Object-Level Semantic Segmentation on the High-Resolution Gaofen-3 FUSAR-Map Dataset

Xianzheng Shi, *Student Member, IEEE*, Shilei Fu, *Student Member, IEEE*, Jin Chen, Feng Wang, *Member, IEEE*, and Feng Xu, *Senior Member, IEEE*

***Abstract*—Land cover classification with SAR images mainly focuses on the utilization of fully polarimetric SAR (PolSAR) images. The conventional task of PolSAR classification is single-pixel-based region-level classification using polarimetric target decomposition. In recent years, a large number of high-resolution SAR images have become available, most of which are single-polarization. This article explores the potential of object-level semantic segmentation of high-resolution single-pol SAR images, in particular tailored for the Gaofen-3 (GF-3) sensor. First, a well-annotated GF-3 segmentation dataset "FUSAR-Map" is presented for SAR semantic segmentation. It is based on four data sources: GF-3 single-pol SAR images, Google Earth optical remote sensing images, Google Earth digital maps, and building footprint vector data. It consists of 610 high-resolution GF-3 single-pol SAR images with the size of 1024 × 1024. Second, an encoder–decoder network based on transfer learning is employed to implement semantic segmentation of GF-3 SAR images. For the FUSAR-Map dataset, an optical image pretrained deep convolution neural network (DCNN) is fine-tuned with the SAR training dataset. Experiments on the FUSAR-Map dataset demonstrate the feasibility of object-level semantic segmentation with high-resolution GF-3 single-pol SAR images. Also, our algorithm obtains fourth place about the PolSAR image semantic segmentation on the "2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation." The new dataset and the encoder–decoder network are intended as the benchmark data and baseline algorithm for further development of semantic segmentation with high-resolution SAR images. The FUSAR-Map and our algorithm are available at github.com/fudanxu/FUSAR-Map/.**

***Index Terms*—2020 GaoFen challenge, encoder–decoder network, FUSAR- map dataset, GaoFen-3 (GF-3) single -polarization SAR images, object-level semantic segmentation.**

I. Introduction

Synthetic Aperture Radar (SAR) can obtain rich information of earth surface under all-time and all-weather conditions. Land cover classification or segmentation using SAR images could be of great importance in a variety of Earth observation applications, such as urban planning, natural disasters monitoring, and environmental resource management [1]–[3].

In the past few decades, land cover classification with SAR images mainly focuses on the utilization of fully polarimetric SAR (PolSAR) images because the PolSAR images with multiple polarizations reveal distinct electromagnetic scattering characteristics of terrain surface [4]. Conventionally, single-pixel-based land cover classification of PolSAR images takes a single pixel as a processing unit, where the polarimetric scattering mechanism and target decomposition theorems could be used to extract physically plausible features of the pixel [5], [6]. However, PolSAR feature extraction with traditional methods is often limited to the intrapixel regime [7]. In recent years, with the rapid development of deep learning method, region-level land cover classification of PolSAR images have emerged, which takes one image patch as input and utilizes convolutional neural networks (CNNs) to extract high-level features and classify the terrain surface [8], [9]. Although these region-level classification methods can improve the land cover classification performance without any hand-crafted features, its accuracy is still on the level of regional mapping applications. Even with the improvement of SAR image resolution, it is not tailored for semantic segmentation, where individual terrain objects can be well segmented. To achieve object-level land cover classification with high-resolution SAR images, which is known as image semantic segmentation in computer vision, deep learning-based segmentation methods have been widely studied [3], [10]–[12].

In the last few years, more and more high-resolution SAR images are available publicly, most of which are single-polarization (single-pol) SAR images. On the one hand, such a large number of data provides a good basis for land cover semantic segmentation using deep learning methods. On the other hand, compared with the physical problem of land cover classification using conventional methods with PolSAR images, the semantic segmentation with single-pol SAR images becomes a pure visual semantic feature extraction problem, where the interpixel spatial pattern becomes the major source of information. Given the fact that human interpreters can easily understand the land semantics on single-pol grayscale SAR images, it's a valid problem that worth to be studied as well.

A. Motivation and Objective

In the existing literature, single-pol SAR image semantic segmentation focuses on the segmentation of a single class of land cover, such as oil spill segmentation [3], building segmentation [11], and road segmentation [12]. Semantic segmentation of single-pol SAR images for multiple types of terrain categories

Manuscript received January 4, 2021; revised January 23, 2021; accepted February 25, 2021. Date of publication March 4, 2021; date of current version March 24, 2021. This work was supported in part by the Natural Science Foundation of China under Grant 61991422 and Grant 61822107. *(Corresponding authors: Feng Wang; Feng Xu.)*

Xianzheng Shi, Shilei Fu, Feng Wang, and Feng Xu are with the Key Lab for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China (e-mail: xzshi19@fudan.edu.cn; fusl17@fudan.edu.cn; fengwang@fudan.edu.cn; fengxu@fudan.edu.cn).

Jin Chen is with the Beijing Institute of Remote Sensing Information, Beijing, China (e-mail: chenjin_wonder@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3063797

is under-explored. The main reason is that there is no benchmark dataset available for land cover semantic segmentation of single-pol SAR images.

Gaofen-3 (GF-3) is the first civilian C-band polarimetric SAR imaging satellite of China for high-resolution earth observations. The highest nominal resolution of GF-3 data is 1 m. GF-3 data are used in [13] to construct large-scale dataset, i.e., FUSAR-Ship for ship detection. However, the single-pol SAR images as an important part of GF-3 are under-explored in the field of land cover semantic segmentation. To this end, it would be beneficial to develop land cover semantic segmentation using high-resolution GF-3 single-pol SAR images. It could open up a wide range of applications for the huge volume of GF-3 single-pol SAR images.

In this article, we aim to develop a deep learning method for land cover semantic segmentation with high-resolution GF-3 single-pol SAR images. The major objectives are as follows:

To construct a large-scale well-annotated land cover semantic segmentation dataset using high-resolution GF-3 single-pol SAR images: The classification capability of both CNN models and FCN models depends heavily on the quantity and quality of the training and validation datasets [14], [15]. To the best of our knowledge, despite the increasingly available high-resolution GF-3 single-pol SAR images, there are no large-scale well-annotated land cover datasets in the research community. The lack of well-annotated large-scale datasets may cause overfitting and limit the generalization ability of deep learning models [16]. Furthermore, this situation to a certain extent will hinder the further development of land cover semantic segmentation with high-resolution SAR images using deep learning techniques. To this end, it is urgent to construct a benchmark dataset for land cover semantic segmentation using high-resolution GF-3 single-pol SAR images. Such a task not only requires professional knowledge but also needs substantial manpower.

