CroFuseNet: A Semantic Segmentation Network for Urban Impervious Surface Extraction Based on Cross Fusion of Optical and SAR Images

Wenfu Wu, Songjing Guo, Zhenfeng Shao, and Deren Li

***Abstract*—The fusion of optical and synthetic aperture radar (SAR) images is a promising method to extract urban impervious surface (IS) accurately. Previous studies have shown that the feature-level fusion of optical and SAR images can significantly improve IS extraction. However, they generally use simple layer stacking for features fusion, ignoring the interaction between optical and SAR images. Besides, most of the features they used are shallow features manually extracted, such as texture and geometric features, lacking the use of high-level semantic features of images. The lack of publicly available IS datasets is considered as an obstacle that prevents the extensive use of deep learning models in IS extraction. Therefore, this study first creates an open and accurate IS dataset based on optical and SAR images, and then proposes a semantic segmentation network based on cross fusion of optical and SAR images features, namely CroFuseNet, for IS extraction. In CroFuseNet, we design a cross fusion module to fuse features of optical and SAR images to achieve better complementarity between the two types of images, and we propose a multimodal features aggregation module to aggregate specific high-level features from optical and SAR images. To validate the proposed CroFuseNet, we compare it with two classical machine learning algorithms and four state-of-the-art deep learning models. The proposed model has the highest accuracy, with OA, MIoU, and F1-Score of 97.77%, 0.9495, and 0.9770, respectively. The quantitative and qualitative experimental results demonstrate that the proposed model is superior to these comparative algorithms.**

***Index Terms*—Cross fusion, multimodal feature aggregation, optical and SAR images fusion, semantic segmentation network, urban impervious surface (IS).**

Manuscript received 19 October 2022; revised 2 January 2023 and 7 February 2023; accepted 25 February 2023. Date of publication 28 February 2023; date of current version 16 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42090012, in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100009, in part by the 03 special research and 5G project of Jiangxi Province in China under Grant 20212ABC03A09, in part by the Zhuhai industry university research cooperation project of China under Grant ZH22017001210098PWC, in part by the Key R & D project of Sichuan Science and technology plan under Grant 2022YFN0031, and in part by the Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence under Grant ZZJJ202202. *(Corresponding author: Zhenfeng Shao.)*

Wenfu Wu and Deren Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: wuwwf09140818@whu.edu.cn; drli@whu.edu.cn).

Songjing Guo is with the School of Geophysics and Geomatics, China University of Geosciences, Wuhan, 430074 , China (e-mail: guosongjing@cug.edu.cn).

Zhenfeng Shao is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: shaozhenfeng@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3250461

I. Introduction

IMPERVIOUS surface (IS) is defined as the surface that prevents water from penetrating underground, which mostly includes artificial structures, such as asphalt roads, building roofs, and parking lots [1]. In the past few decades, the world has been in a wave of rapid urbanization, and a large number of rural people have begun to gather in cities. As of 2018, about 55% of the world's population lived in cities, nearly double the 30% in 1950. By 2050, it is estimated that more than two-thirds of the world's population (nearly 7 billion) will live in cities [2], [3]. In order to meet the living needs of such a dense city population, a large amount of natural surfaces have been replaced by artificial surfaces (i.e., IS), resulting in a series of urban diseases, such as urban heat islands [4] and urban floods [5]. Therefore, IS has become a key indicator to measure the degree of urbanization and the quality of the urban ecological environment. Against this background, the United Nations proposes a sustainable development goal of building inclusive, safe, and disaster resistant sustainable cities and human settlements by 2030. Mapping IS accurately is one of the key measures to achieve the abovementioned goal.

Remote sensing technology has become the most important means to accurately extract and dynamically monitor IS because of its advantages of low cost, high coverage, short data acquisition cycle, large spatial scale, and other advantages. According to a survey by [6], remote sensing-based IS extraction methods can be divided into four categories: spectral mixing analysis, image classification methods, index methods, and multisource data fusion methods. Among them, multisource data fusion is a promising method for IS extraction, particularly for optical and synthetic aperture radar (SAR) images fusion. Different from optical sensors, SAR is an active sensor that detects backscattering information of ground objects and is sensitive to geometric and physical characteristics of ground objects, such as surface roughness, moisture, and complex dielectric constant, which can complement the biochemical characteristics of ground objects detected by optical remote sensing images [7]. In addition, due to the long wavelength of SAR, it has the ability of all-day and all-weather Earth observation. Currently, with the development of SAR sensor technology and the open source of Sentinel-1 data, optical and SAR images fusion has been employed in IS extraction on global and national scales [8], [9].

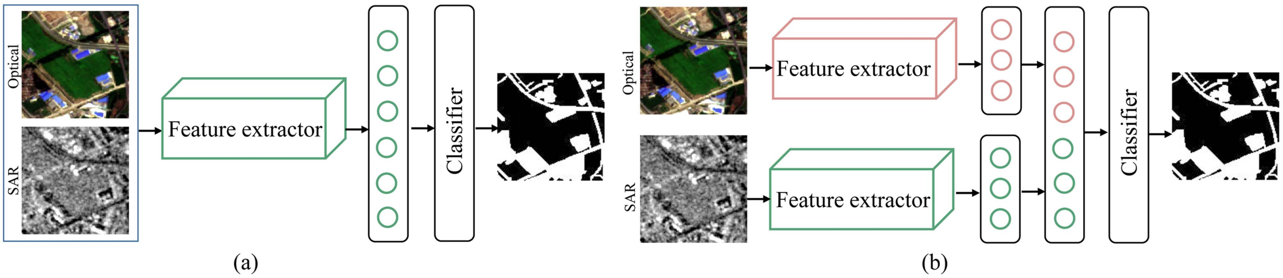

Fig. 1. Illustration for early fusion (a) and middle fusion (b).

In general, the fusion of optical and SAR images has three levels, namely pixel level, feature level, and decision level. Images registration is the premise of images fusion. For heterogeneous sensors, such as optics and SAR, images registration is still a challenging task, and its registration error will affect the fusion effect [32], [33], [34]. According to previous studies, the feature level fusion of optical and SAR images has the lowest sensitivity to geometric registration errors [35]. Therefore, the feature level fusion of optical and SAR images is the most popular and effective fusion level in the extraction of IS [10]. In the process of IS extraction based on optical and SAR images feature level fusion, there are two fusion paradigms, namely early fusion and middle fusion, as shown in Fig. 1. In the early fusion, optical, and SAR images are directly stacked together to form a multichannel image, and then features extraction is performed and input into a classifier. For instance, Lin et al. aligned high spatial resolution optical, SAR, and LiDAR data, and then used the sparse representation method to address the influence of shadows from tall buildings and trees during the extraction of IS based on optical remote sensing images [11]. Besides, Zhang et al. also used early fusion strategy to fuse optical and polarimetric SAR images to generate a multichannel data and input it into a deep convolutional network to extract IS [12]. Different from the early fusion, the middle fusion paradigm first extracts features from optical and SAR images respectively, and then fuses these features according to certain rules and feds them into a classifier. This fusion strategy is widely used in IS extraction based on optical and SAR images fusion. For example, Guo et al. fused the features extracted from polarimetric SAR and optical images by layer stacking and input the fused features into the C5.0 decision tree to extract IS [13]. Zhang et al. first extracted texture features of optical and SAR images, respectively, and then fused them by layer stacking to improve the extraction accuracy of IS [14]. In order to alleviate the influence of shadows in urban IS extraction based on optical remote sensing images, Sun et al. proposed a two-stage hierarchical fusion framework of optical and SAR images fusion for extracting IS based on Sentinel-1 and Sentinel-2 images [15]. Based on the same fusion strategy, many other scholars have contributed to the extraction of IS based on optical and SAR images fusion [8], [16], [17]. Although the abovementioned studies improved the extraction accuracy of IS, they adopted a simple concatenation or layer stacking for features fusion, which made it difficult to capture the interaction between optical and SAR images. In addition, most of the features they used are manually extracted shallow features, such as space, spectral, and texture.

Recently, deep learning has been widely used in the field of remote sensing and has achieved satisfactory results. For example, Sun et al. proposed a series of novel models for accurate classification of hyperspectral images, including successive pooling attention network [38], multistructure kernel extreme learning machine with attention fusion strategy [39], and spatial feature tokenization transformer [40]. Some scholars also started to try to use deep learning-based methods for IS extraction. For example, Zhang et al. proposed a deep convolutional networks based on small patches to extract urban IS using optical and polarimetric SAR images as model inputs [12]. Then, Wang and Li proposed a modified convolutional deep belief network (CDBN) to extract IS and obtained better results compared with other algorithms, such as support vector machine (SVM), CNN, and CDBN [6]. To get rid of the limitations of manually labeled data, Parekh et al. used OpenStreetMap data to automatically generate training and test samples to support the extraction of IS based on deep learning methods [18]. Besides, Feng et al. proposed an IS extraction method that deeply fused features from multispectral and hyperspectral images [19]. Although IS extraction methods based on deep learning can obtain highly competitive results, the application of deep learning in IS extraction is relatively rare compared to other fields. There are twofold limitations in the existing deep learning-based IS extraction methods. On the one hand, due to the limitations of pixel-by-pixel IS annotation dataset, the input of existing methods is mostly based on small image patches or pixels, which lacks the use of image context information. Moreover, the size of patches depends on the spatial resolution of the remote sensing image used, and its optimal value needs to be determined manually according to many experiments. On the other hand, there are few attempts at deep learning in IS extraction based on optical and SAR images fusion.

To address the aforementioned problems of IS extraction based on optical and SAR images fusion, we first construct an optical and SAR images IS dataset, named WHU-IS, and then, based on this dataset, we propose a novel semantic segmentation network CroFuseNet for urban IS extraction, which is based on cross fusion of optical and SAR images features. Fig. 2 provides a typical illustration of IS extraction result using the proposed CroFuseNet, indicating the proposed CroFuseNet achieves satisfactory results. Therefore, the main contribution of this study can be summarized as the following threefold.

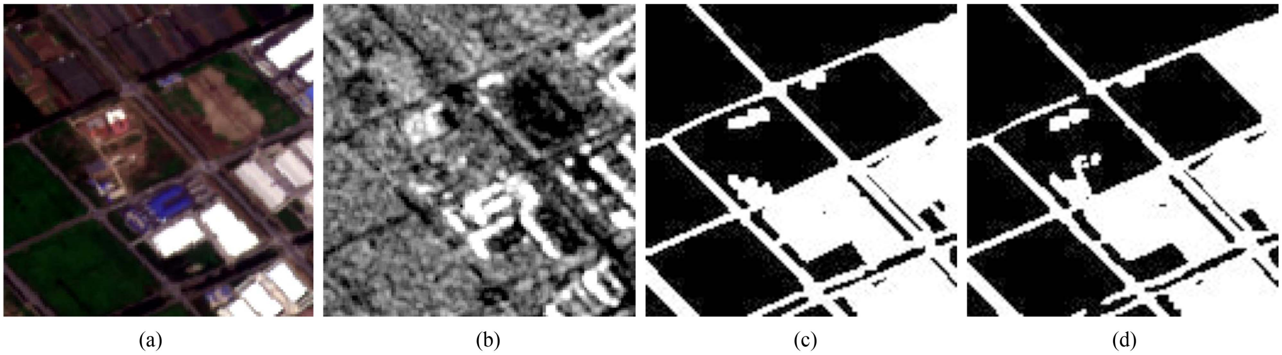

Fig. 2. Schematic illustration of IS extraction. (a) Optical image. (b) SAR image. (c) Ground truth. (d) IS result extracted by the proposed CroFuseNet.

