Joint Network Combining Dual-Attention Fusion Modality and Two Specific Modalities for Land Cover Classification Using Optical and SAR Images

Xiao Liu, Huijun Zou, Shuxiang Wang, Yuzhun Lin, and Xibing Zuo

***Abstract*—Optical and synthetic aperture radar (SAR) images provide various complementary information on land properties, which can substantially improve the accuracy of land cover classification because of the fusion of their multisource information. However, excellent extraction of the discriminative information of optical and SAR images and effective integration of the interpretation information of multisource data remain challenging issues. In this study, we have proposed a novel joint network that combines a dual-attention fusion modality and two specific modalities to achieve superior land cover classification maps, which has two modes of encoding-level fusion (JoiTriNet-e) and decoding-level fusion (JoiTriNet-d). We first proposed a strategy for the fusion modality and specific modalities joint learning, the goal of which was to simultaneously find three mapping functions that project optical and SAR images separately and together into semantic maps. Two architectures were constructed using the proposed strategy, which improved the performance of multisource and single-source land cover classification. To aggregate the heterogeneous features of optical and SAR images more reasonably, we designed a multimodal dual-attention fusion module. Experiments were conducted on two multisource land cover datasets, among which comparison experiments highlighted the superiority and robustness of our model, and ablation experiments verified the effectiveness of the proposed architecture and module.**

***Index Terms*—Convolutional neural network (CNN), data fusion, deep learning (DL), land cover classification, multimodal semantic segmentation, optical images, synthetic aperture radar (SAR) images.**

I. Introduction

With the rapid development of Earth observation technologies, a substantial number of remote sensing satellites equipped with different types of sensors are operated to observe the Earth [1]. Therefore, multisource remote sensing images (RSIs) can be obtained of most regions of Earth [2]. These include optical images with abundant spectral and spatial information and synthetic aperture radar (SAR) images with all-day and all-weather capabilities. These multisource RSIs provide complementary information on land features [3], [4], [5]. Optical images with high spatial resolution can not only provide rich texture information but also present the essential features of ground objects through the multiple different bands contained in them. However, owing to the passive imaging of optical sensors and the influence of natural climate, optical images cannot always be obtained. By contrast, active SAR imaging is not limited by weather conditions and can provide unique scattering and geometric features for ground targets. However, given the inherent speckle noise, the image resolution, signal-to-noise ratio, and semantic interpretation of SAR are lower than those of optics [6]. This has encouraged researchers to combine complementary optical and SAR images for specific applications. Integrating radiation information from optical images and scattering information from SAR images is beneficial for land cover classification [7], [8], [9]. This plays an essential role in agricultural resources statistics [10], [11], ecological environmental monitoring [12], [13], and urban construction planning [14], [15].

Methods for multisource land cover classification using optical and SAR images can be divided into machine learning (ML) and deep learning (DL). Traditional ML methods first fuse optical and SAR images with pixel-, feature-, or decision-level fusion algorithms [16] and then exploit the support vector machine [17], Markov random field [18], subspace [19], decision tree [20], and other classification methods. These methods rely on prior knowledge and artificial descriptors with limited feature expression abilities, which often cannot adequately express complex semantic information. In contrast, DL methods overcome the limitations of artificial features and have become the most advanced tools for a range of tasks in the field of remote sensing image interpretation owing to their powerful feature extraction ability [21], [22], [23]. Various encoder–decoder architectures developed in recent years have considerably improved the accuracy of semantic segmentation, making deep convolutional neural networks (DCNNs) the current mainstream method for land cover classification [24]. In this context, the open multisource land cover datasets SEN12MS [25] and DFC2020 [26], which drive the training of multimodal DL models, have been released successively, providing a new opportunity for more accurate land cover classification using optical and SAR images.

There are two major issues in applying a DCNN to multisource land cover classification using optical and SAR images. One is how to excellently extract the discriminative information

Manuscript received 28 August 2023; revised 18 November 2023; accepted 23 December 2023. Date of publication 27 December 2023; date of current version 18 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 42201443. *(Corresponding author: Huijun Zou.)*

The authors are with the Information Engineering University, Zhengzhou 450001, China (e-mail: liuxiao99919@163.com; 17630025815@163.com; shuxiang1007@163.com; lyz120218@163.com; zuoxibing1015@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3347571

of specific modalities, the quality of which affects the performance of the fusion feature and determines the upper limit of the entire model. However, current research mostly employs a different [27], [28] or pseudo identical [29], [30] two-stream network to extract features from optical and SAR images, where these streams are directly coupled and cannot be used to sufficiently explore the features of specific modalities. The second is how to effectively capture complementary information in multisource RSIs data and integrate multimodal information, the amount of which directly affects the performance of the fusion features and determines the lower limit of the entire model. Most previous studies [30], [31], [32] have designed fusion modules based on attention mechanisms; however, they consider channel attention while excluding the role of spatial attention.

To address these issues, we have proposed a joint network that combines a dual-attention fusion modality and two specific modalities to obtain excellent land cover classification maps, which has two modes of encoding-level fusion (JoiTriNet-e) and decoding-level fusion (JoiTriNet-d). On the one hand, we proposed a strategy for the fusion modality and specific modalities joint learning (FSMJL) to improve the ability to extract modality-specific discriminative information. On the other hand, we designed a new multimodal dual attention fusion module (MDAFM) to improve the ability to integrate multisource complementary information. Specifically, the FSMJL aims to simultaneously find three mapping functions that project optical and SAR images separately and together into semantic maps, enabling the model to improve the performance of both multisource and single-source land cover classifications. The MDAFM aims to perform channel and spatial attention before and after cascading convolution operations to aggregate complementary information of heterogeneous features more effectively. The contributions of this study can be summarized as follows.

1) We first analyzed the multimodal semantic segmentation architecture based on the encoder–decoder structure and summarized two architectures: encoding level feature fusion (ELF) and decoding level feature fusion (DLF). We then proposed the FSMJL strategy to improve the ability to extract discriminative information from specific modalities and accordingly designed fusion modality and specific modalities joint learning architecture of encoding-level feature fusion (FSMJL-ELF) and fusion modality and specific modalities joint learning architecture of decoding-level feature fusion (FSMJL-DLF) architectures.
2) Second, we designed a new multimodal feature fusion module, MDAFM, which comprises two channel attention blocks, spatial attention, and Conv-BN-ReLU blocks.
3) Based on the designed architectures and module, we constructed JoiTriNet-e and JoiTriNet-d. They not only demonstrate exceptional performance compared to other multisource land cover classification methods but also exhibit robustness in generating accurate land cover classification maps when only one modality is available, even surpassing the accuracy of networks trained on a single modality.
4) We demonstrated the superiority and robustness of the proposed network through comparative experiments on two multisource land cover datasets and verified the effectiveness of the proposed strategy and module through ablation experiments.

The rest of this article is organized as follows. Section II presents relevant research on the land cover classification of multisource RSIs. Section III introduces the proposed joint network combining dual-attention fusion modality and two specific modalities. Section IV describes the experiments and discusses the results of the comparative and ablation experiments. Finally, Section V concludes this article.

II. Related Work

A. Land Cover Classification of RSIs

Traditional land cover classification methods for RSIs typically use a limited number of rules to categorize images based on different spatial units, such as pixels and objects [33]. However, artificial feature descriptors in traditional ML approaches often only involve low-level features of the spectral and spatial domains, making it difficult to effectively recognize complex land structures. Recently, DL has been widely applied in land cover classification owing to its advantages in multiscale and multilevel feature extraction and has achieved optimal results [34]. Land cover classification methods based on DL can be broadly divided into two categories according to the spatial representation level of the labels, namely patches, and pixels. Patch-level algorithms are suitable for land cover classification of medium-resolution RSIs because they lack the fine structural information of the image [35]. Sharma et al. [36] proposed a patch-based recurrent neural network for land cover classification in multitemporal and multispectral RSIs. Song et al. [37] designed a lightweight convolutional neural network (CNN) for land cover mapping using Landsat-8 data. Pixel-level algorithms aim to assign land cover labels to each pixel in RSIs using end-to-end DL models, similar to semantic segmentation in natural images. Currently, state-of-the-art semantic segmentation frameworks for RSIs are encoder–decoder structures [38], [39] that capture rich multilevel contextual information over relatively large receptive fields. Ghosh et al. [40] proposed a dilated stacked U-Net architecture for semantic segmentation of RGB RSIs. Mohammadimanesh et al. [41] introduced a fully convolutional network architecture for classifying complex land cover ecosystems using polarized SAR images. Liu et al. [42] proposed a dense dilated convolution merging network for land cover classification using fused local and global contextual information.

Owing to the limitations of remote sensing satellite sensors, it is usually impossible to meet all the requirements for high temporal, spatial, and spectral resolutions. The fusion of complementary information from multisource RSIs for land cover classification has proven to be a promising approach for improving accuracy [9].

B. Multisource Land Cover Classification of RSIs

Land cover classification methods for multisource RSIs can be divided into either ML- or DL-based. Traditional methods

primarily consist of two steps, namely multisource RSIs fusion and ML-based classification. For example, Iervolino et al. [43] fused multispectral, panchromatic, and SAR data based on generalized intensity-hue-saturation transform and the atrous wavelet transform, and used a standard maximum likelihood classifier for land classification. In terms of multisource RSIs fusion methods, salentinig et al. [15] explored the multispectral and SAR data fusion techniques at the raw data, feature, and decision levels. Kulkarni et al. [16] summarized various pixel-level methods for SAR and optical image fusion. In terms of ML-based classification methods, Sukhavatanavit et al. [17] proposed an algorithm based on genetic algorithms and support vector machines for the classification of SAR and multispectral images. Qin et al. [20] proposed a general model for multisource RSIs classification based on the Markov random field. In addition, other methods include subspace [19] and decision tree [20], [44] and other methods. However, these artificial feature descriptor methods often rely on prior knowledge and have limited expressive ability, thereby impeding their ability to fully express complex high-level semantic information.

