Cross-modality Features Fusion for Synthetic Aperture Radar Image Segmentation

Fei Gao, Heqing Huang, Zhenyu Yue, Dongyu Li, Shuzhi Sam Ge, *Fellow, IEEE,* Tong Heng Lee and Huiyu Zhou

***Abstract*—Synthetic Aperture Radar (SAR) image segmentation stands as a formidable research frontier within the domain of SAR image interpretation. The fully convolutional network (FCN) methods have recently brought remarkable improvements in SAR image segmentation. Nevertheless, these methods do not utilize the peculiarities of SAR images, leading to suboptimal segmentation accuracy. To address this issue, we rethink SAR image segmentation in terms of sequential information of transformers and cross-modal features. We first discuss the peculiarities of SAR images and extract the mean and texture features utilized as auxiliary features. The extraction of auxiliary features helps unearth the distinctive information in the SAR images. Afterward, a feature-enhanced FCN with the transformer encoder structure, termed FE-FCN, which can be extracted to context-level and pixel-level features. In FE-FCN, the features of a single-mode encoder are aligned and inserted into the model to explore the potential correspondence between modes. We also employ long skip connections to share each modality's distinguishing and particular features. Finally, we present the connection-enhanced conditional random field (CE-CRF) to capture the connection information of the image pixels. Since the CE-CRF utilizes the auxiliary features to enhance the reliability of the connection information, the segmentation results of FE-FCN are further optimized. Comparative experiments conducted on the Fangchenggang (FCG), Pucheng (PC), and Gaofen (GF) SAR datasets. Our method demonstrates superior segmentation accuracy compared to other conventional image segmentation methods, as confirmed by the experimental results.**

***Index Terms*—Synthetic aperture radar, image segmentation, fully convolutional network, cross-modality features, conditional random field.**

I. Introduction

SYNTHETIC aperture radar (SAR) is a high-resolution ground observation radar with high-penetrating capability [1]. It utilizes the pulse compression and the movement of the radar platform to generate two-dimensional images [2]. The process of SAR image interpretation, which strives to extract valuable insights from the images, primarily revolves around image segmentation, target recognition and detection [3], [4]. SAR image segmentation plays a crucial role in SAR image interpretation by partitioning the images into distinct regions based on their homogeneity [5], [6]. Effective SAR image segmentation methods help improve the performance of subsequent interpretation processes.

Traditional segmentation methods for SAR images involve threshold, edge detection, sparse representation, and region merging methods [7]–[9]. Nonetheless, the efficacy of these methods relies on the efficacy of image features. The convolutional neural network (CNN) is extensively employed in computer vision tasks due to its formidable capability for extracting robust features [10], [11]. As a multi-layer model, CNN automatically extracts the image features using convolution, pooling, and fully connected layers. To enable CNN in image segmentation tasks, Long et al. introduce a significant modification to the traditional architecture by replacing the fully connected layers with deconvolution layers, thus pioneering the concept of the fully convolutional network (FCN) [12]. The FCN incorporates both down-sampling and up-sampling processes, where the former utilizes convolution and pooling layers to extract essential image features. At the same time, the latter improves image features' resolution through the deconvolution layers. Based on the work of Long, various versions of FCN are proposed, effectively enhancing the accuracy of image segmentation tasks [13]–[15]. Inspired by the superiority of FCN models, the experts have applied them in the SAR image segmentation and obtained satisfactory results [16], [17].

Recently, a hot research topic in image segmentation has been enhancing the model's performance. The presence of multi-scale objects increases the difficulty of image segmentation [18]. Multi-scale feature extractors, such as the spatial pyramid pooling module, have been designed to alleviate this problem [19], [20]. Other improved models, such as PSPNet, CCNet, and RefineNet, are also designed to extract contextual information from images [21]–[23]. These approaches can obtain texture and edge features of images but still cannot change the nature of CNNs that have difficulty handling long-range dependencies. Transformer networks [24] possess an inherent advantage in extracting global information due to their self-attention structure. By employing the scaled dot product

This work was supported in part by the National Natural Science Foundation of China, under Grant 61771027, and Grant 61071139. The SAR datasets used in the experiments are provided by the Beijing Institute of Radio Measurement. We are grateful for their support.

Fei Gao is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China, and also with the Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City, Hangzhou 310023, China (e-mail: feigao2000@163.com).

Heqing Huang and Zhenyu Yue are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: huangheqing@buaa.edu.cn; yuezhenyu@buaa.edu.cn).

Dongyu Li is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China (e-mail: dongyuli@buaa.edu.cn).

Shuzhi Sam Ge is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, and also with the Institute for Future, Qingdao University, Qingdao 266071, China (e-mail: samge@nus.edu.sg).

Tong Heng Lee is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleleeth@nus.edu.sg).

Huiyu Zhou is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: hz143@leicester.ac.uk).

Corresponding authors: Dongyu Li.

attention mechanism to process features at different scales, the model gains the ability to effectively capture and handle spatial relationships in SAR images. Furthermore, it is crucial to explore effective strategies for combining the strengths of CNNs and Transformers. Other researchers improve the segmentation performance by further optimizing the segmentation results [25]. The commonly used method is the fully connected CRF model, which mines the connection information of the image pixels [26]. In the realm of SAR image segmentation, particularly in scenarios involving complex scenes and noise, CRF present a valuable approach for enhancing segmentation accuracy. By leveraging the interdependencies between pixels within an image, as well as the features of individual pixels and their neighboring counterparts, CRF effectively captures contextual relationships. Furthermore, CRF offer the potential to enhance the accuracy of segmentation results by effectively modeling the spatial and frequency domains of the image.

The segmentation performance of the SAR image can be significantly improved through several key techniques, including the extraction of multi-scale features, the utilization of context information, and the adoption of fully connected CRF. Enlightened by this, we present a new SAR image segmentation method that mines the rich information in the images from a novel perspective, cross-modality features fusion. In this method, the initial step involves the segmentation of input images into smaller slices. Then we adopt the mean and Gabor filters to extract the SAR images' auxiliary features. The feature-enhanced fully convolutional network (FE-FCN) is employed in this approach, leveraging auxiliary features to enhance the segmentation results. In terms of model structures, although extraordinary symbolic power can be obtained using CNNs, convolutional operations typically exhibit limitations in modeling direct long-range relationships. As a result, these structures usually yield weaker performance, especially for showing significant differences in texture, shape, and dimensions. To overcome this limitation, we provide a better way to assist self-attention by adding Transformer blocks [24] at the encoder for global contextual modeling of the attended features. To guarantee structural integrity, the skip connection is designed to align the single-modal information and explore inter-modal correspondence. Finally, we propose the connection-enhanced CRF (CE-CRF) model for optimizing the segmentation results. These key innovations include:

(1) We propose the cross-modality features extraction for mining the rich information in SAR images. Based on the analysis of the peculiarities of SAR images, the mean and texture features are adopted as auxiliary features.

(2) We present a new FE-FCN model to generate the segmentation results. The FE-FCN contains feature fusion module, hybrid CNN-Transformer architecture, and residual alignment. The representation of SAR image features is enhanced by absorbing cross-modality features.

(3) A novel CE-CRF model is designed to post-process the segmentation results. The CE-CRF simultaneously utilizes the input images and additional features for extracting the connection information of image pixels, which effectively enhances the reliability of the connection information and contributes to a higher segmentation accuracy.

II. Related Work

A. Image Semantic Segmentation Using CNN

The FCN methods have become the mainstream method in image segmentation. Both SegNet and UNet employ encoder-decoder architectures, where the encoder extracts image features and the decoder enhances feature resolution to generate accurate segmentation results [27], [28]. Li et al. present the gated fully fusion network (GFFNet), which adopts the gates to fuse the multi-scale image features in a fully connected manner [29]. During the feature fusion process, the gates can effectively enhance the features' useful information while reducing the noise. The full-resolution residual network (FRRNet) incorporates both pooling and residual streams to effectively leverage abstract features and detailed information simultaneously for image analysis [30]. The conducted experiments serve as compelling evidence of the effectiveness of the FRRNet in enhancing segmentation accuracy, particularly in the challenging region boundaries.

Zhao et al. emphasize the importance of context information in the images and propose the PSPNet [21]. The pyramid pooling model (PPM) is embedded in the PSPNet, which can extract the context information. To effectively extract the context information from input images, the CCNet is designed in [22]. By incorporating recurrent operations, the CCNet allows each pixel to capture dependencies from the entire image, enabling a comprehensive understanding of the contextual information. The RefineNet uses the cascaded architecture to generate high-resolution segmentation maps by combining the multi-scale image features [23]. In addition, the chained residual pooling model is embedded in the RefineNet to fuse the pooling features, thereby capturing the background context information.

These methods focus on the analytical adjustment of the model structure, using convolutional neural networks to extract the delicate spatial information of the image. However, further emphasis should be placed on the comprehensive incorporation of global contextual details pertaining to the image. Unlike the above methods, we employ a hybrid CNN-Transformer structure that embeds the transformer encoder into the FCN model. The CNN's high-resolution spatial information and the transformer's global context information are utilized to achieve accurate localization.