1) We first construct an open and accurate optical and SAR images IS dataset, named WHU-IS, by manual annotation using Sentinel-1 and Sentinel-2 as data sources and the area covering the central urban area and its suburbs of Wuhan, China, as the study area. The dataset contains pixel-by-pixel label image blocks of both pervious surface and IS. There are 5603 patches in total, and the size of each patch is 128 × 128. We believe this dataset can provide data support for the development of advanced IS extraction methods based on deep learning. Our dataset WHU-IS can be available at.[1]
2) We propose a novel semantic segmentation network, named CroFuseNet, for IS extraction based on optical and SAR images fusion. Inspired by the fact that multimodal features contain both specific and shared features [21], we separate and use the shared and specific features of optical and SAR images. Therefore, the proposed CroFuseNet is designed as a network similar to three branches, namely optical specific features learning subnetwork, SAR specific features learning subnetwork, and optical and SAR shared features learning subnetwork. It cannot only learn the shared information between optical and SAR images, but also can make full use of their specific information to improve the extraction accuracy of IS. Considering the superior performance of UNet [20], we use UNet structure as backbone network to construct an encoder–decoder network in the proposed CroFuseNet. The encoders of the optical and SAR specific features learning subnetworks learn the multiscale features of each image and fuse and fed them into the encoder of the optical and SAR shared features learning subnetwork through the cross fusion module (CFM) designed in this study. The CFM can deeply fuse the features of optical and SAR images to achieve better complementarity of the two images. In the decoder, we propose a multimodal features aggregation (MFA) module, which integrates the specific high-level features from decoders of the optical and SAR specific features learning subnetworks into the decoder of the optical and SAR shared features learning subnetwork. The encoder and decoder are connected by skip connections to realize the fusion of shallow and deep information of images, which provides more semantic information and location details for the decoding process. Our source code will be released at.[2]
3) Finally, we construct a comprehensive loss function by considering the common constraints of optical and SAR images on the model, and train the proposed model under the constraints of the comprehensive loss function to improve the performance of the model.

II. Study Area and WHU-IS Dataset

A. Study Area

This study is carried out in Wuhan, Hubei province, China, and the geography of the study area is shown in Fig. 3. Located in 29°58′–31°22′ North latitude and 113°41′–115°05′ East longitude, Wuhan covers an area of 8569.15 km^2. It belongs to a typical subtropical monsoon climate with abundant rainfall all year round. Wuhan is the core city of the Yangtze River Economic Belt with more than 10 million permanent residents and plays a pivotal role in China's economic development. Since 1978, Wuhan has actively responded to the reform and opening up policy implemented by the Chinese government, and its economy has entered a period of rapid development, with the urbanization rate rising from 47.4% in 1978 to 84.3% in 2021. While enjoying the benefits of economic development, Wuhan also faces some urban ecological problems that need to be paid attention to and resolved. For example, Wuhan has suffered two major urban flooding disasters in the past 30 years. For example, in the 2016 urban flooding disaster, a total of 757 000 people in 12 districts of Wuhan were affected, 97 404 hectares of crops were damaged, 5848 houses were collapsed and the direct economic loss reached 2.265 billion CNY, and 14 people died. An important reason for the frequent occurrence of flood disasters in Wuhan is that the paved roads in Wuhan city are predominantly made of impermeable asphalt and cement pavement, which considerably expands the IS area of the city. In addition, the increase of IS area will lead to a significant decrease in the

[1] [Online]. Available: https://pan.baidu.com/s/1jcLrMFZLmvkzf13caKpKOQ?pwd=l6l4

[2] [Online]. Available: https://pan.baidu.com/s/1BTMwdpHFxol5KZ9zzH9r6g?pwd=f5vf

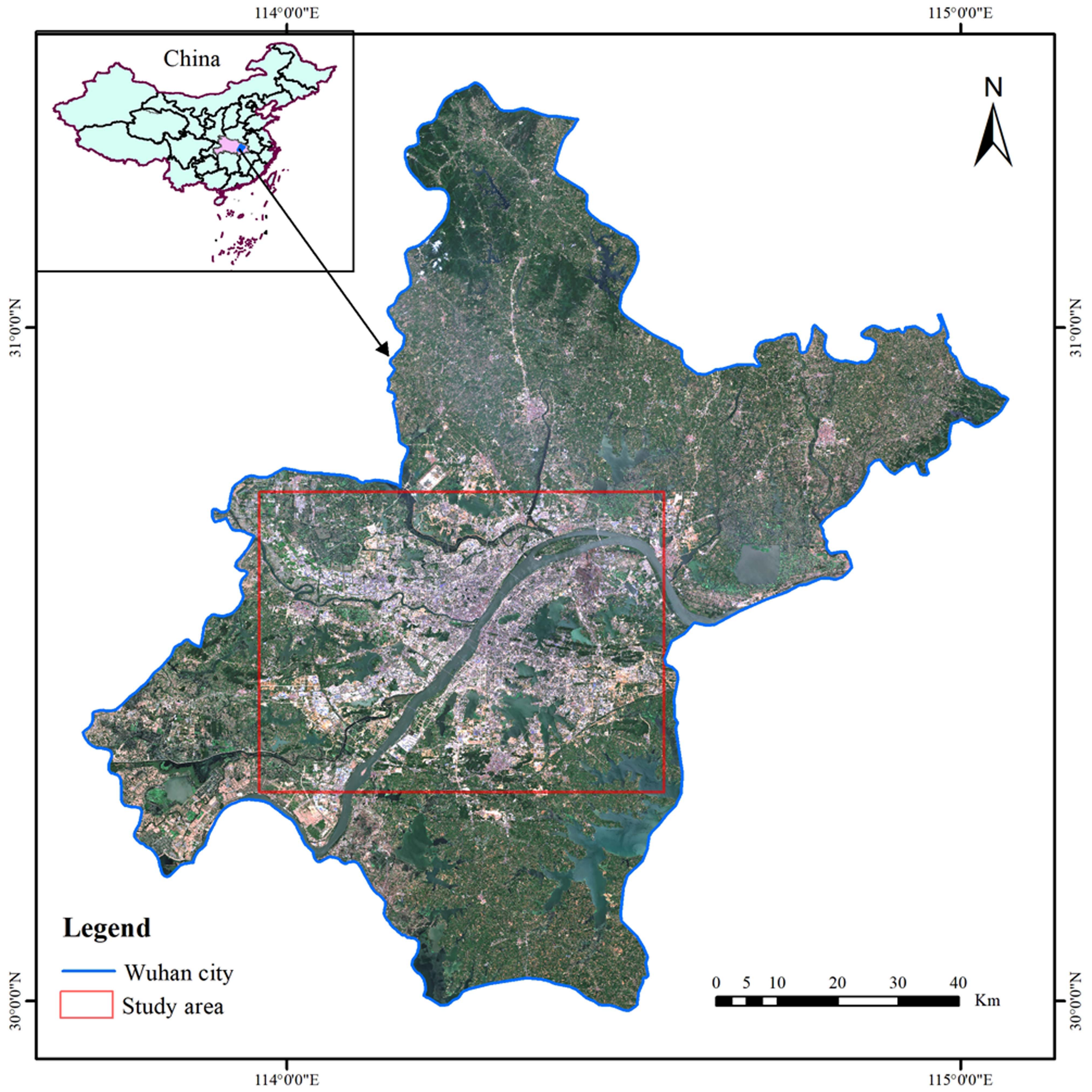

Fig. 3. Geography of the study area.

Manning coefficient of the original ground, hence shortening the rainwater confluence time, thus increasing the rainwater flow and the burden of the pipe network, and greatly increasing the risk of urban rainstorm waterlogging. Therefore, monitoring the distribution of urban IS is the premise and key measure to alleviate the above problems. Considering Wuhan's urgent need for accurate monitoring of IS and its climate characteristics, Wuhan is selected as the study area for IS extraction based on optical and SAR images fusion in this study.

B. WHU-IS Dataset

Developing an accurate IS dataset is a prerequisite for the implementation of supervised deep learning IS extraction models. Hence, in this study, we first construct an IS dataset based on optical and SAR images for the IS extraction task. In consideration of the availability and quality reliability of Sentinel series data, Sentinel-1 and Sentinel-2 multispectral images (MSI) are selected as the data source. The images coverage area is more than 3000 km^2, covering the central urban area of Wuhan and its surrounding suburbs. The Sentinel-2 image bands used in this study are NIR (Band 8), R (Band 4), G (Band 3), and B (Band 2) bands with a spatial resolution of 10 m, and the Sentinel-1 image used includes VV and VH polarization modes. The specific characteristics of Sentinel-1 and Sentinel-2 images used in this study are illustrated in Table I.

Before data annotation, we execute a series of preprocessing operations to improve the image quality. In this study, data collection and preprocessing are completed through Google Earth engine (GEE). The Sentinel-2 data provided by GEE is

TABLE I
SPECIFIC CHARACTERISTICS OF SENTINEL-1 AND SENTINEL-2 IMAGES USED IN THIS STUDY

Sensors	Product	Bands	Spatial resolution	Acquisition date
Sentinel-1	GRD	VV, VH	10 m	June 1, 2019 - June 30, 2019
Sentinel-2	MSI L2A	Band 2: 496.6 nm Band 3: 560.0 nm Band 4: 664.5 nm Band 8: 835.1 nm	10 m	June 1, 2019 - June 30, 2019

Level 2 A (L2A) data that has been atmospheric corrected and no additional preprocessing is required. The Sentinel-1 data are preprocessed through Sentinel-1 toolbox developed by the GEE [22]. The main preprocessing steps of Sentinel-1 data involve thermal noise removal, radiometric calibration, terrain correction, and decibel conversion. It is worth noting that we obtained the final Sentinel-1 and Sentinel-2 data by averaging the data covering the study area from June 1, 2019 to June 30, 2019. This operation can reduce the disturbance of random noise to a certain extent, particularly speckle noise in Sentinel-1 images. Finally, Sentinel-1 and Sentinel-2 images are registered. In this study, in order to achieve the registration between Sentinel-1 and Sentinel-2 images, we manually select road intersections and building corner points that can be well recognized in SAR and optical images as ground control points (GCPs) through visual interpretation, and then use third-order polynomial method to register optical and SAR images based on these GCPs. It is worth noting that in the process of selecting GCPs, we will repeatedly adjust the position of GCPs to ensure that the root mean square error between the corresponding GCPs on the optical and SAR images is less than one pixel. The final spatial resolution of the coregistered Sentinel-1 and Sentinel-2 images used is 10 m.