In recent years, state-of-the-art multisource land cover classification methods have commonly been based on DL. Chen et al. [27] designed a deep neural network framework for land cover classification using multisensory remote sensing data, which consisted of two CNN for extracting features from multi/hyperspectral and light detection and ranging (LiDAR) data, and one fully connected deep neural network for fusing heterogeneous features. Hughes et al. [29] proposed a pseudosiamese architecture with two identical yet separate convolutional streams (PSCNN) to address the task of identifying corresponding patches in very high-resolution optical and SAR imagery of urban scenery, whose information is only fused in a final fully connected decision layer by concatenation and a 1×1 convolutional operation.

However, these patch-based multimodal depth models cannot perform pixel-by-pixel classification of high-resolution multisource RSIs. Xu et al. [28] studied a two-branch CNN for pixel-wise multisource remote sensing data classification (MRSDC) by fusing hyperspectral imagery (HSI) and data from multiple other sensors, such as LiDAR and visible images, which used the spectral and spatial features of hyperspectral data extracted using a two-tunnel CNN framework and other remote sensing data features extracted by a CNN with a cascade block. Audebert et al. [45] explored the advantages and disadvantages of early and late fusion strategies for urban segmentation using paired LiDAR and multispectral data. The results of experiments on deep full convolutional networks (FCNs) indicated that late fusion recovers some critical errors on hard pixels from ambiguous data, whereas early fusion (V-FesuNet) allows for stronger multimodal joint features of learning but higher sensitivity to missing or noisy data. Xu et al. [46] proposed a fusion-FCN framework for classification using three different types of data, that is, LiDAR data, HSI data, and very high-resolution images with RGB channels. Benedetti et al. [47] developed an M^3 fusion (multiscale/modal/temporary fusion) architecture that integrates CNN to manage relatively high spatial resolution information and a recurrent neural network to analyze time-series information.

In summary, land cover classification models based on multimodal DL typically extract features from single modal networks and learn their fusion representations for classification. However, to date, most research works have focused on directly extracting multimodal features through a coupled two-stream network of different or pseudoidentical streams. This has limited the investigation of sufficient features of specific modalities. Therefore, we propose a joint learning strategy that efficiently extracts modality-specific discriminant information by simultaneously learning both fusion and specific modality branches.

C. Multimodal Feature Fusion in DL

Learning the fusion representation of multimodal features is one of the most crucial aspects of multimodal DL methods. However, early models typically obtained fusion features through simple fusion rules, such as addition, multiplication, and concatenation. Therefore, Feng et al. [48] proposed an adaptive feature fusion module that integrates HSI and LiDAR features more naturally based on squeeze-and-excitation networks rather than simply stacking features. Liu et al. [49] proposed a multimodal network for land cover mapping, which included a pyramid attention fusion module to obtain fine-grained contextual representations of each modality and a gated fusion unit for the early merging of features. Hong et al. [50] investigated four plug-and-play fusion modules, namely, early fusion, middle fusion, late fusion, and encoder–decoder fusion, and designed a cross-fusion module.

Optical images record the spectral features of ground objects, whereas SAR images record the scattering features. Given that they are complementary in the interpretation process, researchers have proposed a series of methods to fuse these complex heterogeneous features to improve the accuracy of land cover classification. Li et al. [30] proposed a multimodal bilinear fusion network (MBFNet) for land cover classification that used a bilinear pooling operation to fuse fine channel-attention maps of optical and SAR features. The attention maps were obtained using a second-order attention-based channel selection module that operates on the features extracted from two AlexNets without sharing weights. Subsequently, they [51] explored the inherent complementarity between optical and SAR features and proposed a cooperative attention-based heterogeneous gated fusion network consisting of a dual-stream feature extractor, multimode cooperative attention module, and gated heterogeneous fusion module to improve land cover classification by hierarchical fusion of optical and SAR features. Kang et al. [52] investigated where and how to fuse the optical and SAR images in a modular fully convolutional network model and proposed a cross-gate module with bidirectional information flow to improve the accuracy of land cover classification by preserving important or complementary features. Ren et al. [31] constructed a multimodal land cover classification dataset based on optical images from the Gaofen-2 satellite and SAR images from the Gaofen-3 satellite and designed a dual-stream deep high-resolution network (DDHRNet) that enhances multimodal feature representation by improving the squeeze-and-excitation module. Li et al. [32] studied land use classification based on optical and SAR image fusion and developed a large-scale joint

optical and SAR land use classification dataset. The authors designed a multimodal-cross attention network (MCANet) for semantic segmentation, which included a pseudosiamese feature extraction module for independent feature extraction of optical and SAR images, a multimodal-cross attention module for second-order hidden feature mining, and a low–high-level feature fusion module for multiscale feature fusion.

In summary, these studies were dedicated to designing an attention-based fusion module to combine the heterogeneous complementary features of optics and SAR. The attention mechanism was inspired by the human cognitive system, which allows the network to amplify critical details and focus more on the essential aspects of an image by imitating the selective attention of human vision [53]. However, these attention-based feature fusion modules only consider channel attention, without considering the application of spatial attention. Therefore, we introduced spatial attention into the fusion module to integrate useful multisource information more effectively.

III. Methodology

In this section, we introduced the dual-attention fusion modality and specific modalities for land cover classification using optical and SAR images. First, we summarized two multimodal semantic segmentation architectures based on an encoder–decoder structure, proposed a strategy for the FSMJL to excellently extract the discriminant information of specific modalities, and accordingly proposed two novel architectures. We then provided a detailed description of the dual-attention fusion modality and specific modalities joint network built on the proposed architecture. Finally, we introduced an MDAFM designed to effectively integrate the complementary information from multisource optical and SAR images.

A. Task Description and Possible Architecture

Let $(X_i^{\text{opt}}, X_i^{\text{sar}}) \in \{\mathbf{X}^{\text{opt}}, \mathbf{X}^{\text{sar}}\}$ denote a co-registered optical and SAR image datasets, and $Y_i \in \mathbf{Y}$ denote the semantic labels used for land cover classification. The task of fusing optical and SAR images for land cover classification is to find a mapping function f_{fus} that simultaneously projects optical and SAR images from the same spatial location into semantic labels, denoted as $(X_i^{\text{opt}}, X_i^{\text{sar}}) \rightarrow Y_i$. In the semantic segmentation framework based on DL, the encoder–decoder structure has been shown to significantly improve the performance of segmentation. The encoder extracts multiscale hierarchical features, while the decoder fuses them, and the semantic segmentation head recovers the original resolution by upsampling to generate the semantic segmentation map. Let us represent the encoder, decoder, and semantic segmentation head as $\mathcal{E}$, $\mathcal{D}$, and $\mathcal{H}$, respectively.

According to the location of the multimodal feature fusion, there are two potential encoder–decoder based multimodal semantic segmentation architectures.

ELF architecture is illustrated in Fig. 1(a). The optical and SAR images were input into the corresponding encoder for separate independent encoding. The features extracted at the encoding-level were fed into a fusion module to obtain the fusion features, and the final segmentation was then obtained along a single stream of the decoder and semantic segmentation head. The network proposed by Liu et al. [49] and MCANet [32], as well as the "early fusion" and "late fusion" discussed by Kang et al. [52], fall into this architecture.

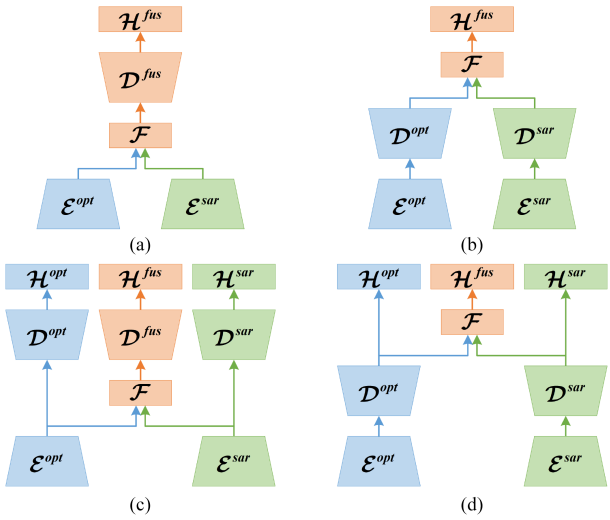

Fig. 1. Four possible architectures for land cover classification using optical and SAR images. (a) ELF architecture. (b) DLF architecture. (c) FSMJL-ELF. (d) FSMJL-DLF. "→" indicates the direction of the data flow.

DLF architecture is shown in Fig. 1(b). The optical and SAR images were processed through two complete encoder–decoder structures to obtain decoding-level features. After fusion by the fusion module, the fusion features were sent to the segmentation head for the final segmentation, as shown in the figure. The "output fusion" discussed by Kang et al. [52] belongs to this architecture.

Given that the quality of optical and SAR features used for fusion determines the capability of the fusion feature, to more effectively extract modality-specific features to generate better fusion results, we proposed a strategy for FSMJL. It enables the network to deeply explore discriminative information specific to each modality through dedicated branches. There are two architectures corresponding to ELF and DLF.

