, respectively, and then two output feature vectors were merged based on element-wise sum to obtain channel attention maps. However, the channel attention maps of CBAM lacked more discriminative second-order statistics.

Meanwhile, we find that some channels of feature maps *X* and *Y* can be ignored, because the corresponding channel attention values are close to zero (shown in Fig. 2). The SACSM contains the channel selection mechanism, which can select high-contribution channels of feature maps to obtain compact and discriminative feature maps, according to the top *q* value of channel attention maps *P*. Taking the SACSM of optical stream as an example, we describe the detailed operation of the proposed SACSM as follows.

1) We obtain the global max-pooling and the global average-pooling features of the normalized feature map *X* as follows:

$$f_{\max} = \mathrm{MaxPool}\left(X\right) \tag{2}$$

$$f_{\mathrm{avg}} = \mathrm{AvgPool}\left(X\right) \tag{3}$$

where $\mathrm{MaxPool}(\cdot)$ denotes the global max-pooling, and $\mathrm{AvgPool}(\cdot)$ denotes the global average-pooling.

2) We compute the outer product between $f_{\max}$ and f_{avg} and obtain the second-order statistics $f' \in \mathbb{R}^{c\times c}$ as follows:

$$f' = f_{\max} \otimes f_{\mathrm{avg}}^{T} \tag{4}$$

where $\otimes$ denotes the matrix multiplication.

3) We implement the average pooling on f' in the form of column, in which the *i*th element of $f'' \in \mathbb{R}^{c}$ is calculated as follows:

$$f''_i = \frac{1}{c}\sum_{j=1}^{c} f'\left(i,j\right). \tag{5}$$

4) The f'' is forwarded to MLP [41] with one hidden layer, and the channel attention map $P \in \mathbb{R}^{c}$ is obtained. The hidden layer size is set to c/s, where *s* is the reduction ratio. The *P* is defined as follows:

$$P = \delta\left(\mathrm{Mlp}\left(f''\right)\right) \tag{6}$$

where δ is the sigmoid function, and the function $\mathrm{Mlp}(\cdot)$ denotes the operation of MLP.

5) The reweighted feature map $\tilde{X}$ is obtained as follows:

$$\tilde{X} = P \odot X \tag{7}$$

where $\odot$ denotes the element-wise multiplication.

6) We sort the values in the *P* in descending order and obtain the sorted channel attention map $P' \in \mathbb{R}^{c}$ and the corresponding index vector $I \in \mathbb{R}^{c}$. According to the top *q* values of *I*, we select the corresponding channels in $\tilde{X}$ to structure compact feature map $X' \in \mathbb{R}^{h'\times w'\times q}$ as follows:

$$X' = \tilde{X}\left(:, :, I\left(1:q\right)\right). \tag{8}$$

Similarly, we can also obtain the compact feature map $Y' \in \mathbb{R}^{h'\times w'\times q}$ of SAR stream. The overall processing of the proposed SACSM is shown in Algorithm 1.

Algorithm 1: The Overall Processing of SACSM.

```
Input: The co-registered optical patches and corresponding SAR patches;
Output: Two compact feature maps $X'$ and $Y'$;
  1: Extract the normalized feature maps X and Y of convolution layer conv5 of optical and SAR streams, respectively;
  2: Compute the global max-pooling and average-pooling features of X and Y based on Eqs. (2) and (3), respectively;
  3: Obtain the second-order statistic vectors of optical and SAR features based on the Eqs. (4) and (5), respectively;
  4: The second-order statistic vectors are fed into MLP, and obtain unselected channel attention maps of optical and SAR stream based on the Eq. (6), respectively;
  5: Compute the compact feature maps $X'$ and $Y'$ based on the Eqs. (7) and (8), respectively.
```

B. Bilinear Fusion Features

Numerous studies [29], [32] have shown that the bilinear pooling model could build the second-order associations between different modality data, which were good for computer vision task. We apply bilinear pooling to aggregate $X' \in \mathbb{R}^{h'\times w'\times q}$ and $Y' \in \mathbb{R}^{h'\times w'\times q}$, and the bilinear fusion feature $Z \in \mathbb{R}^{q\times q}$ with second-order statistics is obtained as follows:

$$Z = \sum_{i\in Q} x_i y_i^T \tag{9}$$

where $\{x_i, y_i | x_i \in X', y_i \in Y', i \in Q\}$ are the local feature vectors, which are formed by splicing the values on the position *i*th of all channels of feature maps X' and Y', respectively. *Q* is the set of spatial locations.

Meanwhile, we can also rewrite (9) in a matrix notation as follows:

$$Z = \sum_{i\in Q} x_i y_i^T = \chi^T \gamma \tag{10}$$

where $\chi = [X'_1, \ldots, X'_q]^T \in \mathbb{R}^{h'w'\times q}$ and $\gamma = [Y'_1, \ldots, Y'_q]^T \in \mathbb{R}^{h'w'\times q}$ are the matrixes by reshaping X' and Y' in terms of the third mode, and X'_i and Y'_i are obtained by flatting *i*th channel of the X' and Y'. Therefore, from (10), we know that *q* determines the size of fusion features *Z*. In order to obtain low dimension but strongly discriminative fusion features, the proposed SACSM infers fine attention channel maps to construct two compact feature maps X' and Y'. The channel number of X' and Y' is small, which is used to construct the low-dimension fusion features Z.

Then, in order to deal with visual burstiness [58], we apply the sign square root and ℓ_2-normalization on the bilinear fusion feature *Z*, to obtain normalized fusion feature vector $Z' \in \mathbb{R}^{q^2}$ as follows:

$$Z_1 = \mathrm{sign}\left(\mathrm{falt}(Z)\right) \times \sqrt{\mathrm{falt}(Z)} \tag{11}$$

$$Z' = \frac{Z_1}{\|Z_1\|_2} \tag{12}$$

```
Algorithm 2: The Processing of the Optical and SAR Fusion.
Input: The co-registered optical and SAR patches;
Output: Bilinear fusion features $Z'$;
  1: Computer the compact feature maps $X'$ and $Y'$ of
     optical and SAR modals based on SACSM;
  2: Compute the bilinear fusion features $Z$ based on the
     Eq. (10);
  3: Deal with visual burstiness of the fusion features $Z$
     based on the Eqs. (11) and (12), and obtain fusion
     features $Z'$.
```

where $\mathrm{falt}(\cdot)$ denotes the operation of flatten, and $\|\cdot\|_2$ denotes ℓ_2-norm. The processing of the optical and SAR fusion is shown in Algorithm 2.

Finally, we will describe the back-propagation of the bilinear pooling layer. Since the architecture of MBFNet is a directed acyclic graph, the hyperparameters can be trained by back-propagating the gradients of the classification loss (*e.g.*, cross-entropy). The bilinear form simplifies the gradients at the pooling layer. Let dE/dZ be the gradient of the loss function E on the fusion features Z, then by the chain rule of gradients, we have

$$\frac{dE}{d\chi} = \gamma\left(\frac{dE}{dZ}\right)^T \tag{13}$$

$$\frac{dE}{d\gamma} = \chi\left(\frac{dE}{dZ}\right). \tag{14}$$

Meanwhile, the gradient of the softmax layer, the sign square root layer, and ℓ_2 -normalization layer is straightforward, which can be computed using the chain rule. Therefore, (13) and (14) are rewritten as follows:

$$\frac{dE}{d\chi} = \gamma\left(\frac{dE}{dZ'}\frac{dZ'}{dZ_1}\frac{dZ_1}{dZ}\right)^T \tag{15}$$

$$\frac{dE}{d\gamma} = \chi\left(\frac{dE}{dZ'}\frac{dZ'}{dZ_1}\frac{dZ_1}{dZ}\right). \tag{16}$$

C. Land Cover Classification

In this article, the loss function is the mean cross-entropy, which is given as follows:

$$E = -\frac{1}{n}\sum_{i=1}^{n} l_i^* \log\left(f\left(Z_f^i\right)\right) \tag{17}$$

where $Z_f^i \in \mathbb{R}^K$ is obtained by feeding the Z' into the fully connected layer, and the size of the fully connected layer is set to K, where K is the number of categories, n is the number of samples in the minibatch, l_i^* is the correct label, E is the mean cross-entropy loss, and $f(\cdot)$ denotes the softmax function.

As described in Section III-B, the MBFNet can be trained by back-propagating the gradients of the classification loss. In this article, we conduct the experiments of MBFNet by Tensorflow library on the platform of Window 10 with Intel(R) Corn(TM) i7-8700k 3.70-GHz CPU, NVIDIA GeForce GTX 1080Ti GPU and 32-GB memory. We fine-tune whole MBFNet by using Adam algorithm [26], in which the minibatch size is 64, the learning rate is 10^{-4}, and the number of epochs is 100.