The WHU-IS dataset constructed in this study is annotated pixel-by-pixel. The specific annotation methods are as follows: First, with the assistance of higher spatial resolution Google Earth images, pixels in images are annotated as pervious surface or IS through visual interpretation, in which pervious surface is marked with 0 and IS is marked with 1. The annotation tool is Adobe Photoshop software. After the completion of the preliminary annotation, manual verification is carried out many times to modify the misclassification and mislabeling on the edge and finally integrated into the IS dataset. After data annotation is completed, we cut the whole image according to the size of 128 × 128 image blocks and the step size of 80 pixels, and finally obtain 5603 patches. Finally, the abovementioned patches are randomly divided into training, test, and validation sets according to the ratio of 8:1:1. In addition, in order to prevent the overfitting and enhance the robustness of the model, we perform data augmentation on the training set, including vertical flip, random rotation, and Gaussian blur, and no data augmentation is executed on the test and validation sets. Fig. 4 shows some examples of the constructed WHU-IS dataset.

III. METHODOLOGY

In this section, the proposed CroFuseNet is mainly introduced. Section III-A shows the overall structure of the proposed CroFuseNet model in detail, and Section III-B presents the constructed loss function in this study.

A. Architecture of the Proposed CroFuseNet

Fig. 5 illustrates the overall framework of the proposed CroFuseNet in this study. Considering the good performance and robustness of UNet, CroFuseNet is designed as an encoder–decoder network with UNet as the backbone network. For multimodal remote sensing data, multimodal features can be divided into shared and specific features, especially for heterogeneous data sources, such as optical and SAR images [21]. Therefore, inspired by this, CroFuseNet network is designed to be something like a three branches network, namely optical image specific features learning subnetwork, SAR image specific features learning subnetwork, and optical and SAR images shared features learning subnetwork. Optical and SAR images are first fed into their own specific features learning subnetworks to learn their specific multiscale features, respectively. Then the learned specific features are cross fused through CFM and input into the encoder of the shared features learning subnetwork. Finally, IS results are generated by decoders of the specific and shared features learning subnetworks, respectively. Note that the final output of CroFuseNet, that is, the IS extraction result, corresponds to the output of the shared features learning subnetwork. At the same time, in order to make better use of the specific high-level features learned by the decoders of the optical and SAR images specific features learning subnetworks, these features are aggregated through MFA and input into the decoder of the optical and SAR images shared features learning subnetwork to further improve the IS extraction accuracy. The encoder and decoder are connected by skip connection, which can better realize the fusion of shallow and deep features. The key parts of CroFuseNet are described in detail as follows.

1) Optical and SAR Images Specific Features Learning Subnetworks: The purpose of specific features learning subnetworks is to enable us to learn effective and powerful individual features for IS extraction by preserving specific properties of optical and SAR images. As shown in Fig. 5, the specific features learning subnetworks of optical and SAR images are the standard UNet structure, which is first proposed to solve the segmentation of biomedical images [20]. UNet is a symmetrical U-shaped encoder–decoder network that can process images of any size. The encoder consists of a series of convolution and max pooling operations to obtain multiscale features of the image, highlighting the global information of the image, and the decoder is correspondingly composed of a series of convolution and upsampling operations, which gradually restore the feature maps output by the encoder to the original image size, and predict the category to which each pixel belongs. However, on the encoder side, the max pooling operation will inevitably lose spatial details information, and only the upsampling operation on the

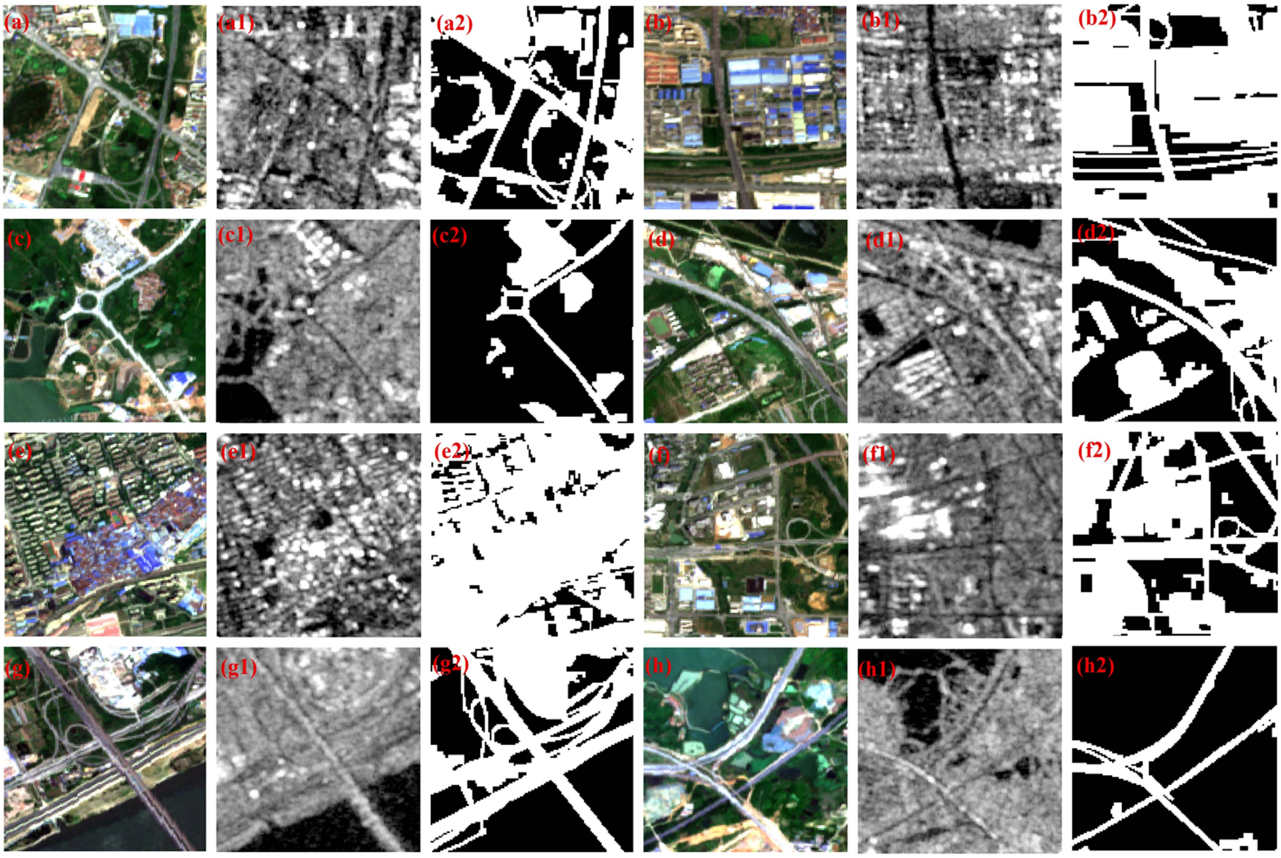

Fig. 4. Some examples of the WHU-IS dataset: (a)–(h) are Sentinel-2 images, (a1)–(h1) are Sentinel-1 images, and (a2)–(h2) are the corresponding ground truth of IS.

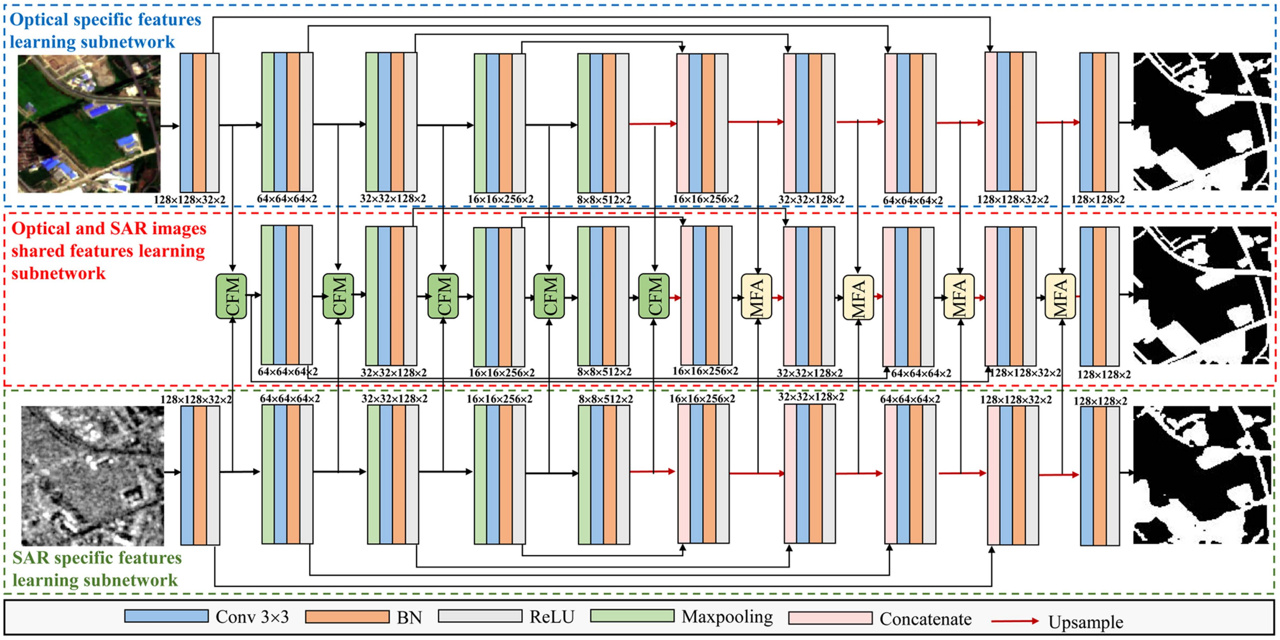

Fig. 5. Architecture of the proposed CroFuseNet.

