A Review of Optical and SAR Image Deep Feature Fusion in Semantic Segmentation

Chenfang Liu, Yuli Sun, Yanjie Xu, Zhongzhen Sun, Xianghui Zhang, Lin Lei, and Gangyao Kuang, *Senior Member, IEEE*

***Abstract*—With the advent of the era of high-resolution remote sensing, semantic segmentation methods for solving pixel-level classification have been widely studied. Deep learning has significantly advanced deep feature extraction methods, becoming widely employed in remote sensing image analysis. Deep feature fusion methods are able to effectively combine features from different sources. Optical and synthetic aperture radar (SAR) images stand out as primary data sources in remote sensing, offering complementary and consistent information. Fusion of deep semantic features of optical and SAR images can alleviate the limitations of single-source images in application and improve semantic segmentation accuracy. Therefore, this article reviews the research on deep fusion of optical and SAR images in semantic segmentation tasks from four aspects. First, we provide a summary of challenges and research methods pertinent to semantic segmentation of remote sensing images. Then the challenges and urgent needs of deep feature fusion of optical and SAR images are analyzed, and current research is summarized from the perspective of structural design by studying various feature fusion strategies. In addition, the compilation and in-depth analysis of open-source optical and SAR datasets suitable for semantic segmentation are undertaken, serving as fundamental resources for future research endeavors. Finally, the article identifies the major challenges summarized from the literature review in this field, outlining expectations and potential future directions for researchers.**

***Index Terms*—Deep feature fusion, optical images, review, semantic segmentation, synthetic aperture radar (SAR) images.**

I. Introduction

SEMANTIC segmentation in remote sensing imagery is to classify geographic spatial data at the pixel level, thereby enhancing the understanding and analysis of the observed landscape [1], [2], [3]. Semantic segmentation extensively applied in various fields of remote sensing, encompassing tasks such as land use and land cover [4], [5], [6], building extraction [7], [8], [9], impervious surface mapping [10], [11], [12], [13], landslide mapping [14], [15], and others. We counted the proportion of these tasks in semantic segmentation, as shown in Fig. 1.

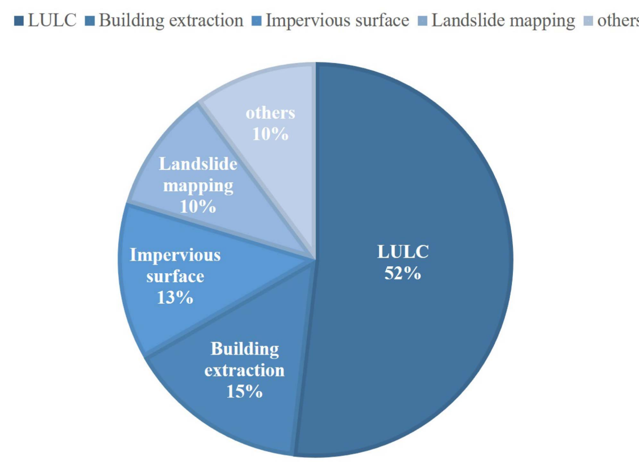

Fig. 1. Distribution of image segmentation study purpose in examined papers.

Remote sensing systems observe image elements across various frequency bands of the electromagnetic spectrum. Optical and synthetic aperture radar (SAR) images are two extensively utilized remote sensing data sources with distinct imaging modalities and representations [16], [17]. Semantic segmentation of remote sensing images heavily relies on the spatial and semantic information embedded in the images. However, the limitations of individual sensors, influenced by factors such as operating environment, wavelength range, and imaging modes, can result in biased scene characterization. Optical sensors passively detect solar radiation reflected from the ground, yielding rich spectral and texture information alongside high spatial resolution. However, they are vulnerable to climatic and lighting effects. SAR, as an active sensor, employs radar waves to penetrate cloud cover, enabling data collection all day and all weather [18], [19], [20], [21]. Nevertheless, SAR is characterized by consistent spot noise and a low signal-to-noise ratio. Fig. 2 shows an example diagram of a pair of optical and SAR images in the same scene [22]. Due to the limitations of information derived from a single source image, it is impractical to comprehensively, accurately, and consistently describe the true state of the scene. Therefore, it becomes imperative to fuse effective information from both optical and SAR images [23], [24], [25], [26]. We collected the number of relevant publications based on the Web of Science and Google Scholar. As depicted in Fig. 3, the number of publications on semantic segmentation for optical and SAR

Manuscript received 23 April 2024; revised 26 May 2024 and 30 June 2024; accepted 4 July 2024. Date of publication 9 July 2024; date of current version 24 July 2024. This work was supported in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20233545, and in part by the Natural Science Foundation of Hunan Province of China under Grant 2024JJ6466. *(Corresponding author: Lin Lei.)*

The authors are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: liuchenfang99@163.com; sunyuli@mail.ustc.edu.cn; 603220383@qq.com; sunzhongzhen14@163.com; zxhzxh0122@163.com; leilin98@nudt.edu.cn; kuanggangyao@nudt.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3424831

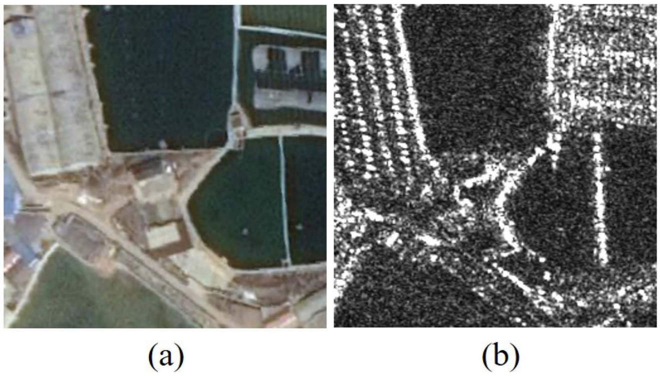

Fig. 2. Example diagram of optical and SAR image pairs. (a) Optical image. (b) SAR image.

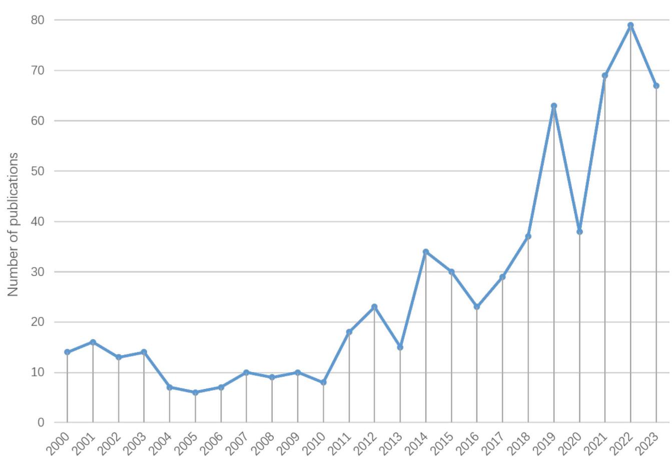

Fig. 3. Number of publications on optical and SAR image segmentation from 2000 to 2023.

image fusion has experienced rapid growth in recent years, attracting widespread attention from researchers.

Multisource data fusion aims to mitigate the heterogeneous differences between modalities while preserving the specific semantic integrity of each modality [27], [28], [29], [30]. One of the critical steps in the data fusion process is image registration, which ensures that the images to be compared are geometrically aligned. This step is particularly challenging and essential for SAR images due to their unique characteristics, such as speckle noise, geometric distortions, and varying acquisition conditions [31], [32], [33], [34]. Xiang et al. [33] proposed a two-stage registration method for large-distortion SAR images based on superpixel segmentation, which can effectively alleviate the difficulty of registration of SAR images with large geometric distortion. Zhang et al.[32] studied the optical flow technology for pixel-by-pixel dense registration of high-resolution optical and SAR images, which can significantly improve the registration accuracy of optical and SAR images. The substantial disparities between optical and SAR images pose challenges in image fusion, and different fusion methods require different registration accuracy. Feature fusion, as a data fusion approach, combines advantageous features from different

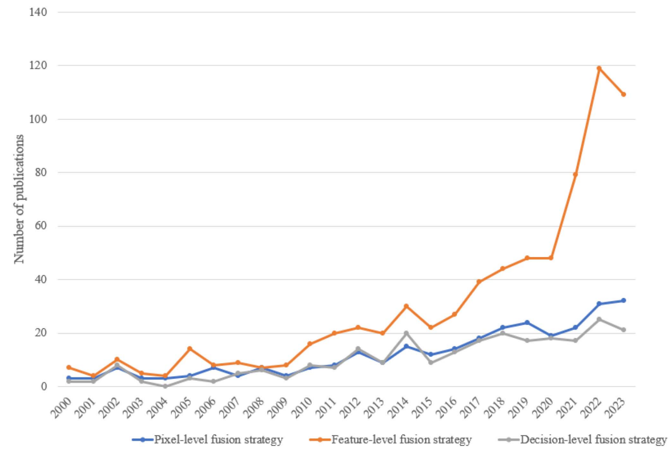

Fig. 4. Number of three image fusion strategies related publications from 2000 to 2023.

modalities, which has gained considerable attention in recent years, as shown in Fig. 4. Moreover, the requirement of feature fusion for image registration accuracy is not as high as that of pixel fusion method. Feature fusion method employing diverse fusion strategies to acquire complementary and consistent features from optical and SAR images helps alleviate the effects of inhomogeneity in remote sensing images and addresses the limitations of single-source data. Since the image structure is invariant to the imaging mode and is insensitive to noise, illumination, and other interference factors, the structural consistency of optics and SAR is also widely used in multisource image feature fusion, polarimetric SAR change detection, multisource image change detection, and other tasks [35], [36], [37], [38].

Traditional methods for extracting image features involve manual design or semiautomated approaches, relying heavily on expert knowledge. However, with the exponential increase in the number of high-resolution remote sensing images resulting from advancements in earth observation technology, manual feature extraction methods struggle to capture the complex features contained in these images. As a result, traditional methods typically extract shallow features and fail to bridge the semantic gap between optical and SAR images. They also lack efficiency in extracting and incorporating advanced semantic information from both categories of images. Consequently, these traditional methods are becoming less suitable for the current era marked by abundant data and advanced remote sensing technology [39], [40], [41], [42]. In response to current demands, remote sensing semantic segmentation techniques based on deep feature fusion continue to evolve. This method can use massive data to drive the model to learn effective features in multimodal images end-to-end, significantly reducing time and labor costs [43], [44], [45]. However, the heterogeneity between optical and SAR images introduces different nonlinear radiometric properties that can impact the extraction and learning of unimodal dominant features. Recent research suggests that modeling the semantic associations between optical and SAR images can enhance semantic segmentation accuracy by learning complementary and coherent features [46], [47].

The articles [48], [49], [50], [51] provide a comprehensive overview of advancements in deep learning for semantic segmentation, specifically summarizing the development of semantic segmentation techniques for natural optical images. In addition, deep learning-based semantic segmentation in remote sensing images is thoroughly reviewed in [52], [53], [54], [55], [56], and [57], with a focus on highlighting the applications of these techniques. The article [28], [58], [59], [60] reviews image fusion and discusses the progress of deep learning methods in this field. However, our survey reveals a gap in the literature regarding advances in fusing SAR and optical image deep features for semantic segmentation. Given the evolving era of multisource big data, our work cannot be delayed. Instead of merely presenting a basic compilation of methods, we delve into the issues and difficulties of image fusion and semantic segmentation to offer a comprehensive overview of the application of optical and SAR image deep feature fusion in semantic segmentation. Our primary contributions are outlined as follows.

1) We synthesized the challenges and methodologies associated with semantic segmentation of remote sensing images, categorizing them based on different core architectures of neural networks.
2) We analyze the challenges and urgent needs of deep feature fusion in optical and SAR images, and meticulously examined the design of feature fusion strategies, presenting a comprehensive compilation of methods for fusing optical and SAR image features in semantic segmentation.
3) We compiled open-source datasets encompassing optical and SAR images for semantic segmentation and conducted a thorough analysis of various datasets.

In this comprehensive review, we present the latest advancements in semantic segmentation utilizing optical and SAR images based on an extensive literature survey. We condense a myriad of technical approaches, laying a robust foundation for future applications of multisource information in semantic segmentation. This review serves as a valuable resource for scholars and practitioners, offering a holistic perspective and identifying open challenges in semantic segmentation of remote sensing images.