B. SAR Image Segmentation

Due to the superiority of FCN models, they have been successfully applied in SAR image segmentation. To accurately distinguish the open water and sea ice in SAR images, Ren et al. propose the dual-attention UNet (DAUNet), which employs UNet as the backbone [31]. The dual-attention mechanism in the DAUNet consists of channel attention and position attention modules, which are utilized to improve the representation of SAR image features. The attention FCN (AFCN) model leverages three strategies to enhance image features: spatial attention, channel attention, and multi-scale feature [32]. Furthermore, the segmentation results of the AFCN model are further refined through the utilization of fully connected CRF. Jing et al. present a new encoder–decoder method for building

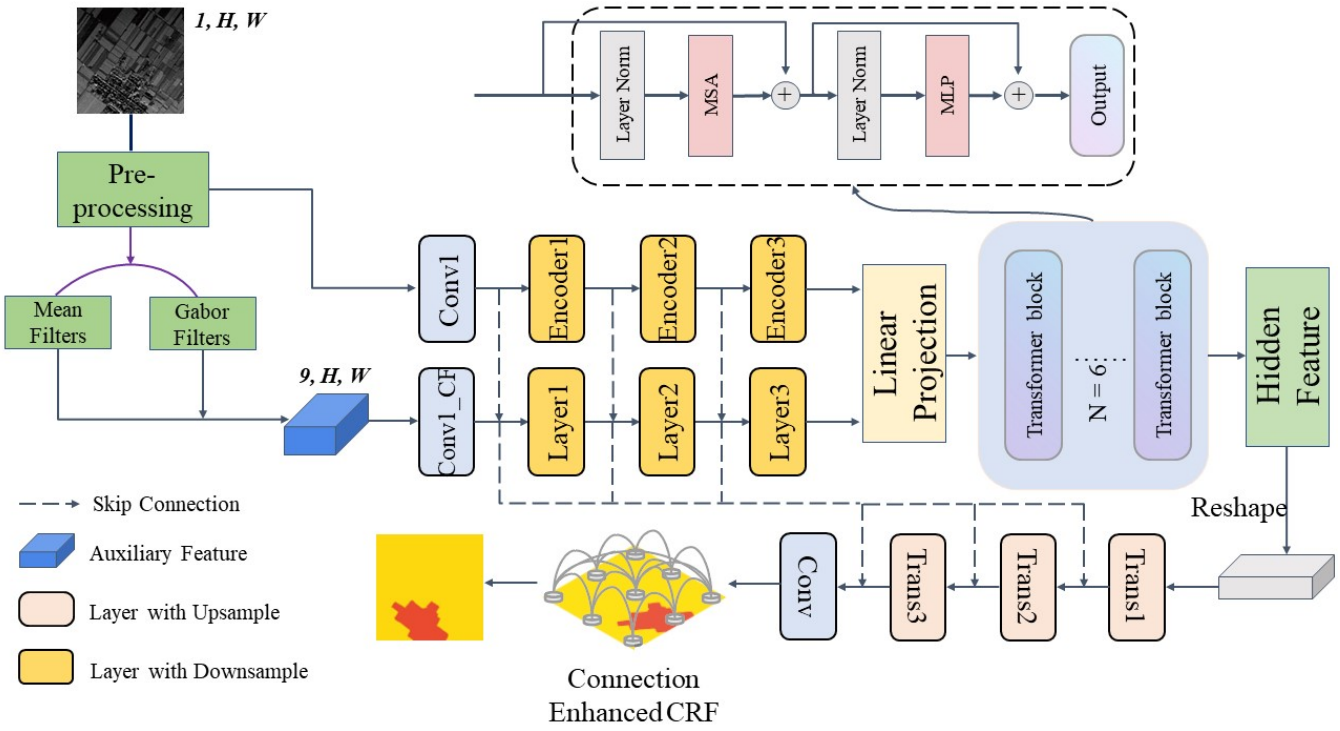

Fig. 1. The flowchart of our method.

segmentation [33]. This method's decoder contains two stages: the selective spatial pyramid dilated (SSPD) network and the context balancing module (CBM) are utilized to fuse and recover the multi-level features. Compared with the traditional FCN models, the multi-path ResNet (MP-ResNet) extracts the context information based on the parallel multi-scale branches [34]. The decoder of MP-ResNet employs the feature fusion mechanism for fusing the image features extracted by different components.

In comparison to the aforementioned methods, our approach explores the information present in SAR images from a novel standpoint, that is, cross-modality feature fusion. We utilize the mean and texture features as auxiliary features and design the FE-FCN and CE-CRF models to promote the segmentation performance. To the best of our knowledge, it is the first time that the mean and texture features have been adopted as cross-modality features for improving the performance of FCN and CRF models. Concretely, the FE-FCN integrates the depth features with the auxiliary features to enhance the SAR image representation. The CE-CRF utilizes the input images and auxiliary features to extract the connection information of the image pixels, hence the segmentation results are further optimized.

C. *Self-Attention Mechanism*

Self-attention mechanism [35], [36] is a technique that enables the learning of correlations between different temporal or spatial locations within an input sequence. It can help the model understand the relationships in the input sequence and capture the critical information, thus improving the model's performance. In CNN, the self-attention mechanism is primarily employed for feature extraction. For example, in SeNet [37], the self-attentive mechanism can automatically learn the correlation between channels across different channels, thus improving the feature representation. Moreover, the self-attention mechanism possesses the remarkable capability to autonomously discern and comprehend the intricate interconnections among features spanning diverse scales. In object detection models such as RetinaNet [38], the self-attentive mechanism can automatically learn feature representations of objects at different scales, thus improving detection accuracy.

The Transformer obtains global contextual information by transforming pixels of an image or image block into a sequence and then applying a self-attentive mechanism. The standard transformer block encompasses essential components, including the multi-headed self-attentive (MSA) mechanism, feed-forward neural network (FFN), and layer normalization (LN). In medical image semantic segmentation, [39], [40] design the swin transformer-based U-shaped encoder-decoder framework. Specifically, the swin transformer-based dual encoder extracts feature representations at different semantic scales. It has also been pointed out that the results of segmenting the network by Transformer alone could be better because Transformer relies too much on the global modeling of the image and needs more detail localization capability. As a result, recent studies such as [41], [42] have attempted to integrate a hybrid structure of CNN and Transformer. This novel model architecture sequentially combines CNN and Transformer layers to construct a new encoder structure, aiming to leverage the strengths of both approaches. SETR [43]

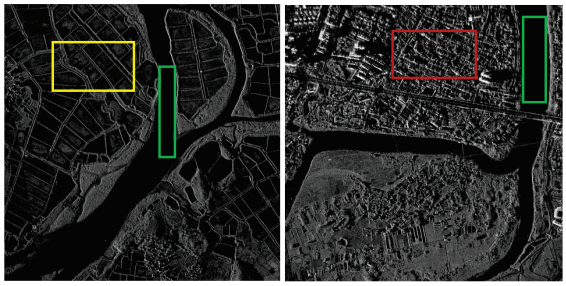

Fig. 2. The SAR images were obtained in Guangxi, China. The areas enclosed by the yellow, green, and red rectangles are farmland, river, and urban regions.

combines ViT [44] feature extraction, multi-level feature fusion, and traditional CNN level-by-level decoding to design a high-precision semantic segmentation framework. SegFormer [45] utilizes Mix-FFN instead of positional embedding and uses Efficient Self-Attention to reduce the time complexity. The authors also propose an overlapping patch segmentation method to preserve local continuity. In recent years, there has also been work to explore applications in remote sensing using self-attentive mechanisms and transformer networks. Zhang et al. [46] introduce a pure transformer network with a Siamese U-shaped structure specifically tailored for addressing the task of remote sensing image transformation detection. [47], [48] attempt to model the global context of remote sensing images by combining CNN with Transformer to enhance feature representation. Hong et al. [49] rethought hyperspectral image classification from the sequential perspective of the Transformer to learn spectral local sequence information from images and reduce the information loss during hierarchical propagation.

Inspired by these excellent works, this study uses an auxiliary encoder of self-attentive mechanism blocks. It provides global contextual information to the CNN-based encoder using multi-level jump connections. In the decoder module, we recover the width and height of the feature vector to match the dimensions of the original image. This is accomplished solely through the application of the conventional CNN.

III. The Proposed Method

We have devised a novel FCN method for SAR image segmentation that incorporates cross-modality features. The flowchart illustrating the workflow of our method is presented in Fig. 1. As can be seen, the proposed method consists of pre-processing, auxiliary features extraction, segmentation results acquisition, and post-processing. We first pre-process the original large-size SAR image and divide it into small slices, facilitating the subsequent stages. In the auxiliary features extraction, we adopt the mean and Gabor filters to extract the mean and texture features. Consequently, the abundance of valuable information encapsulated within SAR images is effectively harnessed. Subsequently, we introduce the FE-FCN model, which serves as a pivotal tool in generating highly accurate segmentation results. Leveraging its capability to assimilate cross-modality features, the FE-FCN significantly enhances the representation of SAR image features. Then, we have developed and implemented the CE-CRF, a sophisticated algorithm meticulously designed to refine and optimize the output. The CE-CRF extracts the connection information of image pixels by simultaneously utilizing the input images and auxiliary features, hence the reliability of the connection information is enhanced. Next, we introduce our method's auxiliary features extraction, FE-FCN, and CE-CRF.