The FSMJL-ELF, shown in Fig. 1(c), consists of two encoders, one fusion module, three consecutive decoders, and segmentation heads. The optical and SAR images were separately fed as inputs into the optical encoder $\mathcal{E}^{\text{opt}}$ and the SAR encoder $\mathcal{E}^{\text{sar}}$. The optical and SAR encoding features are then fused using the feature fusion module $\mathcal{F}$. The fusion features, together with the optical and SAR encode-level features, are modeled through three independent streams: $(\mathcal{H}\circ\mathcal{D})^{\text{opt}}$, $(\mathcal{H}\circ\mathcal{D})^{\text{sar}}$, and $(\mathcal{H}\circ\mathcal{D})^{\text{fus}}$, which connect the decoder and segmentation head to obtain three segmentation results. This process can be summarized as follows:

$$\begin{cases} f_{\text{opt}}\left(X_i^{\text{opt}}\right) = \mathcal{H}^{\text{opt}} \circ \mathcal{D}^{\text{opt}}\left(\mathcal{E}^{\text{opt}}\left(X_i^{\text{opt}}\right)\right) \\ f_{\text{sar}}\left(X_i^{\text{sar}}\right) = \mathcal{H}^{\text{sar}} \circ \mathcal{D}^{\text{sar}}\left(\mathcal{E}^{\text{sar}}\left(X_i^{\text{sar}}\right)\right) \\ f_{\text{fus}}\left(X_i^{\text{opt}}, X_i^{\text{sar}}\right) = \mathcal{H}^{\text{fus}} \circ \mathcal{D}^{\text{fus}}\left(\mathcal{F}\left(\mathcal{E}^{\text{opt}}\left(X_i^{\text{opt}}\right), \mathcal{E}^{\text{sar}}\left(X_i^{\text{sar}}\right)\right)\right). \end{cases} \tag{1}$$

The FSMJL-DLF, as shown in Fig. 1(d), consists of two encoder–decoders, one fusion module, and three segmentation heads. Not only is the architecture composed of one fewer decoder than the first, but the flow of the architecture is also different. The optical and SAR images are input to the complete encoder–decoder streams of optical $(\mathcal{D}\circ\mathcal{E})^{\text{opt}}$ and SAR $(\mathcal{D}\circ\mathcal{E})^{\text{sar}}$, respectively. Therefore, the feature fusion module $\mathcal{F}$ fuses the features at the decoding-level. The optical and SAR decoding features with the fusion features only must be sent to three independent segmentation heads $\mathcal{H}^{\text{opt}}$, $\mathcal{H}^{\text{sar}}$, and $\mathcal{H}^{\text{fus}}$ to obtain three final segmentation results. This process can be summarized as follows:

$$\begin{cases} f_{\text{opt}}\left(X_i^{\text{opt}}\right) = \mathcal{H}^{\text{opt}}\left(\left(\mathcal{D}\circ\mathcal{E}\right)^{\text{opt}}\left(X_i^{\text{opt}}\right)\right) \\ f_{\text{sar}}\left(X_i^{\text{sar}}\right) = \mathcal{H}^{\text{sar}}\left(\left(\mathcal{D}\circ\mathcal{E}\right)^{\text{sar}}\left(X_i^{\text{sar}}\right)\right) \\ f_{\text{fus}}\left(X_i^{\text{opt}}, X_i^{\text{sar}}\right) = \mathcal{H}^{\text{fus}}\left(\mathcal{F}\left(\left(\mathcal{D}\circ\mathcal{E}\right)^{\text{opt}}\left(X_i^{\text{opt}}\right), \left(\mathcal{D}\circ\mathcal{E}\right)^{\text{sar}}\left(X_i^{\text{sar}}\right)\right)\right). \end{cases} \tag{2}$$

The goal of the FSMJL strategy is to simultaneously find three mapping functions: $f_{\text{opt}} : X_i^{\text{opt}} \to Y_i$, $f_{\text{sar}} : X_i^{\text{sar}} \to Y_i$, and $f_{\text{fus}} : (X_i^{\text{opt}}, X_i^{\text{sar}}) \to Y_i$, where f_{opt} and f_{sar} project optical and SAR images, respectively, into the semantic map, and f_{fus} projects optical and SAR images together into the same semantic map. The former encourages modality-specific branches to learn the discriminant information of a specific modality, whereas the latter encourages the modality-fusion branch to learn the detailed information of multiple modalities and integrate high quality multisource information.

The advantages of the proposed strategy lie in the following three aspects.

1) It enables the model not only to deal with multisource RSIs but also to handle single-source RSIs according to the input data, thereby demonstrating excellent flexibility in practical applications. When only one modality is available, other fusion networks experience a rapid decline in classification performance, while the proposed modality-specific branch remains unaffected. The proposed network can consistently produce robust land cover classification maps with accuracies significantly higher than those obtained from the network trained on a single modality.
2) Modality-specific branches provide greater feature extraction capability by learning from their branches, thus enriching the discriminate information needed for the integration of the fusion features.
3) Meanwhile, the learning of the modality-fusion branch not only improves the ability of its fusion feature module to integrate multisource information but also implicitly helps modality-specific branches to further explore distinguishing features. The segmentation performance of the modality-fusion and modality-specific branches is improved simultaneously.

To demonstrate the effectiveness of the proposed strategy, an ablation experiment was conducted in Section IV-D, with FSMJL-DLF identified as the optimal architecture among the two designs.

B. Joint Network Combining Dual-Attention Fusion Modality and Two Specific Modalities

Both the proposed architectures were composed of multiple identical encoders, decoders, segmentation heads, and a feature fusion module. We chose the off-the-shelf ResNet101 [54] initialized with ImageNet [55] as the encoder. The atrous spatial pyramid pooling module exploited in DeepLabV3+ [56] was used as the decoder, followed by a 1×1 convolution and a bilinear interpolation layer with a factor of four as the segmentation head. To integrate as much effective multisource information as possible, we proposed a novel MDAFM, which is described in Section III-C. Three cross-entropy loss functions were simultaneously used to supervise the optical and SAR modality-specific branches and modality-fusion branches, respectively. The overall loss function is calculated as follows:

$$\mathcal{L} = \sum_{m\in\{\text{opt,sar}\}} \mathcal{L}_{\text{ce}}\left(Y \mid X^m; W^m\right) + \mathcal{L}_{\text{ce}}\left(Y \mid X; W\right) \tag{3}$$

where $W = \{W^{\text{opt}}, W^{\text{sar}}, W^{\text{fus}}\}$ represents the weights of the optical, SAR, and fusion modality branches.

Fig. 2 illustrates the proposed encoding-level fusion framework of the joint network combining dual-attention fusion modality and two specific modalities (JoiTriNet-e). The optical and SAR images were input into ResNet101 to obtain five stacked layer features. The second- and fifth-layer features were input into the MDAFM to obtain low- and high-level fusion features. The optical, SAR, and fused encoding-level features were individually entered into the decoder exploited in DeepLabV3+, where the low- and high-level features were input together. We stacked the optical, SAR, and fusion segmentation heads on three parallel branches to obtain the final semantic segmentation results and exploited the triplet segmentation loss for supervision.

Fig. 3 shows the proposed decoding-level fusion framework of the joint network combining dual-attention fusion modality and two specific modalities (JoiTriNet-d). The advanced semantic features of the optical and SAR images were extracted by a stacked ResNet101 encoder and DeepLabV3+ decoder, respectively. These were fed into the MDAFM to obtain fusion features of the same size. We attached the optical, SAR, and fusion segmentation heads to the distinct features obtained in the previous step to achieve the final semantic segmentation results, which were supervised by triplet segmentation loss.

C. Multimodal Dual-Attention Fusion Module

Given that the ability of the feature fusion module to integrate information determines the interpretability of the fusion features, to more effectively capture complementary information in optical and SAR images, we proposed MDAFM, that performs channel attention for a single modality followed by spatial attention after cascading convolution fusion. As shown in Fig. 4, it comprises two channel attention blocks, spatial attention and Conv-BN-ReLU blocks.

The inspiration for this design comes from feature channels representing the discriminative ability of different modalities and feature spaces representing the positional information of

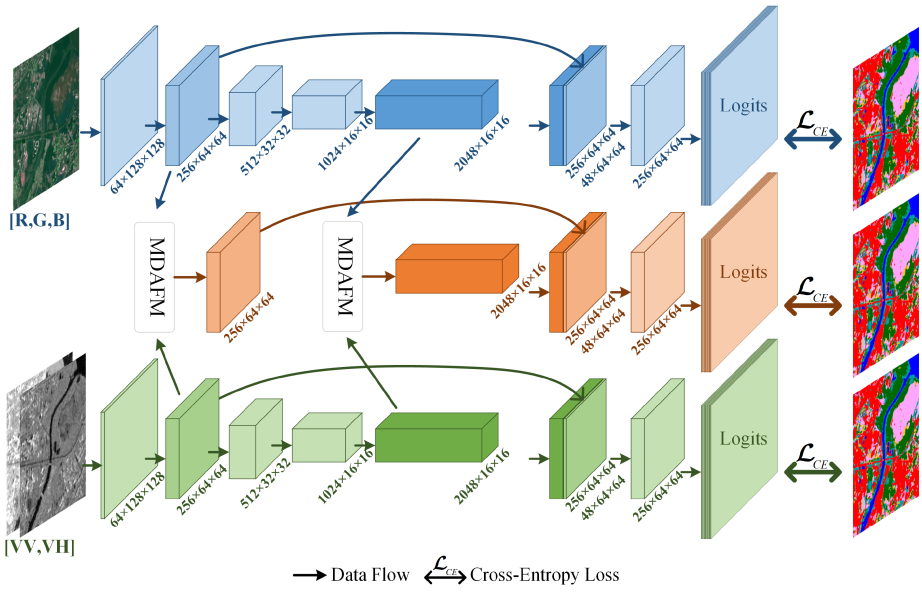

Fig. 2. Encoding-level fusion framework of the joint network combining dual-attention fusion modality and two specific modalities (JoiTriNet-e).