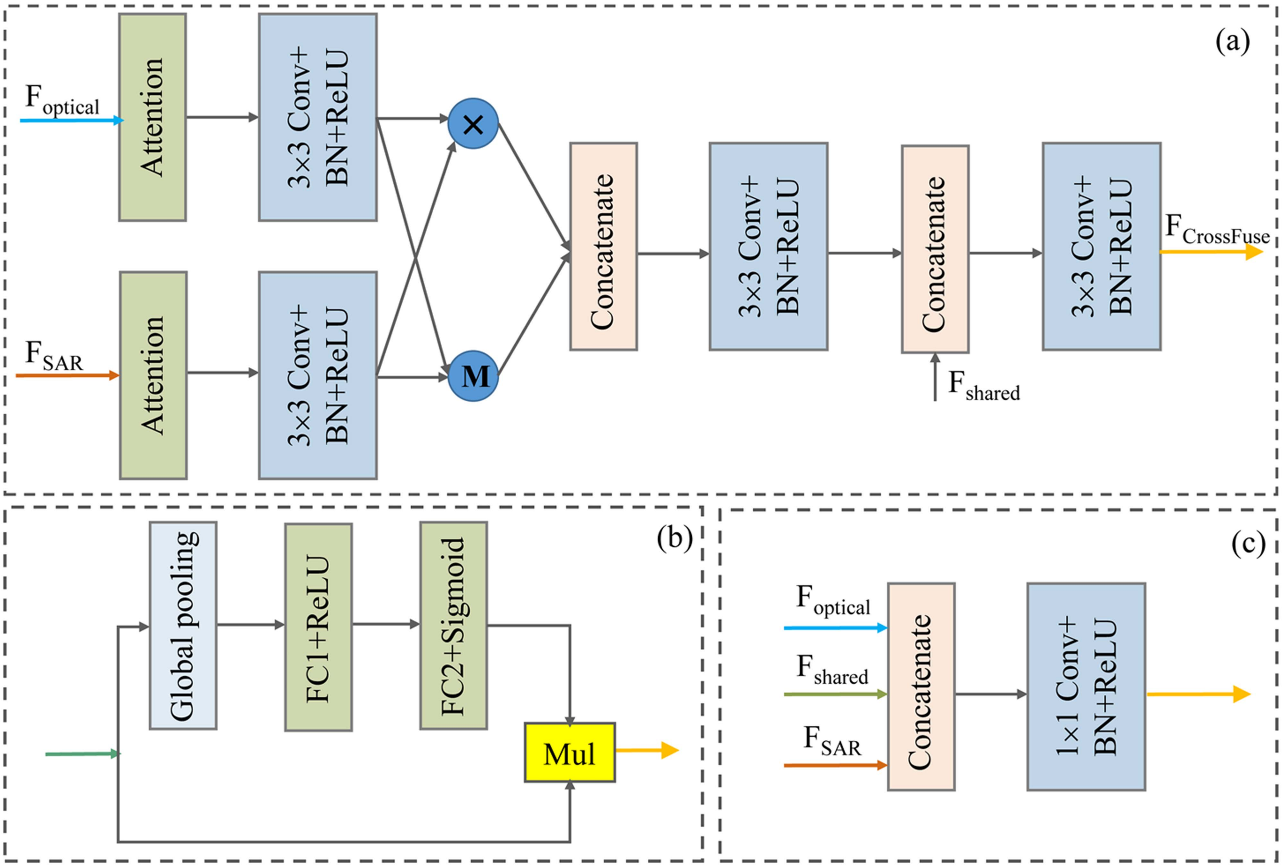

Fig. 6. Diagram of the designed CFM and MFA modules. (a) Diagram of the CFM. Here, "x" and "M" denote the elementwise multiplication and maximization calculation, respectively. (b) Channel attention module used in the CFM. (c) Diagram of the MFA.

decoder cannot better recover the lost spatial details information, and the lost spatial details are conducive to producing good segmentation results. To this end, in order to better reuse the spatial details lost by the max pooling operation in the encoder, the encoder and decoder are connected by skip connection, and the more accurate gradient, point, line, and other information in the encoder of the same level is concatenated directly to the decoder of the same level. Based on this design, the multiscale features and spatial details information are well combined to generate the image segmentation map. Due to its excellent performance, UNet is no longer limited to the segmentation of medical images, and has been widely used in the field of remote sensing [23], [24], [25].

In this study, the sizes of the input optical and SAR images are 128×128, where the channel of optical image is four and that of SAR image is two. The encoder of UNet used in this study consists of five blocks, each of which is composed of two groups of 3×3 convolution, batch normalization (BN) and 2×2 max pooling layers. Therefore, the spatial sizes of the feature maps are reduced from 128×128 to 8×8, and the channels of the feature maps are increased to 512. Correspondingly, the decoder also has four upsampling and four convolution blocks to restore the spatial size of the feature maps to 128×128 and the channel of the output is 32. As this study only involves two categories of IS and pervious surface, the last layer is composed of a 3×3 convolution, BN and a rectified linear unit (ReLU) layer to generate the final segmentation result, which has the same size as the input image, i.e., 128×128×2 (2 represents the number of categories).

2) Optical and SAR Images Shared Features Learning Subnetwork: The main purpose of the shared features learning subnetwork of optical and SAR images is to learn the shared features of optical and SAR images to generate the final IS result. As shown in Fig. 5, like the specific features learning subnetworks, optical, and SAR images shared features learning subnetwork is also on the basis of UNet. The difference is that in the encoder and decoder of the shared features learning subnetwork, the CFM and MFA are designed to fuse and aggregate the learned specific features from the optical and SAR images specific features learning subnetworks, respectively. The detailed structures of CFM and MFA are as follows.

Cross fusion module: The CFM is mainly used to fuse optical and SAR features, and its detailed structure is shown in Fig. 6(a). The F_{Optical} and F_{SAR} represent features learned from optical and SAR images, respectively. To make full use of the key features of optical and SAR images and improve the ability to capture semantic information, a channel attention module is used in the CFM, as shown in Fig. 6(b). After obtaining features with different importance, we then use a 3×3 convolution, BN and ReLU to generate normalized feature maps for features fusion. In deep learning, elementwise multiplication and maximum method are the most commonly used features fusion methods [26]. Generally, it is difficult to determine which of the two fusion methods is better. Therefore, instead of using contenate directly for feature fusion, this study takes advantage of the both features fusion strategies. We first use elementwise multiplication and maximization to fuse the features from optical and SAR images, respectively, and then concatenate the fused features obtained by the two fusion strategies and input them to the next layer of the network. The fusion process can be expressed by the following formula.

$$F_{\text{Optical\_attented}} = \text{Attention}\left(F_{\text{Optical}}\right) \tag{1}$$

$$F_{SAR\_\text{attented}} = \text{Attention}\left(F_{\text{SAR}}\right) \tag{2}$$

$$\begin{aligned} F_{\text{mul}} = {} & \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times 3}\left(F_{\text{Optical\_attented}}\right)\right)\right) \\ & \otimes \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times 3}\left(F_{\text{SAR\_attented}}\right)\right)\right) \end{aligned} \tag{3}$$

$$\begin{aligned} F_{\max} = {} & \text{Max}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times 3}\left(F_{\text{Optical\_attented}}\right)\right)\right)\right. \\ & \left.\text{ReLU}\left(\text{BN}\left(\text{Conv}_{3\times 3}\left(F_{\text{SAR\_attented}}\right)\right)\right)\right) \end{aligned} \tag{4}$$

$$F_{\text{fuse}} = \text{Con}(F_{\text{mul}}, F_{\max}) \tag{5}$$

where, Attention(•) is the channel attention module. $\text{Conv}_{3\times 3}$(•), BN(•), and ReLU(•) represent a 3×3 convolution, BN operations, and activation function, respectively. $\otimes$ and Max(•) indicate elementwise multiplication and maximum calculation. Con(•) is the concatenation operation.

After the fused features F_{fuse} are obtained, a 3×3 convolution, BN, and ReLU layers are followed to capture the relationship between the features obtained by the two fusion strategies, namely elementwise multiplication and maximum calculation. In addition, in order to make full use of the multiscale features learned by the encoder in the shared subnetwork, these features are fused with the features (F_{shared}) of the previous layer of the shared features learning subnetwork encoder through concatenation, and then they are also input into 3×3 convolution, BN, and ReLU layers to generate the final fused features $F_{\text{CrossFuse}}$. According to the fusion process described previously, the CFM can take advantage of the interaction between optical and SAR images through this cross calculation method, and the features of the CFM output are propagated to the next layer of the shared features learning subnetwork to capture information at different scales.

Channel attention is a new and commonly used deep learning strategy that makes feature selection by assigning different global weights to different channels. Assuming that F is the input feature, and its shape is $C \times H \times W$ (C, H, and W are the channel, height, and width of the feature maps, respectively) and M is global weight vector, which can be obtained through global average pooling or global maximum pooling, and its shape is $C \times 1 \times 1$, then the output feature F' of channel attention module can be simplified as

$$F' = F \otimes M \tag{6}$$

where, $\otimes$ indicates the elementwise multiplication of matrixes.

Motivated by SeNet [27], the channel attention module used in this study first codes the entire spatial features of each channel into a global feature through global average pooling, and then uses two fully connected layers to capture the relationship between the global features of each channel. The first full connection layer is used to reduce the channel dimension and is then activated using ReLU. The second fully connected layer is used to restore the channel to its original size, and is activated by the sigmoid function ranging from 0 to 1 to obtain the global weight vector, and then multiply the input features to obtain the features after attention enhancement. The process of the used channel attention module can be expressed as follows:

$$F_{\text{Sq}} = \text{Gpooling}(F) \tag{7}$$

$$F_{\text{Ex1}} = \text{ReLU}\left(\text{FC1}(F_{\text{Sq}})\right) \tag{8}$$

$$F_{\text{Ex}} = \text{Sigmoid}\left(\text{FC2}\left(F_{\text{Ex1}}\right)\right) \tag{9}$$

$$F' = F \otimes F_{\text{Ex}} \tag{10}$$

where, Gpooling(•) represents the global average pooling. FC1(•) is the first fully connection layer, the shape of the input feature is $C \times 1 \times 1$, and the shape of the output feature is $\frac{\text{C}}{r} \times 1 \times 1$. FC2(•) is the second fully connection layer, the shape of the input feature is $\frac{\text{C}}{r} \times 1 \times 1$, and the shape of the output feature is $C \times 1 \times 1$. In this study, r is set to 16 according to the experiment. F' is the attention-enhanced feature with a shape of $C \times H \times W$.

MFA block: The decoder in the encoder–decoder network mainly learns the high-level features of the image. Generally, the correlation between the high-level features of multimodal data is very low. Therefore, in this study, CFM is not used for features fusion for the specific high-level features learned by the decoder of the optical and SAR images specific features learning subnetworks. But, in order to make full use of the learned features of optical and SAR images specific features learning subnetwork decoders, we design a simple but effective MFA module in the shared features learning subnetwork decoder, as shown in Fig. 6(c). We first fuse the specific features learned from the optical and SAR images specific subnetworks with the features learned from the shared subnetworks using a concatenation operation, then reduce the channel dimension of the feature maps through a 1×1 convolution, and finally refine the fusion results by using BN and ReLU layers. Hence, we can obtain a fusion feature map, which can obtain sufficient semantic information from the commonness and specificity between optical and SAR images.

B. Loss Function

Regarding the common constraints of optical and SAR images on the model, this study constructs a comprehensive loss function, which consists of three parts, namely, SAR image specific features learning subnetwork loss L_{SAR}, optical image specific features learning subnetwork loss L_{optical}, and optical and SAR images shared features learning subnetwork loss L_{shared}. The specific expression is as follows:

$$\begin{aligned} \text{Loss}_{\text{total}} = {} & L_{\text{shared}}(S_{\text{shared}}, G) + \alpha L_{\text{optical}}(S_{\text{optical}}, G) \\ & + \beta L_{\text{SAR}}(S_{\text{SAR}}, G) \end{aligned} \tag{11}$$

where, S_{shared}, S_{optical}, and S_{SAR} are the IS prediction results of the shared features learning subnetwork, the optical image specific features learning subnetwork and the SAR image specific features learning subnetwork, respectively. G represents the ground truth of IS. α and β are the weight parameters, which are used to control the tradeoff between different losses. During the IS extraction based on optical and SAR images fusion, optical image generally plays a leading role, while SAR image plays an auxiliary role. Therefore, the loss of optical specific learning subnetwork is given a greater weight in this study, namely $\alpha = 1$, $\beta = 0.8$.