A. Auxiliary Features Extraction

We perform the auxiliary features extraction for mining the rich information in the SAR images. As Fig. 1 shows, the auxiliary features extraction consists of two branches: the mean features extraction based on mean filters and the texture features extraction based on Gabor filters.

1) Mean Features Extraction: The SAR images are acquired based on the radar scattering echoes of ground targets. Due to the distinction of the scattering echoes, the gray values of different areas in the SAR images are different—the greater the disparity in scattering echoes, the more pronounced the variation in grey values. Fig. 2 shows the SAR images obtained in Guangxi, China. The areas enclosed by the yellow, green, and red rectangles in Fig. 2 are the farmland, river, and urban regions. As can be seen, the grey values vary significantly among different categories of areas. The grey values in urban areas are the highest, whereas the grey values in river areas are close to 0. The variation of grey values in different categories of areas is an essential peculiarity of SAR images. Consequently, we investigate the utilization of grey value information within SAR images as a means to enhance the accuracy of segmentation.

Statistical features have the capability to characterize the grey value information present in SAR images. The commonly used statistical features are the mean features extracted by the mean filters. Given the input image $\boldsymbol{I} \in \mathbb{R}^{H\times W}$, where H and W denote the height and width of the image. The mean features are calculated by (1)

$$F_{\text{mean}}(x,y) = \frac{1}{(2w+1)^2} \sum_{i=-w}^{w} \sum_{j=-w}^{w} I(x+i, y+j) \quad (1)$$

where $F_{\text{mean}}(x,y) \in \boldsymbol{F}_{\text{mean}}$, $I(x,y) \in \boldsymbol{I}$. $2w+1$ represents the window size of the mean filter. We can obtain multi-scale mean features by utilizing a set of mean filters with different window sizes.

2) Texture Features Extraction: The texture information in SAR images reflects the structure of ground targets. In Fig. 2, the texture of the river area is smooth and straightforward. However, the texture of the urban area could be more coarse and complicated. Hence, the texture information is conducive to distinguishing different areas in SAR images. The texture features extraction methods for SAR images include Markov random field, gray-level co-occurrence matrix, and Gabor wavelet transform [50], [51]. The Markov random field extracts texture features by capturing the spatial regularity of adjacent pixels. In the gray-level co-occurrence matrix method, the texture information is expressed by the gray-level relationship of different pixels. The Gabor wavelet transform

method utilizes the Gabor filters to process the images, extracting the texture features. The Gabor wavelet transform method effectively extracts multi-orientation texture features by employing a set of Gabor filters that are designed with diverse orientations. Hence we adopt it in our approach.

The function of 2-D Gabor filter is expressed by (2)

$$G\left(x_0, y_0, \theta, \omega_0\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x_0^2+y_0^2}{2\sigma^2}\right)\left[\exp\left(j\omega_0 x_0\right) - \exp\left(-\frac{\omega_0^2\sigma^2}{2}\right)\right] \tag{2}$$

where $x_0 = x\cos\theta + y\sin\theta$, $y_0 = -x\sin\theta + y\cos\theta$. θ and ω_0 denote the orientation and center frequency of the filter, respectively. σ is the standard deviation of Gauss function. To extract the texture features $\boldsymbol{F}_{\text{texture}}$, we perform the convolution operation on the input image and the Gabor filter. $F_{\text{texture}}(x, y) \in \boldsymbol{F}_{\text{texture}}$ is calculated using (3)

$$F_{\text{texture}}\left(x, y\right) = I\left(x, y\right) \otimes G\left(x_0, y_0, \theta, \omega_0\right) \tag{3}$$

where $\otimes$ denotes the convolution operation.

B. Feature Enhanced FCN (FE-FCN)

In the pursuit of generating precise segmentation results, we have developed a state-of-the-art model known as the FE-FCN. This sophisticated architecture is specifically designed to leverage cross-modal features, harnessing their power to greatly enhance the representation of SAR image features. The FE-FCN consists of convolution layers, up-sample layers, down-sample layers, and spatial attention feature fusion modules. The model has two convolutional branches: the original image and auxiliary feature branches. Both model components have the same network structure in the encoder stage, except for the difference in feature channels between the additional features and the original inputs. We first extract the mean features, whose shape is (1, h, w), by concatenating them with the extracted contextual features (8, h, w) as auxiliary features. The shape of this auxiliary feature is (9, h, w). For the original SAR image, the shape of the input is (1, h, w). In the initial step, both components (main and auxiliary features) are individually inputted into a 7 × 7 convolutional layer with a stride of 2. It is important to highlight that the auxiliary features additionally pass through a 3 × 3 max-pooling layer for further processing. Immediately afterward, the features are input into several similar residual modules for down-sampling. We designed three layers of residual units in the encoder and embedded cross-modality features in different layers of branches.

Unlike most existing improvement methods, our approach introduces a self-attention mechanism into the encoder design through the transformer architecture. Once the high-resolution features are extracted using CNNs, they are subsequently reshaped into a sequence of 2D patches. The vectorized patches are mapped into the feature space using trainable linear projections while adding location-specific embeddings to ensure no location information is lost. Then, there are fed into a ViT [44] with a 6-layer transformer block to fully extract the global contextual detail input sequence. This combined CNN-Transformer encoding method enhances the image information with finer details.

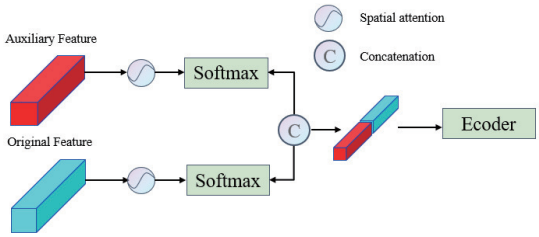

Fig. 3. Structure of the feature fusion module.

In the decoder processes, all the up-sampling layers, except for the final convolution layer, are implemented as residual layers. The final convolution layer is implemented as a single 2 × 2 transpose convolution layer. Meanwhile, the feature maps by the lower layers include detailed information, whereas the higher layers generate feature maps with semantic information. Hence, combining the feature maps using the skip connection method helps optimize the image features.

In the feature fusion module, we propose a feature transfer method. The shared and specific information of the fused features is used efficiently by modeling the spatial attention in the modalities to learn the information of the inter-modal features. The structure of the feature fusion module is illustrated in Fig. 3. First, the original and auxiliary features are computed separately for spatial attention. For each channel, the weight of each position in that channel is calculated. Specifically, the feature matrix of the channel is spread into a vector, then dotted multiplied with a learnable weight vector to obtain a weight matrix of the same size as the feature matrix. The feature matrix in each channel is multiplied by the corresponding weight matrix to obtain the weighted feature matrix. Finally, the weighted feature matrices of all channels are summed by channel direction to get the fused feature matrix. These weight vectors and weight matrices are all learnable parameters, so the back-propagation algorithm can train them. During training, the model automatically learns which positions are more critical for a particular task. We use this soft attention-based approach to better utilize the spatial information in the input images.

C. Connection Enhanced CRF (CE-CRF)

The pixels in SAR images are related to each other. For instance, the pixels with close spatial distance and similar gray values will likely share the same label. Therefore, the connection information of image pixels helps improve the segmentation accuracy. The image pixels are defined as nodes, and the connection of different pixels is represented as edges.

To optimize the segmentation results obtained by FE-FCN, we present the CE-CRF, which can extract the connection information of image pixels. During the extraction of connection information, the traditional CRF only utilizes the input images, whereas the CE-CRF simultaneously utilizes the input images and auxiliary features. Since the auxiliary features also contain rich information, the reliability of the connection information

extracted by the CE-CRF is effectively enhanced. In the CE-CRF, we also define the pixels and the connection of pixels as nodes and edges, respectively. The energy function of CE-CRF is determined by (4)

$$E(\boldsymbol{y}) = \sum_{i} \psi_0(y_i) + \sum_{i,j} \psi_1(y_i, y_j) + \sum_{i,j} \psi_2(y_i, y_j) + \sum_{i,j} \psi_3(y_i, y_j) \tag{4}$$

where $\boldsymbol{y}$ denotes the predicted labels of the image pixels. $\psi_0(y_i)$ represents the unary potential function which is expressed in (5)

$$\psi_0(y_i) = -\log P(y_i) \tag{5}$$

$P(y_i)$ is the label assignment probability of image pixels. $\psi_1(y_i, y_j)$, $\psi_2(y_i, y_j)$, and $\psi_3(y_i, y_j)$ represent the pairwise potential functions that utilize the input images, mean features, and texture features to capture the connection information of image pixels, respectively. The expressions of the three functions are shown in (6)-(8).