To develop a baseline algorithm for land cover semantic segmentation of GF-3 SAR images: For region-level land cover mapping of PolSAR images, CNNs have been proved the high ability to extract high-level features and classifying SAR images [8], [9]. For semantic segmentation of single-pol SAR images, FCN models are widely used to produce pixels-to-pixels results [3], [10]–[12]. Based on the well-annotated land cover semantic segmentation dataset of GF-3 single-pol SAR images we constructed, it is necessary to develop a baseline algorithm for further development and evaluation. Besides, under the "2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation"[17], a supervised deep-learning algorithm has been designed to achieve high performance of Gaofen-3 SAR image semantic segmentation.

B. Related Works

First, for single-pixel land cover classification of PolSAR images, traditional classification methods focus on the polarimetric scattering mechanism [5], [18], [19], statistical distributions of the polarimetric features [20]–[22], and target decomposition theorems [6]. These methods try to establish an accurate and effective description of the polarimetric features of land cover. Although these methods can preserve detailed information of images and have obtained good results for the low-resolution PolSAR images, they often generate the salt-and-pepper-like result with high-resolution images because of the highintraclass and lowinterclass variability of image pixels [7]. Also, these conventional methods cannot deal with single-pol SAR images.

Second, for region-level land cover classification, in [23]–[28], several conventional methods are utilized. CNNs are also used for region-level land cover classification of PolSAR images. Zhang *et al.* [8] proposed a complex-valued CNN specifically for PolSAR image land cover classification, which utilizes both amplitude and phase information of complex SAR imagery. Zhou *et al.* [9] design a four-layer CNN to achieve high performance without any hand-crafted features of land cover classification with PolSAR images. These methods based on CNNs have been proven to better performance than traditional methods. However, with the improvement of SAR image resolution, region-level land cover classification can no longer meet the requirements of advanced applications such as locating individual terrain objects, such as building, road, and vegetation, where object-level classification is necessary.

Finally, to achieve object-level land cover classification with the pixel-wise result of high-resolution SAR images, image semantic segmentation should be pursued. For high-resolution PolSAR images, few works have focused on the land cover semantic segmentation [29]. Wu *et al.* [10] employ FCNs to achieve wetland semantic segmentation from fully polarimetric RADARSAT-2 imagery. Zhang *et al.* [30] utilized FCNs based on transfer learning for PolSAR image semantic segmentation with small training sets. However, semantic segmentation using FCNs with single-pol SAR images is mainly limited to a single class of land cover, such as the oil spill segmentation of single-pol SAR images [3], the building segmentation of high-resolution single-pol TerraSAR-X images [11], and the road segmentation of high-resolution single-pol TerraSAR-X images [12]. In the literature, semantic segmentation of single-pol SAR images based on multiple types of land cover is under-explored. In particular, land cover semantic segmentation using GF-3 single-pol SAR images has not been studied.

C. Contributions

The contributions of this article are as follows.

Large-scale semantic segmentation benchmark dataset for high-resolution GF-3 single-pol SAR images: In order to construct the land cover semantic segmentation datasets, we develop an efficient semiautomated process to annotate single-pol GF-3 SAR images. Four data sources include GF-3 single-pol SAR data, Google Earth optical remote sensing images, Google Earth digital maps and building footprint vector data are utilized. We present a pixel-labeled land cover semantic segmentation dataset of GF-3 single-pol SAR images, which is named as FUSAR-Maps. The dataset consists of 610 high-resolution GF-3 single-pol SAR images with the size of 1024 × 1024, contains eight different areas from six provinces, and covers areas more than 4500 km^2 in China. To the best of our knowledge, the dataset is the first large-scale pixel-labeled land cover semantic segmentation benchmark dataset with high-resolution GF-3 single-pol SAR images. The nominal resolution of these SAR images is 3 m.

Baseline segmentation algorithm for Gaofen-3 SAR image: Based on the large-scale land cover semantic segmentation dataset of GF-3 single-pol SAR images and the PolSAR training datasets from "2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation," we proposed a unified SAR data preprocessing method for GF-3 SAR data, and a deep learning model using encoder–decoder structure based on transfer learning to achieve object-level semantic segmentation for GF-3 images.

D. Organization of This Article

The remainder of this article is organized as follows. First, land cover semantic segmentation dataset construction using high-resolution GF-3 single-pol SAR images is introduced in Section II. Next, Section III describes the semantic segmentation algorithm for GF-3 SAR data. Then, the results of our experiments are presented and discussed in Section IV. Finally, Section V concludes this article.

II. FUSAR-Map: A Benchmark Dataset for SAR Semantic Segmentation

A large-scale well-annotated dataset is important for semantic segmentation of single-pol SAR images using deep learning techniques. However, manual annotation is often time-consuming and requires professional knowledge about the SAR characteristics of different terrain objects. To address this problem, a semiautomated annotation scheme is proposed here. It utilizes multisource auxiliary data including aerial images, digital maps, and building footprint vector data. In digital maps, different types of terrain targets are labeled with different colors, such as water, building, road, vegetation. However, only the water and road labels are complete, the building and vegetation labels are missing in most cities in developing countries. For the regions studied here, we acquired the building footprint data which can be used to complement the digital maps. As for vegetation, we can obtain rather accurate vegetation maps via automatic classification of optical aerial image based on the RGB-value statistical distribution. As for other land cover objects, it is difficult to accurately identify them from the optical images. Therefore, in this article, only four types of terrain objects are considered, i.e., water, road, building, and vegetation. In the proposed semiautomated annotation scheme, we first coregistration GF-3 SAR images with these auxiliary data sources, and then produce merged semantic maps for each pixel.

A. Image Coregistration

The flow-chart of image coregistration is shown in Fig. 1. First, according to the latitude and longitude information of the GF-3 SAR image, the corresponding area can be identified and the optical aerial image and digital map of this area can be obtained from public services such as Google Earth. The nominal resolution of GF-3 SAR data is 3 m, while the resolution of the obtained optical remote sensing image and the digital map is 1.02 m. The building footprint vector data can be acquired from Geographic Information System (GIS) resources. We construct in the source domain four different datasets, i.e., GF-3 single-pol SAR data, the optical remote sensing images, the digital maps, and the building footprint vector data. To derive labels from the heterogeneous auxiliary data, we need to first covert and coregistration them into unified coordinates in a uniform format. The GF-3 single-pol SAR images originally in the L1 format are first geocoded to L2 format using the Pixel Information Expert (PIE) [31] software which is in the latitude–longitude format with the output resolution of 1.02 m to match the rest data sources. Then, we implement a rasterization algorithm that converts the building footprint vector data from shapefiles to raster images to match with the latitude and longitude grid. Due to the conversion errors, the four layers of different data are not always accurately aligned. Then, we perform coregistration by manually picking hundreds of sparse control points. Geometric transformation is conducted to transform the auxiliary data to match the Gf-3 SAR image. Finally, we obtain in the target domain four corresponding layers all well-aligned in the same latitude–longitude grid.