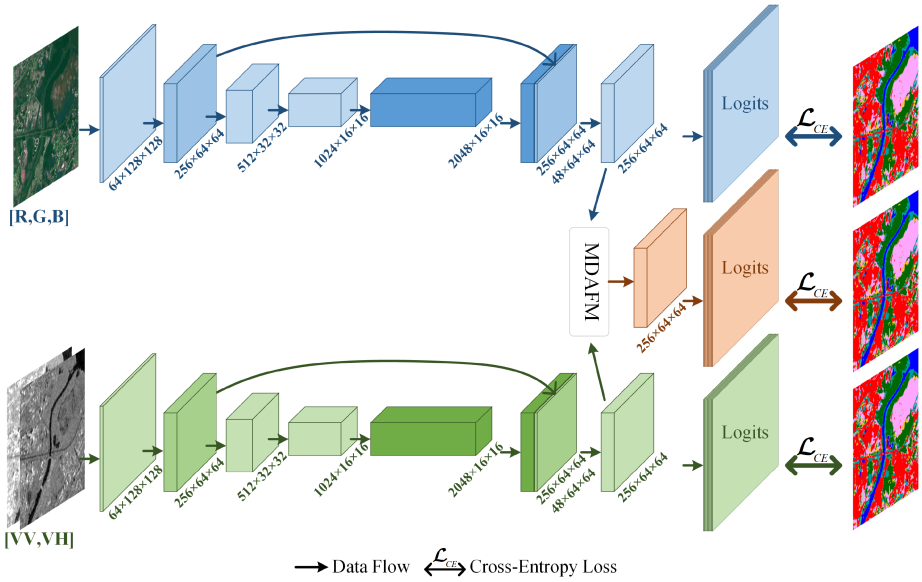

Fig. 3. Decoding-level fusion framework of the joint network combining dual-attention fusion modality and two specific modalities (JoiTriNet-d).

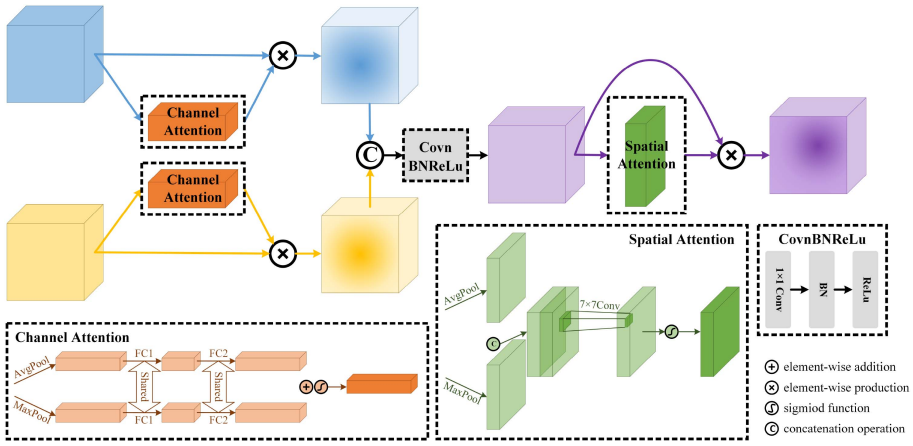

Fig. 4. Structure of the MDAFM.

image features. The channel attention mechanism was used to focus on the band information of a specific modality to improve its feature-extraction ability. After the cascading convolution operation, spatial attention is then used to focus on the position to improve the interpretation ability of the fused features. However, most of the current fusion modules exclude the important role of spatial attention. The detailed operation of the proposed MDAFM module is described as follows.

Given two modality-specific feature maps $\mathbf{F}^{\text{opt}} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}^{\text{sar}} \in \mathbb{R}^{C\times H\times W}$ as inputs, H, W, and C represent the height, width, and channel of the features, respectively. Two channel attention modules are exploited to individually infer domain specific 1-D channel attention maps, $\mathbf{M}_c^{\text{opt}} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathbf{M}_c^{\text{sar}} \in \mathbb{R}^{C\times 1\times 1}$, before merging operations, and broadcast them along the spatial dimension, which can be formulated as $\mathbf{F}_{ca}^{\text{opt}} = \mathbf{M}_c^{\text{opt}}(\mathbf{F}^{\text{opt}}) \otimes \mathbf{F}^{\text{opt}}$ and $\mathbf{F}_{ca}^{\text{sar}} = \mathbf{M}_c^{\text{sar}}(\mathbf{F}^{\text{sar}}) \otimes \mathbf{F}^{\text{sar}}$, respectively, where $\otimes$ denotes the elementwise production operation. Subsequently, the channel attention features of different modalities are combined using a concatenation operation, and one Conv-BN-ReLU block is involved in refining and interactively fusing the features. A spatial attention module is then applied to learn the positional information of the fusion features by inferring the 2-D spatial attention map $\mathbf{M}_s^{\text{fus}} \in \mathbb{R}^{1\times H\times W}$ and broadcasting it along the channel dimension, that is, $\mathbf{F}_{sa}^{\text{fus}} = \mathbf{M}_s^{\text{fus}}(\mathbf{F}^{\text{fus}}) \otimes \mathbf{F}^{\text{fus}}$. This further enhances the useful features and suppresses useless noise to effectively integrate full-modality discriminative information. The channel and spatial attention modules used in this study were referenced from the CBAM [57] and can be summarized as follows:

$$\begin{aligned}\mathbf{M_C} &= \sigma\left(\text{MLP}\left(\text{AvgPool}\left(\mathbf{F}\right)\right) + \text{MLP}\left(\text{MaxPool}\left(\mathbf{F}\right)\right)\right)\\ \mathbf{M_S} &= \sigma\left(f^{7\times 7}\left(\left[\text{AvgPool}\left(\mathbf{F}\right);\left(\text{MaxPool}\left(\mathbf{F}\right)\right)\right]\right)\right) \qquad (4)\end{aligned}$$

where σ denotes the sigmoid function, AvgPool$(\cdot)$ and MaxPool$(\cdot)$ denote the average-pooling and max-pooling operations, respectively, MLP$(\cdot)$ denotes the multilayer perceptron with one hidden layer, and $f^{7\times 7}$ represents a convolution operation with a filter size of 7×7.

IV. Experiment

A. Dataset Configuration

The experiments were conducted on two multimodal land cover classification datasets, namely DFC2020 and Dongying, to quantitatively and qualitatively evaluate the performance of the proposed method.

1) DFC2020: This dataset is the land cover mapping dataset released by the 2020 IEEE Data Fusion Contest [26], which is a subset of the SEN12MS dataset [25] and contains 6114 pairs of labeled optical-SAR image data. The optical image is multispectral data with 13 bands acquired by Sentinel-2, among which we only used the high-resolution RGB bands as input to the optical branch of the proposed network. The SAR image is dual-polarization data composed of VV and VH components obtained by Sentinel-1 during the same season. The ground sampling distance of all the data was 10 m, and the size of each image patch was 256×256 pixels. There are two types of annotations, namely MODIS-derived land cover maps sampled across the globe and semimanually generated high-resolution labels with a resolution of 10 m. We used high-resolution annotations for supervision, which follows a simplified version of the IGBP classification scheme consisting of ten land cover classes. Since the rarity of the savanna and snow/ice categories in this dataset, we excluded them and used eight land cover categories for classification, that is, "forests," "shrublands," "grasslands," "wetlands," "croplands," "urban/built-up," "barren," and "water." Fig. 5 shows an example of DFC2020.

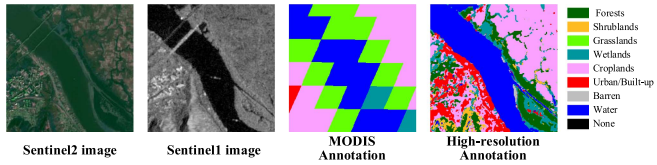

Fig. 5. Example of DFC2020 dataset.

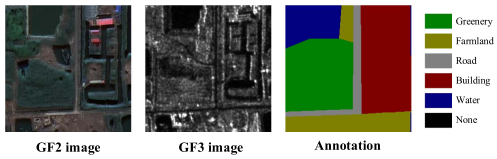

Fig. 6. Example of Dongying dataset.

2) Dongying: Ren et al. [31] published the multimodal RSI dataset including Xi'an, Dongying, and Pohang subsets.Herein, we used the Dongying subset as it is the largest and has a total of 7831 optical–SAR image pairs. The optical image is three-band data fused with panchromatic and multispectral images acquired from the Gaofen-2 satellite. The SAR image is VV polarization data obtained from the Gaofen-3 satellite in slider spotlight mode. The spatial resolution of all data in this dataset is 1 m after preprocessing, while the size of each image patch was 256×256 pixels. The labels were annotated using Labelme software, with five land cover categories, namely, "buildings," "farmland," "greenery," "water," and "roads." Fig. 6 shows an example of Dongying.

In this study, we randomly divided the datasets into training, validation, and testing sets in a numerical ratio of $6:2:2$; the DFC2020 dataset had 3670 samples for training, 1222 for validation, and 1222 for testing; Dongying had 4699 samples for training, 1566 for validation, and 1566 for testing.

B. Experimental Setup

1) Implementation Details: All models were implemented in a PyTorch environment and carried out on an NVIDA Tesla V100 GPU. All experiments were conducted under the same experimental parameter conditions. We used the stochastic gradient descent optimizer to train the model with an initial learning rate of 1×10^{-3} and conducted 200 epochs with a batch size of eight, during which iterative training was terminated early if the validation loss did not decrease for 20 consecutive epochs.