In this stuty, the extraction of IS is actually a dichotomous problem. Therefore, we use cross entropy loss for the L_{shared}, L_{optical} and L_{SAR} in (11). For a dichotomous problem, the cross entropy loss can be expressed as follows:

$$L = \frac{1}{N}\sum_{i} - [y_i \bullet \log(p_i) + (1 - y_i) \bullet \log(1 - p_i)] \quad (12)$$

where, N is the number of samples. y_i indicates the label of the ith sample. p_i represents the probability that the ith sample is predicted to be positive class. In this study, the positive class is IS and labeled with 1, while the negative class is pervious surface and labeled with 0.

IV. Experimental Results

To demonstrate the effectiveness of the proposed CroFuseNet, WHU-IS is used as the dataset for the experiment in this section. Section IV-A introduces the experimental setups of this study in detail, including model training, comparative methods, and evaluation metrics. Section IV-B provides IS extraction results qualitatively and quantitatively. Finally, we will conduct ablation studies in Section IV-C.

A. Experimental Setups

1) Model Training: We train our network for 400 epochs and set the batch size to 32. The Adam optimizer is selected to optimize the object loss function and update the parameters of each layer in the network. During the training of the model, we use the poly learning rate strategy to update the learning rate, and its initial learning rate is set as 0.0001. For each iteration, the learning rate is adjusted by multiplying the initial learning rate by $(1 - \frac{\text{iter}}{\max\_\text{iter}})^{0.9}$. Among which, iter and $\max_$iter represents the current number of iterations and the maximum number of iterations, respectively. In addition, in order to effectively avoid over fitting of the model, this study uses the early stop strategy. If the accuracy of the validation set for 50 consecutive epochs is not improved, the model will stop training in advance. All of the methods, including our proposed method and other comparative deep learning methods, are executed using the PyTorch framework and run on a station with 11 GB of RAM and an NVIDIA GeForce RTX 2080TI.

2) Comparative Methods: In this study, we select six comparative methods, including two traditional machine learning algorithms and four state-of-the-art deep learning methods, to compare with the proposed CroFuseNet. Random forest (RF) [36] and SVM [37] algorithms are the two most widely used and effective machine learning algorithms in IS extraction, so they are selected as the comparison algorithms in this study [10]. Furthermore, many outstanding semantic segmentation networks have been proposed so far. Based on the comprehensive consideration of model performance and open source code, this study finally choose U-Net [20], FCN [28], HRNet [29], and Deeplabv3+ [30] as the comparison methods based on deep learning. To compare them fairly, these comparison models are retrained using the WHU-IS dataset constructed in this study, and their model structures are consistent with those of the original references, and their hyperparameters are consistent with those of the proposed model. Note that the input of these comparison methods is multichannel data formed by directly concatenating optical and SAR images.

3) Evaluation Metrics: In order to comprehensively analyze the performance of the proposed CroFuseNet, we adopt five commonly used evaluation metrics in IS extraction, including the intersection over union (IoU), user's accuracy (UA), producer's accuracy (PA), F1-Score, and overall accuracy (OA). Their calculation formulas are as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \quad (13)$$

$$\text{UA} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (14)$$

$$\text{PA} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (15)$$

$$F1 - \text{Score} = \frac{2 \bullet \text{PA} \bullet \text{UA}}{\text{PA} + \text{UA}} \quad (16)$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (17)$$

where, TP, FP, FN, and TN represent the numbers of pixels that are true-positives, false-positives, false-negatives, and true-negatives for each class, respectively. Normally, there are some conflicts between these metrics, such as PA and UA, so we add F1-Score and IoU for IS evaluation. Compared with PA and UA, F1-Score can more comprehensively evaluate the extraction accuracy of IS. Besides, in order to evaluate the comprehensive performance of these methods, we select the average of F1-Score, OA, and the mean intersection over union (MIoU) to conduct comprehensive analyses. Among them, MIoU is also the metric that we use to judge whether to stop training in advance during model training.

B. Results of IS Extraction

Fig. 7 qualitatively shows the IS extraction results of the proposed CroFuseNet and six comparative methods under different scenarios. From this figure, it can be seen that the IS results extracted by the two traditional machine learning algorithms, RF and SVM, are significantly worse than those based on deep learning methods, especially in places with complex scenes, as shown in Fig. 7(a), (b), and (c). RF and SVM are poor at preserving details of ground objects, such as roads and building edges. In addition, some bare soils are misclassified as IS. However, for some scenarios with low heterogeneity, as shown in Fig. 7(d), RF and SVM can also achieve comparative results. Although FCN can reflect the basic distribution of IS,

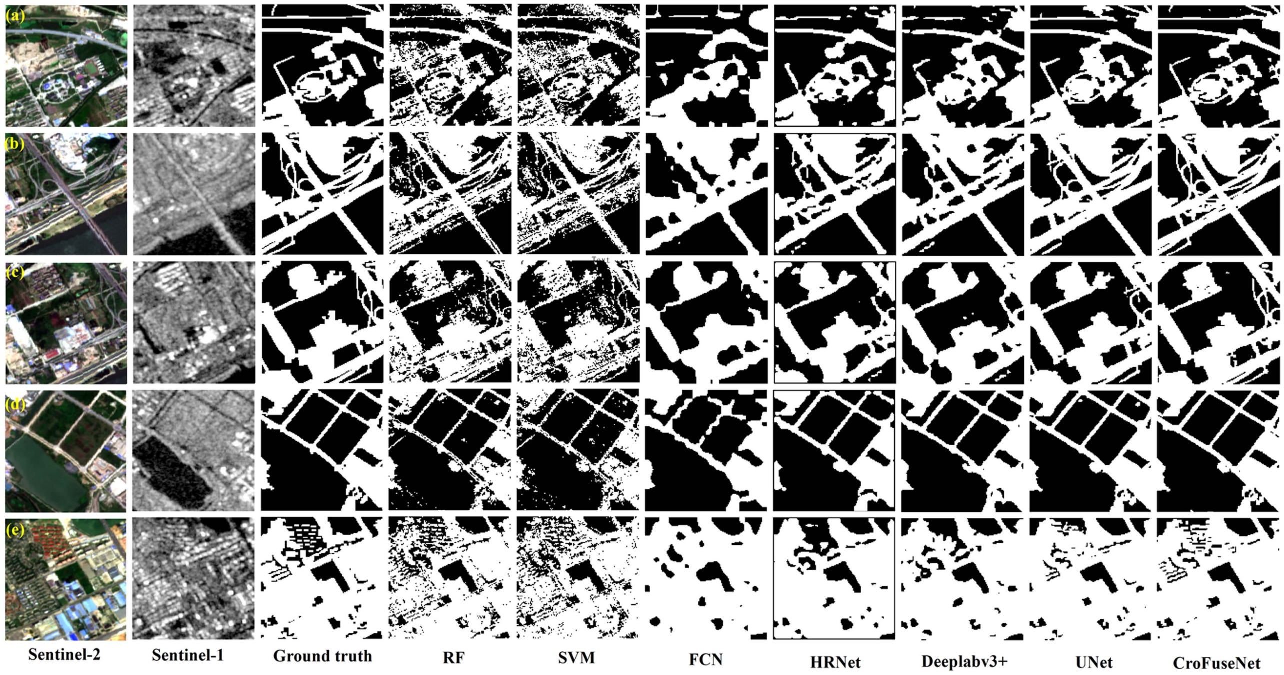

Fig. 7. Qualitative IS extraction results of the proposed CroFuseNet and six comparative methods.

TABLE II
COMPREHENSIVE ACCURACY OF IS EXTRACTION RESULTS EXTRACTED BY THE PROPOSED CROFUSENET AND SIX COMPARATIVE METHODS

Metrics	RF	SVM	FCN	HRNet	Deeplabv3+	UNet	CroFuseNet
MIoU	0.7890	0.7413	0.8741	0.8951	0.9065	0.9355	**0.9495**
OA	89.74%	86.96%	94.17%	95.26%	95.75%	97.12%	**97.77%**
F1-Score	0.8802	0.8487	0.9322	0.9442	0.9506	0.9665	**0.9770**

its results are too smooth and lack reservation of spatial details. For example, it can hardly extract roads. Compared with FCN, HRNet, and Deeplabv3+ provide slightly better visual results than FCN, but also fail to extract tiny IS like roads. UNet and the proposed CroFuseNet have the best results in the extraction of IS. Through careful observation, it was discovered that CroFuseNet has more advantages than UNet. For example, compared with UNet, CroFuseNet has a more complete extraction of roads, as shown in Fig. 7(a), (b), and (c), and a better extraction of single-family buildings, as shown in Fig. 7(e). In addition, CroFuseNet has fewer misclassification than UNet.

To better demonstrate the effectiveness of the proposed method, Table II quantitatively presents three comprehensive accuracy evaluation metrics of IS results extracted by the proposed CroFuseNet and six comparative methods in this study, including MIoU, OA, and the average of F1-Score, among which the best results are highlighted in bold. From the evaluation of MIoU, OA, and F1-Score, it can be seen that the proposed CroFuseNet has the best performance. Its MIoU, OA, and F1-Score are 0.014, 0.65%, and 0.0105 higher than those of the second ranked UNet, respectively. The MIoU, OA, and F1-Score of the CroFuseNet are 0.2065, 10.18%, and 0.1283 higher than that of the SVM, which is the worst, respectively. In conclusion, the proposed CroFuseNet has the highest accuracy, followed by UNet, Deeplabv3+, HRNet, FCN, RF, and SVM, which is consistent with the qualitative results shown in Fig. 7. Table III quantitatively shows the accuracy of each class, and the best results are also highlighted in bold. The results shown in the Table III are consistent with that in Table II, that is, CroFuseNet has the highest accuracy in the extraction of IS and pervious surface. In addition, it has been found that the extraction accuracy of pervious surface is higher than that of IS. Combining the quantitative and qualitative results of IS extraction, the effectiveness and superiority of the proposed CroFuseNet are demonstrated.