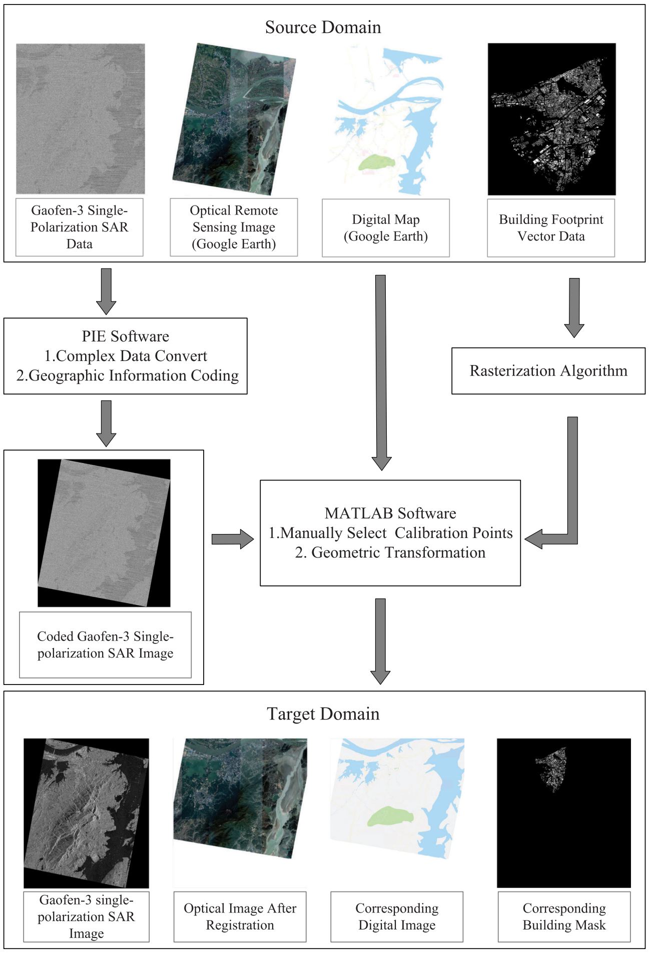

Source Domain
Gaofen-3 Single-Polarization SAR Data
Optical Remote Sensing Image (Google Earth)
Digital Map (Google Earth)
Building Footprint Vector Data
PIE Software
1.Complex Data Convert
2.Geographic Information Coding
Rasterization Algorithm
Coded Gaofen-3 Single-polarization SAR Image
MATLAB Software
1.Manually Select Calibration Points
2. Geometric Transformation
Target Domain
Gaofen-3 single-polarization SAR Image
Optical Image After Registration
Corresponding Digital Image
Corresponding Building Mask

Fig. 1. Image registration.

B. Semantic Map Generation

Fig. 2 shows the process of generating semantic maps of GF-3 single-pol SAR images. The masks of water and road are regarded sufficiently accurate and complete in the digital maps.

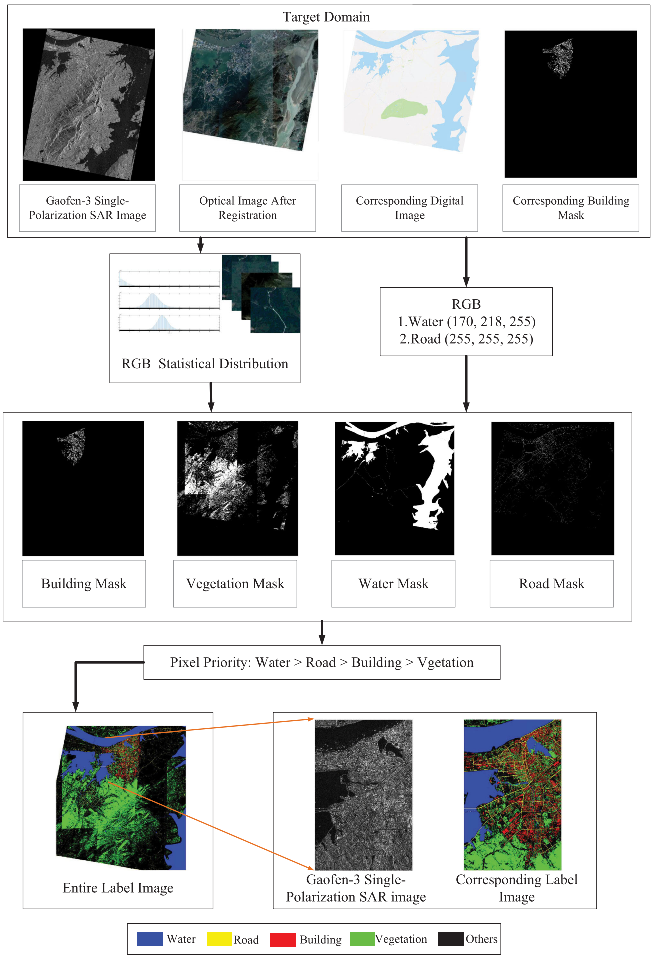
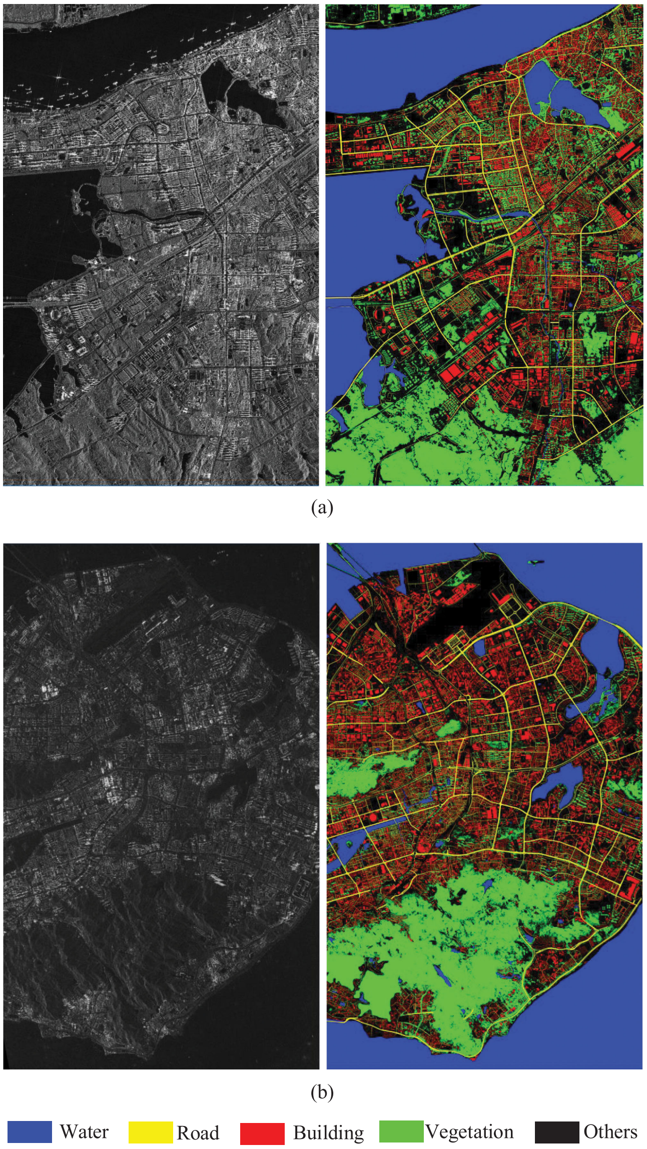