2) Evaluation Metrics: In this study, four indicators were used to evaluate the accuracy of land cover classification, that is, overall accuracy (OA), mean pixel accuracy (mPA), mean

intersection over union (mIoU), and kappa coefficient (Kappa). The OA is expressed as the percentage of correctly predicted pixels over the total number of pixels, which directly reflects the proportion of correct classifications by simple calculations. Given that the number of samples in each category is often unbalanced in practical applications, the OA cannot accurately reflect the performance of the model. We used the mPA, mIoU, and Kappa indicators to penalize the bias of the model to synthetically evaluate whether the predicted results of the model were consistent with the actual classification results. The OA, mPA, mIoU, and Kappa were calculated as follows:

$$\mathrm{OA} = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{5}$$

$$\mathrm{mPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{6}$$

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{7}$$

$$\mathrm{Kappa} = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

$$\text{where}, p_o = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}, p_e = \frac{\sum_{i=0}^{k}\left(\sum_{j=0}^{k} p_{ij} \times \sum_{j=0}^{k} p_{ji}\right)}{\left(\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}\right)^2} \tag{9}$$

where k is the number of land type classes; in this study, $k = 8$. p_{ii} is the number of the pixels of class i correctly predicted to belong to class i, p_{ij} is the number of pixels of class i wrongly predicted to belong to class j, and similarly, p_{ji} is the number of the pixels of class j wrongly predicted to belong to class i.

3) Comparison methods: To evaluate the performance of the proposed method, we compared the JoiTriNet-e and JoiTriNet-d with respect to several state-of-the-art methods for land cover classification tasks in multisource RSIs.

1) PSCNN [29] is a pseudosiamese CNN, which performs land cover classification by dividing large RSIs into small patch scenes. Since we focused on pixel-level classification based on semantic segmentation, its classification head was replaced with a segmentation head.
2) MRSDC [28] is a two-branch CNN whose input is HIS image and visible/LiDAR data. Since we focused on the joint application of optical and SAR images, the input of the HSI branch was replaced with SAR image and the input channel of the first convolutional layer was changed accordingly.
3) V-FesuNet [45] is a deep FCN for segmentation of multimodal remote sensing data, which fuses multispectral and LiDAR data in the encoding phase. As in 2, the input of the LiDAR branch was replaced with SAR image.
4) MBFNet [30] fuses optical and SAR features through a bilinear pooling operation and a second-order attention-based channel selection module. It is also a patch-based land cover classification method operated as in 1.
5) DDHRNet [31] fuses optical and SAR features using multimodal squeeze-and-excitation modules at various stages of feature encoding.
6) MCANet [32] includes a pseudosiamese feature extraction module composed of two ResNet101, a multimodal-cross attention module, and a low–high-level feature fusion module designed with reference to DeepLabV3+.

Among the aforementioned methods, MCANet strictly belongs to the ELF architecture; V-FesuNet and DDHRNet can be regarded as an ELF architecture that integrates fusion modules into the encoder in stages; while PSCNN, MRSDC, and MBFNet can be categorized as DLF architecture after variation.

TABLE I
COMPARISON OF DFC2020 AND DONGYING DATASET

	DFC2020	Dongying
Optical image	RGB bands	Three-band
Optical satellite	Sentinel-2	Gaofen-2
SAR image	VV and VH	VV
SAR satellite	Sentinel-1	Gaofen-3
Acquisition area	Global	Shandong, China
Spatial resolution	10 m	1 m
Image patch size	256 × 256	256 × 256
Dataset size	6114 pairs	7831 pairs
Number of categories	8	5

TABLE II
EVALUATION METRICS (%) CALCULATED BY THE PROPOSED METHOD AND OTHER METHODS FOR LAND COVER CLASSIFICATION OF MULTISOURCE RSIS

Method	DFC2020				Dongying			
	OA	Kappa	mPA	mIoU	OA	Kappa	mPA	mIoU
PSCNN	81.77	77.80	71.74	59.78	91.54	88.79	90.55	83.09
VFesuNet	79.58	75.14	69.54	57.01	91.36	88.54	89.95	82.36
MRSDC	83.97	80.46	72.31	61.58	73.65	64.92	71.85	58.17
MBFNet	76.45	71.07	58.74	47.56	80.19	73.68	78.39	66.20
DDHRNet	81.67	77.76	71.48	59.18	84.66	79.68	82.84	72.25
MCANet	84.55	81.21	75.62	63.92	93.12	90.89	92.11	86.08
JoiTriNet-e	85.73	82.63	76.44	65.77	93.79	91.77	92.90	87.42
JoiTriNet-d	**86.06**	**83.04**	**77.19**	**66.58**	**94.13**	**92.23**	**93.38**	**87.97**

Bold indicates the best value.

C. Comparative Experiments

1) Quantitative Result: We first calculated the metrics of all methods to quantitatively compare the results of the proposed method with other state-of-the-art methods for land cover classification of multisource RSIs, as shown in Table II. The proposed joint network combining dual-attention fusion modality and two specific modalities produced satisfactory results. Both JoiTriNet-e and JoiTriNet-d outperform other methods, in which JoiTriNet-d achieved state-of-the-art classification performance among all the methods compared. Specifically, JoiTriNet-d had the highest OA of 86.06% on the DFC2020

TABLE III
PA (%) AND IoU (%) FOR EACH CATEGORY ON DFC2020 AND DONGYING DATASETS CALCULATED BY DIFFERENT METHODS

Metrics	Method	DFC2020								Dongying				
		Forests	Shrub lands	Grass lands	Wet lands	Crop lands	Urban/ Builtup	Barren	Water	Greenery	Farm land	Road	Building	Water
PA	PSCNN	88.72	52.65	63.47	62.74	81.50	81.39	45.06	98.40	92.25	83.93	87.09	95.87	93.60
	VFesuNet	84.45	55.17	57.04	60.33	81.02	79.00	41.51	97.76	92.91	85.26	83.80	95.34	92.46
	MRSDC	92.58	54.26	72.47	56.41	82.43	86.48	34.54	99.33	77.24	46.31	70.85	85.98	78.85
	MBFNet	86.97	21.30	53.88	47.24	81.55	80.33	0.18	98.43	77.79	64.78	74.48	90.77	84.13
	DDHRNet	91.82	**67.01**	74.06	54.90	71.60	78.09	35.59	98.74	85.45	80.77	74.45	88.92	84.63
	MCANet	90.26	57.29	70.12	62.34	85.29	84.67	**55.84**	99.13	94.79	87.78	87.79	96.09	94.12
	JoiTriNet-e	91.98	62.57	71.12	**67.30**	**86.78**	84.84	47.76	**99.14**	95.78	**89.86**	88.73	95.78	94.37
	JoiTriNet-d	**92.84**	60.45	**74.10**	65.87	85.89	**85.52**	53.72	99.09	**96.24**	88.84	**89.98**	**96.36**	**95.46**
IoU	PSCNN	77.94	38.05	49.66	49.04	63.86	66.68	36.27	96.73	85.94	75.08	77.66	89.93	86.83
	VFesuNet	73.42	37.36	44.23	47.67	61.48	62.60	33.86	95.46	87.42	75.64	73.57	89.12	86.07
	MRSDC	83.46	37.88	57.84	47.44	65.98	73.97	28.30	97.73	56.55	33.34	56.71	72.86	71.41
	MBFNet	74.48	16.19	39.84	38.94	55.08	60.08	0.18	95.65	64.98	47.34	63.99	80.18	74.51
	DDHRNet	81.14	38.48	54.28	46.51	59.96	68.41	27.04	97.61	75.74	60.89	63.91	80.95	79.75
	MCANet	83.99	43.19	56.72	49.35	67.28	72.58	40.44	97.83	88.97	79.75	81.15	91.08	89.47
	JoiTriNet-e	84.74	45.80	59.17	**54.12**	69.96	73.52	40.77	98.07	90.50	81.59	82.76	91.47	90.79
	JoiTriNet-d	**85.08**	**46.21**	**60.42**	53.64	**70.20**	**74.08**	**44.90**	**98.16**	**91.34**	**82.46**	**82.94**	**91.85**	**91.28**

Bold indicates the best value.

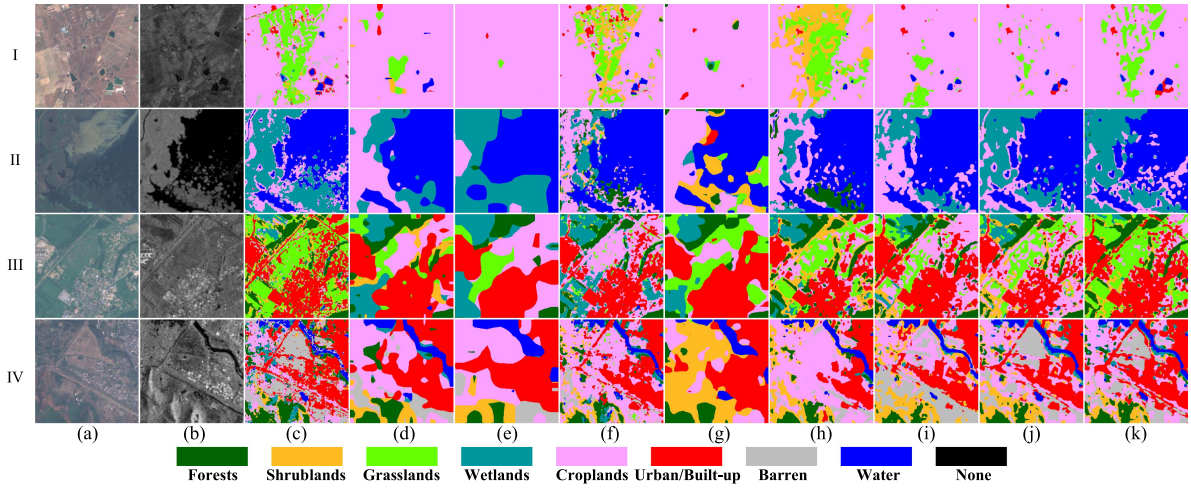

Fig. 7. Land cover classification maps of different methods on the DFC2020 dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) PSCNN. (e) VFesuNet. (f) MRSDC. (g) MBFNet. (h) DDHRNet. (i) MCANet. (j) JoiTriNet-e. (k) JoiTriNet-d.

dataset, achieving an improvement of 1.51% compared with the other best method MCANet. JoiTriNet-d also had the highest OA of 94.13% on the Dongying dataset, 1.01% higher than MCANet. Considering that MCANet is an ELF architecture, we compared it with JoiTriNet-e, another encoding-level feature fusion. JoiTriNet-e could increase the OA by 1.18% to the second highest on the DFC2020 dataset, and the same performance was also achieved on the Dongying dataset. These results highlight the superiority of the proposed method, which improves the performance of multisource RSIs land cover classification algorithm through joint learning of fusion and specific modalities.