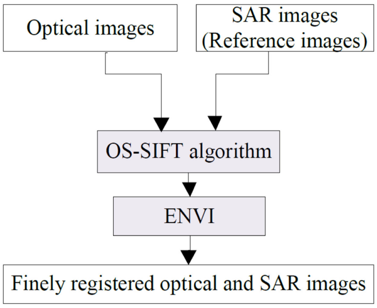

Optical images
SAR images (Reference images)
OS-SIFT algorithm
ENVI
Finely registered optical and SAR images

Fig. 5. Coregistration flowchart of the optical and SAR images.

IV. Experiment Evaluation

In this section, we conduct several experiments to evaluate the land cover classification performance of MBFNet on three broad coregistered optical and SAR datasets. First, the datasets are introduced in Section IV-A. We provide the description of baseline methods in Section IV-B. The configuration of significant hyperparameters, for *e.g.*, the patch size and the reduction ratio, is described in Section IV-C. In Section IV-D, the comparison of fusion feature dimension, classifier computation complexity, and testing time among the proposed model and comparison methods is described. In Section IV-E, we compare the land cover classification results of MBFNet with the state-of-the-art approaches on three coregistered optical-SAR image datasets. In Section IV-F, the qualitative visualization of fusion features is presented to intuitively explain MBFNet. Finally, Section IV-G shows validation curves (accuracy and loss) of PoDelta, ChongMing, and WuHan datasets.

A. Datasets

The first coregistered optical and SAR image dataset is the region of PoDelta, Italy, and its size is 3666 × 1952, with the pixel spacing of 2.5 m. SAR image are COSMO-SkyMed geocoded ground range (GEC) level 1C products in X-band and HH-polar imaging mode, as shown in Fig. 8(b), the optical image is obtained from GoogleEatrh, as shown in Fig. 8(a). And in order to evaluate the MBFNet, we relabel the ground truth of PoDelta provided by the literature [1] to create finer ground truth of PoDelta, according to the visual interpretation of optical images. Ground truth is shown in Fig. 8(c). All the pixels are assigned to five categories: building (Bu), water (W), tree (T), farmland (Fa), and unknown (Un). We refer this dataset as "PoDelta."

The second coregistered optical and SAR image dataset is the area of Chongming County, Yunnan Province, China, and its size is 1772 × 1427, with the pixel spacing of 5 m. The SAR image with HH-polar imaging mode is obtained from Gaofen-3, as shown in Fig. 9(b). The optical image is obtained from GoogleEatrh, as shown in Fig. 9(a). In order to acquire accurate knowledge of Chongming County area, the ground truth

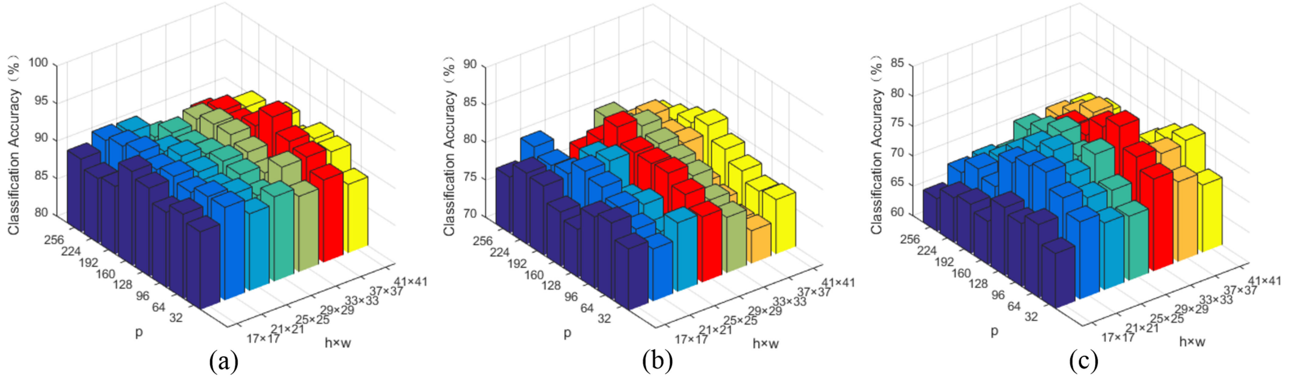


Fig. 6. Histogram to depict the trend between the accuracy and the hyperparameters, the size of image patches ($h \times w$), and the number of selected important channels $\boldsymbol{q}$, on two image datasets. (a) Classification accuracies are given based on different sizes of image patches and the number of selected important channels on PoDelta. (b) Classification accuracies are given based on different sizes of image patches and the number of selected important channels on ChongMing. (b) Classification accuracies are given based on different sizes of image patches and the number of selected important channels on WuHan.

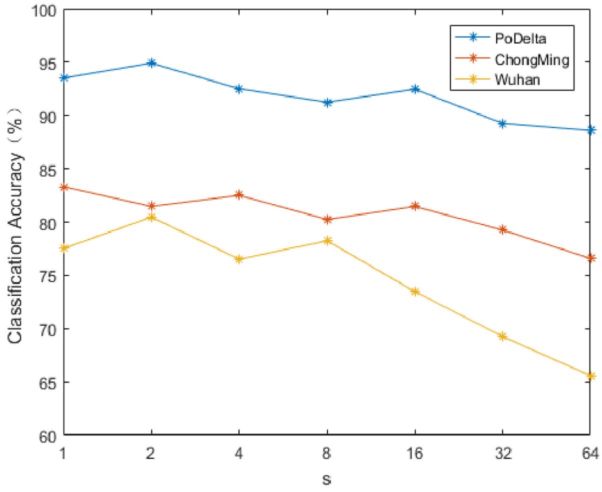


Fig. 7. Curves to depict the trend between the accuracy and the hyperparameters the reduction ratio s on two image datasets. (a) Accuracy curve is given based on different reduction ratios on PoDelta. (b) Accuracy curve is given based on different reduction ratios on ChongMing. (c) Accuracy curve is given based on different reduction ratios on WuHan.

is obtained based on the visual interpretation of optical and SAR images, meanwhile the field studies were carried out to collect ground truth. The ground truth is shown in Fig. 9(c). All the pixels are assigned to five categories: building (Bu), water (W), tree (T), farmland (Fa), and bareland (Ba). We refer this dataset as “ChongMing.”

The third coregistered optical and SAR image dataset is the area of Wuhan city, Hubei Province, China, and its size is 1039 × 886, with the pixel spacing of 3 m. The SAR image with HH-polar imaging mode is obtained from TerraSAR, as shown in Fig. 10(b). The optical image is obtained from GoogleEatrh, as shown in Fig. 10(a). In order to acquire accurate knowledge of Wuhan area, the ground truth is obtained based on the visual interpretation of optical and SAR images and OpenStreetMap [55]. The ground truth images are shown in Fig. 10(c). All the pixels are assigned to eight categories: resident (Re), industrial land (Ind), water (W), tree (T), farmland (Fa), background (Bg), unhardened road (Unr), and asphalt (As). We refer this dataset as “WuHan.”

Meanwhile, we adopt the coarse-to-fine registration strategy to complete the fine registration of optical–SAR images of the abovementioned three datasets. Fig. 5 shows the coregistration flowchart of the optical and SAR images. First, we take the SAR image as a reference image and then use the OS-SIFT algorithm [54] to complete the coarse registration between the optical and SAR images. Then, lots of corresponding points between the optical and SAR images are selected manually to achieve fine registration based on the ENVI software.

B. Baseline Methods

In this article, we use AlexNet as the backbone network in all comparison methods for extracting deep semantic features.

CNN+Optical: We replace the last fully-connected layer of AlexNet with a randomly initialized K-way classification layer and fine-tune. We refer this method as “CNN-Opt.,” and the optical image patches are the input samples.

CNN+SAR: It is similar to the construction of “CNN-Opt.,” but in this method, the SAR image patches are the input samples. We refer this method as “CNN-SAR.”

CNN+Concatenation: We use two AlexNet without weight-sharing, pretrained on ImageNet, as feature extractor.

Two feature maps of the convolution layer *conv5* from optical and SAR streams are flattened and connected to obtain fusion features, and a randomly initialized K-way fully connected layer is used for classification. We refer this method as “CNN-Con.”

CNN+element-wise sum: It is similar to the construction of “CNN-Con.,” but in this method, two feature maps are flattened and element-wise added to obtain the fusion features. We refer this method as “CNN-Sum.”

CNN+element-wise product: It is similar to the construction of “CNN-Con.,” but in this method, two feature maps are flattened and element-wise multiplied to obtain the fusion features. We refer the method as “CNN-Pro.”