Target Domain
Gaofen-3 Single-Polarization SAR Image
Optical Image After Registration
Corresponding Digital Image
Corresponding Building Mask
RGB Statistical Distribution
RGB
1.Water (170, 218, 255)
2.Road (255, 255, 255)
Building Mask
Vegetation Mask
Water Mask
Road Mask
Pixel Priority: Water > Road > Building > Vgetation
Entire Label Image
Gaofen-3 Single-Polarization SAR image
Corresponding Label Image
Water
Road
Building
Vegetation
Others

Fig. 2. Process of generating semantic maps.



(a)

(b)

Fig. 3. Examples of the single-pol SAR images and corresponding label images from FUSAR-Land-Cover. (a) Example of FUSAR-Map, located in Jiujiang, Jiangxi Province. (b) Example of FUSAR-Map, located in Xiamen, Fujian Province.

Thus, we extract the two land cover categories according to the different RGB-value. For the vegetation category, we classify it using the RGB-value statistical distribution of the optical remote sensing images. The classifier is trained using manually picked vegetation areas in the images. It contains different color features of as many types of vegetation in the optical image as possible. Note that one pixel may be included by more than one type of mask. To solve this conflict in semantic map generation, a default priority rule is enforced as, from the highest priority of water, to the road, then building, and lowest priority of vegetation. Such order is decided according to the confidence of the respective mask generation method. The generated masks still have other problems, e.g., missing building data in some areas, invalid areas of the SAR images, etc. These regions are purposely excluded from the final dataset. Finally, the dataset of both SAR and optical images with semantic maps are cut into medium-sized patches of standard size 1024 × 1024, which constructs the benchmark dataset of FUSAR-Map. The details of the dataset are described in Section II-C.

C. *FUSAR-Map Dataset*

The FUSAR-Map dataset consists of 610 high-resolution GF-3 single-pol SAR images with the size of 1024 × 1024. It contains eight different areas from six provinces of more than 4500 km^2 in China. The annotation layer contains four terrain types, i.e., water, road, building, and vegetation. Fig. 3 displays two examples from the datasets.

The GF-3 has 12 observing modes to meet the multiple user requirements. The SAR payload support operation in single-pol (HH or VV), dual-polarization (HH+HV or VH+VV), and quad-polarization (HH+HV+VH+VV) for the observing modes with resolution ranging from 1 m to 500 m and the swath from 10 to 650 km. Since launched in 2016, the GF-3 has been widely used for land and ocean monitoring, disaster mitigation, water conservancy, and other important applications. It's worth noting that the GF-3 single-pol SAR images that we used to establish FUSAR-Map are all on the Ultra-fine strip map (UFS) mode with the nominal resolution of 3 m, the nominal swath width of

TABLE I
INFORMATION OF THE GF-3 SINGLE-POL SAR IMAGES USED TO GENERATE FUSAR-MAP

Properties	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Collection time	2016/08/15	2017/02/07	2016/9/20	2017/02/25	2017/02/17	2017/02/25	2017/02/24	2017/02/07
Center pitch abgle	35.41	36.3	31.57	40.36	35.41	40.36	21.27	36.3
Polarization mode	DH	DH	DH	DH	DH	DH	DH	DH
Nominal resolution	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Imaging mode	UFS	UFS	UFS	UFS	UFS	UFS	UFS	UFS
Width	18403	18975	16111	19469	18506	19408	10293	18825
Height	21510	21134	24916	21000	21805	21001	20358	22047
Wide resolution	1.124	1.124	1.124	1.124	1.124	1.124	1.124	1.124
High resolution	1.754	1.731	1.539	1.730	1.731	1.730	1.731	1.728
Upper left latitude	29.926477	24.680311	30.781008	28.961019	28.213232	28.677635	39.066645	40.854942
Upper left longitude	115.9135	118.0375	114.3946	121.3789	120.5619	121.3236	117.4109	114.6032
Upper right latitude	29.87409	24.734097	30.72653	28.91173	28.2661	28.62824	39.00798	40.9045
Upper right longitude	116.2454	118.3558	114.7296	121.7051	120.8884	121.649	117.775	114.9869
Lower left latitude	29.51855	24.27946	30.36475	28.55729	27.80298	28.2731	38.69106	40.44118
Lower left longitude	115.8184	118.1196	114.3408	121.3002	120.6478	121.2449	117.3109	114.6981
Lower right latitude	29.46618	24.33337	30.30973	28.50783	27.85594	28.22416	38.63249	40.4908
Lower right longitude	116.1482	118.4368	114.6767	121.6251	120.973	121.569	117.673	115.0794
Proximal pitch angle	39.510	40.573	34.937	45.722	39.491	45.726	23.051	40.594
Far pitch angle	41.254	42.292	36.796	47.142	41.247	47.141	25.112	42.296
Lifting rail	DEC	ASC	DEC	DEC	ASC	DEC	DEC	ASC
Shooting direction	R	R	R	R	R	R	R	R
Province	Jiangxi	Fujian	Hubei	Zhejiang	Zhejiang	Zhejiang	Tianjin	Hebei