C. Ablation Studies

In the proposed CroFuseNet, we design two key modules, namely CFM and MFA, to fuse and aggregate the learned features from the specific features learning subnetworks of optical and SAR images. To demonstrate the effectiveness of these two modules, four ablation experiments are carried out and compared with the proposed model. Case 1: CFM and MFA modules are removed from CroFuseNet. Case 2: CFM module is only removed from CroFuseNet. Case 3: MFA module is only removed from CroFuseNet. Case 4: CFM module in CroFuseNet is replaced by concatenation operation. Case 5: the proposed CroFuseNet. Fig. 8 qualitatively shows the IS extraction results of the ablation experiments. Visually, the differences between these results are

TABLE III
QUANTITATIVE COMPARISON OF EACH TYPE OF EXTRACTION RESULTS BETWEEN THE PROPOSED CROFUSENET AND SIX COMPARATIVE METHODS

Land cover type	Impervious surface				Pervious surface			
Metrics	IoU	PA	UA	F1-Score	IoU	PA	UA	F1-Score
RF	0.7165	83.67%	83.31%	0.8349	0.8614	92.46%	92.64%	0.9255
SVM	0.6563	78.58%	79.93%	0.7925	0.8264	90.85%	90.14%	0.9079
FCN	0.8297	90.18%	91.2%	0.9069	0.9185	96%	95.51%	0.9575
HRNet	0.8563	93.78%	90.19%	0.9226	0.9339	95.89%	97.28%	0.9658
Deeplabv3+	0.873	92.73%	93.71%	0.9322	0.94	97.14%	96.68%	0.9691
UNet	0.9122	**94.67%**	96.15%	0.9541	0.9588	98.25%	97.55%	0.979
CroFuseNet	**0.931**	96.14%	**96.72%**	**0.9641**	**0.968**	**98.51%**	**98.24%**	**0.9838**

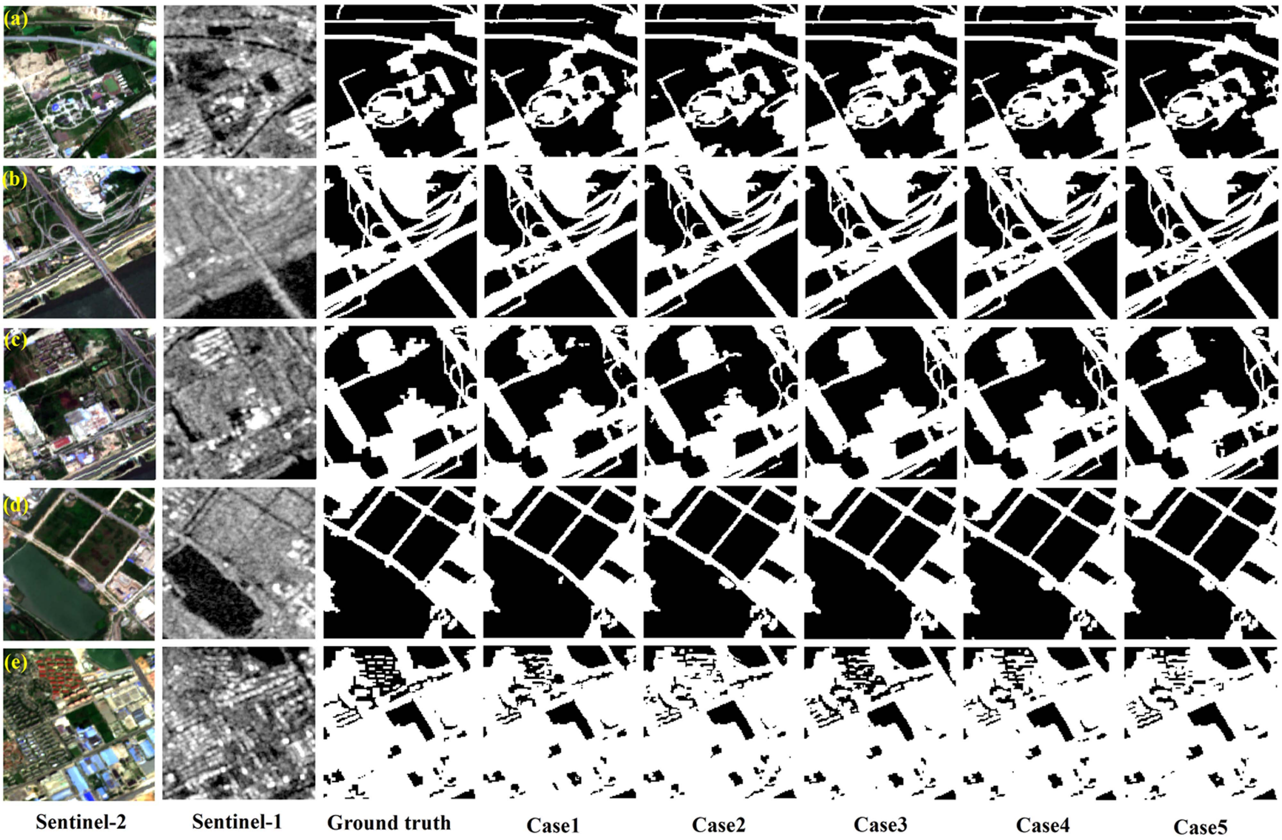

Fig. 8. Qualitative IS extraction results for ablation studies.

not significant. However, some differences can be discovered through careful comparison. For Case 1, the larger scale IS is better extracted, but some smaller scale IS, such as roads, have not been extracted [see Fig. 8(c)], which shows that the design of CFM fusing multiscale features from optical and SAR images is effective and reasonable. In addition, the phenomenon that bare soils are mistakenly classified as IS in Case 1 is relatively obvious [see Fig. 8(a)]. Similarly, due to the lack of CFM, some small roads are not extracted in Case 2, as shown in Fig. 8(b) and (c). However, compared with Case 1, the misclassification of bare soils and IS has been improved, indicating the effectiveness of MFA. For Case 3, which lacks MFA, small roads have been basically extracted, but compared with Case 2, there is a significant increase in the misclassification of other ground objects as IS. Case 4 has the best results in ablation experiments and is the closest to the proposed CroFuseNet. Compared with the other three ablation experiments, the misclassification and extraction of small ground objects have been significantly improved. However, when compared with Case 5, it is found that Case 5 is better at extracting small ground objects, and IS is likewise fewer misclassified, which indicates that the cross fusion strategy used in this study is more effective than the concatenation fusion strategy.

In addition, Table IV also quantitatively presents the accuracy evaluation of the abovementioned ablation experimental results. According to the results of the abovementioned qualitative analysis, the model proposed in this study has the best results, followed by Case 4, Case 2, Case 3, and Case 1. The abovementioned experimental results again prove the effectiveness of CFM and MFA designed in this study, and show that the cross fusion by considering the interaction between optical and SAR images is more effective than simple concatenate fusion, which is consistent with the motivation of this study.

TABLE IV
QUANTITATIVE EVALUATION FOR ABLATION STUDIES

	Impervious surface				Pervious surface				Comprehensive analysis		
Metrics	IoU	PA	UA	F1-Score	IoU	PA	UA	F1-Score	MIoU	OA	F1-Score
Case1	0.9153	95.21%	95.95%	0.9558	0.9606	98.16%	97.82%	0.9799	0.9379	97.23%	0.9678
Case2	0.9287	96.02%	96.59%	0.963	0.967	98.45%	98.19 %	0.9832	0.9479	97.69%	0.9731
Case3	0.922	95.53%	96.36%	0.9594	0.9637	98.35%	97.96%	0.9815	0.9429	97.46%	0.9705
Case4	0.929	**96.2%**	96.44%	0.9632	0.9672	98.39%	**98.28%**	0.9833	0.9481	97.7%	0.9732
Case5	**0.931**	96.14%	**96.72%**	**0.9641**	**0.968**	**98.51%**	98.24%	**0.9838**	**0.9495**	**97.77%**	**0.974**

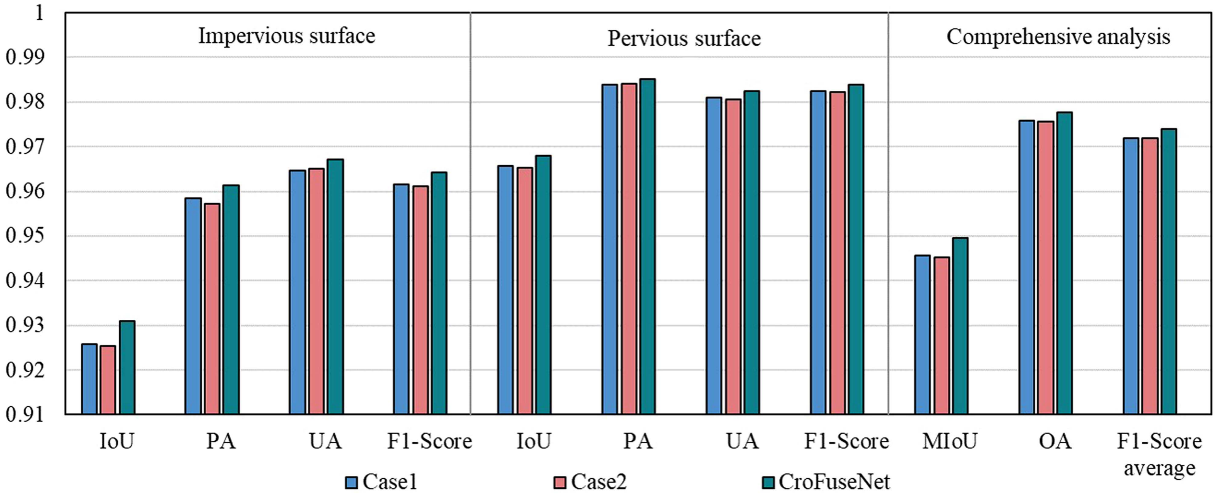

Fig. 9. Performance of the comprehensive loss function designed in this study.

V. DISCUSSION

In this section, we first discuss the performance of the comprehensive loss function designed in this study, and then compare the IS extraction results based on the fusion of optical and SAR images using CroFuseNet with the results using optical and SAR images alone to discuss their roles in IS extraction. Finally, we qualitatively discuss the spatial and temporal transferability of the proposed CroFuseNet.

A. Performance of the Comprehensive Loss Function Designed

Considering the common constraints of optical and SAR images, designing a comprehensive loss function is another important strategy to improve the extraction accuracy of IS in this study. In order to prove the effectiveness and superiority of the designed comprehensive loss function, we carry out two groups of comparative experiments. Case 1: the common constraints of optical and SAR images are not considered, and only the loss of optical and SAR images in the shared features learning subnetwork of CroFuseNet is considered. Case 2: the common constraints of optical and SAR images are considered, but optical and SAR images are regarded as equally important and given the same weight, that is, α and β in (11) are given the same value. Fig. 9 shows the accuracies comparison of Case 1, Case 2, and the proposed CroFuseNet. We find that the performance of the proposed CroFuseNet is significantly better than that of Case 1 and Case 2 in all metrics. Interestingly, the extraction accuracy of Case 2 is even lower than that of Case 1. The reason may be that when the loss of optical and SAR images is given the same weight, the constraint of SAR image will be too strong. However, due to the imaging mechanism of SAR image itself, the inherent quality of SAR image will affect the extraction accuracy of IS. Therefore, according to the abovementioned results, it is reasonable to comprehensively consider the common constraints of optical and SAR images, and give different weights according to the difference between their primary and secondary roles, which can significantly improve the extraction accuracy of IS based on optical and SAR image fusion.