$$\psi_1(y_i, y_j) = \mu(y_i, y_j)(\omega_1 e^{(-\frac{|p_i - p_j|^2}{2\theta_1^2} - \frac{|I_i - I_j|^2}{2\theta_2^2})} + \omega_2 e^{(-\frac{|p_i - p_j|^2}{2\theta_3^2})}) \tag{6}$$

$$\psi_2(y_i, y_j) = \mu(y_i, y_j)(\omega_3 e^{(-\frac{|p_i - p_j|^2}{2\theta_4^2} - \frac{|\boldsymbol{M}_i - \boldsymbol{M}_j|^2}{2\theta_5^2})} + \omega_4 e^{(-\frac{|p_i - p_j|^2}{2\theta_6^2})}) \tag{7}$$

$$\psi_3(y_i, y_j) = \mu(y_i, y_j)(\omega_5 e^{(-\frac{|p_i - p_j|^2}{2\theta_7^2} - \frac{|\boldsymbol{T}_i - \boldsymbol{T}_j|^2}{2\theta_8^2})} + \omega_6 e^{(-\frac{|p_i - p_j|^2}{2\theta_9^2})}) \tag{8}$$

As can be seen, the pairwise potential functions contain two Gaussian kernel functions. p_i represents the pixel positions, I_i represents the gray value of the pixels. $\boldsymbol{T}_i$ and $\boldsymbol{M}_i$ respectively denote the texture and mean feature vectors. $\{\omega_1, \omega_2, \cdots, \omega_6\}$ are the weight coefficients, and $\{\theta_1, \theta_2, \cdots, \theta_9\}$ denote the parameters that control the scale of Gaussian kernel functions. $\mu(y_i, y_j)$ is the penalty function which is calculated by (9).

$$\mu(y_i, y_j) = \begin{cases} 1, \text{if} y_i \neq y_j \\ 0, \text{if} y_i = y_j \end{cases} \tag{9}$$

IV. Experiments

In this section, the datasets, the detailed structure of the proposed method, and the evaluation measures are introduced first. Afterward, the segmentation performance of our approach and other state-of-the-art methods is compared. Finally, several ablation studies are designed in this study to evaluate the effectiveness of the different components.

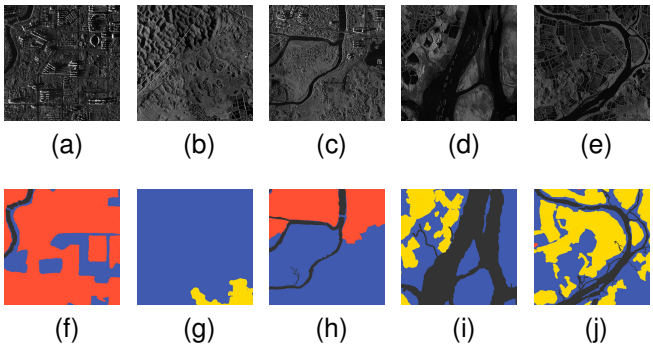

Fig. 4. The SAR images and corresponding ground-truth in FCG dataset. (a)-(e) SAR images. (f)-(j) Ground-truth. The red, black, yellow, and blue colors denote urban, river, farmland, and background areas, respectively.

TABLE I
Numbers of Pixels in the FCG Training and Testing Sets

	Urban	Farmland	River	Background
Training set	0.94 M	0.63 M	1.02 M	2.23 M
Testing set	2.40 M	2.21 M	3.75 M	15.72 M

TABLE II
Numbers of Pixels in the PC Training and Testing Sets

	Urban	Farmland
Training set	0.27 M	2.95 M
Testing set	4.58 M	24.32 M

A. Preliminary

1) Datasets: We adopt three SAR datasets in the experiments: Fangchenggang (FCG), Pucheng (PC), and Gaofen (GF) datasets. The SAR images in the FCG dataset are collected in Fangchenggang, China. This dataset comprises a total of 36 images, each having a resolution of 2 m and dimensions of 875×883 pixels. Fig. 4 shows the FCG dataset, wherein the red, yellow, black, and blue colors denote the urban, farmland, river, and background areas, respectively. A total of 6 images from the FCG dataset were selected as the training set, while the remaining 30 images were allocated to form the testing set. Each category should be included and contain sufficient image pixels when selecting the training images. As Table I shows, the number of pixels in different categories is imbalanced (M denotes the abbreviation of million). For instance, the urban area in the training set contains 0.94 million pixels, whereas the background area contains 2.23 million pixels. This imbalance increases the difficulty of the segmentation tasks.

The PC dataset is composed of 40 SAR images which are acquired in Pucheng, China. The image resolution is 1 m, and the image size is 850×850 pixels. Fig. 5 shows that two categories of areas are contained in this dataset: the urban and farmland areas. We choose 4 images in the PC dataset to form the training set and the other 36 images are adopted as the testing set. Table II shows the number of pixels contained in the two categories. As can be seen, the farmland area has much more pixels than the urban area.

The GF dataset is an open-access spaceborne SAR dataset

TABLE III
NUMBERS OF PIXELS IN THE GF TRAINING AND TESTING SETS

	River	Vegetation	Residential-area	Industrial-area	Bare-land	Non-image	Others
Training set	11.27 M	36.49 M	47.23 M	23.51 M	2.00 M	7.79 M	2.79 M
Testing set	3.01 M	7.49 M	34.71 M	22.90 M	1.63 M	5.24 M	3.66 M

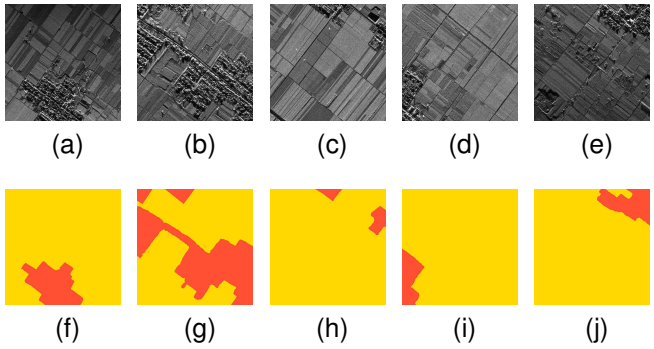

Fig. 5. The SAR images and corresponding ground-truth in PC dataset. (a)-(e) SAR images. (f)-(j) Ground-truth. The red and yellow colors denote urban and farmland areas, respectively.

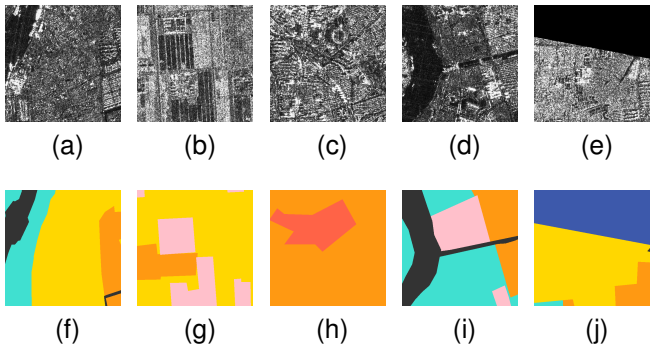

Fig. 6. The SAR images and corresponding ground-truth in GF dataset. (a)-(e) SAR images. (f)-(j) Ground-truth. The black, yellow, orange, cyan, pink, blue, and red colors denote river, vegetation, residential-area, industrial-area, bare-land, non-image, and other areas, respectively.

released in the 4th High-Resolution Remote Sensing Image Interpretation Software Competition. The resolutions of the images vary from 10 m to 30 m, and the image size is 512×512 pixels. The training and testing sets consist of 500 and 300 images. As Fig. 6 shows, the GF dataset includes seven categories of areas: river, vegetation, residential-area, industrial-area, bare-land, non-image, and others. The pixels in each category are shown in Table III.

2) Detailed Structure: In the pre-processing, we divide each image in the FCG and PC datasets into 16 small slices with a step of 200 pixels and a size of 224×224 pixels. The images in the GF dataset are partitioned into four smaller patches with a step of 256 pixels and a size of 256×256 pixels. We adopt two means and eight Gabor filters to extract the mean and texture features in the auxiliary features extraction. The two mean filters' window sizes are set to 3 and 5. The orientations of the eight Gabor filters range from 0 to π with a step of $\pi/8$.