The rest of this article is organized as follows. In Section II, we introduce the semantic segmentation method of remote sensing images. First, we introduce the difficulties of semantic segmentation, and comprehensively introduce the semantic segmentation method from four aspects: Convolutional neural network (CNN), fully convolutional network (FCN), recurrent neural network (RNN), and other model architectures. Section III introduces the concept of image fusion, mainly introduces the advantages and disadvantages of pixel-level fusion, feature-level fusion, and decision-level fusion, and discusses the reasons for using feature-level fusion for optical and SAR semantic segmentation. Section IV summarizes the semantic segmentation methods based on optical and SAR feature fusion and categorizes them according to the fusion strategy. In Section V, we collect existing publicly available optical and SAR image datasets for semantic segmentation. Section VI exemplifies commonly used semantic segmentation evaluation indicators. Section VII summarizes the current difficulties and development trends faced by optical and SAR images. Finally, Section VIII concludes this article.

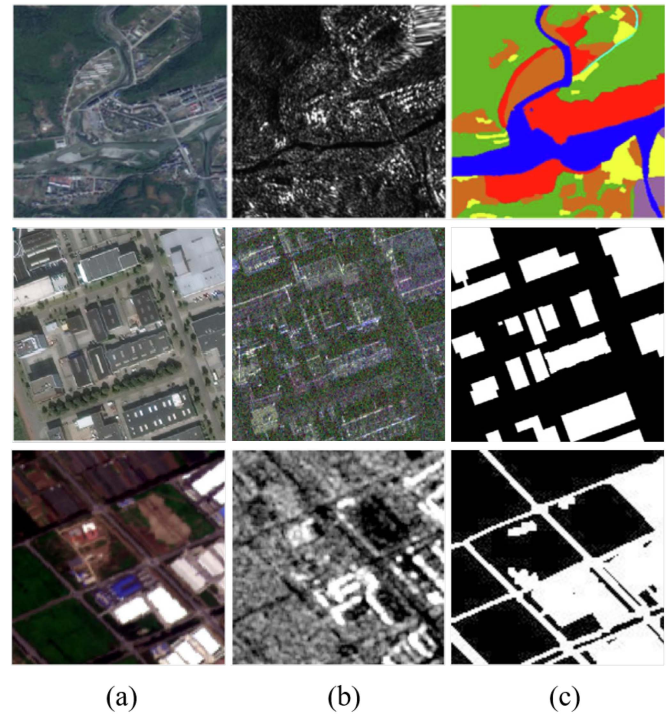

Fig. 5. Examples of image segmentation of optical and SAR images. (a) Optical image. (b) SAR image. (c) Ground truth. The first row is the segmentation example for the land use classification task. The second row is a segmentation example image of building extraction. The third row is an example of IS segmentation.

II. Remote Sensing Image Semantic Segmentation Methods

With the continuous improvement of the resolution of remote sensing images, the images contain a large amount of information with high spatial, spectral, and temporal resolution, retaining complex spatial texture details. As shown in Example 5, we exemplify three semantic segmentation example images, among which the first row is the segmentation example image for the land use classification task [61]. The second row is a segmentation example image of building extraction [62]. The third row is an example image of IS segmentation [13]. From left to right are optical images, SAR images, and segmentation ground truth examples. The features of remote sensing image objects undergo changes over time and in response to climate variations. Objects of the same type often exhibit heterogeneity across different scenes, leading to overlapping features and unclear object edges when viewed. In summary, the distribution characteristics of remote sensing images can be encapsulated in the following three key points.

1) *Complexity of object class types in remote sensing images:* Due to the elevated imaging height, expansive coverage area, and large image width, remote sensing images encompass a diverse array of object types.

2) *Variability in remote sensing image object categories sizes:* Similar objects within the same scene exhibit substantial scale variations.
3) *Heterogeneity of remote sensing image classes:* The imaging mechanism complexity results in distinct appearances of object classes at different moments, creating strong variability within the same class. In addition, intercategory similarity occurs, where different categories present indistinguishable appearances due to spectral similarity and other factors, such as various types of buildings or crops.

Based on the complex characteristics of remote sensing image scenes, the main difficulties faced in semantic segmentation of remote sensing images lie in the following points.

1) How to segment the same object at different scales in large scene images.
2) How to overcome the situation of foreign objects in the image having the same spectrum, and at the same time alleviate the confusion and blurred boundaries between different objects.
3) How to solve the segmentation difficulties caused by similarities between classes and differences within classes in remote sensing images.

Before the advent of deep learning, semantic segmentation primarily relied on manually designed features and traditional machine learning classification methods. Manually designed methods included texture, structural, spectral, and scattering features [63], [64], [65], [66], [67]. Machine learning methods involved random forests, support vector machines (SVMs), decision trees, and Bayesian classifiers [68], [69], [70], [71], [72]. The wavelet energy is a good feature to describe the texture of objects in images. Akbarizadeh et al.[65] proposed skewness wavelet energy utilizes the local statistical intensity information of each region. Therefore, it can solve the nonlinear intensity nonuniformity existing in SAR images. Akbarizadeh [66] proposed a new energy kurtosis wavelet energy as the texture discriminant feature of each region based on the segmentation of wavelet coefficient energy kurtosis values. The kurtosis value of the wavelet energy feature of the SAR image forms a feature vector, which can extract more texture statistical information and segment the SAR image. Tirandaz et al. [67] proposed a two-stage SAR image segmentation technology. In the first and second stages, a new parameter estimation algorithm based on curvelet coefficient energy to design the optimal kernel function and an unsupervised spectral regression method were proposed, respectively, for SAR image segmentation. Tirandaz et al. [71] proposed a polarization synthetic aperture radar (PolSAR) image segmentation method that does not require any parameter initialization based on lossy minimum description length, zero-fill weighted neighborhood filter bank, and hidden Markov random field expectation maximization. This method has efficient and good performance in the detection of different regions and boundaries. However, traditional methods required expert knowledge to design classification features based on task requirements, making the process time consuming, labor-intensive, and lacking in generalization and robustness. Consequently, any replacement of test data or external interference noise could impact classification performance.

In recent years, the continuous development of deep learning has significantly enhanced the performance of semantic segmentation. Deep CNNs prove effective in extracting image features, thereby improving the accuracy of semantic segmentation. In the current era of big data remote sensing, we are able to obtain a large number of remote sensing images. Traditional manual feature extraction methods fail to efficiently process such a large amount of remote sensing data. At the same time, the features of the same object in remote sensing images may also be different. Traditional methods are not only time consuming and labor-intensive in designing features, but also have weak generalization and robustness. Deep learning methods can fully and automatically extract features of a large number of images end-to-end, which not only saves manpower and material resources, but also achieves better robustness and generalization. This section provides a comprehensive overview of their application in the semantic segmentation of remote sensing images.

A. Convolutional Neural Network Methods for Remote Sensing Image Semantic Segmentation

The sliding window method is a common approach in CNN-based semantic segmentation, where the network classifies each image block, considering the classification result as the classification result for the pixel at the center of the block. Fig. 6 illustrates the basic sequence of CNN image classification. The final segmentation result is obtained by annotating the classification results of each pixel with the original image. Längkvist et al. [73] employed a CNN-based sliding window classification strategy for image patch segmentation, utilizing the classification results to enhance advanced segmentation results. Alshehhi et al. [74] introduced a CNN for classifying features in high-resolution remote sensing images using single image patches, demonstrating the effectiveness of CNN feature extraction in segmentation tasks. For the extraction of road and building data, Saito et al. [75] introduced a CNN-based model that automatically creates a feature extractor and classifier. This approach can extract numerous object features by training a single CNN. Yu et al. [76] achieved high-precision segmentation of hyperspectral remote sensing images by incorporating techniques such as data augmentation and adjusting the convolution kernel size. Paisitkriangkrai et al. [77] combined conditional random fields with a CNN classification framework to improve segmentation accuracy. Addressing the heterogeneity of remote sensing images, Feng et al. [78] introduced a CNN framework with two branches, incorporating a decision-based regionalization strategy to effectively differentiate between homogeneous and heterogeneous regions. Different architectures were developed to learn region-specific spectral features, improving segmentation accuracy. Kampffmeyer et al. [79] introduced a CNN for image segmentation, addressing the issue of feature class imbalance by measuring pixel-level uncertainty, thereby improving segmentation accuracy. Sharifzadeh et al. [80] proposed a neural network

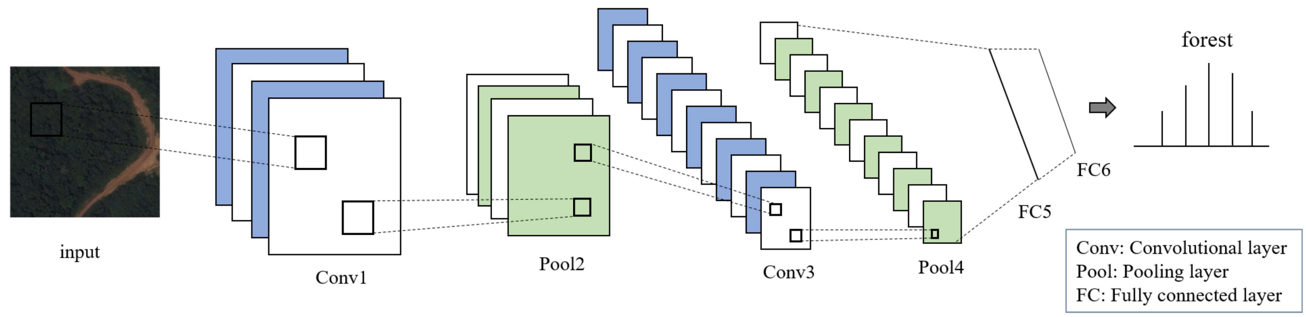

Fig. 6. Simple chart form of CNN architecture for image classification.

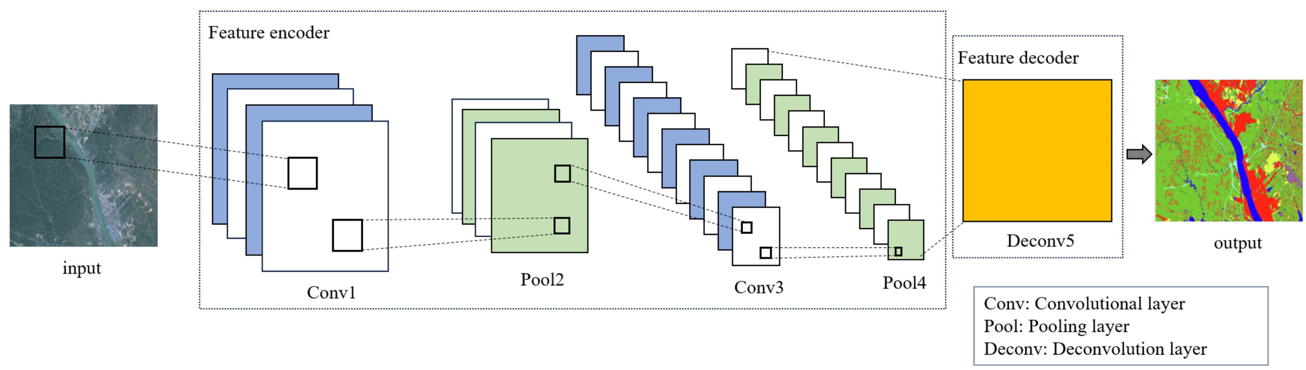

Fig. 7. Simple chart form of FCN architecture for image segmentation.

that combines CNN and multilayer perceptron (CNN-MLP) for SAR ship image pixel classification.

However, the sliding window method has limitations, as it may lead to the repeated use of some pixels and may lose spatial context association information between pixels. In addition, it cannot process images end-to-end, which can impact segmentation speed. Researchers are exploring alternative methods to overcome these limitations and enhance the efficiency of semantic segmentation in CNN-based approaches.

B. *Fully Convolutional Network Methods for Semantic Segmentation*

The FCN method [81] was proposed as a solution to address the challenges faced by traditional CNN in achieving end-to-end semantic segmentation of images, while also improving segmentation accuracy and speed. FCN extends the architecture of CNN into an encoder–decoder framework and replaces the last fully connected layer of CNN with a convolutional layer, enabling it to learn visual features in an end-to-end manner. The encoder mainly includes pooling and convolution modules, which continuously reduce the spatial size of feature maps to capture high-level semantic information. During the decoding process, an upsampling strategy is used to map the classification results to the size of the original image, generate pixel-level labels and obtain the segmentation results at the original pixel size of the input image. This design enables the network to comprehensively learn image features from beginning to end, thereby facilitating the task of assigning semantic labels to every pixel in the image. The key to the network architecture is to strategically build the encoder and decoder structures to maximize the receptive field, learn multiscale features, and prevent the reduction of feature map resolution during upsampling. Fig. 7 illustrates the basic sequence of FCN image segmentation.