Full Bilinear Pooling: We use two AlexNet [21] without weight-sharing, pretrained on ImageNet, to replace Vgg-16 in

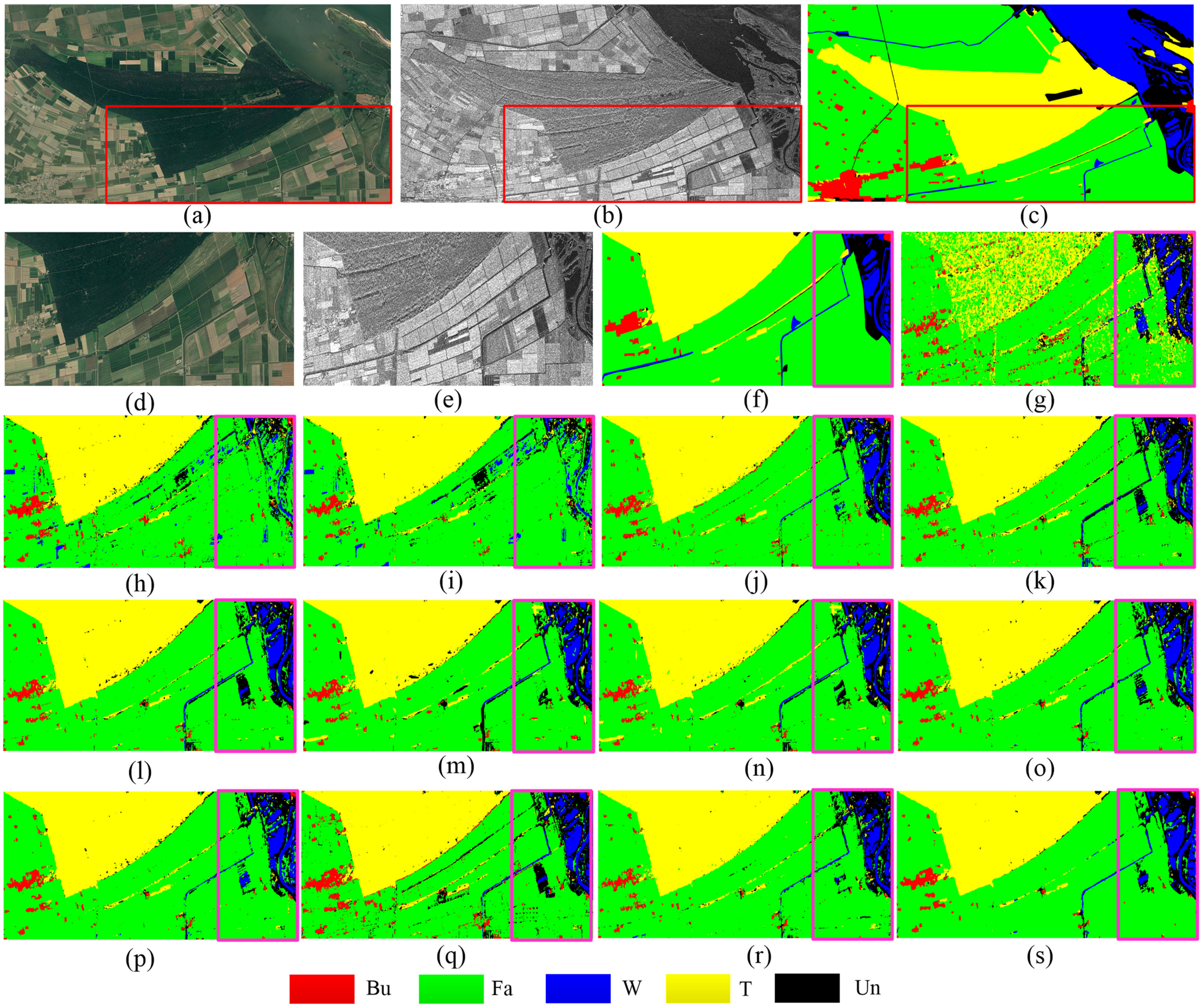


Fig. 8. Classification maps of methods on PoDelta. (a) Original optical image. (b) Original SAR image. (c) Ground truth. (d) Test optical image. (e) Test SAR image. (f) Test ground truth. (g) CNN-SAR. (h) CNN-Opt. (i) CNN-Con. (j) CNN-Sum. (k) CNN-Pro. (l) FBP. (m) CBP-RM. (n) CBP-TS. (o) FBP-SE. (p) FBP-CBAM. (q) 2D-CNN. (r) MRSDC. (s) MBFNet.

BCNN [31]. Then, we obtain bilinear fusion features between the optical and SAR features and apply element-wise sign square root normalization and ℓ_2-normalization prior to deal with visual burstiness [58] of the bilinear fusion features. Finally, the softmax classifier completes classification. We refer this method as "FBP."

Compact Bilinear Pooling: Referencing MCB [29], we use the same network construction of FBP to extract feature maps of the convolution layer *conv5* from optical and SAR streams. Then, two feature maps are fused by using two approximate methods via RM and TS, and we obtain compact fusion features with the dimension $d = 16\,000$. We refer this method as "CBP-RM" and "CBP-TS."

Full Bilinear Pooling+SE: The SE block [40] is embedded into each stream of FBP, simultaneously. The optical and SAR feature maps processed by SE block are fused by bilinear pooling. Finally, bilinear fusion features are used for land cover classification. We refer this method as "FBP-SE."

Full Bilinear Pooling+CBAM: It is similar to the "FBP-SE," but the SE block is replaced by the CBAM block [42], which can integrate the global average-pooling and global max-pooling information to obtain channel attention maps. We refer this as "FBP-CBAM."

D-CNN [56]: A multimodal deep network was proposed to fuse the optical and SAR images for crop type classification, which demonstrated the high-quality performance of fusion features in the RS field. To adapt such model to our multisource scenario, we connect optical and SAR patches to obtain four-channel inputs. Then, we add one convolution layer before the AlexNet pretrained on ImageNet. Finally, we replace the last fully-connected layer of AlexNet with a randomly initialized K-way classification layer and fine-tune for land cover classification. Also in this case, this competitor is learned end-to-end. We refer the method as "2D-CNN."

MRSDC: Xu *et al.* [57] proposed a two-brand convolution network for multisource RS data classification, in which the two-tunnel convolution network is developed to extract the spectral–spatial features from HIS, and the CNN with cascade block is designed for the feature extraction from LiDAR. To adapt such model to our multisource scenario, we use the