3) Evaluation Measures: The pixel accuracy (PA), mean pixel accuracy (MPA), mean intersection over union (MIoU), and frequency-weighted intersection over union (FWIoU) are utilized to evaluate the segmentation methods. The PA, MPA, MIoU, and FWIoU are calculated based on the confusion matrix of the segmentation results. Suppose N denotes the number of image pixels, and K denotes the number of area categories. PA is obtained by calculating the proportion of pixels that are correctly classified:

$$\mathrm{PA} = \frac{1}{N}\sum_{i=0}^{K-1} p_{ii} \tag{10}$$

p_{ii} denotes the element of confusion matrix at coordinate (i, i). MPA is the averaged PA of different categories:

$$\mathrm{MPA} = \frac{1}{K}\sum_{i=0}^{K-1} \frac{p_{ii}}{\sum_{j=0}^{K-1} p_{ij}} \tag{11}$$

MIoU represents the averaged IoU of different categories:

$$\mathrm{MIoU} = \frac{1}{K}\sum_{i=0}^{K-1} \frac{p_{ii}}{\sum_{j=0}^{K-1} p_{ij} + \sum_{j=0}^{K-1} p_{ji} - p_{ii}} \tag{12}$$

FWIoU takes into account the weights for different categories:

$$\mathrm{FWIoU} = \frac{1}{N}\sum_{i=0}^{K-1} \frac{\left(\sum_{j=0}^{K-1} p_{ij}\right) p_{ii}}{\sum_{j=0}^{K-1} p_{ij} + \sum_{j=0}^{K-1} p_{ji} - p_{ii}} \tag{13}$$

B. Segmentation Performance Comparison

1) Experiments on the FCG Dataset: In this segment, a rigorous evaluation is conducted to assess the segmentation performance of various methodologies on the FCG dataset. As part of this analysis, we carefully select state-of-the-art techniques as the benchmark for comparison, including SegNet [27], UNet [28], GFFNet [29], FRRNet [30], PSPNet [21], CCNet [22], RefineNet [23], Swin-UNet [39], TransUNet [41], Segnext [52] and PidNet [53]. The comparison methods employ the same pre-processing manner as our method.

Table IV showcases the segmentation performance of various methods on the FCG dataset. Notably, our proposed method achieves the highest scores in terms of MIoU, FWIoU, PA, and MPA. Although some comparison methods (such as GFFNet, RefineNet, and CCNet) extract the multi-scale features or the context information in the images, their segmentation performance falls short compared to our approach. This is because our method takes into account the peculiarities of SAR images, which is conducive to mining the rich information in the images. Besides, the FE-FCN model in our method

TABLE IV
THE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE FCG DATASET

Methods	MIoU	FWIoU	PA	MPA
SegNet	66.13%	76.23%	86.58%	74.25%
UNet	68.03%	77.21%	87.22%	75.76%
GFFNet	68.27%	77.61%	87.46%	75.89%
FRRNet	68.51%	77.66%	87.46%	76.14%
PSPNet	56.76%	70.40%	82.60%	65.12%
CCNet	58.33%	71.08%	83.09%	67.17%
RefineNet	65.54%	75.70%	86.25%	73.62%
Swin-UNet	67.42%	75.62%	87.58%	80.43%
TransUNet	69.33%	74.84%	85.91%	78.75%
Segnext	70.01%	76.87%	84.95%	79.47%
PidNet	67.14%	76.88%	87.36%	77.32%
Ours	**72.36%**	**79.51%**	**89.53%**	**82.37%**

TABLE V
THE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE PC DATASET

Methods	MIoU	FWIoU	PA	MPA
SegNet	82.20%	90.14%	94.62%	89.26%
UNet	84.74%	91.51%	95.39%	91.74%
GFFNet	81.74%	89.85%	94.44%	89.15%
FRRNet	82.73%	90.63%	95.00%	88.01%
PSPNet	79.69%	88.42%	93.46%	89.38%
CCNet	79.90%	88.73%	93.73%	88.27%
RefineNet	84.12%	91.29%	95.31%	90.15%
Swin-UNet	83.47%	89.12%	95.32%	90.01%
TransUNet	84.99%	90.61%	95.70%	91.12%
Segnext	86.33%	91.78%	94.87%	90.83%
PidNet	82.69%	89.09%	94.24%	88.77%
Ours	**87.58%**	**93.46%**	**96.21%**	**92.67%**

enhances the representation of SAR image features by fusing FCN features and auxiliary features. The comparison methods utilize the CRF to optimize the segmentation results. At the same time, the CE-CRF in this paper extracts more reliable connection information by simultaneously utilizing the input images and auxiliary features. Specifically, our method can achieve 72.36% MIoU and 89.53% PA on the FCG dataset. Compared to PSPNet, the method improves 15.6% MIoU and 6.93% PA, respectively. Compared to Segnext, our method achieves 79.51% FWIoU and 82.37% MPA, which is 3.28% and 8.12% improvement, respectively. The experiment was also tested on Transformer-based methods, and the results showed that none of them could achieve the same metrics as the method in this paper, further proving the effectiveness of cross-modal features.

Next, we will use the results visualization to compare the differences in segmentation performance between the different methods more visually and intuitively. For the purpose of visual analysis, we have selectively chosen two images from the FCG testing set. In Figures 7 and 8, we present the visualization images that showcase the segmentation results obtained from our methodology. Our method's visual figures match the ground truth better than the comparison methods. Due to the superiority of FE-FCN and CE-CRF, our approach can effectively distinguish the pixels of different categories. Nevertheless, upon careful examination of the visual figures from the comparison methods, it becomes apparent that numerous misclassified pixels are present. These inaccuracies stem from the comparatively weaker segmentation performance, particularly evident in the visual figures associated with the PSPNet and CCNet methods.

2) Experiments on the PC Dataset: Next, we compare the performance of different methods on the PC dataset. As Table V shows, our method demonstrates superior segmentation performance when compared to the other methods. For instance, the PA of our approach reaches 96.21%, which is superior to that of the comparison methods. Compared to Segnext, our method improves the PA by 1.34%. Moreover, our process yields an improvement of 7.68% MIoU compared to CCNet (87.58% vs. 79.90%). Since our method mines the rich information in the SAR images through the auxiliary features extraction, the performance of FE-FCN and CE-CRF is effectively enhanced, thereby contributing to a higher segmentation accuracy.

In Figs. 9 and 10, we present the visual figures depicting the segmentation results obtained by our method. It is evident from the figures that our method effectively and accurately classifies the pixels belonging to farmland and urban areas. The number of misclassified pixels in the visual figures of FRRNet, PSPNet, and CCNet methods are obviously more significant than that of our approach. Especially in Figs. 9 (e)-(h), the farmland and urban areas must be clarified.

3) Experiments on the GF Dataset: The performance of different methods on the GF dataset is shown in Table VI. Compared with the PC dataset, the GF dataset contains seven categories and the image scenes are more complex. Nevertheless, the performance of our method still ranks first among these methods. The FWIoU of our approach is 87.03%, which is much higher than that of the SegNet (74.12%) and UNet (74.32%). In particular, when our method uses an encoder with a self-attentive mechanism, this allows our method to perform very well when processing high-resolution images from GF datasets. For example, FE-FCN obtains 5.69% MIoU over PidNet (82.73% vs. 77.04%). Although TransUNet has a better and excellent global modeling capability, more than simply stacking similar structures is required. Our method combines features from different modalities of SAR images, which will significantly improve the segmentation performance. Compared with TransUNet, the segmentation accuracy of PA is improved by 2.92%. In conclusion, our method demonstrates competitive segmentation results even in complex SAR image scenes.

We utilize visual figures to demonstrate the superiority of our method. As shown in Figs. 11-13, it is clear that our method classifies most of the pixels into correct categories. Because of the complexity of image scenes, the segmentation results of the comparison methods contain numerous incorrectly classified pixels. For example, as Figs. 12 (c)-(i) shows many pixels in the industrial area are incorrectly classified as the residential area. In Fig. 13, the pixels in

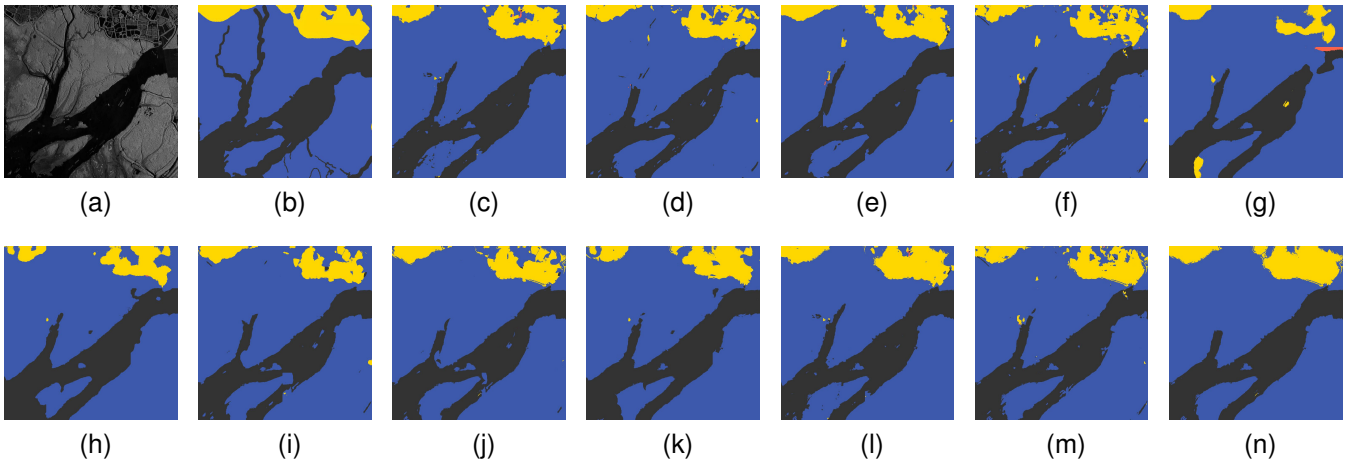

Fig. 7. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The red, black, yellow, and blue colors denote urban, river, farmland, and background areas, respectively.