In the context of remote sensing image segmentation, FCN plays a crucial role in extracting spatial context information between pixels, learning multiscale features, improving segmentation accuracy, and refining segmentation boundaries. Zhao and Du [82] proposed a multiscale convolutional neural network using a pyramid structure to extract deep spatial features from remote sensing images. This method leverages the multiscale spatial spectral information to obtain different levels of contextual information, enhancing the accuracy of classification. Zheng et al. [83] verified the use of convolution and pooling in the FCN encoding stage in remote sensing image segmentation, as well as the role of deconvolution and upsampling in the decoding stage. They used three variants of FCN (FCN-32s, FCN-16s, and FCN-8s) for semantic segmentation on unmanned aerial vehicle (UAV)-borne remote sensing images, demonstrating the significant efficiency improvement brought by FCN. Henry et al. [84] evaluated the effectiveness of three segmentation network structures—FCN, UNet [85], and DeepLabv3+ [86] in road segmentation of SAR images. This study provides a reliable method for road feature extraction in SAR images and demonstrates that modifying the regression loss function can improve category balance and segmentation accuracy.

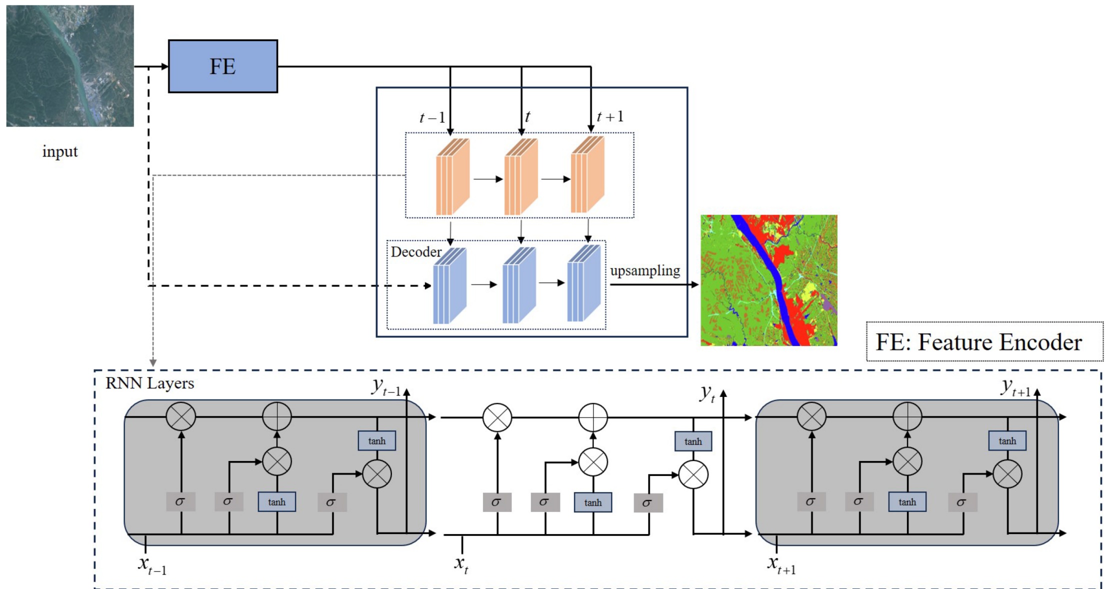

Fig. 8. Flowchart of RNN-based remote sensing image semantic segmentation method.

Niu et al. [87] utilized the capabilities of DeepLab to extract complex remote sensing features at different levels, avoiding spatial resolution decrease during the decoding process. Li et al. [88] extended the functionality of UNet by integrating specific convolutional layers, linking the ReLU layer module of downblock in UNet and the convolutional layers of the sampling layer module of upblock in the expansion path. This modification resulted in improved segmentation accuracy for remote sensing images. Marmanis et al. [89] enhanced the clarity of class boundaries by combining semantic segmentation with edge detection. The integration of boundary detection into SegNet [90] architecture and FCN-type models significantly improved the segmentation accuracy of remote sensing images. Diakogiannis et al. [91] proposed ResUnet network to segment high-resolution remote sensing images. The network can effectively learn multiscale contextual features by combining residual blocks, astrous convolution, and pyramid pooling modules, and combined with the improved loss function can solve the problem of category imbalance in remote sensing images. Li et al. [92] proposed an attention mechanism network MANet for semantic segmentation of remote sensing images. The network uses ResNet as a feature extractor and embeds the dot product attention mechanism into the network model. Finally, upsampling is used to obtain more refined segmentation results. Ghara et al.[93] used UNet and DeeplabV3 neural networks to segment SAR images with the smallest possible number of images and the highest accuracy respectively. Aghaei et al. [94] proposed an end-to-end SAR image segmentation network. They introduce ShuffleNet network block in SAR image segmentation, which is able to effectively suppress the phenomena of speckle noise, heterogeneous background, and edge blur in SAR image segmentation.

These studies collectively highlight the versatility and effectiveness of FCN in remote sensing image segmentation, showcasing their capability to enhance accuracy and handle complex spatial features. These methods have proven that the semantic segmentation network can achieve better accuracy on SAR images and optical images, respectively, but it is still very difficult to face situations such as the same object in optics having different color features. At the same time, when segmenting SAR images, problems such as coherent speckle noise and edge blur still cannot be well solved. It can be seen that the mechanical defects of single source data are difficult to solve through the design of network structure. Multisource data fusion is able to make up for the regrets of single-source data, so it is crucial to make full use of the information from multisource data to solve the difficulties faced by single-source data.

C. *RNNs for Remote Sensing Image Semantic Segmentation*

Since the RNN can process the temporal information of the image, this architecture is used to process temporal remote sensing image segmentation. However, in remote sensing image processing, the RNN model in natural optical images is improved so that it can learn the spatial, temporal, and spectral information of remote sensing images. Fig. 8 shows the semantic segmentation method of remote sensing images based on RNN.

Wu and Prasad [95] used a convolutional RNN to obtain the combined features of hyperspectral data, and uses the recurrent layer to extract spectral context information from the multiscale features generated by the convolutional layer. Ienco et al. [96] evaluated the effectiveness of RNN and LSTM [97] models in remote sensing multitemporal image feature classification, and experimentally demonstrated that RNN and LSTM networks are more effective in both pixel-based and object-based classification methods. Campos-Taberner et al. [98] evaluated the effectiveness of deep recurrent network-based methods for classifying time series features in Sentinel-2 data. The bidirectional LSTM network achieved an overall accuracy of 98.7%, exceeding the performance of all tested classification techniques. Sun et al. [99] built a hybrid LSTM-RNN model to improve accuracy and reduce model complexity by leveraging the multitemporal properties of features in image time series. Lin et al. [100] introduced LSTM-MTL, a multitask spatiotemporal deep learning model, for large-scale rice mapping by leveraging time series Sentinel-1 SAR data. The model is able to learn multiple timing consistency features and region-specific features simultaneously, significantly improving the performance of SAR feature classification. Chen et al. [101] designed a novel interpolation BiLSTM model, Im-BiLSTM, to effectively solve the interaction between interpolation and classification tasks and minimize the errors and uncertainties caused by the separation process between interpolation and classification.

The RNN network is able to process time series data well and use multiscale time information to improve the accuracy of remote sensing image segmentation. The core module of RNN can make full use of contextual information, but it is still difficult to solve the stubborn problems caused by single-source data.

D. Other Models for Remote Sensing Image Semantic Segmentation

In recent years, the transformer architecture, originally proposed for natural language processing tasks, has gained significant attention and demonstrated remarkable performance across various domains. By leveraging the self-attention mechanism, transformers can effectively capture long-range dependencies within remote sensing images, enabling them to grasp intricate spatial relationships and context information crucial for accurate semantic segmentation. The forefront methods of transformer-based semantic segmentation now relies on vision transformer [102] or Swin transformer [103] for feature extraction as the backbone. Zhang et al. [104] proposed an encoder-decoder structure by using Swin-transformer to extract features and CNN-based models strategies to upsample features. Dong et al. [105] introduced a cross-model knowledge distillation framework designed to boost the segmentation performance of both CNN-based and transformer-based segmenters by incorporating and distilling their complementary strengths. Liu et al. [106] analyzed some limitations of existing transformer-based semantic segmentation methods for remote sensing images and proposed a global local transformer framework to obtain consistent feature representation by using transformers for both encoding and decoding.

However, the transformer model requires high computational complexity, especially when faced with the task of semantic segmentation of high-resolution remote sensing images. The high complexity brings about the problem of large memory requirements, causing certain difficulties. Gu and Dao [107] built upon the state space model [108] is a network model that establishes long-distance dependencies while maintaining linear computational complexity. Liu et al. [109] and Zhu et al. [110] used Mamba to process computer vision tasks and achieve superior results, proving the potential of the Mamba architecture for image processing. He et al. [111] introduced Mamba into the pan-sharpening task of remote sensing images and implemented it by designing a channel swapping Mamba module and a cross-modal Mamba module. Chen et al. [112] designed the RSMamba network based on the Mamba model, which combines the advantages of global receptive field and linear modeling complexity, and is used in remote sensing image scene classification. Ma et al. [113] combined Mamba with a dual-branch structure for the task of semantic segmentation of remote sensing images. This was the first attempt of Mamba in the semantic segmentation task of remote sensing images, proving the effectiveness and potential of Mamba for semantic segmentation of remote sensing images.

In the realm of machine learning, large models represent a paradigm shift towards handling vast amounts of data and intricate tasks. Characterized by their extensive parameter counts and complex architectures, these models have become key to advancing the field. Initially rooted in the foundations of deep neural networks, large models have evolved over the years to accommodate the growing demands of diverse applications. The design purpose of large models is to improve the expressive ability and predictive performance of the model, and to be able to handle more complex tasks and data. The research on large models has become the current development trend of image processing. However, there are still relatively few applications in semantic segmentation of remote sensing images. Wang et al. [114] first proposed a large-scale basic vision model suitable for remote sensing image processing tasks, including remote sensing object detection, remote sensing scene classification, remote sensing semantic segmentation. Hong et al. [115] introduce SpectralGPT, a universal RS foundation model tailored for processing spectral RS images, utilizing a novel 3-D generative pretrained transformer (GPT). SpectralGPT has demonstrated superior performance in downstream tasks such as change detection and semantic segmentation.

The development of large models has also brought convenience to semantic segmentation of remote sensing images. However, the disadvantages of large models are that they consume huge amounts of memory and calculations, and require a huge amount of data. At the same time, the inherent shortcomings of single source data cannot be well compensated. Multisource data fusion is still a development trend. Effective use of multisource data information is of guiding significance in solving the inherent defects of single-source data and effectively improves accuracy.

III. Image Fusion Strategy

Although semantic segmentation methods develop rapidly, most of them are based on single-source remote sensing images. It is difficult to distinguish areas with similar characteristics using only single-source remote sensing image data. With the advent of the era of multisource big data, multisource remote sensing images can provide richer and more complete information. Image fusion strategies can effectively utilize the information of multisource remote sensing images to improve the accuracy of semantic segmentation. According to the different fusion levels, image fusion strategies can be divided into three types: 1) pixel-level fusion; 2) decision-level fusion; and 3) feature-level

TABLE I
IMAGE FUSION STRATEGIES, THEIR STRENGTHS AND WEAKNESS AND APPLICATION EXAMPLES

Image fusion strategy	Strengths	Weakness	Application examples
Pixel-level fusion	Fusing images prior to network input minimizes information loss and effectively preserves complex details in the image scene.	The input images are required to be subjected to overly strict image registration and preprocessing, along with the presence of significant memory consumption.	[58], [116], [117], [125]
Decision-level fusion	For the fusion of different modal output results, the requirement for input image registration is low, with real-time and fault-tolerant characteristics.	Very poor information interaction ability between modalities, and large information loss.	[118]–[124]
Feature-level fusion	Fusion of features extracted from different modalities reduces the data dimensions, realizes information compression, improves robustness and interpretability.	There will be a partial loss of information, and the input image is required to have a high accuracy of registration.	The relevant details of the summary are shown in Table II

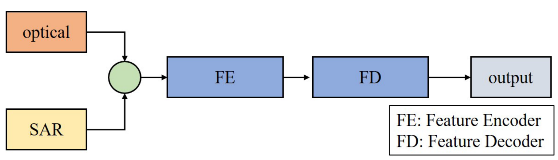

Fig. 9. Structure of pixel-level fusion method.

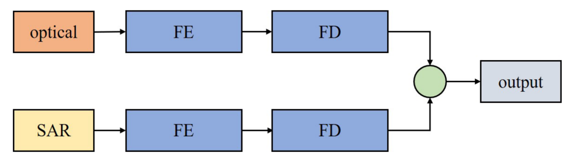

Fig. 10. Structure of decision-level fusion method.

fusion. We summarize the three image fusion methods as shown in Table I.