The improvement of the Kappa, mPA, and mIoU is particularly noteworthy as they better reflect the model performance in the case of class-imbalanced samples. Notably, the mIoU value of JoiTriNet-d showed a significant increase of 2.66% and 1.89% on the DFC2020 and Dongying datasets, respectively. Furthermore, we presented the PA and IoU values for each category in Table III to verify the prediction ability of our method across different categories. Our method achieved the highest PA and IoU values for all categories in the Dongying dataset and also improved the accuracy to varying degrees in the DFC2020 dataset with varying degrees of accuracy improvement for all categories in the DFC2020 dataset.

2) Qualitative Analysis: To visually assess the segmentation performance of the proposed method, we present comparisons of the land cover classification maps generated using different methods for several samples of the test set. The samples shown in Fig. 7 are selected from the DFC2020 dataset. Overall, compared with PSCNN, VFesuNet, and MBFNet methods, MRSDC, DDHRNet, MCANet, and the proposed methods showed finer segmentation granularity, among which JoiTriNet-d performed particularly well. The other methods showed varying degrees of confusion between "croplands" and "grasslands," "wetlands," "barren" in different scenarios. For example, most of the other methods in Groups I failed to separate the "grasslands" from the "croplands," and in Groups II they also misidentified the "grasslands" as "croplands." The Groups III and IV misclassified

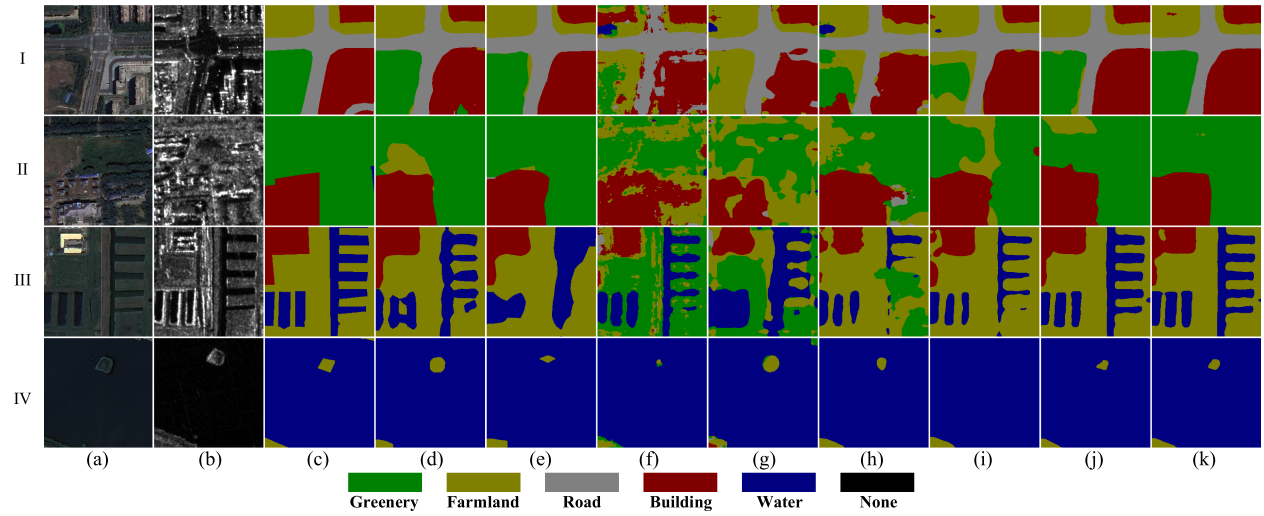

Fig. 8. Land cover classification maps of the different methods used on the Dongying dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) PSCNN. (e) VFesuNet. (f) MRSDC. (g) MBFNet. (h) DDHRNet. (i) MCANet. (j) JoiTriNet-e. (k) JoiTriNet-d.

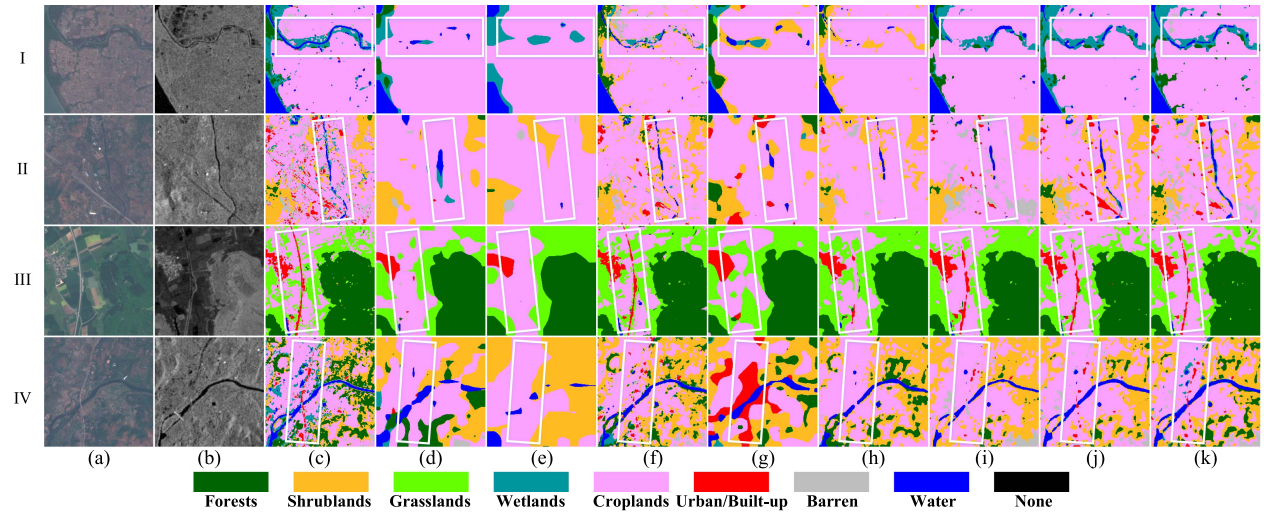

Fig. 9. Detailed comparison of the classification results generated using different methods on the DFC2020 dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) PSCNN. (e) VFesuNet. (f) MRSDC. (g) MBFNet. (h) DDHRNet. (i) MCANet. (j) JoiTriNet-e. (k) JoiTriNet-d. The main differences are highlighted with white rectangles.

large areas of "wetlands" and "barren" as "croplands." This is primarily because these categories are inherently difficult to distinguish in this scenario, and optical and SAR images have varying degrees of confusion with these categories. The fusion process does not completely distinguish between confusion categories, because the interference between the discriminant features of specific modalities weakens the ability of the fusion features. In the proposed JoiTriNet, this situation was improved because the feature-extraction ability of the modality-specific branches was enhanced during the joint learning process, thereby enhancing the discriminative ability of the modality-fusion branch and alleviating the phenomenon of low separability between categories. Furthermore, the superiority of our method was also demonstrated by the land cover classification map obtained from the Dongying dataset shown in Fig. 8. Overall, our method outperformed others by effectively distinguishing between "greenery" and "farmland" while maintaining clear boundaries between "farmland" and "water."

In addition, the proposed method yielded more accurate predictions for some thin-strip-shaped surface features, as shown in Figs. 9 and 10. Compared with other methods in the white box selection area shown in Fig. 9, our method, especially JoiTriNet-d, outlined the river completely and without truncation (in Groups I and II) and successfully predicted the forests in the urban/built-up areas (in Groups III and IV). The samples from the Dongying dataset illustrated in Fig. 10 more obviously demonstrate that the proposed method enhanced continuity for recognizing and classifying narrow ground objects, such as roads and rivers.