B. Benefits of Optical and SAR Images Fusion

Improving the extraction accuracy is the original intention of extracting IS by fusing optical and SAR images. Therefore, in order to better illustrate the benefits of optical and SAR images fusion, we separately use optical and SAR images alone for IS extraction and compare their results with that extracted from the proposed CroFuseNet. The extraction results are shown in Fig. 10. It is worth noting that the model proposed in this study takes UNet as the backbone network, so when using optical and SAR images alone to extract IS, the optical and SAR images are input into the UNet, respectively, and other parameters remain unchanged. It can be seen from Fig. 10 that for IS of building types, they can be basically extracted using SAR images alone [see Fig. 10(c) and (d)], but the edges of roads and buildings can not be well preserved [see Fig. 10(a), (b), (d), and (e)]. Compared with the results extracted by using the SAR image alone, the extraction results of using optical images

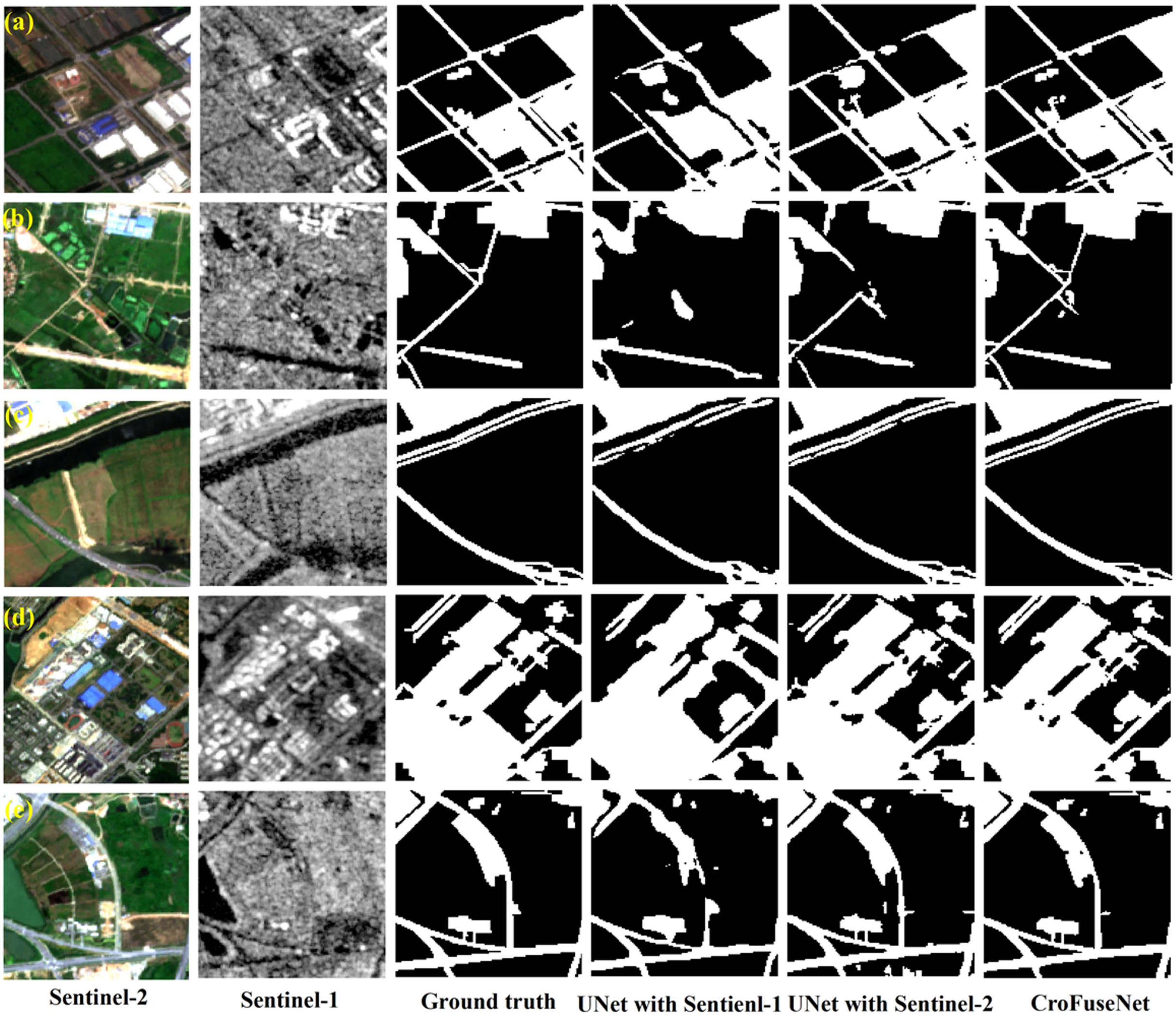

Fig. 10. Qualitative IS extraction results of UNet based on Sentinel-1 alone, UNet based on Sentinel-2 alone and the proposed CroFuseNet based on Sentinel-1 and Sentinel-2 images fusion.

alone are obviously better. The edges of roads and buildings are well preserved, which is the advantage of optical images. However, when extracting IS based on optical images alone, even using data-driven deep learning methods can not solve the limitation of spectral similarity on the extraction of IS. The IS extraction results based on the fusion of optical and SAR images using the proposed model in this study is significantly better than that of using optical or SAR images alone, whether in terms of preserving the details of ground objects or in terms of misclassification.

Fig. 11 quantitatively shows the comparison of the IS results of the UNet model based solely on Sentinel-1, the IS results of the UNet model based solely on Sentinel-2, and the IS extraction results of the CroFuseNet model based on the fusion of Sentinel-1 and Sentinel-2 images. Consistent with the qualitative results provided in Fig. 10, the IS extraction results of optical and SAR images fusion are the best, followed by the results of using optical images alone and SAR images alone, which is also consistent with previous studies [10], [31]. The MIoU, OA, and the average of the F1-Score of the CroFuseNet extraction results are 0.0136, 0.63%, and 0.0073 higher than those of extraction results of using optical images alone, and are 0.0937, 4.54%, and 0.0525 higher than those of extraction results of using SAR images alone, respectively. The abovementioned results show that the fusion of optical and SAR images can significantly improve the accuracy of IS extraction. In addition, the IS result of using SAR image alone is worse than that of using optical image alone, which indicates that optical image plays a greater role in the extraction of IS. Therefore, it is reasonable to give greater weight to optical image when designing the comprehensive loss function in this study.

C. Spatial and Temporal Transferability of the Proposed CroFuseNet

Because pixel-by-pixel data annotation is time-consuming and labor-intensive, the transferability of the model is most concerned by researchers and users, including spatial and temporal transfer. Therefore, due to the lack of labeled data, we qualitatively discuss the temporal and spatial transfer

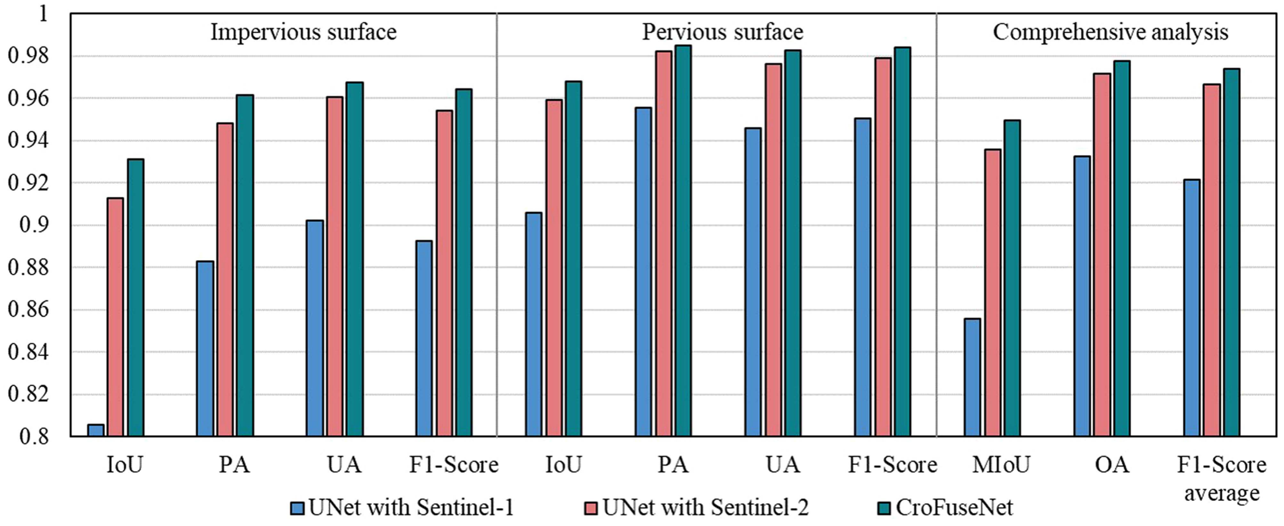

Fig. 11. Quatitative IS extraction results of UNet based on Sentinel-1 alone, UNet based on Sentinel-2 alone and the proposed CroFuseNet based on Sentinel-1 and Sentinel-2 images fusion.

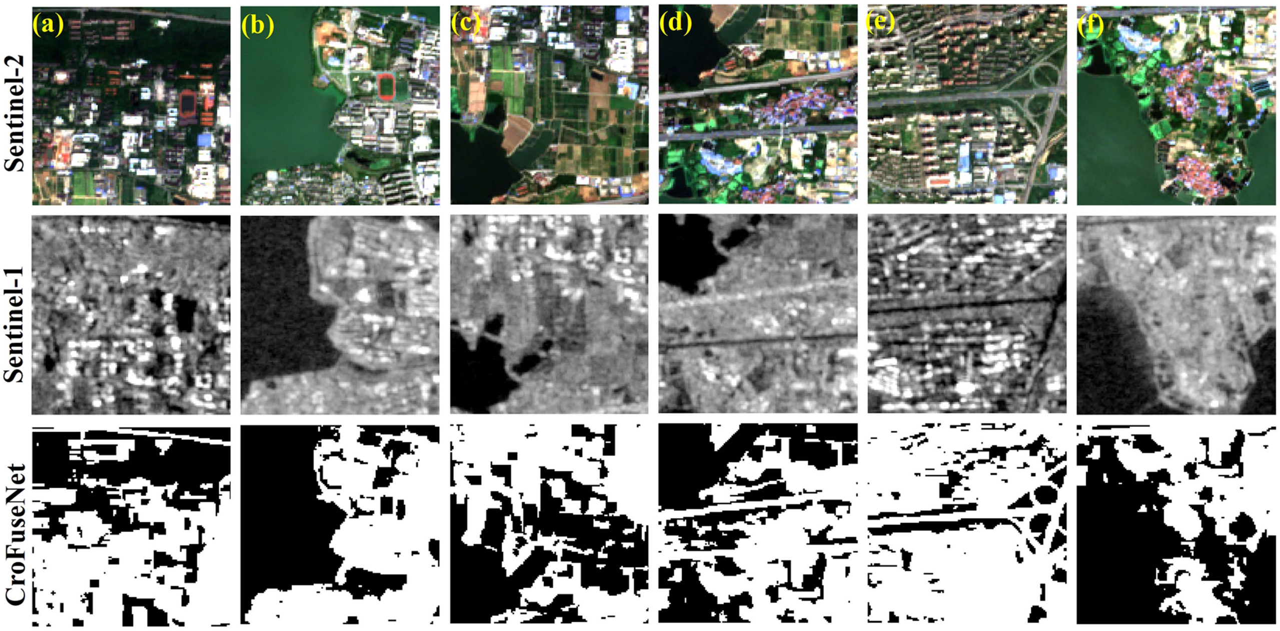

Fig. 12. Qualitative results of IS in Wuhan in 2020 extracted by the proposed CroFuseNet.

capabilities of the proposed CroFuseNet in this study. In the temporal transfer experiment, we use Sentinel-1 and Sentinel-2 data covering the same area as this study from June 1 to June 30, 2020 as the data source. The IS extraction results are shown in Fig. 12. We find that the proposed CroFuseNet can reflect the overall spatial distribution of IS and has a certain time transfer ability. However, by comparing with the corresponding optical images, we find that the misclassification between other ground objects and IS is relatively obvious [see Fig. 12(b), (c), and (d)], especially the misclassification of bare soils and IS. Interestingly, roads and other small ground objects have still been well extracted, which again shows the effectiveness of the model proposed in this study. In the spatial transfer experiment, considering the influence of weather, we take Nanjing as the study area, and Sentinel-1 and Sentinel-2 from May 1 to August 30, 2019 as the data source. The results are shown in Fig. 13. Similarly, although the proposed model can reflect the overall spatial distribution of IS and has a certain spatial transfer ability, the phenomenon of misclassification is quite obvious, and the extracted results are not as fine as those extracted under samples supervision. The spatial and temporal transfer capabilities of the proposed model is still a difficult point that limits the popularity and application of the deep learning model. In the future research, we can consider using the idea of the generative adversarial networks (GAN) to improve the transferability of the model.