TABLE II
DETAILED INFORMATION OF THE EXAMPLE PATCH IMAGES IN FIG. 4

Patch Image	Size (Width × Height)	SAR image source	Latitude and Longitude (/°) (min-lon, min-lat; max-lon, max-lat)	Pixel Location (min-row, min-col; max-row, max-col)
(a)	1024×1024	(5)	[120.662468326791,27.9770194135509; 120.673806716993,27.9884550418547]	[24837,9073;25860,10096]
(b)	1024×1024	(1)	[115.992563804380,29.8758818305501; 116.004158501314,29.8870996151940]	[3591,15370;4614,16393]
(c)	1024×1024	(5)	[120.651118853118,27.9770194135509; 120.662457243320,27.9884550418547]	[24837,8049;25860,9072]
(d)	1024×1024	(1)	[115.992563804380,29.8646530803297; 116.004158501314,29.8758708649737]	[4615,15370;5638,16393]
(e)	1024×1024	(5)	[120.692050109245,27.8854449589418; 120.703388499447,27.8968805872456]	[33029,11742;34052,12765]
(f)	1024×1024	(5)	[120.726098530263,27.8854449589418; 120.737436920466,27.8968805872456]	[33029,14814;34052,15837]
(g)	1024×1024	(1)	[115.980957773431,29.8871105807704; 115.992552470365,29.8983283654144]	[2567,14346;3590,15369]
(h)	1024×1024	(5)	[120.730565168828,27.9999130272032; 120.741903559030,28.0113486555070]	[22789,15217;23812,16240]

30 km, and the center incident angle ranging from 20° to 50°. The time of acquisition is between August 15, 2016, and February 25, 2017. The detailed information of 8 GF-3 single-pol SAR scenes is summarized in Table I. In Table I, the polarization mode DH means HH polarization. ASC and DEC mean ascending and descending trajectories, respectively. *R* means the right-looking direction.

In FUSAR-Map, each class is labeled with different colors, i.e., water in the blue, road in yellow, building in red, and vegetation in green. Pixels of the undetermined or unknown class are colored in black. Fig. 4 shows some patch images and the corresponding semantic maps. The advantages of the dataset include:

Accurate latitude and longitude coordinates are provided for each sample: Each sample image in the FUSAR-Map are provided with accurate latitude and longitude information, as shown in Table II, which could be used to match with more sources of data in the future. The bounding box of the latitude and longitude for each sample image is provided. The min-lon and max-lon in the Table II are the minimum longitude and maximum longitude, respectively. The min-lat and max-lat are that of latitude. Furthermore, all sample images contain their coordinates in the original GF-3 scenes, which are shown in Table II. The bounding box of the pixel location for each sample images are represented by the min-row, max-row, and min-col, max-col in Table II.

Accurate masks of water, road, and building: It provides accurate masks and outline of the water, road, and building areas. Thus, it can be used in other fields such as road segmentation, river and land segmentation, and building segmentation with high-resolution GF-3 single-pol SAR images. Fig. 4(a)–(d) display the accurate outline of these object types.

Difference in features of the same category in the same or different patch images: Due to the different collection time and geographic location with GF-3, the same types in the FUSAR-Map show the weak or strong difference of intensity and characterization. For example, in Fig. 4, the (a)–(c), (e)–(f), and (g)–(h) show the feature difference of building, vegetation,

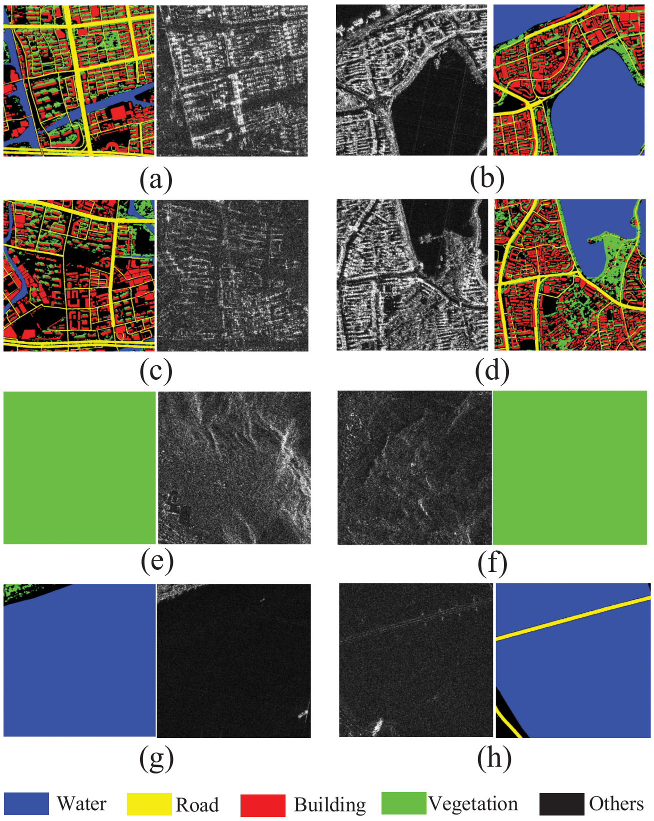


Fig. 4. Examples of the patch images and corresponding label images.

and water, respectively. Such variety can enhance the robustness of the trained semantic segmentation networks.

III. Semantic Segmentation Method

With the FUSAR-Map dataset, we propose a supervised deep-learning algorithm with the encoder–decoder structure for SAR image semantic segmentation. For the task of semantic segmentation of RGB optical images, various CNN network structures have been proposed with breakthroughs. Based on these pretrained CNN network structures, a unified SAR data preprocessing method is first proposed to deal with the L2 format SAR data. Then, a supervised encoder–decoder network based on the structure of DeepLabv3+ [32] is proposed to achieve Gaofen-3 SAR image semantic segmentation with fine-tuning.

A. Unified Data Preprocessing for SAR Data

Different from RGB optical images, SAR L2 format data may appear very different if using different preprocessing methods. To fine-tune the pretrained DCNN network weights from the optical image using the SAR dataset, the L2 format SAR data needs to be uniformly preprocessed. In this article, we utilize *N* times the *m*ean of non-zero data to *t*runcate and stretch the original SAR data, which abbreviated as NMT, detailed in (1). For the FUSAR-Map dataset, the single-channel SAR data is processed by the NMT method. For the PolSAR data on "2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation," NMT is individually used for each channel

$$I = \mathrm{uint8}\left[\frac{I(I \le \mathrm{N} \cdot \mathrm{mean}(I))}{\mathrm{N} \cdot \mathrm{mean}(I)} \times 255\right] \in [0, 255]. \tag{1}$$

In (1), I represents the data value of each channel, $\mathrm{mean}(I)$ represents the average value of all non-zero data for each channel I, and N is an adjustable float value belongs to [2, 3] in this article.