A. Pixel-level Fusion Strategy

The pixel-level fusion structure diagram is shown in Fig. 9. The original input pixels are directly processed and analyzed, which requires the input multisource images to undergo fine registration. This method has little loss of information and retains the rich detailed information in the image scene to a great extent. However, it processes a large amount of data and has high requirements on memory and equipment. For tasks such as ground object classification and fire monitoring, image preprocessing such as image augmentation and noise reduction must be performed before real-time processing. Kussul et al. [116] performed classified crop classification using spectral and spatial features of Landsat-8 and Sentinel-1A images by fusing them at pixel level as inputs to 1D-CNN and 2D-CNN models, respectively. Lin et al. [117] presents a method to classify optical, SAR, and LIDAR data using a sparse representation dictionary to improve classification accuracy by integrating SAR data and airborne LIDAR data.

Pixel-level fusion methods perform fusion on the input side and cannot overcome the semantic differences between multisource images. Therefore, the pixel-level fusion method cannot effectively extract the complementary and consistent information of multisource images, and it is easy to obtain image features that are irrelevant to the ground object classification task, resulting in a decrease in classification performance.

B. Decision-level Fusion Strategy

The decision-level fusion method represents the highest-level fusion strategy, fusing single-modal segmentation results through different decision-making criteria, as depicted in Fig. 10. This approach demands minimal accuracy in image registration, exhibits commendable real-time performance, and boasts robust fault tolerance and openness. Paisitkriangkrai et al. [118] combined CNN and Random Forest results by multiplying their posterior probabilities. Chen et al. [119] applied Bayes rule, while Kahou et al. [120] integrated outcomes from various unimodalities identified by CNN. In a different vein, Audebert et al. [121] utilized a residual network to derive coefficients for rectifying the average fusion outcomes for multimodal data. Maggiolo et al. [122] introduced a Bayesian decision fusion strategy for classifying optical and SAR images, incorporating a swift version of the iterative conditional mode Markov optimization algorithm, relying on convolutional operations. This method demonstrating the scalability and effectiveness in handling large-scale applications. Moreover, Waske and Benediktsson [123] categorized each data source using SVMs, utilized the initial output of each SVM discriminant function in the subsequent fusion procedure, and then employed a new SVM to fuse the segmentation results. In addition, Vohra and Tiwari [124] introduced a technique enhancing the accuracy of feature classification by integrating classifier decisions with auxiliary information obtained from spectral and spatial data.

However, a drawback of decision-level fusion is its exclusive focus on output results rather than features, limiting the exchange of radiation information between single-source data and resulting in a significant loss of information. Furthermore, it impedes the synergy and consistency between optical and SAR images.

C. Feature-level Fusion Strategy

The feature-level fusion technique involves extracting feature information from the original image. Multisource images of

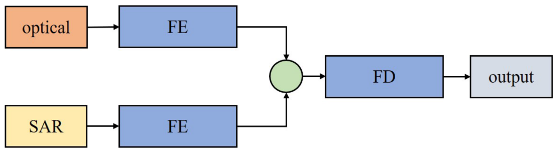

Fig. 11. Structure of feature-level fusion method.

the same scene are then subjected to feature extraction, association, and fusion processing using various feature association methods, as illustrated in Fig. 11. This strategy is able to reduce data dimensions, compress information, as well as analyze and fuse multimodal data features from different perspectives. In addition, it can convert multidimensional features into 1-D representations. Feature-level fusion method enhances the resilience and stability of the network structure, improving model interpretability by fusing different source features based on task requirements. In the context of feature classification tasks, diverse fusion strategies should be designed for different scenarios. The primary objective is to minimize semantic differences between optical and SAR images, extract complementary and consistent features from both image types, and ultimately enhance classification performance. The feature fusion methods of optical and SAR images are introduced in detail in Section IV.

IV. Optical and SAR Image Fusion for Remote Sensing Image Segmentation

A. Difficulties in Optical and SAR Image Feature Fusion

Within the realm of remote sensing, multisource images refer to images of the same scene or object captured by various sensors. Optical images are collected by passive sensors, which mainly rely on solar radiation reflected by ground objects to obtain image information. As a result, they exhibit rich textures, colors, and other radiant properties. However, optical sensors are susceptible to climate and lighting constraints when acquiring images. SAR is an active sensor whose imaging band is microwave. SAR image information is formed by backscattering from ground objects and is sensitive to building facades. SAR images offer extensive structural details and can be obtained at any time of the day. SAR images therefore produce clearer contour information than optical images. The purpose of optical and SAR feature fusion is to use complementary information in optical and SAR images to make up for the respective deficiencies of optical and SAR images in image expression. For example, there are problems such as the situation of same objects with different spectra and different objects with the same spectrum in optics, the lack of detailed texture information in SAR images, and poor image quality. And this complementarity has been confirmed in the literature [126], [127], [128], [129], [130]. At the same time, the structural consistency information of optical and SAR images is used to effectively correlate the optical and SAR images, which can fully improve the segmentation accuracy. The studies [18], [62], [131] have confirmed the robust structural coherence present in both optical and SAR images.

The purpose of optical and SAR feature fusion is to combine complementary and structurally consistent information in optical and SAR images, thereby improving segmentation accuracy. When training the network to extract fusion features, the heterogeneous nonlinear radiation information between optical images and SAR images destroys the relevant complementary information. This ultimately greatly reduces the effectiveness of the fusion features and decreases classification accuracy. At the same time, many studies focus too much on the learning of complementary information and ignore the role of structural consistency, leading to the loss or discarding of associated information and degradation of classification performance. Therefore, the key to fusion of optical and SAR image features lies in how to learn complementary information and structural consistency information that is beneficial to the classification task without destroying the integrity of the semantic information of each modality. Another problem that has come to the fore in the field of multisource image feature categorization is the lack of large publicly available optical SAR datasets. The fusion methods of optical and SAR image features based on feature classification face the following three main challenges.

1) How to effectively learn the complementary information and consistency information of multisource images.
2) How to eliminate the impact of feature shifts caused by appearance differences in multisource images.
3) Lack of publicly available datasets.
   Based on these problems, we summarized relevant solutions in a large number of literature, and classified them into the following four design methods according to the module design perspective:

a) linear fusion module;
b) attention mechanism fusion module;
c) gated fusion module;
d) feature alignment module.

At the same time, we summarize the advantages and disadvantages of these methods, as shown in Table II.

B. Optical and SAR Fusion Strategies for Remote Sensing Image Semantic Segmentation

To address the aforementioned challenges pertaining to the fusion of optical and SAR images, numerous approaches have been suggested to resolve these issues. Prior to using various fusion algorithms, the two-branch network architecture is utilized to independently extract optical and SAR image features. This approach helps transfer the issue of feature extraction bias caused by the disparate radiometric information of the images prior to fusion. Various fusion algorithms are developed based on the distinctive characteristics of optical and SAR images, as well as the specific task requirements. The objective is to get valuable and coherent features that effectively address the challenge of image fusion.

1) Linear feature fusion strategy: The linear fusion model is a widely used approach for combining features from several sources. It employs linear fusion algorithms such as feature concatenating, as shown in Fig. 12, feature summation, as shown

TABLE II
FEATURE FUSION STRATEGIES, THEIR KEY TECHNOLOGIES, CHARACTERISTICS, AND APPLICATION EXAMPLES

Feature fusion strategy	Key technologies	Characteristics	Weakness	Application examples
Linear fusion strategy	Feature summation Feature concatenation Feature dot product	The features from different modalities can be directly fused, and this method is simple and straightforward to operate.	It cannot realize the multi-source information interaction and has more limitations.	[12], [129], [132]–[136]
Attention-based fusion strategy	Channel attention Spatial attention Self-attention	Allocating different attentional resources to different modal data highlights the focus on regional features that can learn complementary information across multiple modalities.	It cannot screen redundant information in real time and is easily affected by differences in optical and SAR appearance.	[61], [62], [130], [137]–[140]
Gate-based fusion strategy	Independent gate Complementary gate Interactive gate	Features from different modalities can be filtered and weighted to learn complementary and consistent features.	It is easy to overlook the connection between categories and is susceptible to differences in the appearance of optical and SAR images.	[9], [138], [141]–[144]
Feature alignment fusion strategy	Spatial alignment methods	Eliminate appearance differences by aligning feature distribution.	Features cannot be filtered and are easily affected by image heterogeneity.	[62], [145]–[148]

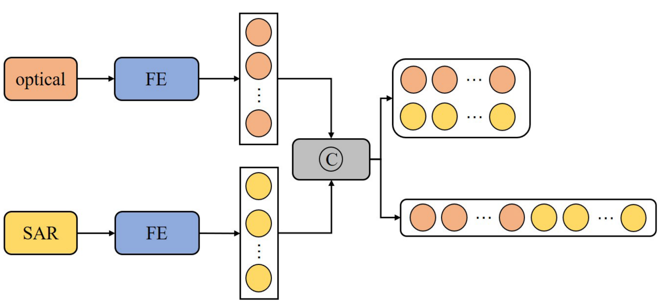

Fig. 12. Structure of feature concatenating method.

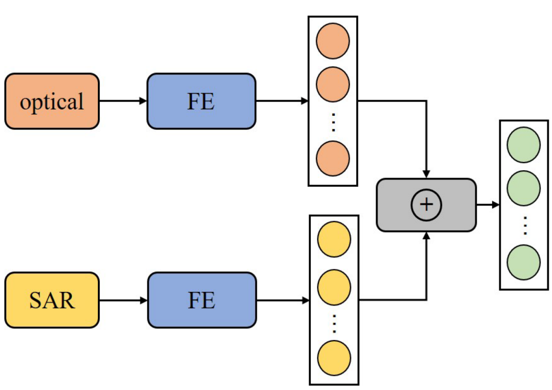

Fig. 13. Structure of feature summation method.

in Fig. 13, and feature dot product shown in Fig. 14 to merge the distinct data features.

Xu et al. [132] achieved feature fusion by linking radiation and structural features obtained from a two-branch CNN, they proved that the dual-branch network structure can learn multisource remote sensing image features better than the single-branch structure. Hughes et al. [133] adopted a two-branch pseudo-siamese CNN to independently learn the characteristics of optical and SAR images, and then linear fusion methods are used to concate the obtained optical features and SAR features. Zhang et al. [129] introduced a block-based deep convolutional network to extract features from optical and polarization SAR images as input to the model. Guo et al. [134] adopted the fusion method of linear layered superposition to fuse the features of SAR and optical images, and inputted the fused features into the decision tree for feature extraction. Several other scholars have also contributed to feature extraction using optical and SAR image fusion, following the linear fusion methods [135], [136].

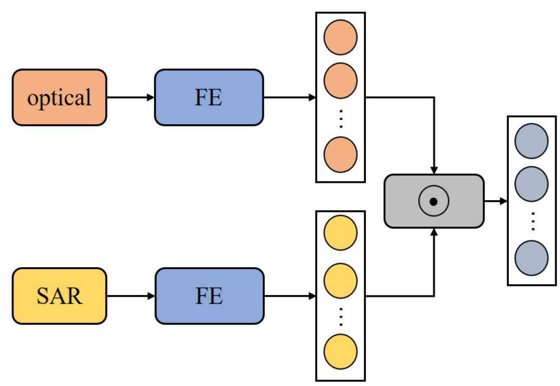

Fig. 14. Structure of feature dot product method.

Although linear fusion is simple and convenient to operate, its ability to fully integrate the interaction of multisource information and the importance of different modes in classification is limited. This limitation stems from the fact that the method fuses data from different modalities using the same weights. Therefore, linear fusion methods have inherent limitations in practical applications.

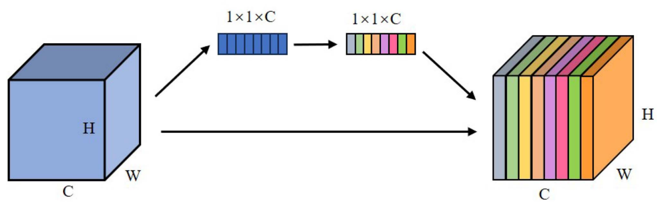

Fig. 15. Simple form of channel attention structure.

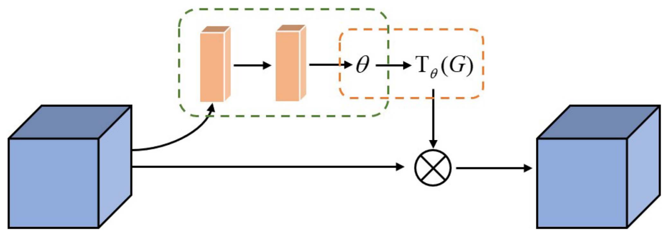

Fig. 16. Simple form of spatial attention structure.