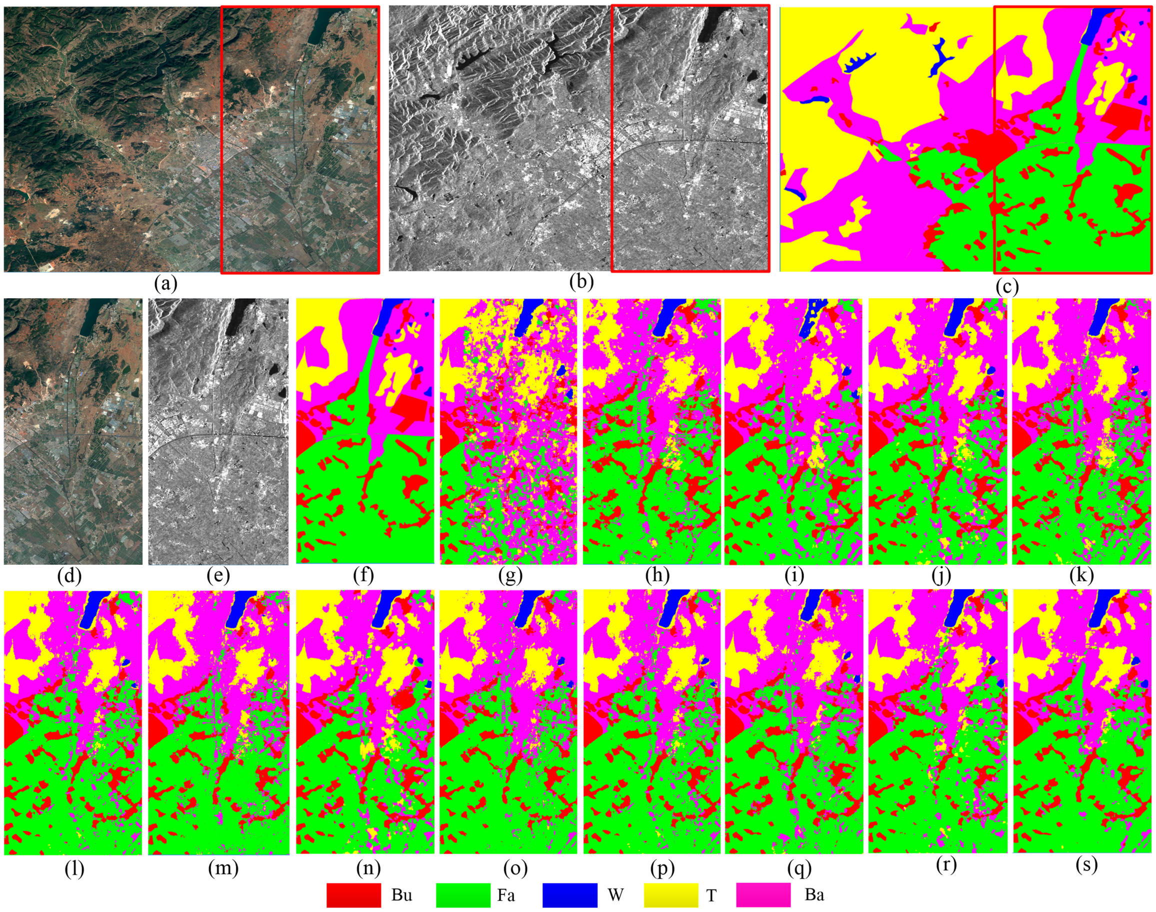


Fig. 9. Classification maps of methods on PoDelta. (a) Original optical image. (b) Original SAR image. (c) Ground truth. (d) Test optical image. (e) Test SAR image. (f) Test ground truth. (g) CNN-SAR. (h) CNN-Opt. (i) CNN-Con. (j) CNN-Sum. (k) CNN-Pro. (l) FBP. (m) CBP-RM. (n) CBP-TS. (o) FBP-SE. (p) FBP-CBAM. (q) 2D-CNN. (r) MRSDC. (s) MBFNet.

two-tunnel convolution network to deal with optical images, and the CNN with cascade block is designed for the feature extraction from SAR images. We refer this method as “MRSDC.”

In a word, we present 11 comparison methods, which can be roughly divided into five categories. The CNN-Opt. and CNN-SAR belong to the first category methods, namely the unimodal deep learning model, which only explore the impact of single-source features on land cover classification. The second category methods include the CNN-Con., CNN-Sum, and CNN-Pro., whose fusion rules may not have expressive enough to capture the complex associations between the optical and SAR features. The FBP, CBP-RM, and CBP-TS belong to the third categories, which obtain the fusion features by using bilinear pooling. The fourth final categories are FBP-SE and FBP-CBAM, which are obtained by embedding different channel attention modules into the FBP. The final categories are the 2D-CNN and MRSDC, which are state-of-the-art land cover classification methods.

C. *Configuration of Hyperparameters*

For the MBFNet, we find that the following three crucial hyperparameters have the obvious impact on the land cover classification experiments: the size of patches $h \times w$, the number of selected important channels q, and the reduction ratio s in the SACSM. The specific analysis of hyperparameters based on the validation sets (shown in Tables II, IV and VI) of PoDelta, ChongMing, and WuHan datasets is elaborated as follow.

Frist, we simultaneously analyze the size of image patches $h \times w$ and the number of selected important channels q. In our experiments, every optical–SAR patch pair is regarded as a basic processing unit, whose size determines whether the spatial information of image patch pair is good for extracting discriminative fusion features in our proposed model. As a significant hyperparameter in the SACSM, the q guides the selection of high-contribution channel features, which are used to structure low-dimension bilinear fusion features with better discrimination. The parameters $h \times w$ and q are determined by the classification accuracy on PoDelta dataset in Fig. 6(a). The horizontal axis denotes the parameters $h \times w$, and their value ranges from 17×17 to 41×41, with a span of 4; the vertical axis denotes the parameter q, and its value ranges from 32 to 256, with a span of 32. It can be viewed that we can achieve the highest accuracy, when the parameter $h \times w$ is 37×37 and the parameter $q = 128$, where the 37×37 patches can contain richer spatial information than the small patches, which can be used to generate more robust features. Meanwhile compared with the 41×41 patches, the 37×37 patches contain the homogeneous spatial information and have less computational complexity

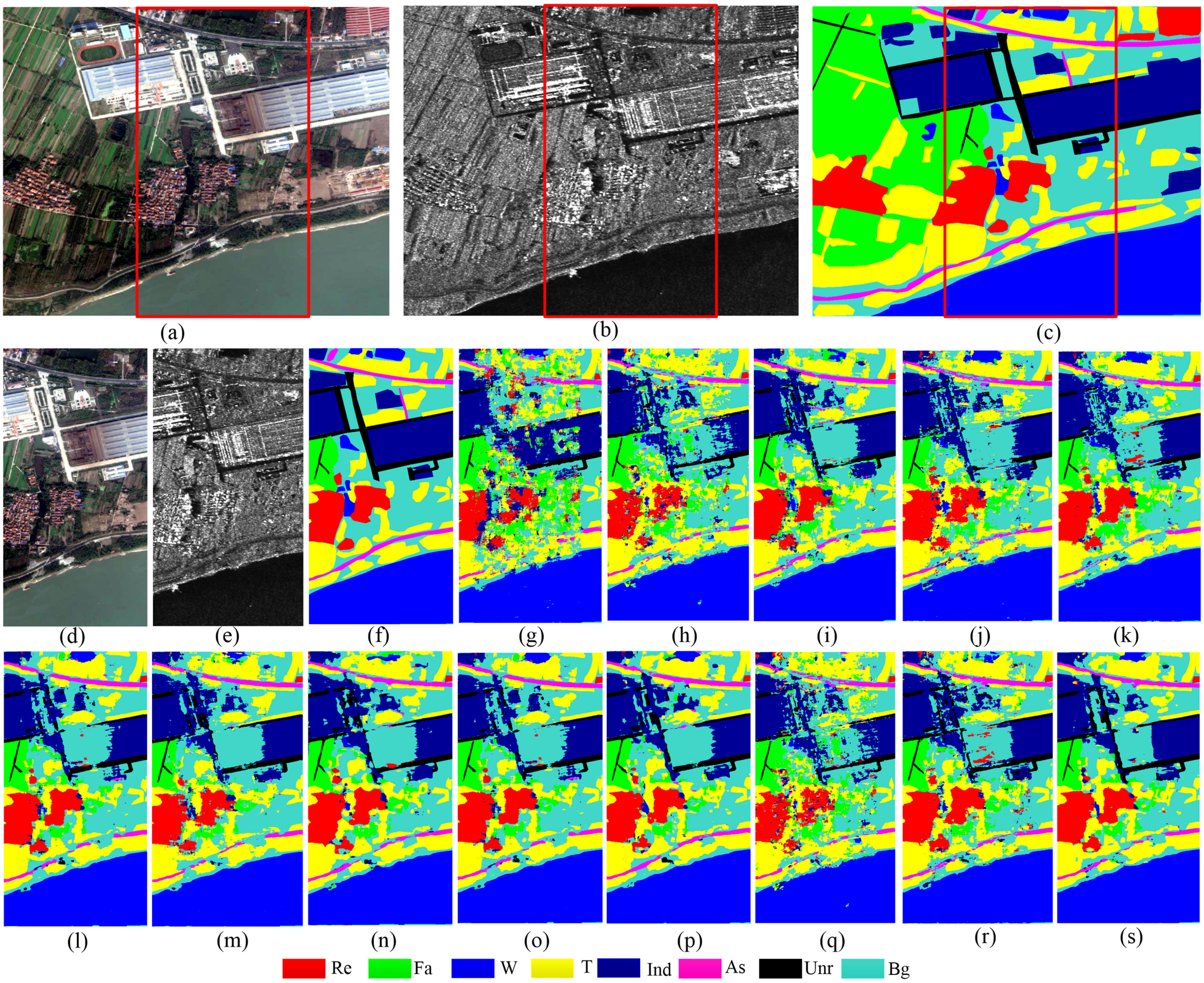


Fig. 10. Classification maps of methods on PoDelta. (a) Original optical image. (b) Original SAR image. (c) Ground truth. (d) Test optical image. (e) Test SAR image. (f) Test ground truth. (g) CNN-SAR. (h) CNN-Opt. (i) CNN-Con. (j) CNN-Sum. (k) CNN-Pro. (l) FBP. (m) CBP-RM. (n) CBP-TS. (o) FBP-SE. (p) FBP-CBAM. (q) 2D-CNN. (r) MRSDC. (s) MBFNet.

during the network training. For $q = 128$, we can know that the top 128 channels of $\tilde{X}$ have the high contribution, which can better improve network performance and reduce network parameters. Meanwhile, in this experiment, the reduction ratio $s = 2$.