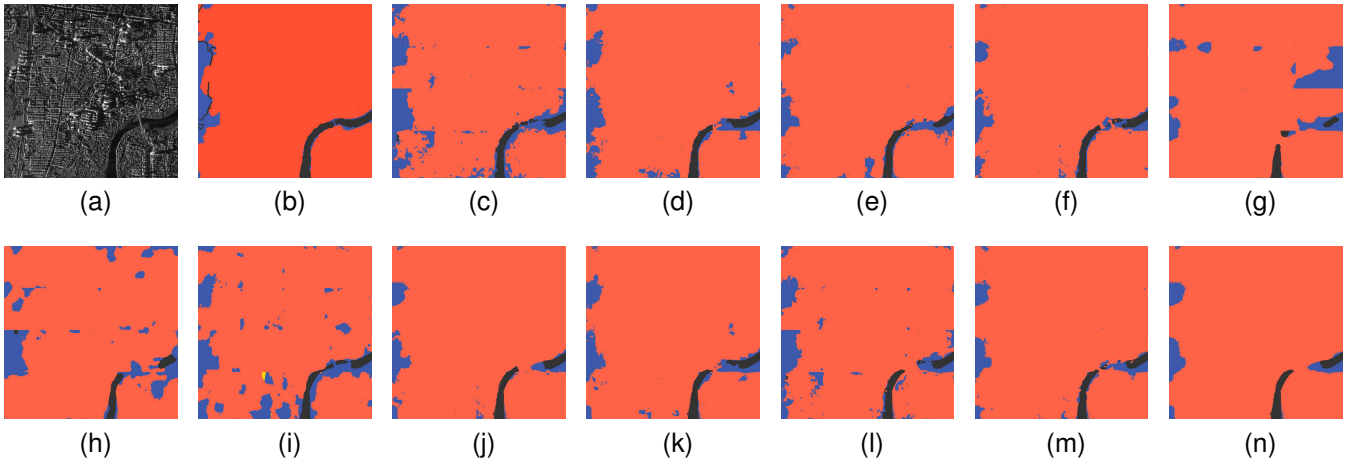

Fig. 8. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The red, black, yellow, and blue colors denote urban, river, farmland, and background areas, respectively.

Fig. 9. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The red and yellow colors denote urban and farmland areas, respectively.

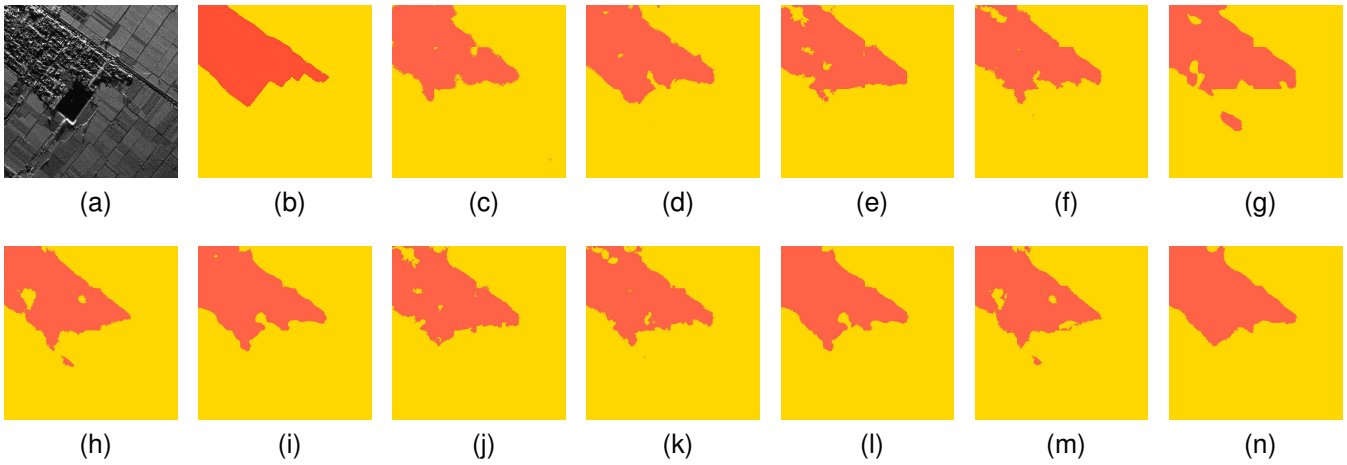

Fig. 10. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The red and yellow colors denote urban and farmland areas, respectively.

TABLE VI
THE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE GF DATASET

Methods	MIoU	FWIoU	PA	MPA
SegNet	65.86%	74.12%	84.74%	76.94%
UNet	67.32%	74.32%	84.90%	77.96%
GFFNet	76.68%	81.31%	89.55%	84.94%
FRRNet	72.40%	78.98%	88.09%	80.21%
PSPNet	79.30%	83.56%	90.94%	86.86%
CCNet	80.91%	85.63%	92.14%	87.65%
RefineNet	72.12%	78.41%	87.63%	79.85%
Swin-UNet	74.63%	79.75%	88.15%	85.15%
TransUNet	76.91%	81.48%	89.63%	85.35%
Segnext	81.33%	85.12%	90.15%	86.88%
PidNet	77.04%	83.55%	89.93%	85.51%
Ours	**82.73%**	**87.03%**	**92.55%**	**89.15%**

the others category are hard for the comparison methods to classify. This is because the training pixels of this category are limited, which increases the classification difficulty. However, our method correctly identifies most of the pixels in the other category, proving its effectiveness despite the deficiency of training pixels.

C. Discussion

1) Evaluation of the FE-FCN: The FE-FCN module employed in our method exhibits remarkable capabilities in enhancing SAR image features through the fusion of mean and texture features. To validate the effectiveness and reliability of the FE-FCN, a series of experiments are conducted on the FCG dataset. As shown in Table VII, the FCN without cross-modality feature fusion module is adopted as the "Baseline1". "FE-FCN (mean)" and "FE-FCN (texture)" respectively denote the models which fuse the mean and texture features, while the FE-FCN model fuses both the two types of features. As can be seen, the MIoU, FWIoU, PA, and MPA of "FE-FCN (mean)" and "FE-FCN (texture)" are superior to those of the "Baseline1", which verifies the effectiveness of the mean and texture features. Moreover, the MIoU and MPA of FE-FCN are 2.20% and 2.72% higher than those of the "Baseline1", respectively. Hence, simultaneously utilizing the two types of auxiliary features further improves the segmentation performance.

TABLE VII
THE EFFECTIVENESS EVALUATION OF FE-FCN

Models	MIoU	FWIoU	PA	MPA
Baseline1	69.18%	77.46%	86.99%	79.12%
FE-FCN (mean)	70.07%	78.23%	87.48%	80.88%
FE-FCN (texture)	70.93%	78.66%	87.91%	81.01%
FE-FCN	**71.38%**	**78.83%**	**88.53%**	**81.74%**

2) Evaluation of the CE-CRF: To optimize the segmentation results, we present the CE-CRF in the post-processing stage. Compared with the CRF, the CE-CRF utilizes the input images, mean features, and texture features to enhance the reliability of the connection information of image pixels. This section presents a meticulously designed set of experiments conducted on the GF dataset to rigorously evaluate the effectiveness of the CE-CRF algorithm. The segmentation results generated by the FE-FCN are adopted as the "Baseline2". The CRF only uses the input images to capture the connection information. Apart from the use of input images, the "CE-CRF (mean)" and "CE-CRF (texture)" respectively adopt the mean and texture features. As Table VIII shows, compared with the "Baseline2", the CRF post-processing effectively optimizes the segmentation results. In addition, the MIoU of "CE-CRF (mean)" reach 81.59%, which is higher than that of CRF (80.76%). Therefore, the mean and texture features are conducive to improving the segmentation performance and generating better segmentation results. In addition, CE-CRF obtains the best MIoU (82.73%) by utilizing both the mean and texture features, which respectively outperforms "Baseline2" and "CRF" by 2.80% and 1.97%.