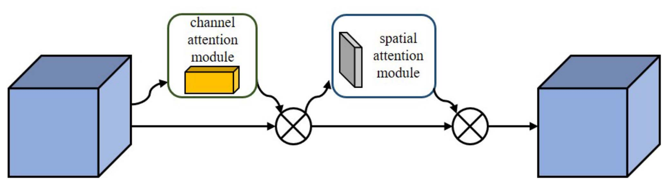

Fig. 17. Simple form of mixed attention structure.

2) Attention-based feature fusion strategy: The attention mechanism in computer vision mimics the visual attention mechanism of the human brain. It comprehensively scans the input image and allocates different attention resources to different areas, and selects key areas to collect more detailed information while suppressing attention to nonkey areas. The attention mechanism in multisource feature learning mainly includes channel attention mechanism, spatial attention mechanism, and self-attention mechanism.

The channel attention mechanism achieves fusion by calculating the importance of each channel, as shown in Fig. 15. SENet [149] selectively enhances important features by explicitly modeling relationships between channels, dynamically adjusting channel feature responses, and collecting overall statistics for each channel through average pooling.

The channel attention mechanism aims to quantify the importance level of each channel, while the spatial attention mechanism allows the model to dynamically learn the attention weights of various regions by incorporating attention modules, as shown in Fig. 16. In this approach, the model can focus more on key image areas and ignore uninteresting parts. The convolutional block attention module (CBAM) [150] uses a combination of channel attention and spatial attention to improve the attention capability of the convolutional neural network, as shown in Fig. 17. It captures the overall statistics of each channel by utilizing global average pooling and global max pooling, and then learns channel weights through two fully connected layers.

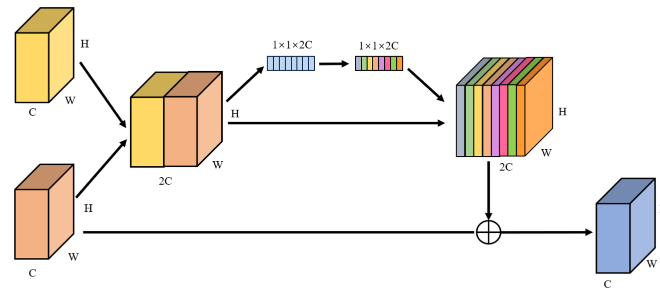

Fig. 18. Structural diagram of DDHRNet fusion module MSE.

The two results generated by the processing are then added together and adjusted for each channel by using the sigmoid function to normalize the weights to a range of 0 to 1. Finally, the scaled channel features are ultimately multiplied with the original features to enhance the importance of each channel within the features.

Multisource fusion approaches utilize the attention mechanism to assign weights to the attention region, allowing the network model to acquire additional information from several data sources. Fu et al. [137] concurrently incorporated spatial and channel attention, merging the results of both attention methods to augment feature representation. This method proves that the dual attention module can effectively capture long-range contextual information and give more accurate segmentation results. Ren et al. [151] proposed a dual-stream deep high-resolution network (DDHRNet) to deeply fuse SAR and optical data at the feature level of each branch. The network designs a multimodal extrusion and excitation (MSE) module. As shown in Fig. 18, MSE uses the channel attention mechanism to integrate model features and can effectively utilize complementary information in heterogeneous images. Li et al. [138] introduced CHGFNet, a hierarchical fusion network that combines optical and SAR information for land cover classification using a collaborative attention-based heterogeneous gated fusion method. This method automatically realizes the weighted fusion of optical and SAR features, proving that the collaborative attention mechanism is able to effectively learn the complementary features of optical and SAR images. Yang et al. [139] introduced AFNet, a hybrid attention fusion network, the block diagram of the network is shown in Fig. 19. This method utilizes a channel attention module to calculate feature weights along the channel dimension, and a spatial attention module to calculate feature weights along the spatial dimension. It effectively fuse different types of features and allow the network to learn effective information in different types of data to improve the segmentation accuracy of high-resolution remote sensing images. Li et al. [130] proposed a multimodal bilinear fusion network (MBFNet) to learn complementary features of optical and SAR images to enhance feature classification. This method proposes the second-order attention-based channel selection module (SACSM) module for feature fusion, and the SACSM structure is shown in the Fig. 20. The SACAM attention

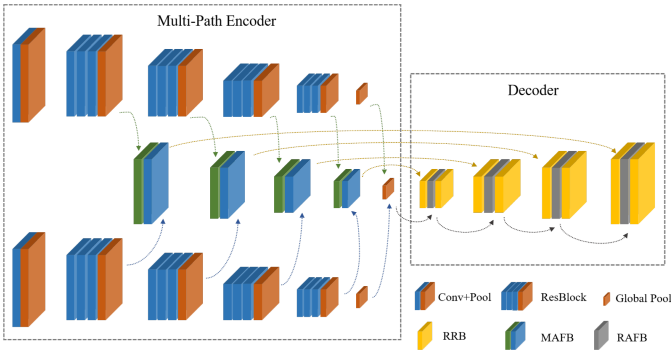

Fig. 19. Main encoder–decoder structure block diagram in AFNet.

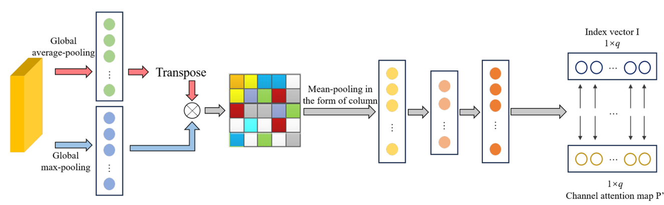

Fig. 20. Structural diagram of MBFNet fusion module SACSM.

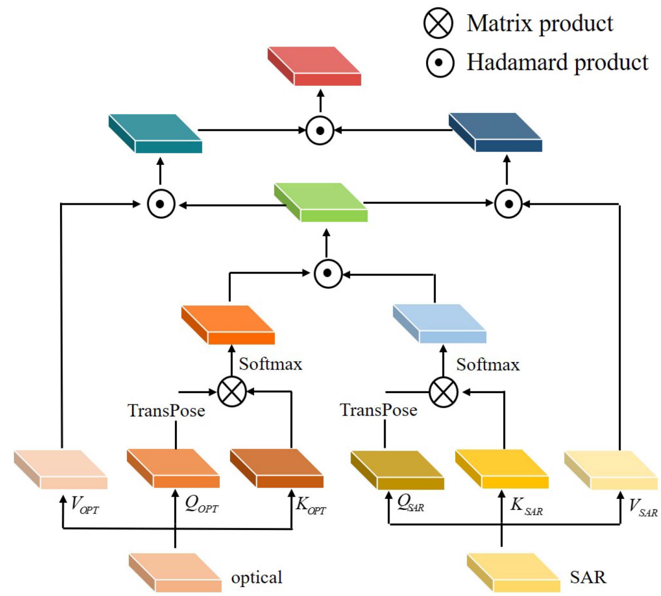

Fig. 21. Simple form of self-attention structure in MCANet.

module in this network model can effectively utilize the relationship between channels of the feature map, automatically emphasize important channels in the feature map, and reconfigure the compact feature map, thereby enhancing the representation of the network and improving its discriminative performance. Li et al. [62] utilized knowledge of phase coherence to guide the acquisition of structural coherence properties from optical and SAR images. They designed a multistage progressive feature fusion framework that utilizes structural consistency information to correlate optical and SAR image features, and learn complementary features of optical and SAR images through a channel attention mechanism. This method proves that consistency information can effectively correlate optical and SAR image information, improving the accuracy and robustness of the model. Liu et al. [152] designed a multimodal dual-attention fusion module, which enables the network to more reasonably fuse the heterogeneous features of optical and SAR images, thereby enhancing the robustness of the model.

The self-attention mechanism seeks to form connections between input vectors, enabling the network to learn associations between distinct components of the input. The crucial aspect of self-attention is that the internal parameters are all structurally similar and finally acquire knowledge about the fundamental connections between different locations inside the image. Zhou et al. [140] introduced a three-branch self-attention module that consists of two input branches with different characteristics and a cross-modal distillation branch. This module utilizes self-attention blocks to merge features from different modalities. Li et al. [61] introduced a multimodal cross-attention network called MCANet. This network incorporates a multimodal cross-attention module that utilizes a self-attention mechanism to capture positional relationships among feature maps from different data sources, as shown in Fig. 21. This module enables effective interaction between optical and SAR image features in a 2-D space.

In general, the attention mechanism can assign higher weights to the features that one wants to focus on, and by designing the corresponding attention module, complementary features of optical and SAR images can be learnt in image fusion. However, the attention mechanism method cannot filter redundant information in real time and is easily affected by the difference in appearance of optical and SAR images.

3) Gated-based feature fusion strategy: Gating mechanisms are frequently employed in recurrent neural networks to address challenges related to multitemporal features. These mechanisms possess a strong capability to selectively filter features, enabling them to regulate the transmission of valuable information from many modalities. Gated fusion mechanisms have the ability to selectively combine information from different levels. They are commonly employed for the purpose of screening and fusing features from several sources, with the aim of enhancing the performance of classification, this approach has been supported by studies [153], [154], [155], [156]. The gating fusion methods can be classified into three types: 1) independent gates shown in Fig. 22; 2) complementary gates shown in Fig. 23; and 3) interactive gates shown in Fig. 24 [9]. The independent gating mechanism is a network with two branches that extract features individually. Each modal image is screened independently for relevant characteristics, and then the screened features are fused together. However, this fusion approach lacks information sharing, making it challenging to acquire meaningful complementary and coherent features. The complementary gate integrates the optical and SAR images by initially fusing them at the input. It then selects complementary characteristics from both modalities for fusion by configuring the gate function. However, the limitation of using only a single gate module hinders the

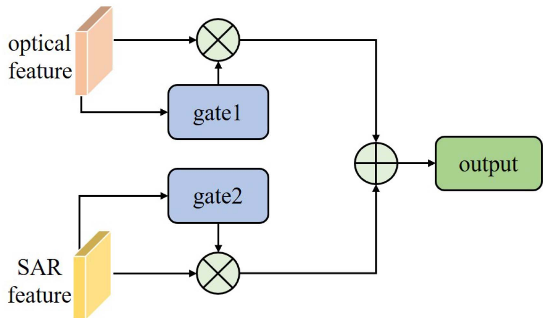

Fig. 22. Structure of independent gating method.

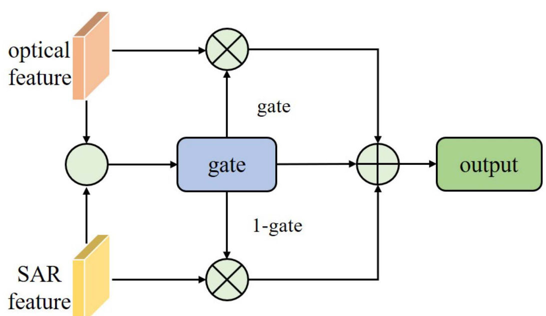

Fig. 23. Structure of complementary gating method.

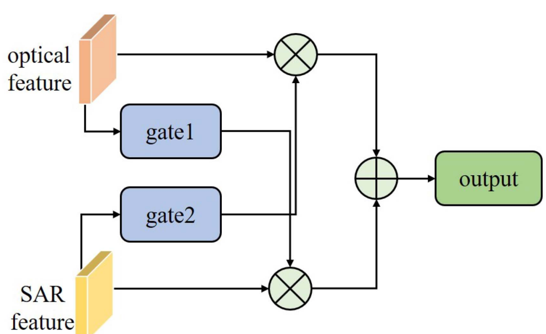

Fig. 24. Structure of interactive gating method.

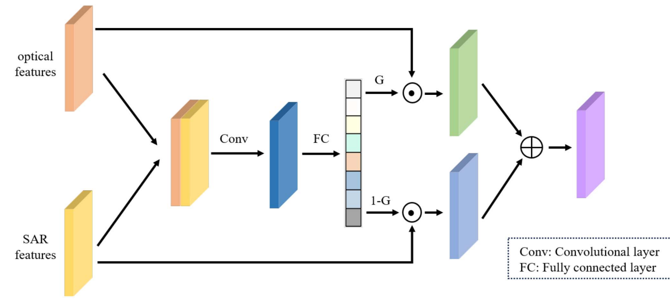

Fig. 25. Structural diagram of CHGFNet fusion module GHFM.