Then, we analyze the reduction ratio $\boldsymbol{s}$, which provides an effective tradeoff between the module complexity and the performance [41]. In our experiments, we obtain the results with varied parameter s from $\{1, 2, 4, 8, 16, 32, 64\}$. The parameter selection results are described in Fig. 7, and we can know that the land cover classification effect is the optimal, when the parameter $s = 2$, where the reduction ratio $s = 2$ can balance well the model complexity and the land cover classification performance. Meanwhile, in this experiment, the parameters $h \times w$ are 37×37 and the parameter $q = 128$.

Similarly, the process of determining the hyperparameters on the ChongMing and WuHan datasets is same as the hyperparameter configuration on the PoDelta dataset. And the theoretical analysis of configurations is also same as the parameter description on the PoDelta. For ChongMing, the optimal hyperparameters $h \times w$, q, and s are described in Figs. 6(b) and 7, respectively. It can be viewed that we can achieve the highest accuracy, when the parameters $h \times w$ are 29×29, the parameter $q = 192$, and the parameter $s = 1$. For WuHan, the optimal hyperparameters $h \times w$, q, and s are described in Figs. 6(c) and 7, respectively. It can be viewed that we can achieve the highest accuracy, when the parameters $h \times w$ are 33×33, the parameter $q = 96$, and the parameters $s = 2$.

D. Analysis of Computational Efficiency

We study the computational complexity of MBFNet and the seven comparison methods described in Section IV-B. The fusion features are computed over two feature maps of convolution layer *conv*5, whose dimension is $h' \times w' \times c$, for a K-way classification problem. In this section, taking PoDelta as an example, the size of input image patches $h \times w$ is 37×37, so the sizes of features for the convolution layer *conv5* are $h' = w' = 9$ and

TABLE I
COMPARISON OF DIFFERENT METHODS IN TERMS OF FUSION FEATURE DIMENSION, CLASSIFICATION COMPUTATION COMPLEXITY AND TESTING TIME ON PODELTA[1]

Methods	Fusion feature dimension	Classification computation	Testing time
CNN-Pro.	$h'w'c$ (20K)	$O(K\,h'w'c\,)$	2.58 sec.
CNN-Sum	$h'w'c$ (20K)	O($K\,h'w'c$)	2.56 sec.
CNN-Con	$2\,h'w'c$ (41K)	O($2K\,h'w'c$)	3.53 sec.
FBP	c^2 (65K)	O(Kc^2)	4.56 sec.
MCB-RM	d (16K)	O(Kd)	3.16 sec.
MCB-TS	d (16K)	O(Kd)	3.39 sec.
FBP-SE	c^2 (65K)	O(Kc^2)	3.88 sec.
FBP- CBAM	c^2 (65K)	O(Kc^2)	4.04 sec.
MBFNet	q^2 (4K)	O(Kq^2)	2.22 sec.

[1] *Note:* The hyperparameters $h' = 9,\ w' = 9,\ c = 256$, $d = 16\,000$, $q = 64$, and $K = 5$ on PoDelta dataset.

$c = 256$. Table I provides a detailed comparison for different methods in terms of fusion feature dimension, computational complexity of producing classifier scores, and testing time. In particular, the conventional setup in MCB-TS and MCB-MR is that the dimension of bilinear fusion features $d = 16\,000$, as reported in [29]. MBFNet achieves similar classification performance when $q = 64$ in $X' \in \mathbb{R}^{h' \times w' \times q}$ and $Y' \in \mathbb{R}^{h' \times w' \times q}$, and compared with MCB-TS and MCB-MR, the dimension of bilinear fusion features in MBFNet is a quarter. From Table I, it can be seen that MBFNet is most appealing in terms of the dimension of fusion features and classification computational complexity. It is worth noting that the feature dimension of MBFNet is about 16 times smaller than the FBP, and about 4 times smaller than MCB-TS and MCB-MR. Meanwhile, considering the speed of the proposed MBFNet is also a very important index, we provide the computational time for 100 patches of size 37 × 37 in Table I. From the table, we can see that the speed of the proposed MBFNet is faster than comparison methods.

E. Classification on Three Datasets

We use four quantitative evaluation indexes to evaluate the classification performance: user's accuracy (UA), producer's accuracy (PA), overall accuracy (OA), and kappa coefficient (Kappa). The UA and PA provide the detailed classification performance of each category, whereas the OA and Kappa describe the global classification performance. The definition of quantitative evaluation indexes UA, PA, OA, and Kappa are as follows.

Let $M = \{m_{i,j}\}, i, j = [1, \ldots, K]$, be the size of $K \times K$ confusion matrix, where K is the number of the category. When $i \neq j, m_{i,j}$ indicates the number that class i wrongly classified to class j. The $m_{i,i}$ indicates the number that the class i is correctly classified. The jth class PA_j of PA, the ith class UA_i of UA, OA, and Kappa are calculated as follows:

$$\mathrm{PA}_j = \frac{m_{j,j}}{\sum_{i=1}^{K} m_{i,j}} \times 100\% \tag{18}$$

$$\mathrm{UA}_i = \frac{m_{i,i}}{\sum_{j=1}^{K} m_{i,j}} \times 100\% \tag{19}$$

TABLE II
SUMMARY STATISTIC OF PODELTA DATASET

Class	Label	All pixel	Training set	Validation set	Test set
1	Bu	213797	136032	34008	43757
2	Fa	3833295	1800672	450167	1582456
3	W	825539	578817	144704	102018
4	T	1786298	883255	220813	682230
5	Un	496170	287691	71922	136557
Total	-	7155099	3686467	921614	2547018

$$\mathrm{OA} = \frac{\sum_{i=1}^{K} m_{i,i}}{\sum_{i=1}^{K} \sum_{j=1}^{K} m_{i,j}} \times 100\% \tag{20}$$

$$\begin{cases} p_e = \sum_{i}^{K} \left(N_{\bullet,i} \times N_{i,\bullet}\right) / \left(\sum_{i=1}^{K} \sum_{j=1}^{K} m_{i,j}\right)^2 \\ \mathrm{Kappa} = (\mathrm{OA} - p_e) \,/\, (1 - p_e) \times 100\% \end{cases} \tag{21}$$

where $N_{\bullet,i} = \sum_{j=1}^{K} m_{j,i}$ and $N_{i,\cdot} = \sum_{j=1}^{K} m_{i,j}$.

1) Classification Results on PoDelta: In the experiment of PoDelta, the following are the three crucial hyperparameters: the size of patches $h \times w$ is 37 × 37, the number of selected channel $q = 128$, and reduction ratio $s = 2$. The detailed description is shown in experiments presented in Section IV-C. The areas enclosed by the red box in Fig. 8(a)–(c) are taken as the test samples, which are shown in Fig. 8(d)–(f), and the remaining area are taken as the training and validation samples; the number of each land category for the training, validation, and test sets is shown in Table II. To objectively evaluate the performance of our method, four quantitative metrics of our model and several comparison methods are shown in Table III. The classification maps of the proposed model and several comparison methods are shown in Fig. 8.

From Table III, we first discover that the proposed model have much better accuracies than the first category comparison methods, such as the CNN-SAR and CNN-Opt., with more than 15% improvement, which verifies MBFNet can fully capture the complementary information between optical and SAR features for improving land cover classification. Then, by comparing with the second category comparison methods, such as CNN-Con., CNN-Sum, and CNN-Pro., it is advisable to use bilinear pooling [30] for fusing both optical and SAR features in MBFNet, which can generate the stronger discrimination fusion features to promote better land cover classification accuracies. Then, different from the method of obtaining compact fusion features in CBP-RM and CBP-TS, MBFNet introduces a novel channel attention mechanism, namely the SACSM, which can automatically select the high-contribution channels to solve the high-dimension issue of bilinear fusion feature. Meanwhile, we discover that the proposed SACSM bilinear integrates global average-pooling and global max-pooling features to infer finer channel-wise attention than FBP-CBAM and FBP-SE. Finally, we compare the proposed MBFNet with the state-of-the-art methods, 2D-CNN [56] and MDSDC [57], which show the