Based on the NMT method, the L2 format SAR data can be converted to a uint8 data format belonging to [0, 255]. Owing to the pretrained DCNN network trained by the normal optical images of ImageNet [33], the images are all RGB format with three channels, so the SAR data needs to be transformed into three channels to form a pseudocolor image. For FUSAR-Map data, single-channel data is copied three times to form a three-channel image. For PolSAR data on "2020 Gaofen Challenge on Automated High-Resolution Earth Observation Image Interpretation," the four-channel data is fused into a three-channel pseudocolor image. According to the reciprocity principle of backscattering SAR, its scattering intensity satisfies

$$I_{\mathrm{hv}} = I_{\mathrm{vh}}. \tag{2}$$

Empirically, the difference in intensity value of HV and VH channels is small. Therefore, we consider fusing the HV channel data and the VH channel data into one channel. To minimize the noise impact of cross-polarization, we use the method of averaging to merge the intensity values with HV and VH channels, like the I_{m} in (3)

$$I_{\mathrm{m}} = \frac{I_{\mathrm{hv}} + I_{\mathrm{vh}}}{2}. \tag{3}$$

In the process of synthesizing the pseudocolor image, we choose HH channel, the average I_{m} of HV and VH channel, VV channel as R, G, B channel, respectively.

$$\mathrm{R} = \mathrm{uint8}\left[\frac{I_{\mathrm{hv}}(I_{\mathrm{hv}} \le \mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{hv}}))}{\mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{hv}})} \times 255\right] \in [0, 255] \tag{4}$$

$$\mathrm{G} = \mathrm{uint8}\left[\frac{I_{\mathrm{m}}(I_{\mathrm{m}} \le \mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{m}}))}{\mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{m}})} \times 255\right] \in [0, 255] \tag{5}$$

$$\mathrm{B} = \mathrm{uint8}\left[\frac{I_{\mathrm{vh}}(I_{\mathrm{vh}} \le \mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{vh}}))}{\mathrm{N} \cdot \mathrm{mean}(I_{\mathrm{vh}})} \times 255\right] \in [0, 255]. \tag{6}$$

Data visualization comparison based on NMT and other preprocessing methods will be given in detail in Section V.

B. Encoder–Decoder Network for SAR Image Semantic Segmentation

An encoder–decoder model based on the structure of DeepLabv3+ [32] is designed to achieve the SAR image semantic segmentation. As shown in Fig. 5, the network is divided into two parts, which are the encoder part and decoder part. The purpose of the encoder module is to gradually capture the high semantic information from the input SAR images, and the decoder module is to gradually use the spatial information from the feature maps to classify each pixel of the input images.

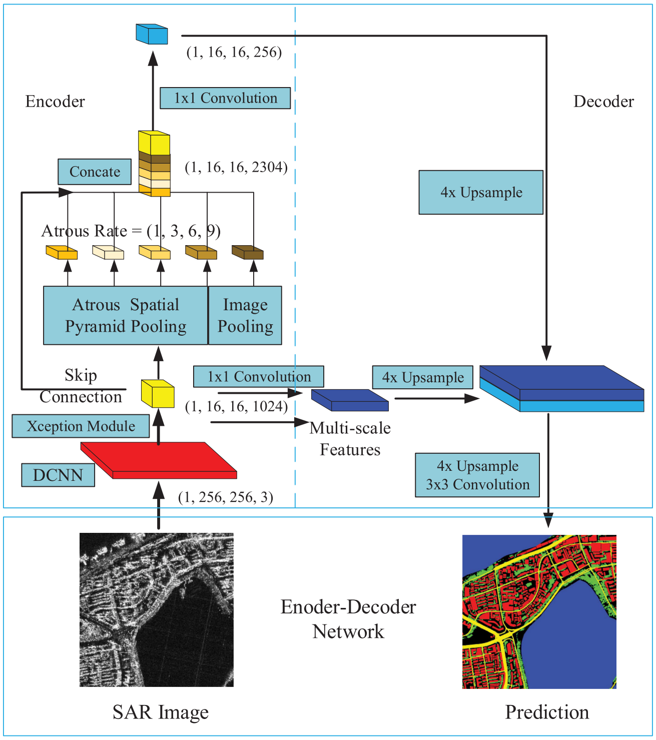


Fig. 5. Structure of the encoder-decoder network.

In the encoder part, DCNN is employed to extract low-level and high-level features. To reduce the computational complexity and improve the high-level feature extraction ability of the network, we use the Xception [34] module in the DCNN. In Xception, we employ the atrous convolution [32] to replace the standard convolution. For atrous convolution, the output $y[i]$ for each location i of a 2-D input $x[i]$ with a filter $w[k]$ of length K is defined as

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k]. \tag{7}$$

In (7), the r corresponds to the input sample stride, which is equivalent to convolving the input x with upsampled filters produced by inserting $r - 1$ zeros between two consecutive filter values along each spatial dimension [35]. In a DCNN, the size of field-of-view (FOV) can roughly indicates how much we use context information. Consider the nth convolution layer, where the size of the convolution kernel size is K_n, the stride size is S_n, and the receptive field size FOV_n is

$$\mathrm{FOV}_n = (\mathrm{FOV}_{n+1} - 1) \cdot S_n + K_n. \tag{8}$$

The receptive field size of the network is

$$\mathrm{FOV}_0 = \sum_{n=1}^{\mathrm{N}} \left((K_n - 1) \prod_{n=1}^{\mathrm{N}-1} S_n \right) + 1. \tag{9}$$

For atrous convolution, the kernel size K_a is

$$K_a = K_n + (K_n - 1) \cdot (r - 1). \tag{10}$$

Equations (9) and (10) proves that atrous convolution can effectively enlarge FOV without increasing the number of parameters or the amount of computation. We follow the DCNN with atrous spatial pyramid pooling (ASPP) [36] module and image global average pooling to capture multiple scales context information and image-level features. To avoid only the center weights effect of the 3 × 3 convolution, the convolutions of the ASPP module have $r = (3, 6, 9)$ with 3 × 3 convolution. Learn from PSPNet [37], we concatenate the output feature maps of DCNN, ASPP, and image pooling. With 1 × 1 convolution, the encoder output feature map contains 256 channels and rich semantic information.

In the decoder part, the encoder output feature maps are bilinearly upsampled by a factor of 4. As for the low-level features, we choose multiscale DCNN feature maps. Empirically, using different sizes of feature maps from low-level features can enrich the spatial information for image semantic segmentation. To balance the weight ratio between the low-level features and high-level features, we choose the number of low-level features to 32 with 1 × 1 convolution. Next, a bilinear upsampling layer is applied to generate features with the same size of network inputs. Finally, we use two 3 × 3 convolution layers to obtain sharper segmentation results.

IV. Experiments and Discussion

In this section, we present and discuss the experimental results of the encoder–decoder model on FUSAR-Map and 2020 Gaofen Challenge. There are two main purposes in our experiments: one is to verify the efficacy of the FUSAR-Map dataset we constructed, and the other is to test the performance of semantic segmentation using our algorithm on GF-3 SAR images. In order to compare the performance of the experimental results, several baseline encoder–decoder network such as U-Net [38], SegNet [39] and DeepLabv3+ [32] models are tested in our work. The experimental setup is described specifically in Section IV-A. The experimental results of GF-3 single-pol SAR images are presented in Section IV-B. And the experimental results of the 2020 Gaofen Challenge are described in Section IV-C.