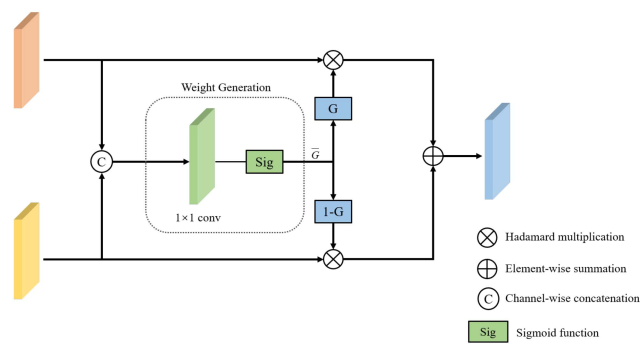

Fig. 26. Structural diagram of CMGFNet fusion module GFM.

utilization of shared features in a logical manner. The interaction gate establishes two gate modules that acquire features from distinct modalities and mutually regulate the selection of features from the other modality.

Zhang et al. [157] introduced a novel adaptive collaborative attention network that employs a gated multimodal fusion module to combine textual and visual data. This approach demonstrates the ability of the gating mechanism to adaptively adjust the amount of multimodal information to be considered at each time step. Cheng et al. [153] introduced a locally sensitive DecovNet that incorporates a novel gated fusion layer to effectively merge spectral and depth information. Wang et al. [154] introduced three novel dynamic fusion techniques to enhance multimodal word representations. These techniques involve utilizing modality-specific gates, category-specific gates, and sample-specific gates to determine distinct weights for each input modal representation.

Image feature fusion involves the adaptive learning of discriminative features by including a gated fusion module to assign weights to each modality and eliminate irrelevant components. Nevertheless, the task of effectively combining complementary information for the purpose of resolving scenes in multimodal remote sensing images continues to be a challenge. Reducing the resolution of the feature map using neural networks might cause a loss of spatial information, perhaps leading to blurred object borders and incorrect classification of small objects. Furthermore, the sizes of objects in remote sensing images exhibit significant variability, resulting in a decline in categorization accuracy.

Li et al. [138] introduced CHGFNet, which incorporates a gated heterogeneous fusion module (GHFM) in the form of complementary gates, as shown in Fig. 25. This module enables adaptive weighted fusion of optical and SAR information, resulting in enhanced segmentation accuracy. Hosseinpour et al. [141] introduced the gated fusion module (GFM) as a method to learn cross-modal features using complementary gates. The GFM then combines high-level semantic features with underlying features using a multilevel feature fusion strategy. This method is able to combine information from different modalities based on the functional quality of each region of interest. Zhou et al. [142] introduced the CEGFNet, an end-to-end network that combines conventional extraction and gate fusion techniques to effectively collect both high-level semantic characteristics and low-level spatial data for scene parsing of remote sensing images. This network proposes a complementary gating fusion module GFM, the structure of GFM is shown in Fig. 26. The gating module can eliminate redundant features in the data by setting gating thresholds and extract complementary features from the data

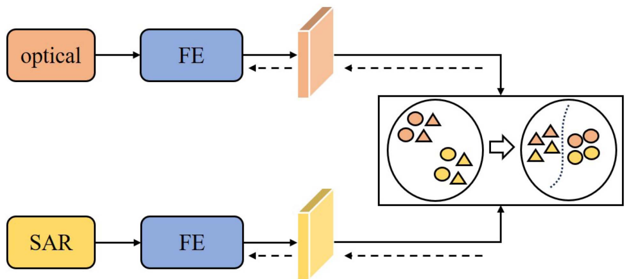

Fig. 27. Simple form of feature space alignment structure.

to enhance multimodal feature fusion. Furthermore, the global context module and the multilayer aggregation decoder, respectively, address the issues of scale differences between objects and spatial feature loss caused by downsampling. Kang et al. [9] introduced a cross-modal interaction gating method that enables the extraction of optical and SAR image features by utilizing a two-branch network topology with bidirectional information flow. Simultaneously, the network establishes a cross-modal transmission gate to enable independent feature learning while facilitating bidirectional information flow between optical and SAR images. This allows for comprehensive learning of complementary information and structural consistency, resulting in enhanced segmentation accuracy. Huang et al. [143] introduced GRRNet, a gated residual refinement network that integrates raw data from several modalities into numerous channels. The encoder unit of GRRNet incorporates an enhanced residual network, and employs gated feature labeling to enhance the accuracy of segmentation outcomes. Geng et al. [144] utilized the complementary gate module to fully explore the interaction patterns between heterogeneous features, effectively integrating information from different sources. They also emphasize the significance of high-level information interaction across data sources.

The gating mechanism can effectively screen the redundant information in optical and SAR images. By designing the tricks of different gating modules, the complementary and consistent features of optical and SAR images can be effectively learned. However, the connection between the specific categories of optical and SAR images is still easily overlooked, and the feature learning process is still easily affected by the appearance differences between optical and SAR images.

4) Optical and SAR image feature space alignment methods: Due to the substantial appearance differences, multisource data may experience semantic misalignment during feature fusion. This causes the network to excessively concentrate on the appearance differences, which in turn leads to the loss and corruption of semantic information. Spatial alignment eliminates the noticeable variations in the data by aligning the feature distributions of distinct modalities, the structure is shown in Fig. 27.

Hong et al. [145] employed mathematical models, including CoSpace [158] and L1 CoSpace [159], to integrate multimodal data aspects. Zhang et al. [146] utilized the image transfer technique in the preprocessing stage of image alignment to enhance the issue of inter-image disparity. They then aligned the images using the scale-invariant feature transform algorithm [160], which was employed for spatial alignment between multimodal image pairings. Quan et al. [147] introduced a two-channel convolutional neural network with distinct parameters to extract potential correlation features from multimodal image pairings. The objective was to determine if these pairs are a match or not, in order to accomplish spatial alignment between optical and SAR image pairs. Li et al. [62] proposed a progressive fusion learning framework that utilizes modality-invariant features to correlate optical and SAR image features to address the impact of appearance variations in building extraction. Li et al. [148] introduced a spatially-aware circular module to generate cross-modal receptive fields. In addition, they utilized a feature transformation technique to map the optical and SAR high-level features extracted by the network into a shared latent space, thereby mitigating the impact of modal appearance disparities. The approach enhances segmentation accuracy by aligning the semantic distribution of complimentary information from each modality.

The feature alignment method is the most effective method to solve the difficulty in feature extraction caused by the difference in appearance of optical and SAR images. However, redundant features cannot be filtered and are easily affected by image heterogeneity, resulting in the inability to effectively obtain complementary features of optical and SAR images.

In summary, each module has its corresponding advantages and disadvantages. In the semantic segmentation task, the characteristics of the features required for downstream tasks should be analyzed, and different feature modules can be mixed and used to achieve the purpose of effective feature learning.

V. Datasets

Deep learning methods rely heavily on data, and having a high-quality dataset that is closely related to the current task is crucial to effectively train the model. Simultaneously, having a unified and large-scale dataset is extremely important for advancing subject research. However, in the field of remote sensing, data annotation is challenging, and remote sensing data are iteratively updated very quickly. Due to the inconsistent data requirements of different segmentation tasks and significant variations in ground feature characteristics across different areas, constructing a unified large-scale dataset poses a substantial challenge. This section summarizes the currently proposed open-source optical and SAR-registered datasets suitable for semantic segmentation of multisource remote sensing images in the context of feature classification, as shown in Table III.

SEN12MS dataset consisting of 180 662 triplets of SAR image patches, multispectral image patches, and MODIS land cover maps, covering all meteorological seasons. It is suitable for the tasks of scene classification or semantic segmentation for land cover mapping. This dataset can be downloaded from https://mediatum.ub.tum.de/1474000.

MS-SAR LCZ dataset consisting of multispectral data and dual-polarimetric SAR data, it contains ten categories.

TABLE III
OPTICAL AND SAR DATASETS IN SEMANTIC SEGMENTATION

Dataset name	Image size(pixels)	Resolution(m)	Optical data source	SAR data source
SEN12MS [161]	256×256	10.0×10.0	Sentinel-1	Sentinel-2
MS-SAR LCZ [17]	–	10.0×10.0	Sentinel-1	Sentinel-2
LandCoverNet [162]	256×256	10.0×10.0	Sentinel-1	Sentinel-2
MSAW [163]	900×900	0.5×0.5	Worldview-2	UAV
GFB [164]	256×256	1.0×1.0	Google Earth	Gaofen-3
optical-SAR [151]	256×256	1.0×1.0	Gaofen-2	Gaofen-3
WHU-OPT-SAR [61]	5556×3704	5.0×5.0	Gaofen-1	Gaofen-3
Hunan [165]	256×256	10.0×10.0	Sentinel-1	Sentinel-2
MDAS [166]	1371×888	10.0×10.0	Sentinel-1	Sentinel-2
YYX-OPT-SAR [167]	512×512	0.5×0.5	Google Earth	UAV

(MS-SAR LCZ dataset contains images of two areas: Berlin and Hong Kong. The optical and SAR image size of Berlin is 643×626 pixels, and the image size of the Hong Kong area is 528×529 pixels.)

This dataset can be downloaded from https://github.com/danfenghong/IEEE_TGRS_MDL-RS.

LandCoverNet dataset covers all images from Africa, with main categories including water, two types of bare ground, and three types of vegetation. This dataset can be downloaded from www.landcover.net.

MSAW dataset collects very high-resolution optical and SAR data from the port of Rotterdam, the Netherlands, features buildings, vehicles, and boats of various sizes. This dataset can be downloaded from www.spacenet.ai.

GFB dataset is a building segmentation dataset, which cover nine cities from seven countries. The dataset can be accessed in: 10.11878/db.202104.000008.

Ren et al. [151] proposed an optical-SAR multimodal dataset for land cover classification, which covers three areas from two countries: Xi'an city in Shanxi Province, China; Dongying city in Shandong Province, China; and Pohang city in South Korea. This dataset can be accessed in: https://github.com/XD-MG/DDHRNet.

WHU-OPT-SAR dataset covers 100 pairs of optical images and SAR images from Hubei province, China. This dataset mainly contains seven types of land objects suitable for land use classification: Water, farmland, city, village, forest, road, and others. The dataset can be downloaded from https://github.com/AmberHen/WHU-OPT-SAR-dataset.git.

Hunan dataset covers optical and SAR images from Hunan province, China. This dataset mainly contains seven types of land objects suitable for land cover classification: Water, cropland, built-up area, grassland, forest, wetland, and unused land. The dataset can be downloaded from https://github.com/LauraChow/HunanMultimodalDataset.

MDAS dataset has five modalities remote sensing data: SAR data, multispectral image, hyperspectral image, DSM, and GIS data. This dataset mainly contains six types of land objects suitable for land cover classification: pavement, soil, roof, low vegetation, tree, and water. The dataset can be downloaded from https://doi.org/10.14459/2022mp1657312.

YYX-OPT-SAR dataset comprises 150 pairs of optical and SAR images covering urban, suburban, and mountain settings. The dataset can be downloaded from https://github.com/yeyuanxin110/YYX-OPT-SAR.

The datasets we have collected fall into two categories: 1) spaceborne data; and 2) airborne data. Among them, there are four satellite-borne datasets and two SAR image datasets acquired by UAVs. The satellite sources primarily include the Sentinel-1 and Sentinel-2 series, as well as the Gaofen-1 and Gaofen-3 series. However, a resolution issue arises: The resolution of the Sentinel series datasets is typically only 10 m, whereas the resolution of the Gaofen series datasets is 5 m. Therefore, very high-resolution (< 1 m) optical and SAR datasets are currently lacking. At the same time, the field of multisource remote sensing has always lacked a large-scale and high-quality dataset that can be used as a benchmark dataset for semantic segmentation methods in remote sensing tasks. This is very important and will promote the development of multisource remote sensing image research in the future.

VI. Evaluation Indicators for Semantic Segmentation

In the field of semantic segmentation, the main evaluation indicators of algorithm performance can be summarized as: Running time, running memory, and accuracy.

A. Running Time

Running speed is a very important indicator that reflects the performance of an algorithm, and is usually reflected by running time. With the continuous development of deep learning, model complexity continues to increase, network levels continue to deepen, and many algorithm structures are overly dependent on hardware and time. In some cases, the accuracy provided by the algorithm is not enough. Therefore, the running time can be used as an evaluation criterion for the algorithm in practical applications, and it can be tested which algorithm is more efficient under the same conditions.

In the semantic segmentation network, the timeliness of the network can be described by the speed of processing the number of images per second, which is usually expressed by frames per second.

B. Running Memory

Memory usage is one of the important indicators for evaluating algorithms. Many algorithms achieve faster time and improved accuracy by continuously expanding memory capacity. However, in practical applications, ideal memory conditions are often difficult to achieve, and high-performance GPUs generally cannot be equipped with large memories. Therefore, memory usage can also be used as one of the evaluation criteria for algorithm performance.