TABLE III
UA, PA, OA, Kappa Coefficient of Different Methods on PoDelta

Methods	UA					PA					Kappa	OA
	Bu	Fa	W	T	Un	Bu	Fa	W	T	Un		
CNN-SAR	54.15	87.16	62.12	56.25	62.33	26.23	82.17	63.58	76.11	48.12	54.66	76.68
CNN-Opt.	**74.12**	93.26	21.92	95.41	41.04	50.95	90.32	35.01	98.13	44.92	76.68	87.72
CNN-Con.	71.92	92.69	66.12	96.46	45.57	58.92	89.92	34.19	97.01	58.41	77.14	87.51
CNN-Sum	67.17	95.39	63.02	92.65	56.12	41.01	92.98	72.15	98.28	56.02	83.01	91.90
CNN-Pro.	71.34	92.97	60.94	96.07	53.54.	39.73	**95.70**	**74.57**	97.16	47.18	83.57	90.81
FBP	59.89	95.05	68.81	96.26	64.09	55.76	95.61	75.02	98.28	54.81	85.31	91.56
CBP-RM	63.28	93.83	68.67	**96.82**	63.42	46.89	95.64	63.23	97.81	54.21	85.34	91.23
CBP-TS	66.78	94.12	68.76	95.95	63.31	50.61	96.17	66.20	97.51	56.21	85.23	91.61
FBP-SE	64.54	96.01	70.53	95.87	62.01	62.78	95.39	71.31	97.95	59.12	86.11	92.27
FBP-CBAM	66.21	96.42	67.01	95.34	58.06	62.33	95.01	73.89	97.98	61.17	86.29	92.41
2D-CNN[57]	74.02	93.98	30.22	91.42	44.47	56.15	91.31	67.12	95.18	44.21	78.31	88.17
MRSDC[58]	67.19	95.71	52.31	95.23	50.43	56.22	94.89	67.35	96.32	55.97	83.82	91.32
MBFNet	70.66	**97.11**	**70.91**	95.83	**64.41**	**70.69**	95.52	69.61	**99.09**	**67.26**	**87.86**	**93.61**

The bold values denotes the maximum of the columns.

TABLE IV
Summary Statistic of ChongMing Dataset

Class	Label	All pixel	Training set	Validation set	Test set
1	Bu	196408	61106	15276	120026
2	Fa	651011	125045	31261	494705
3	W	35544	13996	3498	18050
4	T	826627	557710	139427	129490
5	Ba	822598	420784	105196	296618
Total	-	2528644	1178641	294658	1055345

MBFNet achieves more effective land cover classification. In a word, MBFNet not only introduces a novel SACSM, which can infer finer channel-wise attention to structure compact fusion feature, but also is the first to use bilinear pooling [30] to explore the second-order associations between optical and SAR features.

Fig. 8 shows the classification maps of our method and several contrast methods. We reveal that the classification map derived from our method is better than other contrast method. As shown in the upper right corner area (purple box) of the different classification maps, the unknown areas are misclassified into farmland and tree areas. In contrast, our method produces fewer misclassification. This confirms that the fusion features extracted by the MBFNet have stronger discrimination performance compared with other methods.

2) Classification Results on ChongMing: In the experiment of the ChongMing dataset, the following are the three crucial hyperparameters: the size of image patches $h \times w$ is 29×29, the number of selected channel q is 192, and the reduction ratio $\boldsymbol{s}$ is 1. The areas enclosed by the red box in Fig. 9(a)–(c) are taken as test samples, which are shown in Fig. 9(d)–(f), and the remaining area are taken as training and validation samples. The number of each category for the training, validation, and test sets is shown in Table IV. Four quantitative metrics of the proposed model and several comparison methods are shown in Table V.

As shown in Table V, we discover that UA and PA of most categories are optimal in MBFNet, and the Kappa and OA are better than all comparison methods. Although the UA of bareland (Ba) and tree (T) are not optimal, they are not far from the best result. The PA of farmland (Fa), tree (T), and water (W) also is not far from the best result. In a word, our proposed model has much better accuracies than comparison methods. Meanwhile, the classification maps of our method and several contrast methods are shown in Fig. 9.

From Fig. 9, it is noticed that the classification map of the proposed MBFNet is better than several contrast methods, which demonstrates the superiority of our method. However, it is worth noting that the classification maps of the farmland and bareland are easily confused.

3) Classification Results on WuHan: In the experiment of the WuHan dataset, the following are the three crucial hyperparameters: the size of image patches $h \times w$ is 33×33, the number of selected channel q is 96, and the reduction ratio $\boldsymbol{s}$ is 1. The areas enclosed by the red box in Fig. 10(a)–(c) are taken as test samples, which are shown in Fig. 10(d)–(f), and the remaining areas are taken as training and validation samples.

The number of each land category for the training, validation, and test sets is shown in Table VI. Four quantitative metrics of the proposed model and several comparison methods are shown in Table VII.

As shown in Table VI, we discover that MBFNet's UA of the resident (Re), industrial land (Ind), water (W), tree (T), unhardened road (Unr), and background(Bg) is optimal. And the PA of tree (T), unhardened road (Unr), industrial land (Ind), and background (Bg) is better than other methods. Meanwhile we also know that Kappa and OA of the MBFNet are superior to other comparison methods. In a word, the proposed model can extract the complementary information between the optical and SAR features to obtain discriminative fusion features, which are used to improve the land cover classification. Meanwhile, the classification maps of different methods are shown in Fig. 10.

From Fig. 10(g)-(h), it can be seen that the classification maps based on single-modal data are worst. Fig. 10(i)–(s) shows the classification maps based on multimodal data, which are better than the classification maps shown in Figs. 10(g) and (h). Therefore, the optical and SAR have complementary information that is good for land cover classification. Especially, the classification map of our method is better than several contrast methods, which demonstrates the superiority of our method.

TABLE V
UA, PA, OA, KAPPA COEFFICIENT OF DIFFERENT METHODS ON CHONGMING

Methods	UA					PA					Kappa	OA
	Bu	Fa	W	T	Ba	Bu	Fa	W	T	Ba		
CNN-SAR	56.78	41.82	83.79	77.19	55.93	63.81	90.96	80.82	41.65	35.71	35.58	52.59
CNN-Opt.	67.92	77.54	73.13	72.02	76.07	56.01	88.31	90.24	68.91	67.81	64.28	75.98
CNN-Con.	69.91	73.21	69.37	80.66	83.49	71.17	**92.85**	82.81	72.48	63.13	66.38	76.62
CNN-Sum	69.98	76.89	82.91	82.70	86.01	70.11	92.67	89.95	73.15	68.98	69.92	79.32
CNN-Pro.	74.26	75.34	85.71	**86.08**	80.17.	69.51	92.73	89.06	68.18	69.14	68.98	78.29
FBP	71.58	82.03	83.99	82.43	83.78	72.57	92.31	85.02	77.61	70.01	71.81	80.73
CBP-RM	67.27	81.35	74.12	81.04	81.25	73.02	90.98	90.25	74.45	69.31	70.61	79.78
CBP-TS	71.33	80.48	76.46	82.67	82.89	71.02	92.13	89.71	76.91	68.99	71.67	80.21
FBP-SE	69.97	84.63	61.57	83.45	80.54	72.68	89.61	87.12	77.23	72.23	73.43	81.21
FBP-CBAM	70.82	84.81	71.23	83.45	81.39	72.81	89.24	**91.02**	78.78	73.23	73.51	81.75
2D-CNN[57]	75.01	63.53	84.99	83.84	78.54	72.56	74.11	88.49	77.45	74.23	71.78	81.63
MRSDC[58]	74.21	72.61	81.13	83.18	**84.28**	71.63	91.58	**88.64**	78.01	70.17	72.57	81.15
MBFNet	**75.49**	**85.67**	**86.42**	82.89	82.75	**74.42**	92.48	85.56	**78.91**	**75.05**	**75.13**	**82.61**

The bold values denotes the maximum of the columns.