A. Experimental Setup

For object-level semantic segmentation of GF-3 single-pol SAR images, to obtain the experimental dataset, first of all, we implement the semiautomated process to generate a well pixel-labeled dataset, FUSAR-Map. Next, the 610 high-resolution GF-3 SAR images with the size of 1024 × 1024 are divided into the training dataset with 530 patch images and the test dataset with 80 patch images. For image augmentation, the training dataset is augmented synthesizing seven new rotations (-135, -90, -45, 45, 90, 135, 180). And flip left and right, flip up and down of images are randomly used.

The experiment fully polSAR dataset of this article is provided by the 2020 Gaofen Challenge. The polSAR images are collected from the GF-3 satellite. Their resolution is from 1 to 4 m. It should be mentioned that only the amplitude information of the four channels (HH, HV, VH, and VV) is provided in the polSAR images. Also, the ground truth maps are humanly annotated according to six land-cover types: building area, industry area, vegetation, water, bare soil, and others. In this challenge, 500 pairs of PolSAR images with the size of 512 × 512 are provided as the training set and the testing data is not visible

TABLE III
SPECIFIC EXPERIMENTAL PARAMETERS FOR FUSAR-MAP DATASET

Model	Pre-trained	Trainable weights	Initial Learning Rate	Training Samples	Validation Samples	Input Size	Batch Size	Loss Function
U-Net [38]	No	18M	1e-3	17560	4390	256×256	10	Weighted Categorical Crossentropy
SegNet [39]	No	12M						
VGG-SegNet [39]	Yes	22M						
DeepLabv3+ [32]	No	36M						
PT-DeepLabv3+ [32]	Yes	36M						
ours	No	34M						
PT-ours	Yes	34M						

to participants. We applied the same data augmentation method described above.

As for a semantic segmentation system, its performance must be evaluated with standard and well-known metrics. Therefore, in this article, two metrics of the execution time and classification accuracy are considered. Meanwhile, the classification accuracy is of more importance than execution time in our work.

For the execution time assessment, we employ the training time and test time of each encoder–decoder network to provide the conditions for the benchmark. Besides, our implementation is built on TensorFlow1.15 [40] and our experiments are implemented on a single NVIDIA GeForce RTX 2080 Ti GPU with the memory of 10 GB in Linux system.

In the semantic segmentation accuracy aspect, we assess the experimental results with perclass accuracy (PA), overall accuracy (OA), and frequency weighted intersection over union (FWIoU) in this article. To express the mathematical formulas of these evaluation metrics, we assume p_{ij} is the amount of pixels of class i predicted to class j and T_i is the total number of pixels labeled to class i. And the k is the total number of the classes. Thus, the accuracy metrics are defined as follows:

Perclass Accuracy (PA): it simply computes a ratio between the number of properly classified pixels and the total number of pixels for each class.

$$\text{PA} = \frac{p_{ii}}{T_i} \tag{11}$$

where $i = 1$, 2,..., k.

OA: it's a percentage of the properly classified pixels and the total number of pixels in the entire image

$$\text{OA} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} T_i}. \tag{12}$$

FWIoU: this is a standard metric to measures the similarity between the prediction results and the ground truth.

$$\text{FWIoU} = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}} \sum_{i=1}^{N} \frac{\sum_{j=1}^{N} S_{ij} S_{ii}}{\sum_{j=1}^{N} (S_{ij} + S_{ji}) - S_{ii}}. \tag{13}$$

Of all metrics described above, the FWIoU is the most used one. The higher value of these accuracy metrics indicates the better semantic segmentation performance.

B. Experiments on GF-3 Single-Pol SAR Images

To verify the efficacy of FUSAR-Map and test the effectiveness of the encode–decoder model based on transfer learning, we compare the performance using encoder–decoder models with and without pretrained. For the compared methods, the basic encoder–decoder models are U-Net, SegNet, and Deeplabv3+. Furthermore, the SegNet with VGG-16 and DeepLabv3+ pretrained on Cityscapes [41] are fine-tuned in our experiments, which are named as VGG-SegNet and PT-DeepLabv3+, respectively. In these comparison methods, including U-Net, SegNet, DeepLabv3+, and ours that use the normal distribution to initialize network parameters randomly. Also, include VGG-SegNet pretrained with ImageNet, DeepLabv3+ pretrained with Cityscapes and ours pretrained with ImageNet. The specific parameters of these models are shown in Table III. The network optimizer is Adam with an initial learning rate of 1e-3, which multiply by 0.93 in every two epochs. And the loss function is weighted categorical cross-entropy [40], described in (14).

$$\text{Loss} = -\sum_{i=1}^{k} W_i \cdot [y_i \log \bar{y}_i + (1 - y_i) \log (1 - \bar{y}_i)] \tag{14}$$

where y_i is the ground truth with one-hot coding for class i, $\bar{y}_i$ is the softmax function output for class i, k is the total number of output class, and W_i is the balanced weights of class i. Besides, we choose the batch size of 10 with an input size of 256×256 because of the limitation by the GPU memory. And the specific trainable weights of networks are presents in Table III. These models are trained or fine-tuned on the training datasets with 17 560 patch images and validated with 4390 patch images. In experiments, we take the other category as the vegetation category for training to reduce the interference from the invalid samples, as Fig. 6(c). For the generalization test, we utilize these encoder–decoder models to classify the test image from FUSAR-Map, as Fig. 6(a). The experimental results of these methods are shown in Table IV.

Through the results given in the Table IV, our network achieves the highest OA and FWIoU of 0.758 and 0.666, respectively. The results of SegNet and DeepLabv3+ indicate that the pretrained models for fine-tuning on FUSAR-Map can effectively reduce the training time and achieve better object-level semantic segmentation performance than the raw models. Also, from the perspective of test time consumption, the encoder–decoder network can quickly achieve efficient pixels-to-pixels classification from the input image. Moreover, these encoder–decoder models all achieve high perclass classification accuracy on the water and vegetation categories. However, the pixel-wise classification accuracy of the road and building categories is very low, especially the road categories. Considering the trainable parameters in Table III and the classification results in Table IV, the