Floating-point operations (FLOPs) can represent the calculation amount of forward propagation in CNNs, and is used to estimate the computing resource usage of the segmentation network, thereby measuring the complexity of the algorithm. Multiply-accumulate operations (MACs) can also be used to represent the amount of operation. 1 MAC usually corresponds to two FLOPs. In addition, the most direct indicator of memory usage is to look at the number of network parameters and storage space usage. The number of network parameters is represented by parameters, and the memory space occupied unit is generally MB. Statistics on the amount of operations and the number of parameters can indicate the complexity of the network. The higher the values of these parameters, the more complex these network models are.

C. Accuracy

The evaluation indicators used in semantic segmentation of remote sensing images generally include pixel accuracy (PA), mean pixel accuracy (MPA), intersection over union (IoU), mean Intersection over union (MIoU), and frequency weighted intersection over union (FWIoU). For ease of understanding, we define the specific parameters as follows: K represents the category of pixels, t_i represents the total number of pixels in the ith category, n_{ii} represents the total number of pixels actually predicted to be ith category and is also ith category, n_{ij} represents the total number of pixels actually predicted to be jth category for the ith category.

PA: PA is an evaluation criterion for the accuracy of predicting pixels.

$$\mathrm{PA} = \sum_{i=1}^{K} n_{ii} / \sum_{i=1}^{K} t_i. \tag{1}$$

MPA: MPA represents the average PA of image pixels across all categories.

$$\mathrm{MPA} = \left(\sum_{i=1}^{K} n_{ii} / \sum_{i=1}^{K} t_i \right) \Big/ K. \tag{2}$$

IoU: IoU represents the degree of coincidence between the segmented image and the true value of the original image, and the value range is between 0–1.

$$\mathrm{IoU} = \sum_{i=1}^{K} \frac{n_{ii}}{t_i + \sum_{j=1}^{K} n_{ji} - n_{ii}}. \tag{3}$$

MIoU: MIoU represents the average IoU of image pixels across all categories.

$$\mathrm{MIoU} = \left(\sum_{i=1}^{K} \frac{n_{ii}}{t_i + \sum_{j=1}^{K} n_{ji} - n_{ii}} \right) \Big/ K. \tag{4}$$

FWIoU: FWIoU aims to weight the category of each pixel according to its frequency of occurrence.

$$\mathrm{FWIoU} = \frac{1}{\sum_{i=1}^{K} t_i} \sum_{i=1}^{K} \frac{\sum_{j=1}^{K} n_{ij} n_{ii}}{\sum_{j=1}^{K} n_{ij} + \sum_{j=1}^{K} n_{ji} - n_{ii}}. \tag{5}$$

VII. Challenges and Future Directions

Semantic segmentation problems typically manifest in two forms: 1) oversegmentation; and 2) undersegmentation. Oversegmentation involves segmenting the same object into different categories, while undersegmentation involves segmenting multiple objects into a single category. The challenge of remote sensing image segmentation stems from the inherent heterogeneity of remote sensing images, which significantly increases the complexity of semantic segmentation and leads to problems such as confusing image categories and unclear boundaries.

Optical and SAR images are complementary, and the complementary properties of multisource images can be used to make up for the shortcomings of single-source data. At the same time, optical and SAR images have structural consistency, and the consistency can be used to build correlations between heterogeneous data. Many methods have been proposed to learn and exploit the complementarity and structural consistency of optical and SAR images, but the following difficulties still exist.

1) *Feature fusion contribution challenge:* The disparities in imaging mechanisms lead to complementary and consistent representations of optical and SAR images across various object categories. Current research on feature fusion primarily emphasizes overall accuracy, resulting in high accuracy for some categories and very low accuracy for others. This is mainly due to the fact that the model does not adaptively adjust the weights of optical and SAR images for different categories. In addition, there is a failure to analyze the relationship between utilization and object categories.
2) *Data demand challenge:* Despite the abundance of remote sensing image data today, the availability of effective data suitable for training is limited. This constraint arises from challenges in labeling remote sensing images and the necessity for preprocessing operations such as alignment before feature fusion. The absence of benchmark datasets introduces a bias in evaluating modeling algorithms.

The above challenges can serve as future research priorities and development trends.

1) *Dynamic assignment of feature contribution:* Among the strategies for designing feature fusion networks, a critical aspect requiring exploration is the construction of a model training network capable of assigning distinct weights to optical and SAR images in different feature categories. There are semantic correlations between different categories, and at the same time, the contribution of measuring multimodal features includes assigning high weights to the dominant modes in various feature categories and guiding the weak modes to learn complementary features. This technique stands as a key aspect of optical SAR feature learning that warrants further investigation.
2) *Development of large-scale benchmark datasets:* Creating large-scale, high-quality benchmark datasets comparable to those in natural optical images is imperative. These datasets can be made available to researchers for comparing model algorithms, thereby expediting the research process in remote sensing image segmentation.

3) *Feature fusion in small samples or low-quality data:* Although small-sample algorithms have attracted attention in target recognition in optical and SAR images, there are currently very few relevant studies in the field of remote sensing image segmentation. Since segmentation has relatively low real-time requirements, historical data can be used to train the model to extract features. Consequently, a pressing question arises concerning the optimal utilization of low-quality, unlabeled historical data. Therefore, training models to extract effective features from low-quality data remains a pivotal focus for future research.
4) *Research on large models based on semantic segmentation of multisource remote sensing images:* The development of large models is the mainstream trend in current research. How to make large model learning fully utilize multisource information to improve the task performance of downstream remote sensing image processing is also a hot topic to be studied.

VIII. Conclusion

Deep neural network extraction of features for semantic segmentation has achieved great success in the field of computer vision. These findings have inspired researchers to apply it to the field of remote sensing image segmentation. The complexity, heterogeneity, and scale differences of remote sensing images pose many problems for semantic segmentation. Due to the differences in imaging modalities, single-source images have certain limitations. Optical and SAR, as commonly used data sources in remote sensing, have complementarity and consistency. How to effectively fuse the features of optical and SAR images has always been a research hot spot and a research difficulty. In this article, we summarize the research progress in semantic segmentation of remote sensing images for deep feature fusion of optical and SAR images. We also outline the challenges of the topic and provide a comprehensive overview for scholars and practitioners from the technical point of view of network module design. Specifically, we summarize the research progress in semantic segmentation of remote sensing in detail, and then summarize the technical approaches of optical and SAR image feature fusion in semantic segmentation from the perspective of feature fusion module design.

References

[1] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[2] X. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[3] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 20–33, 2021.

[4] B. L. Turner and W. B. Meyer, "Global land-use and land-cover change: An overview," *Changes Land Use Land Cover: Glob. Perspective*, vol. 4, no. 3, 1994.

[5] E. F. Lambin et al., "The causes of land-use and land-cover change: Moving beyond the myths," *Glob. Environ. Change*, vol. 11, no. 4, pp. 261–269, 2001.

[6] N. Zang, Y. Cao, Y. Wang, B. Huang, L. Zhang, and P. T. Mathiopoulos, "Land-use mapping for high-spatial resolution remote sensing image via deep learning: A review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5372–5391, 2021.

[7] U. Weidner and W. Förstner, "Towards automatic building extraction from high-resolution digital elevation models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 50, no. 4, pp. 38–49, 1995.

[8] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 1, pp. 43–63, 2007.

[9] 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.

[10] D. Lu, E. Moran, and S. Hetrick, "Detection of impervious surface change with multitemporal landsat images in an urban–rural frontier," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 298–306, 2011.

[11] Q. Weng, "Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends," *Remote Sens. Environ.*, vol. 117, pp. 34–49, 2012.

[12] Y. Wang and M. Li, "Urban impervious surface detection from remote sensing images: A review of the methods and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 3, pp. 64–93, Sep. 2019.

[13] W. Wu, S. Guo, Z. Shao, and D. Li, "CroFuseNet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and sar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2573–2588, 2023.

[14] M. Galli, F. Ardizzone, M. Cardinali, F. Guzzetti, and P. Reichenbach, "Comparing landslide inventory maps," *Geomorphol.*, vol. 94, no. 3-4, pp. 268–289, 2008.

[15] C. Zhong et al., "Landslide mapping with remote sensing: Challenges and opportunities," *Int. J. Remote Sens.*, vol. 41, no. 4, pp. 1555–1581, 2020.

[16] E. Kuruoglu and J. Zerubia, "Modeling sar images with a generalization of the rayleigh distribution," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 527–533, Apr. 2004.

[17] 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.

[18] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.

[19] Z. Huang, M. Datcu, Z. Pan, and B. Lei, "Deep sar-net: Learning objects from signals," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 179–193, 2020.

[20] X. Zhang, S. Feng, C. Zhao, Z. Sun, S. Zhang, and K. Ji, "MGSFA-Net: Multi-scale global scattering feature association network for sar ship target recognition," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4611–4625, 2024.

[21] Z. Sun, X. Leng, X. Zhang, B. Xiong, K. Ji, and G. Kuang, "Ship recognition for complex SAR images via dual-branch transformer fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[22] M. Huang et al., "The QXS-saropt dataset for deep learning in sar-optical data fusion," 2021, *arXiv:2103.08259*.

[23] 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.

[24] J. Ling and H. Zhang, "WCDL: A weighted cloud dictionary learning method for fusing cloud-contaminated optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2931–2941, 2023.

[25] Z. Du, X. Li, J. Miao, Y. Huang, H. Shen, and L. Zhang, "Concatenated deep learning framework for multi-task change detection of optical and sar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 719–731, 2024.

[26] Z. Lv, H. Huang, W. Sun, T. Lei, J. A. Benediktsson, and J. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[27] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, Dec. 2016.

[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[29] M. Reichstein et al., "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[30] J. Hu, D. Hong, and X. X. Zhu, "MIMA: MAPPER-induced manifold alignment for semi-supervised fusion of optical image and polarimetric sar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9025–9040, Nov. 2019.

[31] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and sar image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[32] H. Zhang et al., "Optical and sar image dense registration using a robust deep optical flow framework," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1269–1294, 2023.

[33] D. Xiang, X. Pan, H. Ding, J. Cheng, and X. Sun, "Two-stage registration of sar images with large distortion based on superpixel segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[34] D. Xiang, H. Ding, X. Sun, J. Cheng, C. Hu, and Y. Su, "PolSAR image registration combining siamese multiscale attention network and joint filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[35] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–21, 2022.

[36] Y. Sun, L. Lei, Z. Li, and G. Kuang, "Similarity and dissimilarity relationships based graphs for multimodal change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 208, pp. 70–88, 2024.

[37] Z. Lv, T. Yang, T. Lei, W. Zhou, Z. Zhang, and Z. You, "Spatial-spectral similarity based on adaptive region for landslide inventory mapping with remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[38] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[39] Q. Ding, Z. Shao, X. Huang, O. Altan, and Y. Fan, "Improving urban land cover mapping with the fusion of optical and sar data based on feature selection strategy," *Photogramm. Eng. Remote Sens.*, vol. 88, pp. 17–28, 2022.

[40] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sens. Environ.*, vol. 80, pp. 185–201, 2002.

[41] H. Lou, S. Li, and Y. Zhao, "Detecting community structure using label propagation with weighted coherent neighborhood propinquity," *Physica A.*, vol. 392, pp. 3095–3105, 2013.

[42] R. Girma, C. Fürst, and A. Moges, "Land use land cover change modeling by integrating artificial neural network with cellular automata-Markov chain model in gidabo river basin, main ethiopian rift," *Environ. Challenges*, vol. 6, 2022, Art. no. 100419.

[43] G. Sisay, B. Gessesse, C. Fürst, M. Kassie, and B. Kebede, "Modeling of land use/land cover dynamics using artificial neural network and cellular automata Markov chain algorithms in goang watershed, ethiopia," *Heliyon*, vol. 9, 2023, Art. no. e20088.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[45] N. Davari, G. Akbarizadeh, and E. Mashhour, "Corona detection and power equipment classification based on GoogleNet-AlexNet: An accurate and intelligent defect detection model based on deep learning for power distribution lines," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 2766–2774, Aug. 2022.

[46] 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.

[47] Z. Zhu, C. E. Woodcock, J. Rogan, and J. Kellndorfer, "Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using landsat and sar data," *Remote Sens. Urban Environ.*, vol. 117, pp. 72–82, 2012.

[48] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[49] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retr.*, vol. 7, pp. 87–93, 2018.

[50] H. Yu et al., "Methods and datasets on semantic segmentation: A review," *Neurocomputing*, vol. 304, pp. 82–103, 2018.

[51] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[52] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417.