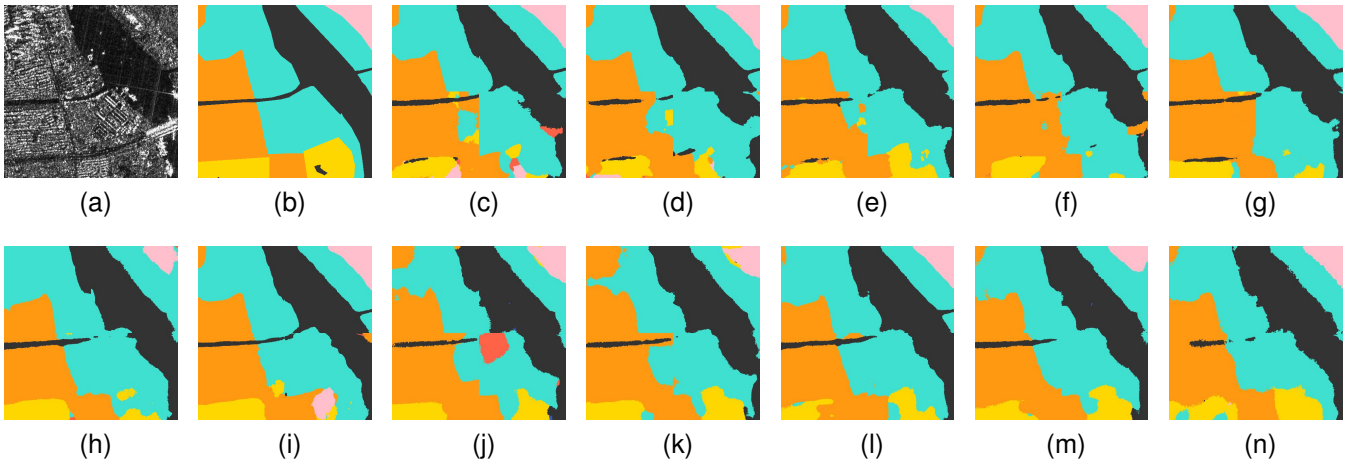

Fig. 11. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The black, yellow, orange, cyan, pink, blue, and red colors denote river, vegetation, residential-area, industrial-area, bare-land, non-image, and other areas, respectively.

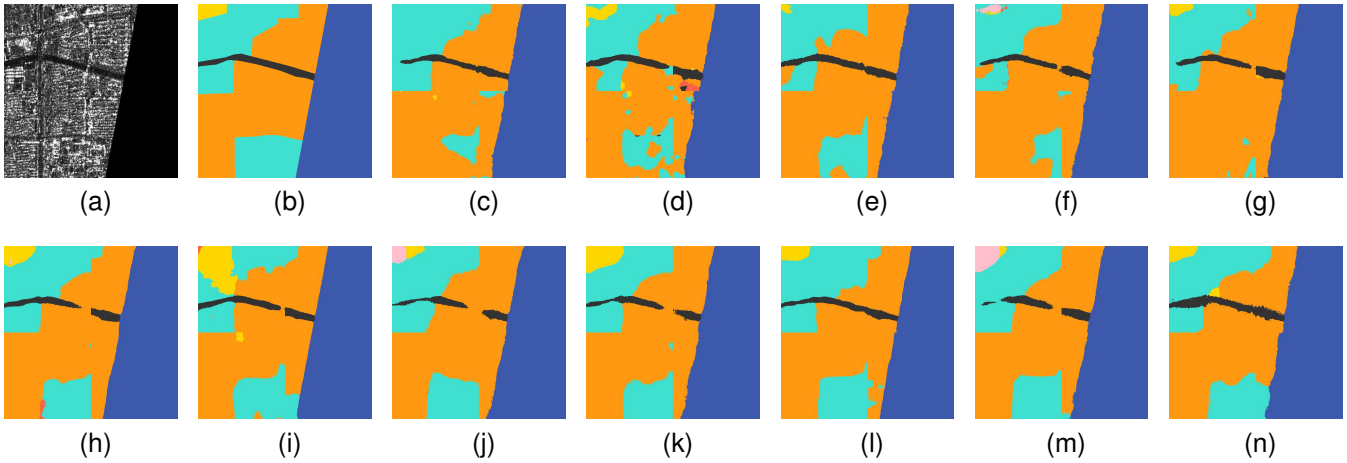

Fig. 12. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The black, yellow, orange, cyan, pink, blue, and red colors denote river, vegetation, residential-area, industrial-area, bare-land, non-image, and other areas, respectively.

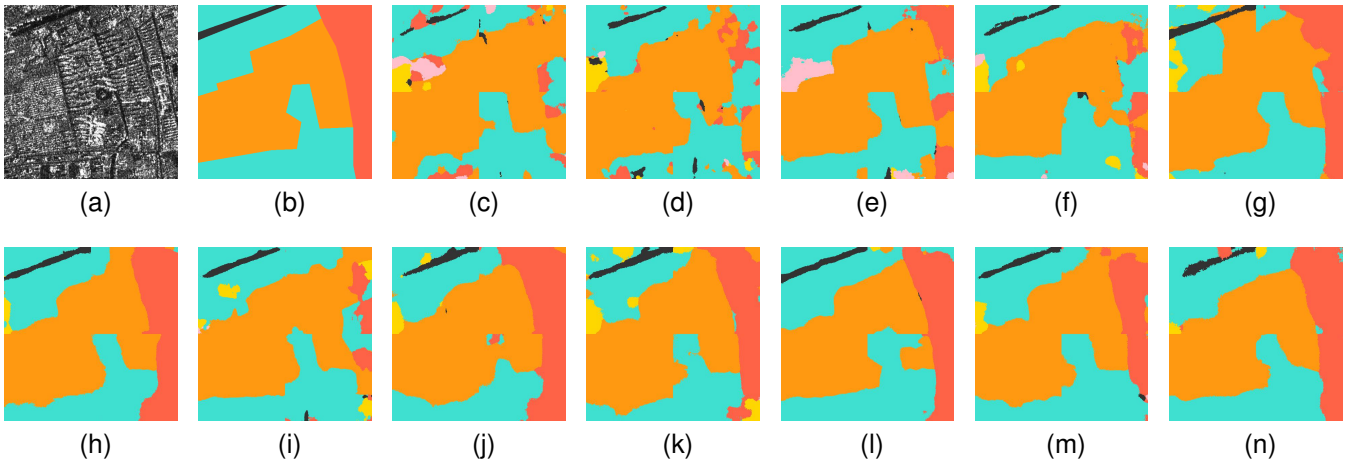

Fig. 13. Visual figures of the segmentation results obtained by different methods. (a) Test image. (b) Ground-truth. (c) SegNet. (d) UNet. (e) GFFNet. (f) FRRNet. (g) PSPNet. (h) CCNet. (i) RefineNet. (j) Swin-UNet. (k) TransUNet. (l) Segnext. (m) PidNet. (n) Our method. The black, yellow, orange, cyan, pink, blue, and red colors denote river, vegetation, residential-area, industrial-area, bare-land, non-image, and other areas, respectively.

TABLE VIII
THE EFFECTIVENESS EVALUATION OF CE-CRF

Models	MIoU	FWIoU	PA	MPA
Baseline2	79.93%	84.98%	91.61%	88.37%
CRF	80.76%	85.12%	91.87%	88.82%
CE-CRF (mean)	81.59%	85.47%	92.11%	88.97%
CE-CRF (texture)	81.53%	85.65%	92.23%	89.06%
CE-CRF	**82.73%**	**87.03%**	**92.55%**	**89.15%**

TABLE IX
THE EFFECTIVENESS EVALUATION OF MODEL STRUCTURE

Models	MIoU	FWIoU	PA	MPA
Variant 0	68.92%	76.22%	87.19%	79.76%
Variant 1	70.38%	78.54%	87.99%	81.32%
Variant 2	**72.36%**	**79.51%**	**89.53%**	**82.37%**

3) Ablation Study on the Model Structure: This section is dedicated to conducting meticulous ablation experiments on the FCG dataset, to evaluate and assess the effectiveness of the FE-FCN model design. ” variant 0” is the model without adding transformer encoder and long skip connection between different resolution features. The ”variant 1” is the model with the transformer encoder added. The ”variant 2” is the complete FE-FCN. As Table IX shows, where the MIoU of ”variant 1” reaches 70.38%, which is higher than that of ”variant 0” (68.92%). This proves the effectiveness of the hybrid CNN-Transformer encoder in the structure. Moreover, FE-FCN, based on the CNN-Transformer design, obtains the best MIoU (72.36%) by using long skip connections, which is 1.52% higher than that of ”variant 1”. We firmly believe that the integration of rich skip connections plays a pivotal role in enhancing the finer segmentation details by effectively recovering low-level spatial information.

V. CONCLUSION

This paper presents the FCN method based on cross-modality features fusion for SAR image segmentation. In light of the peculiarities of SAR images, we extract the auxiliary features, including mean and texture features. Extracting the auxiliary features is conducive to mining the information in SAR images. To generate precise segmentation results for input images, we introduce the FE-FCN framework as a novel alternative to the traditional FCN-based SAR image segmentation method. The key innovation of the FE-FCN lies in its capability to enhance SAR image features through the seamless integration of auxiliary features and features extracted by the CNN. It encodes powerful global context by treating image features as sequences through Transformer and applies fusion structures to the encoder to merge depth information. The CE-CRF is designed to optimize the segmentation results by capturing the connection information of image pixels. Since the auxiliary features are utilized in the CE-CRF, the reliability of the connection information is enhanced, further promoting our method’s segmentation performance. We meticulously conduct a series of comprehensive experiments on three prominent datasets, namely FCG, PC, and GF. The segmentation results demonstrate the clear superiority of our approach. For example, the pixel accuracy of our approach in the GF testing set attains 92.55%, which outperforms that of RefineNet, GFFNet, and FRRNet. Our future work includes increasing the focus on the categories with fewer pixels to alleviate the effect of category imbalance and improve the overall segmentation accuracy.

REFERENCES

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

[2] K. Ouchi, “Recent trend and advance of synthetic aperture radar with selected topics,” *Remote Sens.*, vol. 5, no. 2, pp. 716–807, 2013.