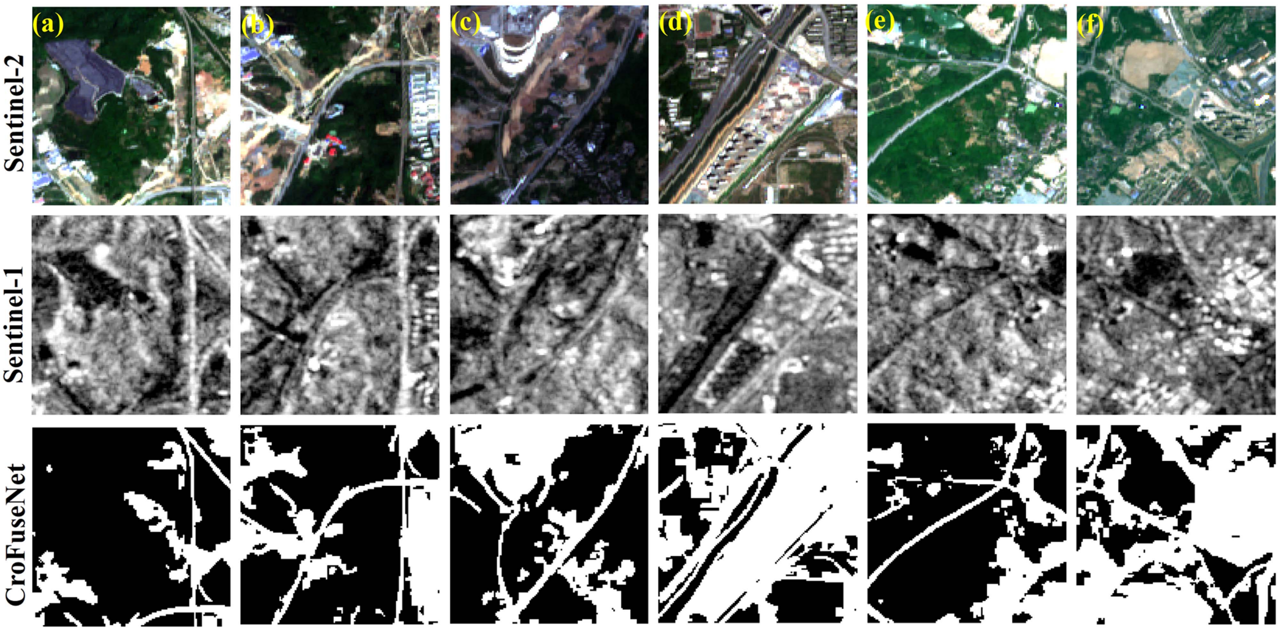

Fig. 13. Qualitative results of IS in Nanjing in 2019 extracted by the proposed CroFuseNet.

VI. Conclusion

To extract urban IS accurately, this study proposes a semantic segmentation network based on optical and SAR images fusion, namely CroFuseNet. In the proposed CroFuseNet, we design a CFM to deeply fuse the multiscale features of optical and SAR images. This module considers the interaction between optical and SAR images, which can achieve better complementarity between the two images, and make up for the shortcomings of concatenation features fusion methods commonly used in the past. Besides, we also propose an MFA module to further integrate the specific high-level features from the decoders of the optical and SAR images specific features learning subnetworks into the decoder of the optical and SAR images shared features learning subnetwork. At the same time, to make up the gap of IS dataset, this study construct an open and accurate IS dataset based on optical and SAR images, called WHU-IS. Then, based on this dataset, we validate the proposed CroFuseNet and compare it with two classical machine learning algorithms and four state-of-the-art deep learning models. The quantitative and qualitative experimental results demonstrate that the proposed model in this study is superior to these comparative algorithms. This study proves the great superiority of deep learning in IS extraction, and can provide data support for developing more advanced IS extraction methods based on deep learning. In the future, we will try to use GAN to further improve the transferability of the proposed model and free people from the heavy work of data annotation.

References

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

[2] United Nations, Department of Economic and Social Affairs, Population Division. World Urbanization Prospects: The 2018 Revision, 2018. [Online]. Available: https://esa.un.org/unpd/wup/Publications

[3] Z. Shao, W. Wu, and D. Li, “Spatio-temporal-spectral observation model for urban remote sensing,” *Geo-Spatial Inf. Sci.*, vol. 24, no. 3, pp. 372–386, 2021.

[4] P. Fu and Q. Weng, “A time series analysis of urbanization induced land use and land cover change and its impact on land surface temperature with Landsat imagery,” *Remote Sens. Environ.*, vol. 175, pp. 205–214, 2016.

[5] Z. Shao et al., “Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation,” *Remote Sens. Environ.*, vol. 232, 2019, Art. no. 111338.

[6] 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.

[7] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassio, “A tutorial on synthetic aperture radar,” *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013.

[8] Y. Lin et al., “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.

[9] Z. Sun et al., “Global 10-m impervious surface area mapping: A big earth data based extraction and updating approach,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 109, 2022, Art. no. 102800.

[10] H. Zhang and R. Xu, “Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the Pearl River Delta,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 64, pp. 87–95, 2018.

[11] Y. Lin et al., “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.

[12] 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.

[13] H. Guo et al., “Synergistic use of optical and PolSAR imagery for urban impervious surface estimation,” *Photogrammetric Eng. Remote Sens.*, vol. 80, no. 1, pp. 91–102, 2014.

[14] Y. Zhang, H. Zhang, and H. Lin, “Improving the impervious surface estimation with combined use of optical and SAR remote sensing images,” *Remote Sens. Environ.*, vol. 141, pp. 155–167, 2014.

[15] G. Sun et al., "Hierarchical fusion of optical and dual-polarized SAR on impervious surface mapping at city scale," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 264–278, 2022.

[16] L. Yang et al., "Quantifying sub-pixel urban impervious surface through fusion of optical and InSAR imagery," *GIScience Remote Sens.*, vol. 46, no. 2, pp. 161–171, 2009.

[17] W. Wu et al., "Extraction of impervious surface using Sentinel-1A time-series coherence images with the aid of a Sentinel-2A image," *Photogrammetric Eng. Remote Sens.*, vol. 87, no. 3, pp. 161–170, 2021.

[18] J. R. Parekh et al., "Automatic detection of impervious surfaces from remotely sensed data using deep learning," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3166.

[19] X. Feng et al., "Integrating Zhuhai-1 hyperspectral imagery with Sentinel-2 multispectral imagery to improve high-resolution impervious surface area mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2410–2424, 2022.

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

[21] D. Hong et al., "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.

[22] A. Mullissa et al., "Sentinel-1 SAR backscatter analysis ready data preparation in Google Earth engine," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1954.

[23] L. Yang et al., "Semantic segmentation based on temporal features: Learning of temporal–spatial information from time-series SAR images for paddy rice mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403216.

[24] F. Wu et al., "Built-up area mapping in China from GF-3 SAR imagery based on the framework of deep learning," *Remote Sens. Environ.*, vol. 262, 2021, Art. no. 112515.

[25] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[26] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving RGB-D saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4661–4671.

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

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[29] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.

[30] L. C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[31] Z. Shao, W. Wu, and S. Guo, "IHS-GTF: A fusion method for optical and synthetic aperture radar data," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2796.

[32] H. Zhang et al., "Explore better network framework for high-resolution optical and SAR image matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4704418.

[33] Y. Xiang, N. Jiao, F. Wang, and H. You, "A Robust two-stage registration algorithm for large optical and SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5218615.

[34] 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, 2022, Art. no. 5235913.

[35] W. Wu et al., "Quantifying the sensitivity of SAR and optical images three-level fusions in land cover classification to registration errors," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 112, 2022, Art. no. 102868.

[36] L. Breiman and R. A. Cutler, "Random forests machine learning," *J. Clin. Microbiol.*, vol. 2, pp 199–228, 2001.

[37] V. Cherkassky and Filip M. Mulier, "Statistical learning theory," *Learn. Data: Concepts, Theory, Methods*, pp. 99–150, 2007, doi: 10.1002/9780470140529.ch4.

[38] L. Sun, S. Cheng, Y. Zheng, Z. Wu, and J. Zhang, "SPANet: Successive pooling attention network for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp 4045–4057, 2022.

[39] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multi-structure KELM with attention fusion strategy for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539217.

[40] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–Spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

Wenfu Wu received the master's degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

His research interests include synthetic aperture radar (SAR) image despeckling, the fusion of optical and SAR images, and the applications of deep learning in remote sensing.

Songjing Guo received the master's degree in surveying and mapping engineering from the China University of Geosciences, Wuhan, China, in 2020, where she is currently working toward the Ph.D. degree in spatial information detection.

Her research interests include ecological carrying capacity assessment and geological hazard detection using remote sensing technology.

Zhenfeng Shao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2004.

Since 2009, he has been a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. He has authored or coauthored more than 50 peer-reviewed articles in international journals. His research interests include high-resolution image processing, pattern recognition, and urban remote sensing applications.

Dr. Shao was a recipient of the Talbert Abrams Award for the Best Paper in Image Matching from the American Society for Photogrammetry and Remote Sensing, in 2014, and the New Century Excellent Talents in University from the Ministry of Education of China, in 2012. Since 2019, he has been an Associate Editor for the photogrammetric engineering and remote sensing specializing in smart cities, photogrammetry, and change detection.

Deren Li received the Ph.D. degree in photogrammetry and remote sensing from the University of Stuttgart, Germany, in 1985.

He is a Scientist in photogrammetry and remote sensing, and also enjoys dual membership of both the Chinese Academy of Sciences, Beijing, China, and the Chinese Academy of Engineering, Beijing, China. He is a Professor and a Ph.D. Supervisor with Wuhan University, Wuhan, China.

Mr. Li is a "National Level Young and Middleaged Expert with Outstanding Contribution," Member of the Euro-Asia International Academy of Science, Member of the 9th National Committee of the Chinese People's Political Consultative Conference.