[53] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 309–322, 2021.

[54] Y. Wang, H. Lv, R. Deng, and S. Zhuang, "A comprehensive survey of optical remote sensing image segmentation methods," *Can. J. Remote Sens.*, vol. 46, no. 5, pp. 501–531, 2020.

[55] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 115–134, 2019.

[56] K. R. Hasan, A. B. Tuli, M. A.-M. Khan, S.-H. Kee, M.A. Samad, and A.-A. Nahid, "Deep learning-based semantic segmentation for remote sensing: A bibliometric literature review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1390–1418, 2024.

[57] L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, "Deep learning-based semantic segmentation of remote sensing images: A survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8370–8396, 2024.

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

[59] S. Mahyoub, A. Fadil, E. Mansour, H. Rhinane, and F. Al-Nahmi, "Fusing of optical and synthetic aperture radar (sar) remote sensing data: A systematic literature review (SLR)," *ISPRS Arch.*, vol. 42, pp. 127–138, 2019.

[60] S. Agrawal and G. Khairnar, "A comparative assessment of remote sensing imaging techniques: Optical, SAR and LiDAR," *ISPRS Arch.*, vol. 42, pp. 1–6, 2019.

[61] X. Li et al., "MCANet: A joint semantic segmentation framework of optical and sar images for land use classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, 2022, Art. no. 102638.

[62] X. Li et al., "Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and sar images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 178–191, 2023.

[63] Z. Qiu, L. Yue, and X. Liu, "Void filling of digital elevation models with a terrain texture learning model based on generative adversarial networks," *Remote Sens.*, vol. 11, pp. 2829–2850, 2019.

[64] E. Gumus and P. Kirci, "Selection of spectral features for land cover type classification," *Expert Syst. Appl.*, vol. 102, pp. 27–35, 2018.

[65] G. Akbarizadeh, G. A. Rezai-Rad, and S. B. Shokouhi, "A new region-based active contour model with skewness wavelet energy for segmentation of sar images," *IEICE Trans. Inf. Syst.*, vol. 93, no. 7, pp. 1690–1699, 2010.

[66] G. Akbarizadeh, "A new statistical-based kurtosis wavelet energy feature for texture recognition of sar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4358–4368, Nov. 2012.

[67] Z. Tirandaz and G. Akbarizadeh, "A two-phase algorithm based on kurtosis curvelet energy and unsupervised spectral regression for segmentation of SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1244–1264, Mar. 2016.

[68] J. Paneque-Gálvez et al., "Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 23, pp. 372–383, 2013.

[69] F. Zhang and X. Yang, "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection," *Remote Sens. Environ.*, vol. 251, 2020, Art. no. 112105.

[70] F. Samadi, G. Akbarizadeh, and H. Kaabi, "Change detection in sar images using deep belief network: A new training approach based on morphological images," *IET Image Process.*, vol. 13, no. 12, pp. 2255–2264, 2019.

[71] Z. Tirandaz, G. Akbarizadeh, and H. Kaabi, "Polsar image segmentation based on feature extraction and data compression using weighted neighborhood filter bank and hidden Markov random field-expectation maximization," *Measurement*, vol. 153, 2020, Art. no. 107432.

[72] N. Aghaei, G. Akbarizadeh, and A. Kosarian, "GreywolfLSM: An accurate oil spill detection method based on level set method from synthetic

aperture radar imagery," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 181–198, 2022.

[73] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sens.*, vol. 8, no. 4, Art. no. 329, 2016.

[74] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[75] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *Electron. Imag.*, vol. 2016, no. 10, pp. 1–9, 2016.

[76] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017.

[77] S. Paisitkriangkrai, J. Sherrah, P. Janney, and V.-D. Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 36–43.

[78] J. Feng, L. Wang, H. Yu, L. Jiao, and X. Zhang, "Divide-and-conquer dual-architecture convolutional neural network for classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 484.

[79] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–9.

[80] F. Sharifzadeh, G. Akbarizadeh, and Y. Seifi Kavian, "Ship classification in sar images using a new hybrid CNN–MLP classifier," *J. Indian Soc. Remote Sens.*, vol. 47, pp. 551–562, 2019.

[81] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[82] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 113, pp. 155–165, 2016.

[83] G. Zheng, Z. Jiang, H. Zhang, and X. Yao, "Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network," *Front. Earth Sci.*, vol. 11, 2023, Art. no. 1115805.

[84] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in sar satellite images with deep fully convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1867–1871, Dec. 2018.

[85] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[86] 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.

[87] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "Deeplab-based spatial feature extraction for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 251–255, Feb. 2019.

[88] R. Li et al., "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.

[89] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[90] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[91] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[92] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[93] F. M. Ghara, S. B. Shokouhi, and G. Akbarizadeh, "A new technique for segmentation of the oil spills from synthetic-aperture radar images using convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8834–8844, 2022.

[94] N. Aghaei, G. Akbarizadeh, and A. Kosarian, "Osdes_Net: Oil spill detection based on efficient_shuffle network using synthetic aperture radar imagery," *Geocarto Int.*, vol. 37, no. 26, pp. 13539–13560, 2022.

[95] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 298.

[96] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

[97] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[98] M. Campos-Taberner et al., "Understanding deep learning in land use classification based on sentinel-2 time series," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 17188.

[99] Z. Sun, L. Di, and H. Fang, "Using long short-term memory recurrent neural network in land cover classification on landsat and cropland data layer time series," *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 593–614, 2019.

[100] Z. Lin et al., "Large-scale rice mapping using multi-task spatiotemporal deep learning and sentinel-1 sar time series," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 699.

[101] B. Chen et al., "A joint learning im-bilstm model for incomplete time-series sentinel-2a data imputation and crop classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102762.

[102] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[103] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.

[104] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.

[105] Z. Dong, G. Gao, T. Liu, Y. Gu, and X. Zhang, "Distilling segmenters from CNNs and transformers for remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[106] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[107] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[108] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

[109] Y. Liu et al., "VMamba: Visual state space model," 2024. *arXiv:2401.10166*.

[110] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," 2024, *arXiv:2401.09417*.

[111] X. He et al., "Pan-Mamba: Effective pan-sharpening with state space model," 2024, *arXiv:2402.12192*.

[112] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "RSMamba: Remote sensing image classification with state space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[113] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual state space model for remote sensing images semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[114] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[115] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[116] N. Kussul, G. Lemoine, F. J. Gallego, S. V. Skakun, M. Lavreniuk, and A. Y. Shelestov, "Parcel-based crop classification in Ukraine using landsat-8 data and Sentinel-1a data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2500–2508, Jun. 2016.

[117] Y. Lin, H. Zhang, G. Li, T. Wang, L. Wan, and H. Lin, "Improving impervious surface extraction with shadow-based sparse representation from optical, SAR, and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2417–2428, Jul. 2019.

[118] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van Den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2868–2881, Jul. 2016.

[119] Y.-T. Chen, J. Shi, C. Mertz, S. Kong, and D. Ramanan, "Multimodal object detection via bayesian fusion," *In Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 139–158.

[120] S. E. Kahou et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[121] 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.

[122] L. Maggiolo, D. Solarna, G. Moser, and S. B. Serpico, "Optical-SAR decision fusion with Markov random fields for high-resolution large-scale land cover mapping," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 5508–5511.

[123] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858–3866, Dec. 2007.

[124] R. Vohra and K. Tiwari, "Comparative analysis of SVM and ann classifiers using multilevel fusion of multi-sensor data in urban land classification," *Sens. Imag.*, vol. 21, pp. 1–21, 2020.

[125] Y. Zhouping, "Fusion algorithm of optical images and SAR with SVT and sparse representation," *Int. J. Smart Sens. Intell. Syst.*, vol. 8, no. 2, pp. 1123–1141, 2015.

[126] I. R. Farah, W. Boulila, K. S. Ettabaa, and M. B. Ahmed, "Multiapproach system based on fusion of multispectral images for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4153–4161, Dec. 2008.

[127] M. Alonzo, B. Bookhagen, and D. A. Roberts, "Urban tree species mapping using hyperspectral and LiDAR data fusion," *Remote Sens. Environ.*, vol. 148, pp. 70–83, 2014.

[128] 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.

[129] 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.

[130] 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.

[131] C. Liu, H. Sun, Y. Xu, and G. Kuang, "Multi-source remote sensing pretraining based on contrastive self-supervised learning," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4632.

[132] 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.

[133] 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.

[134] H. Guo, H. Yang, Z. Sun, X. Li, and C. Wang, "Synergistic use of optical and polsar imagery for urban impervious surface estimation," *Photogramm. Eng. Remote Sens.*, vol. 80, no. 1, pp. 91–102, 2014.

[135] W. Wu, J. Teng, Q. Cheng, and S. Guo, "Extraction of impervious surface using sentinel-1a time-series coherence images with the aid of a sentinel-2a image," *Photogramm. Eng. Remote Sens.*, vol. 87, no. 3, pp. 161–170, 2021.

[136] Y. Lin, H. Zhang, H. Lin, P. E. Gamba, and X. Liu, "Incorporating synthetic aperture radar and optical images to investigate the annual dynamics of anthropogenic impervious surface at large scale," *Remote Sens. Environ.*, vol. 242, 2020, Art. no. 111757.

[137] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[138] 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.

[139] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 238–262, 2021.

[140] W. Zhou, J. Yuan, J. Lei, and T. Luo, "TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 73–78, 2020.

[141] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 96–115, 2022.

[142] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

[143] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and lidar data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, 2019.

[144] X. Geng et al., "Multisource joint representation learning fusion classification for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[145] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.

[146] J. Zhang, W. Ma, Y. Wu, and L. Jiao, "Multimodal remote sensing image registration based on image transfer and local features," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1210–1214, Aug. 2019.

[147] D. Quan et al., "Deep generative matching network for optical and sar image registration," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6215–6218.

[148] W. Li et al., "Aligning semantic distribution in fusing optical and sar images for land use classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 199, pp. 272–288, 2023.

[149] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[150] 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.

[151] 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. Geoinf.*, vol. 112, 2022, Art. no. 102896.

[152] X. Liu, H. Zou, S. Wang, Y. Lin, and X. Zuo, "Joint network combining dual-attention fusion modality and two specific modalities for land cover classification using optical and sar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3236–3250, 2024.

[153] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3029–3037.

[154] S. Wang, J. Zhang, and C. Zong, "Learning multimodal word representation via dynamic fusion methods," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[155] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," 2017, *arXiv:1702.01992*.

[156] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*.

[157] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[158] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[159] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-LiDAR and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[160] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[161] 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. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 4, no. 2-W7, pp. 153–160, 2019.

[162] H. Alemohammad and K. Booth, "Landcovernet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.

[163] J. Shermeyer et al., "SpaceNet 6: Multi-sensor all weather mapping dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 196–197.

[164] J. Xia, N. Yokoya, B. Adriano, L. Zhang, G. Li, and Z. Wang, "A benchmark high-resolution GaoFen-3 SAR dataset for building semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5950–5963, 2021.

[165] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen, "DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 186, pp. 170–189, 2022.

[166] J. Hu et al., "MDAS: A new multimodal benchmark dataset for remote sensing," *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 113–131, 2023.

[167] J. Li, J. Zhang, C. Yang, H. Liu, Y. Zhao, and Y. Ye, "Comparative analysis of pixel-level fusion algorithms and a new high-resolution dataset for sar and optical image fusion," *Remote Sens.*, vol. 15, no. 23, 2023, Art. no. 5514.

Chenfang Liu received the M.S. degree in electronic information, in 2022, from the National University of Defense Technology, Changsha, China, where she is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects.

Her research interests include multisource image interpretation, feature extraction, and machine learning.

Yuli Sun received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2023.

He has authored or coauthored papers in IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, ISPRS *Journal of Photogrammetry and Remote Sensing*, *Pattern Recognition*, etc. He is currently a Postdoctor with the College of Aerospace Science and Engineering, NUDT. His research interests include remote sensing image processing and multitemporal image analysis.

Yanjie Xu received the B.S. and M.S. degrees in information and communication engineering, in 2018 and 2021, respectively, from the National University of Defense Technology, Changsha, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects.

His research interests include SAR ATR, incremental learning, and deep learning.

Zhongzhen Sun received the B.S. and M.S. degrees in information and communication engineering, in 2018 and 2021, respectively, from the National University of Defense Technology, Changsha, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects.

His research interests include AI for SAR, ship detection, and recognition.

Xianghui Zhang received the B.S. degree in information engineering, in 2022, from the National University of Defense Technology, Changsha, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects.

His research interests include SAR image interpretation, feature extraction, and deep learning.

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

She is currently a 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.

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, 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.