3) Robustness Test: The proposed networks can not only output the fusion modality classification results but also retain the output of two modality-specific branches, namely optical

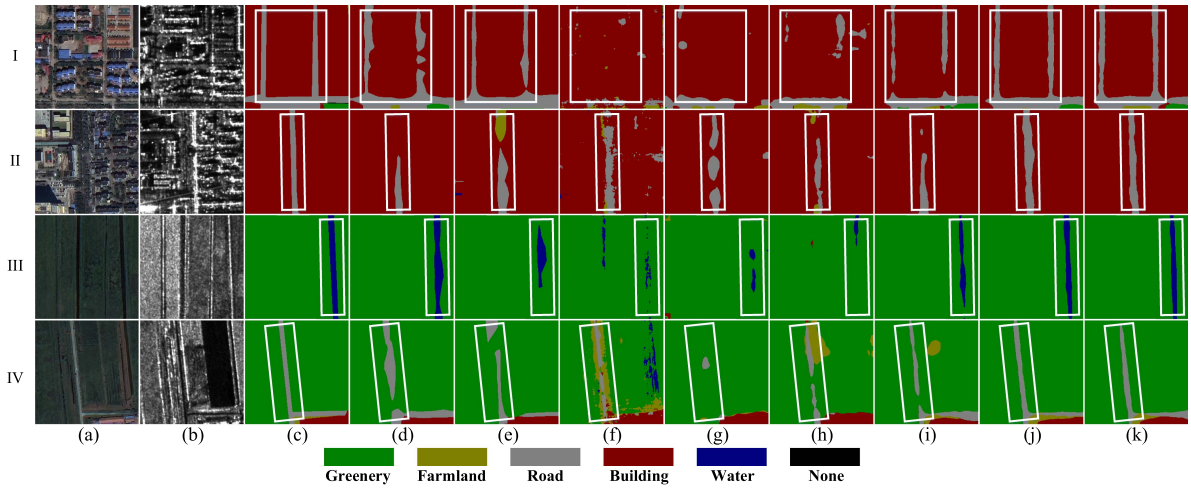

Fig. 10. Detailed comparison of the classification results generated using different methods on the Dongying dataset. (a) Optical images. (b) SAR images. (c) Annotations. (d) PSCNN. (e) VFesuNet. (f) MRSDC. (g) MBFNet. (h) DDHRNet. (i) MCANet. (j) JoiTriNet-e. (k) JoiTriNet-d. The main differences are highlighted with white rectangles.

TABLE IV
EVALUATION METRICS (%) CALCULATED USING VARIOUS MULTISOURCE RSIS LAND COVER CLASSIFICATION METHODS WHEN ONLY ONE MODAL DATA WAS AVAILABLE

Input	Method	DFC2020				Dongying			
		OA	Kappa	mPA	mIoU	OA	Kappa	mPA	mIoU
Optical image + all black image	PSCNN	25.12	1.68	13.53	4.05	50.41	36.09	51.68	30.90
	VFesuNet	24.05	0.19	12.81	3.30	67.80	58.63	69.57	49.19
	MRSDC	25.73	2.79	15.30	5.45	70.51	60.70	67.02	54.11
	MBFNet	23.93	0.01	12.50	3.00	17.20	1.79	25.57	7.73
	DDHRNet	26.85	4.42	15.08	5.62	56.07	42.83	53.55	37.72
	MCANet	24.12	0.34	13.05	3.52	57.51	41.71	53.72	35.46
	JoiTriNet-e	84.57	81.22	75.38	64.33	93.40	91.26	92.57	86.72
	JoiTriNet-d	84.80	81.49	75.78	64.64	93.68	91.64	93.10	87.22
all black image + SAR image	PSCNN	44.64	27.56	23.70	14.95	30.89	15.66	32.46	14.01
	VFesuNet	47.18	32.90	27.45	18.21	36.43	5.18	24.49	11.86
	MRSDC	46.95	30.43	24.65	15.67	35.38	0.00	20.00	7.08
	MBFNet	57.78	47.26	36.62	24.80	25.14	4.04	22.25	7.54
	DDHRNet	29.75	7.95	15.74	6.39	33.42	19.96	38.70	21.42
	MCANet	45.80	33.34	29.95	16.06	16.78	7.23	29.47	8.71
	JoiTriNet-e	81.34	77.21	69.49	58.16	89.86	86.58	88.56	79.84
	JoiTriNet-d	81.36	77.25	69.58	58.16	90.13	86.93	88.82	80.24

modality and SAR modality. Even when only one modality of data is available, the respective branch can still generate a robust land cover classification map.

To evaluate its superiority, we compared the performance of various multisource RSIs land cover classification networks using only optical or SAR images. When only one modality of data is available, the other modality branch utilizes all-black data as the input to ensure the normal operation of the fusion network. As shown in Table IV, while other fusion networks rapidly declined in their classification performance and failed to generate accurate land cover maps, our proposed method maintained a high level of accuracy. For instance, on the DFC2020 dataset with only optical images available, JoiTriNet-d achieved an OA of 84.80%, whereas other methods struggle to reach even half of that accuracy level. Similar results were observed when only SAR images were used for classification. These findings demonstrate the excellent robustness of our method as it remains unaffected by the limited availability of modality data while achieving consistently high accuracy.

TABLE V
EFFECTIVENESS OF THE FSMJL STRATEGY AND THE MDAFM MODULE IN THE DONGYING DATASET

Architecture	FSMJL	MDAFM	Metrics (%)			
			OA	Kappa	mPA	mIoU
ELF	✗	✗	93.14	90.91	92.15	86.17
	✓	✗	93.79	91.77	92.90	87.39
	✓	✓	93.79	91.77	92.90	87.42
DLF	✗	✗	93.13	90.91	92.35	86.17
	✓	✗	93.72	91.68	92.69	87.22
	✓	✓	94.13	92.23	93.38	87.97

D. Ablation Study

1) Effectiveness of the FSMJL Strategy and MDAFM Module: To verify the effectiveness of FSMJL strategy and the contribution of MDAFM module, we conducted extensive ablation experiments. As shown in Table V, the results of the ablation experiments demonstrated how the proposed method

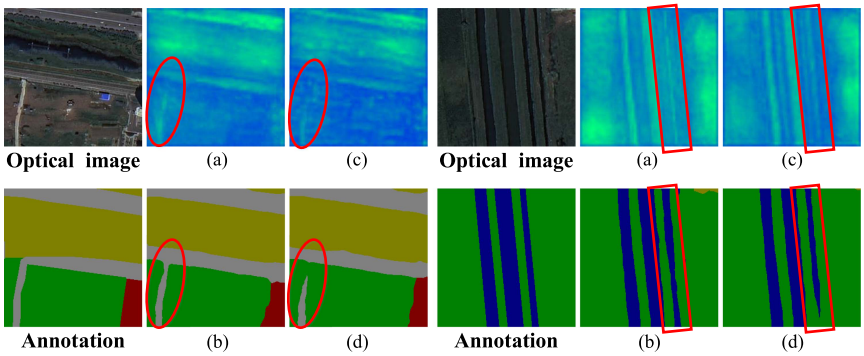

Fig. 11. Visualization results of dual attention decoupling for the proposed MDAFM on FSMJL-DLF architecture. (a) Fusion features and (b) the final land cover classification map generated using MDAFM with spatial attention blocks; (c) and (d) are generated without spatial attention blocks. The main differences are highlighted with red boxes.

TABLE VI
EFFECTIVENESS OF FSMJL STRATEGY IN ENHANCING THE PERFORMANCE OF MODALITY-SPECIFIC BRANCHES

Modality	Method	OA (%)	Kappa (%)	mPA (%)	mIoU (%)
DeepLabV3+	OPT	84.73	81.41	75.65	64.48
FSMJL-ELF	OPT	93.40	91.26	92.57	86.72
FSMJL-DLF	OPT	93.68	91.64	93.10	87.22
DeepLabV3+	SAR	81.39	77.29	69.84	58.40
FSMJL-ELF	SAR	89.86	86.58	88.56	79.84
FSMJL-DLF	SAR	90.13	86.93	88.82	80.24

TABLE VII
EFFECTIVENESS OF SPATIAL ATTENTION BLOCK IN MDAFM MODULE

Architecture	MDAFM		OA(%)	Kappa(%)	mPA(%)	mIoU(%)
	Channel	Spatial				
ELF	✓		93.02	90.75	92.10	85.97
	✓	✓	93.18	90.96	92.19	86.22
FSMJL-ELF	✓		93.96	92.01	93.22	87.67
	✓	✓	93.79	91.77	92.90	87.42
DLF	✓		92.77	90.43	92.09	85.50
	✓	✓	92.91	90.61	92.10	85.75
FSMJL-DLF	✓		93.99	92.04	93.08	87.70
	✓	✓	94.13	92.23	93.38	87.97

successively improved the accuracy. Specifically, we utilized the ELF and DLF architectures designed with ResNet101 encoder and DeepLabV3+ decoder as baselines, which fused optical and SAR features with the fusion rule of pixelwise addition. When the FSMJL strategy was introduced into ELF and DLF architecture, the performance of the fusion network significantly improved. The metrics of ELF architecture are increased by 0.65% (OA), 0.86% (Kappa), 0.75% (mPA), and 1.22% (mIoU), respectively, while the DLF architecture achieves enhancements of 0.59% (OA), 0.77% (Kappa), 0.34% (mPA), and 1.05% (mIoU). This improvement can be attributed to enhancing feature extraction capabilities in modality-specific branches through the FSMJL strategy, thereby boosting overall modality-fusion branch performance. In addition, by introducing the MDAFM, compared with the slight improvement in ELF architecture, DLF architecture demonstrated significant enhancement with improvements of 0.41% (OA), 0.55% (Kappa), 0.69% (mPA), and 0.75% (mIoU), respectively. This indicates that the dual attention fusion implemented by MDAFM further enhances the performance of the fusion network.

2) Contribution of Modality-Specific Branches: To illustrate the effectiveness of the FSMJL strategy in enhancing the performance of modality-specific branches, we compared it with the DeepLabV3+ network trained solely on single-modal images. As shown in Table VI, both ELF and DLF architectures exhibited significant improvements. Taking the optical modality branch of JoiTriNet-d as an example, the segmentation accuracy was improved by 8.95% (OA), 10.23% (Kappa), 17.45% (mPA), and 22.74% (mIoU). Combined with Table IV, our method surpasses the single-modal network in classification performance when only one modality was available. These ablation results demonstrate the effectiveness of the FSMJL strategy, achieving simultaneous improvement of modality-fusion branch and modality-specific branches performance.