TABLE VI
SUMMARY STATISTIC OF WUHAN DATASET

Class	Label	All pixel	Training set	Validation set	Test set
1	Re	63730	28367	7091	28272
2	Ind	118496	48251	12062	58183
3	W	154211	50279	12569	91363
4	T	178188	78804	19700	79684
5	Fa	158980	117049	29262	12669
6	Bg	191202	69939	17484	103779
7	Unr	34174	9168	2292	22714
8	As	21573	7125	1780	12668
Total	-	920554	408982	102240	409332

TABLE VII
UA, PA, OA, KAPPA COEFFICIENT OF DIFFERENT METHODS ON WUHAN

Methods	UA								PA								Kappa	OA
	Re	Fa	W	T	Ind	As	Unr	Bg	Bu	Fa	W	T	Ind	As	Unr	Bg		
CNN-SAR	65.1	88.1	87.9	58.5	**76.1**	33.8	38.3	48.2	68.1	27.5	98.6	58.4	68.5	52.4	75.1	52.0	56.97	64.47
CNN-Opt.	74.2	85.8	90.2	59.4	64.8	65.8	41.5	52.4	70.1	35.1	97.9	65.8	67.8	65.4	76.6	58.9	61.13	68.88
CNN-Con	84.1	80.5	92.2	78.3	66.1	60.3	49.1	65.5	91.7	44.7	98.4	78.1	66.3	70.3	82.5	60.5	69.06	74.73
CNN-Sum	87.6	85.5	94.0	78.7	61.9	55.5	36.5	64.1	85.2	49.4	97.8	77.2	63.8	73.9	76.1	59.3	67.94	73.58
CNN-Pro.	89.5	85.6	**94.3**	79.5	66.6	45.8	38.9	65.0	86.1	38.7	97.9	81.2	68.0	79.1	87.4	60.4	69.29	74.88
FBP	86.4	77.7	92.2	78.7	63.1	41.8	51.7	72.9	96.6	53.1	98.4	81.0	70.5	76.3	81.3	58.6	70.27	75.88
CBP-RM	82.2	78.1	91.1	80.7	64.4	55.6	48.8	69.5	94.0	56.6	98.5	77.5	66.6	**79.8**	84.1	59.4	69.67	75.37
CBP-TS	88.9	83.5	91.6	82.5	63.5	62.0	52.9	66.5	90.8	49.9	**99.1**	78.7	67.9	79.5	79.4	60.9	70.54	75.96
FBP-SE	89.9	86.4	91.7	82.7	64.7	57.8	65.1	68.8	91.8	54.1	98.4	76.9	75.2	77.4	85.9	62.7	71.48	76.79
FBP-CBAM	87.1	80.6	92.6	84.6	64.3	**67.5**	50.6	69.5	**94.4**	**65.1**	98.7	77.2	68.5	74.2	78.8	63.2	71.55	76.82
2D-CNN[57]	76.8	**90.3**	88.7	75.6	61.4	44.7	38.4	58.5	79.0	34.9	97.4	69.3	66.8	68.1	77.5	57.3	63.05	69.71
MRSDC[58]	87.6	86.5	94.0	78.7	61.9	55.5	36.5	65.8	85.2	49.5	97.8	**77.3**	63.8	73.9	76.0	59.3	67.91	73.85
MBFNet	**90.6**	86.4	92.8	**82.8**	64.9	57.8	**69.5**	**69.7**	92.1	54.1	98.4	**78.7**	**76.7**	77.4	**86.7**	**63.9**	**73.29**	**78.22**

The bold values denotes the maximum of the columns.

F. *Qualitative Visualization*

To validate the discrimination of fusion features of MBFNet, taking PoDelta as an example, we use t-SNE [54] to visualize the features of each category. As shown in Fig. 11, we find that fusion feature space of each category in MBFNet is more disjoint than other methods, which means the fusion features extracted by our method have strong individuality or discrimination for improving land cover classification.

G. *Validation Curves*

Figs. 12 and 13 show the validation curves (accuracy and loss) of PoDelta, ChongMing, and WuHan datasets. Fig. 12 shows the accuracy curves (training and validation) of three datasets. As the number of iterations increases, the accuracy curves of training and validation tend to be consistent. Fig. 13 shows that the loss information for three datasets continued to decrease with training iteration. The validation curves indicate that the MBFNet shows no sign of overfitting in the processing.

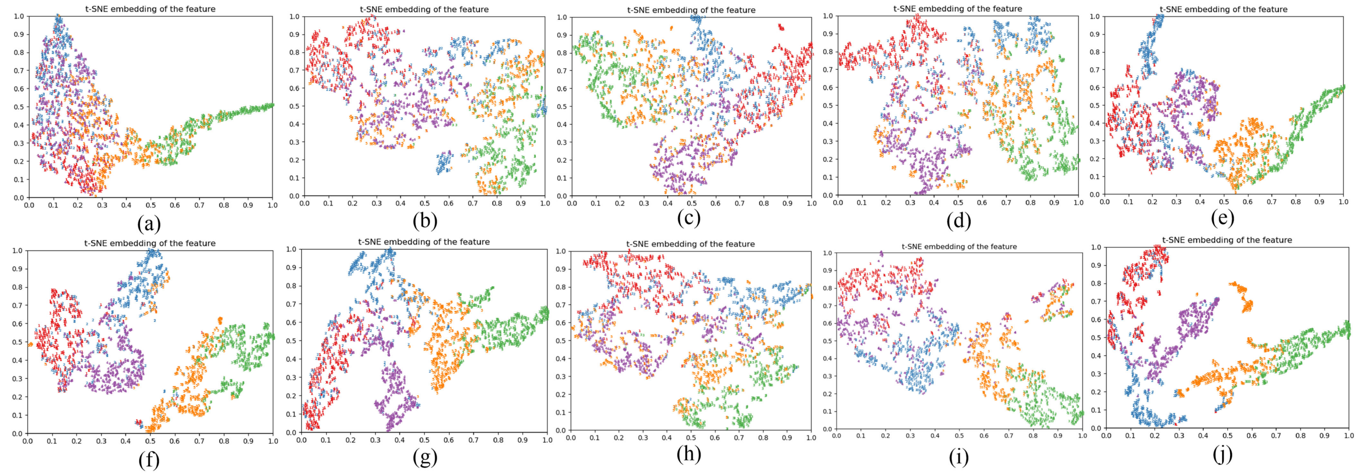


Fig. 11. Feature visualization of different methods on PoDelta dataset. (a) CNN-SAR. (b) CNN-Opt. (c) CNN-Con. (d) CNN-Pro. (e) FBP. (f) CBP-TS. (g) FBP-CBAM. (h) 2D-CNN. (i) MRSDC. (j) MBFNet. Different colors denote different categories: “red” for building, “green” for water, “blue” for farmland, “purple” for tree, and “yellow” for unknown.

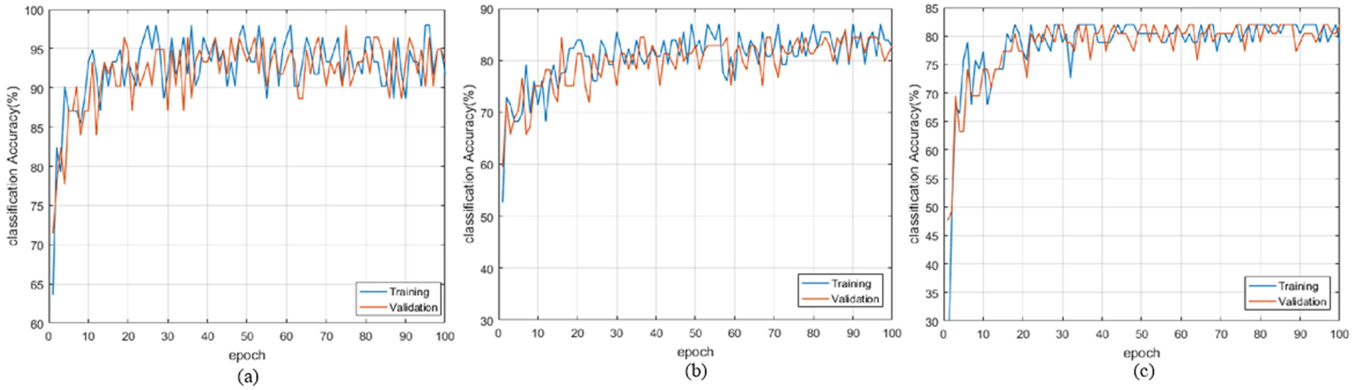


Fig. 12. MBFNet accuracy curves of different datasets. (a) MBFNet accuracy curves on PoDelta. (b) MBFNet accuracy on ChongMing. (c) MBFNet accuracy curves on WuHan.

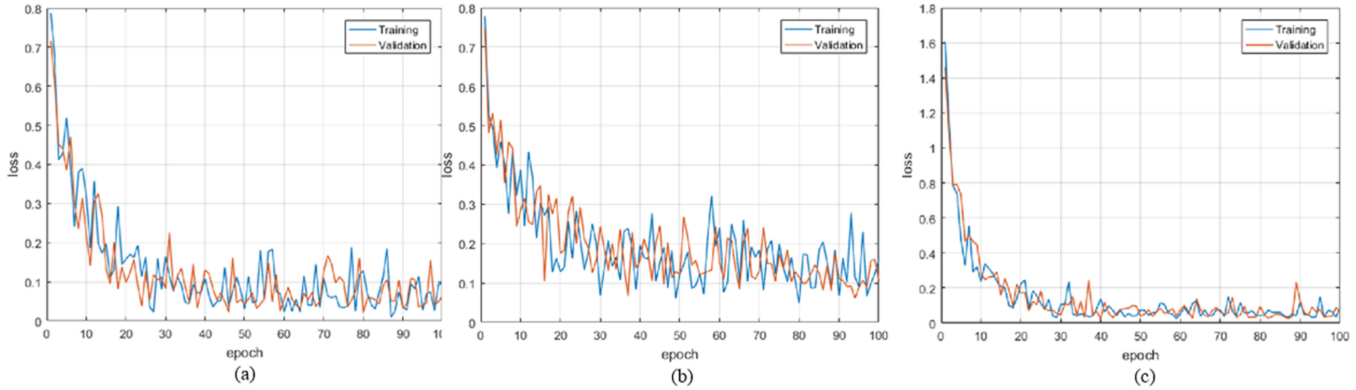


Fig. 13. Training and validation loss curves of different datasets. (a) Training and validation loss curves on PoDelta. (b) Training and validation loss curves on ChongMing. (c) Training and validation loss curves on WuHan.