[3] F. Ma, X. Sun, F. Zhang, Y. Zhou, and H.-C. Li, “What catch your attention in sar images: Saliency detection based on soft-superpixel lacunarity cue,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2022.

[4] Z. Yue, F. Gao, Q. Xiong, J. Wang, T. Huang, E. Yang, and H. Zhou, “A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition,” *Cognitive Computation*, vol. 13, pp. 795–806, 2021.

[5] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, “Sar image segmentation based on convolutional-wavelet neural network and markov random field,” *Pattern Recognit.*, vol. 64, pp. 255–267, 2017.

[6] F. Gao, F. Ma, Y. Zhang, J. Wang, J. Sun, E. Yang, and A. Hussain, “Biologically inspired progressive enhancement target detection from heavy cluttered sar images,” *Cognitive Computation*, vol. 8, pp. 955–966, 2016.

[7] X. Qin, S. Zhou, and H. Zou, “Sar image segmentation via hierarchical region merging and edge evolving with generalized gamma distribution,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1742–1746, 2014.

[8] J. Gu, L. Jiao, S. Yang, F. Liu, B. Hou, and Z. Zhao, “A multi-kernel joint sparse graph for sar image segmentation,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1265–1285, 2015.

[9] H. Yu, X. Zhang, S. Wang, and B. Hou, “Context-based hierarchical unequal merging for sar image segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 995–1009, 2012.

[10] F. Zhang, Y. Liu, Y. Zhou, Q. Yin, and H.-C. Li, “A lossless lightweight cnn design for sar target recognition,” *Remote Sens. Lett.*, vol. 11, no. 5, pp. 485–494, 2020.

[11] F. Gao, Y. Huo, J. Sun, T. Yu, A. Hussain, and H. Zhou, “Ellipse encoding for arbitrary-oriented sar ship detection based on dynamic key points,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–28, 2022.

[12] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2015, pp. 3431–3440.

[13] W. Sun and R. Wang, “Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, 2018.

[14] A. Buslaev, S. Seferbekov, V. Iglovikov, and A. Shvets, “Fully convolutional network for automatic road extraction from satellite imagery,” in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2018, pp. 207–210.

[15] L. Mou, Y. Hua, and X. X. Zhu, “A relation-augmented fully convolutional network for semantic segmentation in aerial scenes,” in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2019, pp. 12 416–12 425.

[16] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, “Polsar image semantic segmentation based on deep transfer learning—realizing smooth classification with small training sets,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 977–981, 2019.

[17] 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, 2018.

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[19] Y. Wang, B. Liang, M. Ding, and J. Li, “Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery,” *Remote Sens.*, vol. 11, no. 1, p. 20, 2019.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.
[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2017, pp. 2881–2890.
[22] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[23] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2017, pp. 1925–1934.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[25] H. Zhou, J. Zhang, J. Lei, S. Li, and D. Tu, "Image semantic segmentation based on fcn-crf model," in *Int. Conf. Image, Vis. Comput.* IEEE, 2016, pp. 9–14.
[26] J. Zhang, M. Cui, and B. Wang, "Sar image change detection method based on neural-crf structure," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 3797–3800.
[27] 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, 2017.
[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2015, pp. 234–241.
[29] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang, "Gated fully fusion for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 11 418–11 425.
[30] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2017, pp. 4151–4160.
[31] Y. Ren, X. Li, X. Yang, and H. Xu, "Development of a dual-attention u-net model for sea ice and open water classification on sar images," *IEEE Geosci. Remote Sens. Lett.*, 2021.
[32] Z. Yue, F. Gao, Q. Xiong, J. Wang, A. Hussain, and H. Zhou, "A novel attention fully convolutional network method for synthetic aperture radar image segmentation," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4585–4598, 2020.
[33] H. Jing, X. Sun, Z. Wang, K. Chen, W. Diao, and K. Fu, "Fine building segmentation in high-resolution sar images via selective pyramid dilated network," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6608–6623, 2021.
[34] L. Ding, K. Zheng, D. Lin, Y. Chen, B. Liu, J. Li, and L. Bruzzone, "Mp-resnet: Multipath residual network for the semantic segmentation of high-resolution polsar images," *IEEE Geosci. Remote Sens. Lett.*, 2021.
[35] B. Zhao, X. Li, and X. Lu, "Cam-rnn: Co-attention model based rnn for video captioning," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5552–5565, 2019.
[36] X. Li, B. Zhao, X. Lu *et al.*, "Mam-rnn: multi-level attention model based rnn for video captioning." in *IJCAI*, vol. 2017, 2017, pp. 2208–2214.
[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[39] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 2023, pp. 205–218.
[40] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.
[41] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[42] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.
[43] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.
[46] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
[47] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
[48] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[49] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[50] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution sar image classification via deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2351–2355, 2015.
[51] H. Yu, L. Jiao, and F. Liu, "Crim-fcho: Sar image two-stage segmentation with multifeature ensemble," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2400–2423, 2015.
[52] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *arXiv preprint arXiv:2209.08575*, 2022.
[53] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired from pid controller," *arXiv preprint arXiv:2206.02066*, 2022.

Fei Gao received the B.S. degree in industrial electrical automation, and the M.S. degree in electromagnetic measurement technology and instrument from the Xi'an Petroleum Institute, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in signal and information processing from the Beihang University, Beijing, China, in 2005. He is currently a Professor with the School of Electronic and Information Engineering, Beihang University. His research interests include target detection and recognition, image processing, deep learning for applications in remote sensing.

Heqing Huang received the B.S. degree in computer science and technology from the Zhengzhou University, Zhengzhou, China, in 2019, and received the M.S. degree in agricultural engineering and information technology from the South China Agricultural University, Guangzhou, China, in 2022. He is currently working toward the Ph.D. degree in signal and information processing at the Beijing University of Aeronautics and Astronautics. His research interests include radar signal processing, incremental learning, machine learning and image processing.

Zhenyu Yue received the B.S. degree in electronic and electrical engineering for civil aviation from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2016. He is currently a Ph.D. student with the School of Electronic and Information Engineering, Beihang University, and a joint Ph.D. student with the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include radar signal processing, machine learning, image segmentation, and image recognition.

Dongyu Li received the B.S. and Ph.D. degree from Control Science and Engineering, Harbin Institute of Technology, China, in 2016 and 2020. He was a joint Ph.D. student with the Department of Electrical and Computer Engineering, National University of Singapore from 2017 to 2019, and a research fellow with the Department of Biomedical Engineering, National University of Singapore, from 2019 to 2021. He is currently an Associate Professor with the School of Cyber Science and Technology, Beihang University, China. His research interests include networked system cooperation, adaptive systems, and robotic control.

Shuzhi Sam Ge received the B.Sc. degree from the Beihang University, Beijing, China, in 1986 and the Ph.D. degree from the Imperial College London, London, U.K., in 1993.

He is the Director with the Social Robotics Laboratory of Interactive Digital Media Institute, Singapore and the Centre for Robotics, Chengdu, China, and a Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, on leave from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu. He has co-authored four books and over 300 international journal and conference papers. His current research interests include social robotics, adaptive control, intelligent systems, and artificial intelligence. Dr. Ge is the Editor-in-Chief of the International Journal of Social Robotics (Springer). He has served/been serving as an Associate Editor for a number of flagship journals, including IEEE Transactions on Automation Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, and Automatica. He serves as a Book Editor for the Taylor and Francis Automation and Control Engineering Series. He served as the Vice President for Technical Activities from 2009 to 2010 and Membership Activities from 2011 to 2012, and a member of the Board of Governors from 2007 to 2009 at the IEEE Control Systems Society. He is a fellow of the International Federation of Automatic Control, the Institution of Engineering and Technology, and the Society of Automotive Engineering.

Tong Heng Lee received the B.A. degree with First Class Honours in the Engineering Tripos from Cambridge University, England, in 1980; the M.Engrg. degree from NUS in 1985; and the Ph.D. degree from Yale University in 1987. He is a Professor in the Department of Electrical and Computer Engineering at the National University of Singapore (NUS); and also a Professor in the NUS Graduate School, NUS NGS. He was a Past Vice-President (Research) of NUS.

Dr. Lee's research interests are in the areas of adaptive systems, knowledge-based control, intelligent mechatronics and computational intelligence. He currently holds Associate Editor appointments in the IEEE Transactions in Systems, Man and Cybernetics; and in Control Engineering Practice (an IFAC journal); and is a Consulting Editor (Advisory) in the International Journal of Systems Science(Taylor and Francis, London). In addition, he is a past Deputy Editor-in-Chief of IFAC Mechatronics journal.

Huiyu Zhou received the B.Eng. degree in radio technology from Huazhong University of Science and Technology, Wuhan, China, the M.S. degree in biomedical engineering from University of Dundee, Dundee, U.K., and the Ph.D. degree in ratio technology, biomedical engineering, and computer vision from Heriot-Watt University, Edinburgh, U.K. He is currently a Professor with the School of Informatics, University of Leicester, Leicester, U.K. His research interests include medical image processing, computer vision, intelligent systems, and data mining.