3) Decoupling of Dual Attention: The multimodal fusion module designed herein incorporates spatial attention in addition to the channel attention mechanism for fusing multimodal features. It applies channel attention to each modality individually, followed by cascade convolution fusion, and finally adds spatial attention. To validate the effectiveness of this structure, we conducted ablation experiments by removing the spatial attention block from the fusion module under four architectures. As shown in Table VII, the MDAFM with spatial attention block performed better in ELF, DLF, and FSMJL-DLF architectures. This indicates that the spatial attention block more effectively integrates complementary information of multimodal features and improves classification accuracy.

To visually assess the effectiveness of the spatial attention block, we presented the fusion features aggregated by the MDAFM and the final land cover classification map obtained by the FSMJL-DLF architecture, in Fig. 11. The visualization results indicated that the fusion features generated by the MDAFM with spatial attention block were more sensitive to the spatial structure of ground objects and responded to the whole road and the whole river, and the segmented road was more complete. This indicates that spatial attention further focuses on location

information and improves the interpretation ability of fusion features.

E. Discussion

In this study, we illustrated JoiTriNet and the effectiveness of FSMJL strategy and MDAFM module from various perspectives. Compared with other methods for land cover classification in multisource RSIs, JoiTriNet achieved the best performance. The quantitative indicators showed significant improvement and the visualization results of the land cover classification showed enhanced refinement. Moreover, robust output was obtained even when only one modality was available. Through ablation studies, we validated that the FSMJL strategy enhanced both modality-fusion branch and modality-specific branches segmentation performance, while the MDAFM further improved fusion features by introducing dual attention formed through spatial attention. However, the SAR modality branch demonstrated a smaller accuracy improvement compared with its optical counterpart, and the impact of MDAFM on FSMJL-ELF architecture was not as pronounced as in FSMJL-DLF architecture. There remains ample room for improvement in these aspects. Overall, our proposed method exhibits great potential.

V. Conclusion

In this study, we investigated multisource land cover classification using optical and SAR images. First, we analyzed the multimodal semantic segmentation architecture based on an encoder–decoder structure and summarized the encoding and decoding level feature fusion architectures. Subsequently, we proposed a strategy for FSMJL, leading to the design of two corresponding architectures. The designed architectures improved the accuracy of multisource and single-source RSIs land cover classification by encouraging modality-specific branches to learn unique discriminative information while encouraging modality-fusion branch to learn detailed interpretation information. Furthermore, based on the proposed architectures, we designed a joint network combining dual-attention fusion modality and two specific modalities, which has two modes of encoding-level fusion (JoiTriNet-e) and decoding-level fusion (JoiTriNet-d). They included a novel MDAFM to efficiently aggregate optical and SAR features. Finally, we evaluated the proposed method through a series of experiments. Comparative experimental results showed that the proposed method outperformed other multisource RSIs land cover classification methods, among which JoiTriNet-d achieved the highest level of accuracy. Ablation experiments demonstrated the effectiveness of the proposed architecture and module from different perspectives. The experimental findings highlight the significant potential of our method in land cover classification tasks involving multisource and single-source RSIs.

References

[1] X. He, S. Zhang, B. Xue, T. Zhao, and T. Wu, "Cross-modal change detection flood extraction based on convolutional neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, 2023, Art. no. 103197.

[2] Y. Ma et al., "Remote sensing Big Data computing: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, 2015.

[3] H. Zhang and R. Xu, "Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the pearl river delta," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 87–95, 2018.

[4] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[5] 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.

[6] J. Kang et al., "DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[7] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7797–7808, 2022.

[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] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE Proc. IRE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[10] Y. Alebele et al., "Estimation of crop yield from combined optical and SAR imagery using Gaussian kernel regression," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10520–10534, 2021.

[11] W. Zhao, Y. Qu, J. Chen, and Z. Yuan, "Deeply synergistic optical and SAR time series for crop dynamic monitoring," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111952.

[12] L. Gao, X. Li, F. Kong, R. Yu, Y. Guo, and Y. Ren, "AlgaeNet: A deep-learning framework to detect floating green algae from optical and SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2782–2796, 2022.

[13] M. Cutler, D. S. Boyd, G. M. Foody, and A. Vetrivel, "Estimating tropical forest biomass with a combination of SAR image texture and landsat TM data: An assessment of predictions between regions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 70, pp. 66–77, 2012.

[14] H. Zhang et al., "A manifold learning approach to urban land cover classification with optical and radar data," *Landscape Urban Plan.*, vol. 172, pp. 11–24, 2018.

[15] A. Salentinig and P. Gamba, "Combining SAR-based and multispectral-based extractions to map urban areas at multiple spatial resolutions," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 100–112, Sep. 2015.

[16] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Inf. Fusion*, vol. 59, pp. 13–29, 2020.

[17] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 284–288, Mar. 2017.

[18] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.

[19] H. Bagan, T. Kinoshita, and Y. Yamagata, "Combination of AVNIR-2, PALSAR, and polarimetric parameters for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1318–1328, Apr. 2012.

[20] Y. Qin et al., "Forest cover maps of China in 2010 from multiple approaches and data sources: PALSAR, landsat, modis, FRA, and NFI," *ISPRS J. Photogrammetry Remote Sens.*, vol. 109, pp. 1–16, 2015.

[21] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag,*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[22] X. X. Zhu et al., "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 143–172, Dec. 2021.

[23] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[24] W. Zhang, P. Tang, and L. Zhao, "Fast and accurate land-cover classification on medium-resolution remote-sensing images using segmentation models," *Int. J. Remote Sens.*, vol. 42, no. 9, pp. 3277–3301, 2021.
[25] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS–a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019, doi: 10.5194/isprs-annals-IV-2-W7-153-2019.
[26] N. Yokoya, P. Ghamisi, R. Hänsch, and M. Schmitt, "2020 IEEE GRSS data fusion contest: Global land cover mapping with weak supervision [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 1, pp. 154–157, Mar. 2020.
[27] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.
[28] 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.
[29] 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.
[30] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, 2020.
[31] B. Ren et al., "A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, 2022, Art. no. 102896.
[32] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 106, 2022, Art. no. 102638.
[33] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 277–293, 2017.
[34] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.
[35] A. Sharma, X. Liu, X. Yang, and D. Shi, "A patch-based convolutional neural network for remote sensing image classification," *Neural Netw.*, vol. 95, pp. 19–28, 2017.
[36] A. Sharma, X. Liu, and X. Yang, "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks," *Neural Netw.*, vol. 105, pp. 346–355, 2018.
[37] H. Song, Y. Kim, and Y. Kim, "A patch-based light convolutional neural network for land-cover mapping using landsat-8 images," *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 114.
[38] Y. Lin, F. Jin, D. Wang, S. Wang, and X. Liu, "Dual-task network for road extraction from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 66–78, 2023.
[39] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.
[40] A. Ghosh, M. Ehrlich, S. Shah, L. S. Davis, and R. Chellappa, "Stacked U-Nets for ground material segmentation in remote sensing imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 252–2524.
[41] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 223–236, 2019.
[42] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.
[43] P. Iervolino, R. Guida, D. Riccio, and R. Rea, "A novel multispectral, panchromatic and SAR data fusion for land classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 3966–3979, Oct. 2019.
[44] J. Reiche, C. M. Souza, D. H. Hoekman, J. Verbesselt, H. Persaud, and M. Herold, "Feature level fusion of multi-temporal ALOS PALSAR and Landsat data for mapping and monitoring of tropical deforestation and forest degradation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 5, pp. 2159–2173, Oct. 2013.
[45] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, 2018.
[46] Y. Xu, B. Du, and L. Zhang, "Multi-source remote sensing data classification via fully convolutional networks and post-classification processing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3852–3855.
[47] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "m^3Fusion: 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.
[48] Q. Feng, D. Zhu, J. Yang, and B. Li, "Multisource hyperspectral and LiDAR data fusion for urban land-use mapping based on a modified two-branch convolutional neural network," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 1, p. 28, 2019.
[49] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Multi-modal land cover mapping of remote sensing images using pyramid attention and gated fusion networks," *Int. J. Remote Sens.*, vol. 43, no. 9, pp. 3509–3535, 2022.
[50] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[51] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3829–3845, May 2021.
[52] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A cross fusion network for joint land cover classification using optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1562–1574, 2022.
[53] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," 2022, *arXiv:2204.07756*.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. Comput. Vis. pattern Recognit.*, 2009, pp. 248–255.
[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

Xiao Liu received the B.S. degree in remote sensing from the Shandong University of Science and Technology, Qingdao, China, in 2021. She is currently working toward the M.S. degree in surveying and mapping from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China.

Her research interests include remote sensing image interpretation, deep learning, and image processing.

Huijun Zou received the M.S. degree in management from the Nanjing University of Science and Technology, Nanjing, China, in 2006.

She is currently an Associate Professor with Information Engineering University, Zhengzhou, China. Her research interests include vocational education and occupation skill appraisal.

Shuxiang Wang received the B.S. degree from Information Engineering University, Zhengzhou, China, in 2005, and the M.S. degree from Hohai University, Nanjing, China, in 2009, both in photogrammetry and remote sensing. She is currently working toward the Ph.D. degree in remote sensing image processing and machine learning with Information Engineering University.

She is currently an Associate Professor with Information Engineering University. Her research focuses on remote sensing image processing.

Yuzhun Lin received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2015 and 2018, respectively, where he is currently working toward the Ph.D. degree in remote sensing image processing and machine learning.

He is currently a Lecturer with Information Engineering University. His research interests include remote sensing image processing and machine learning.

Xibing Zuo received the M.S. degree in surveying and mapping engineering from the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, in 2022, where he is currently working toward the Ph.D. degree in surveying and mapping science and technology.

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