V. Conclusion

In this article, we present a novel MBFNet to efficiently fuse deep features of optical and SAR modals for improving land cover classification. In MBFNet, a pseudo-siamese CNN is taken as the feature extractor, which captures deep semantic features of the optical and SAR images, respectively. Then, the SACSM is proposed and embedded into each stream, which can effectively exploit the interchannel relationship of feature maps based on fine channel attention maps with high-order statistics.

Meanwhile, according to the top q values in channel attention maps, we select the high-contribution channels to reconfigure discriminative and compact feature maps of optical and SAR streams, respectively. Finally, we use the bilinear pooling model to effective fuse the different modal compact feature maps and obtain discriminative bilinear fusion features for land cover classification. Experiment results on two broad coregistered optical and SAR datasets show that the land cover classification of the proposed MBFNet outperforms state-of- the-art methods.

Acknowledgment

The authors would like to thank the support of Google Earth with the optical remote sensing images used for this research.

References

[1] Z. Ren, B. Hou, Z. Wen, and L. Jiao, "Patch-sorted deep feature learning for high resolution SAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3113–3126, Sep. 2018.

[2] D. Lu and Q. Weng, "Use of impervious surface in urban land-use classification," *Remote Sens. Environ.*, vol. 102, no. 2, pp. 146–160, May 2006.

[3] D. Lu, P. Mausel, M. Batistella, and E. Moran, "Comparison of land-cover classification methods in the Brazilian Amazon Basin," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 6, pp. 723–731, Jun. 2004.

[4] L. Xu, H. Zhang, C. Wang, and M. Liu, "Crop classification based on temporal information using sentinel-1 SAR time-series data," *Remote Sens.*, vol. 11, no. 1, pp. 723–731, 2019.

[5] A. Langner *et al.*, "Spectral normalization of SPOT 4 data to adjust for changing leaf phenology within seasonal forests in Cambodia," *Remote Sens. Environ.*, vol. 143, no. 5, pp. 122–130, Mar. 2014.

[6] S. Quan, B. Xiong, D. Xiang, and G. Kuang, "Derivation of the orientation parameters in built-up areas: With application to model-based decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4714–4730, Aug. 2018.

[7] J. Fil *et al.*, "Multi^{3}Net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery," in *Proc. 33rd AAAI Conf. Art. Intell.*, 2019, vol. 33, pp. 702–709.

[8] H. Zhang, L. Wan, T. Wang, Y. Lin, H. Lin, and Z. Zheng, "Impervious surface estimation from optical and polarimetric SAR data using small-patched deep convolutional networks: A comparative study," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2374–2387, Jul. 2019.

[9] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[10] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[11] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[12] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[14] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information." in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2141–2149.

[15] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. Conf. Compu. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 4353–4361.

[16] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion*, Xi'an, China, 2017, pp. 1–7.

[17] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[18] M. Cen and C. Jung, "Fully convolutional siamese fusion networks for object tracking," in *Proc. IEEE Int. Conf. Image Process.*, Athens, Greece, 2018, pp. 3718–3722.

[19] H. Schilling, D. Bulatov, R. Niessner, W. Middelmann, and U. Soergel, "Detection of vehicles in multisensor data via multibranch convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4299–4316, Nov. 2018.

[20] L. H. Hughes, M. Schmitt, and X. X. Zhu, "Mining hard negative samples for SAR-optical image matching using generative adversarial networks," *Remote Sens.*, vol. 15, no. 10, pp. 1552–1568, Sep. 2018.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[22] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[23] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Nov. 2016, pp. 213–228.

[24] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Nov. 2018.

[25] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep CNNs for action recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Lake Placid, NY, USA, 2016, pp. 1–8.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2014.

[27] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[28] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M^3 fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018.

[29] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*.

[30] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.

[31] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1449–1457.

[32] J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha, and B. T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*.

[33] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Greece, 2017, pp. 1839–1848.

[34] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. Conf. Compu. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 317–326.

[35] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Honolulu, HI, USA, 2017, pp. 7025–7034.

[36] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Greece, 2017, pp. 2631–2639.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[38] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2018, pp. 714–722.

[39] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3141–3149.

[40] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[41] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmospheric Environ.*, vol. 32, no. 14/15, pp. 2627–2636, 1998.

[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module” in *Proc. Euro. Int. Conf. Comput. Vis.*, 2018, pp. 3–19.

[43] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[44] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” in *Proc. Euro. Int. Conf. Comput. Vis.*, Oct. 2016, pp. 646–661.

[45] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, “Semi-supervised manifold alignment of multimodal remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7708–7720, Dec. 2014.

[46] N. Srivastava and R. Salakhutdinov, “Learning representations for multimodal data with deep belief nets,” in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, vol. 79.

[47] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, “Multimodal deep learning for robust RGB-D object recognition,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 681–687.

[48] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[49] F. Wang *et al.*, “Residual attention network for image classification,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6450–6458.

[50] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[51] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in *Proc. BMVC*, 2018, p. 285.

[52] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” in *Proc. Euro. Int. Conf. Comput. Vis.*, Berlin, Germany, 2010, pp. 143–156.

[53] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in *Proc. Conf. Compu. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 3304–3311.

[54] Y. Xiang, F. Wang, and H. You, “OS-SIFT: A robust sift-like algorithm for high-resolution optical-to-SAR image registration in suburban areas,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.

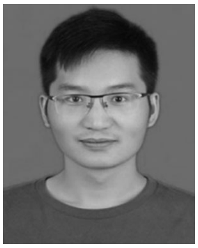
[55] M. Haklay and P. Weber, “OpenStreetMap: User-Generated Street Maps,” *IEEE Perv. Comput.*, vol. 7, no. 4, pp. 12–18, Oct.–Dec. 2008.

[56] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep learning classification of land cover and crop types using remote sensing data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[57] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, “Multisource remote sensing data classification based on convolutional neural network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[58] H. Jegou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, 2019, pp. 1169–1176, doi: 10.1109/CVPR.2009.5206609.

[59] L. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

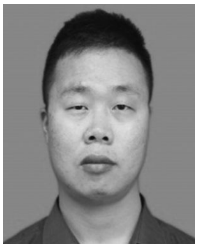


Xiao Li received the M.S. degrees in control science and engineering from Xiangtan University, Xiangtan, China, in 2018. He is currently working toward the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China.

His research interests include image processing and pattern recognition, sparse representation, deep learning, and remote sensing image applications.

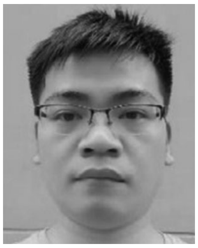
Lin Lei received the Ph.D. degree in information and communication engineering from National University of Defense Technology, Changsha, China, in 2008.

She is currently an Associate Professor with the school of Electronic Science, National University of Defense Technology. Her research interests include computer vision, remote sensing image interpretation and data fusion.



Yuli Sun received the M.S. degree from the University of Science and Technology of China, China, in 2014. He has been working toward the Ph.D. degree at the College of Electronic Science, National University of Defense Technology, since 2019.

His research interests include machine learning and remote sensing image processing.



Ming Li received the bachelor’s and the M.S. degrees from Central South University, Changsha, China, in 2017 and 2013, respectively. He has been working toward the Ph.D. degree in the College of Electronic Science, National University of Defense Technology, since 2017.

His research interests include remote sensing object detection, image retrieval, and change detection.

Gangyao Kuang (Senior Member, IEEE) received the B.S. and M.S. degrees in geophysics from the Central South University of Technology, Changsha, China, in 1988 and 1991, respectively, and the Ph.D. degree in communication and information from the National University of Defense Technology, Changsha, China, in 1995.

He is currently a Professor with the School of Electronic Science, National University of Defense Technology. His research interests include remote sensing, SAR image processing, change detection, SAR ground moving target indication, and classification with polarimetric SAR images.