TABLE IV
COMPARISON OF DIFFERENT NETWORKS FOR SEMANTIC SEGMENTATION RESULTS ON FUSAR-MAP TEST DATASET

Methods	OA (%)	FWIoU (%)	Per-class Accuracy (%)				Time of Training (/h)	Total Test Time of Figure 6(a) (/s)
			water	road	building	vegetation&others		
U-Net [38]	66.92	57.34	85.84	18.40	35.63	86.31	19	25
SegNet [39]	64.78	54.76	83.19	9.1	19.05	85.94	13	11
VGG-SegNet [39]	68.26	59.21	86.14	11.10	46.12	87.51	8	11
DeepLabv3+ [32]	72.02	62.23	89.97	19.39	52.67	88.74	14	25
PT-DeepLabv3+ [32]	74.31	65.07	91.99	30.2	59.24	90.61	8	25
Ours	73.16	63.22	90.17	23.53	56.24	90.0	15	24
PT-Ours	**75.84**	**66.59**	**92.49**	35.69	63.25	**92.17**	**8**	24

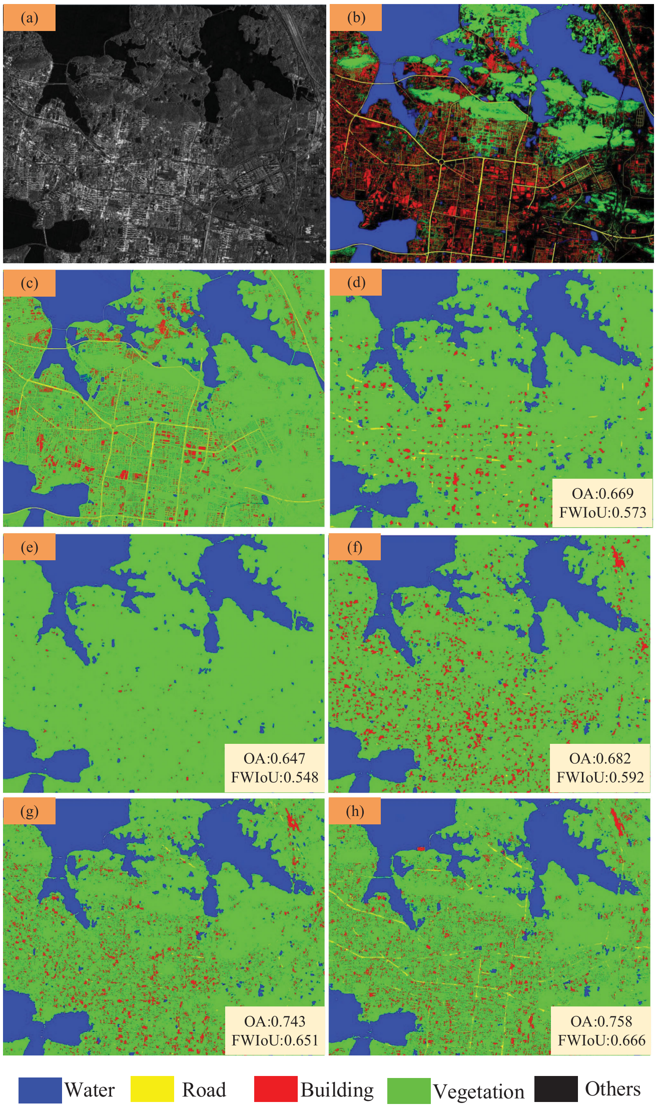


Fig. 6. (a) The GF-3 single-pol SAR image located in Wuhan, Hubei Province in China. (b) Label mask base on our pixel labeling process. (c) Changed label with vegetation. (d)–(h) Semantic segmentation results of U-Net, SegNet, VGG-SegNet, PT-DeepLabv3+, and PT-ours.

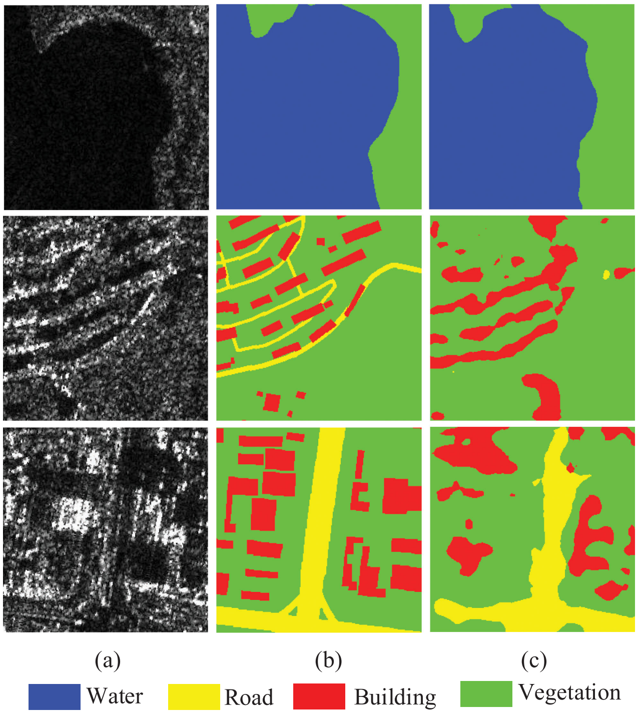


Fig. 7. The test results of our network. (a) The SAR patch images with size of 256 × 256. (b) The ground truch. (c) The semantic segmentation results of our network.

deeper networks often can achieve better object-level semantic segmentation performance.

To compare the segmentation performance of different models more intuitively, we display the entire segmentation maps combined with the 1280 test patch images in FUSAR-Map, as shown in Fig. 6. Fig. 6(a) is a high-resolution GF-3 single-pol SAR image with the size of 8192 × 10240, located in Wuhan, Hubei Province in China, and obtained on February 25, 2017. Fig. 6(b) is the corresponding label mask. And Fig. 6(c) is the changed test label image with vegetation category from Fig. 6(b). Fig. 6(d)–(h) are the classification results generated by U-Net, SegNet, VGG-SegNet, pretrained DeepLabv3+ (PT-DeepLabv3+), and pretrained ours (PT-ours). From the images of the segmentation results, the water category can be easily identified. However, it's difficult to distinguish the building and vegetation categories, not to mention the road category. Overall, compared to these object-level semantic segmentation results, the results predicted by our network shows a clearer distribution of building and the main road structure.

Compared with the region-level classification, the object-level semantic segmentation can generate more accurate edges for targets. To verify the pixel-to-pixel capability of the encoder–decoder network, we display several test patch images with the size of 256 × 256 and the corresponding segmentation results predicted by our network, which are shown in Fig. 7. Fig. 7(a) shows the GF-3 single-pol SAR test patch images

TABLE V
SPECIFIC EXPERIMENTAL RESULTS ON 2020 GAOFEN CHALLENGE

SAR Images	Pre-trained Backbone	K-fold OA		K-fold FWIoU		Prelimimary stage testing scores	Final stage testing scores
		ours	DeepLabv3+	ours	DeepLabv3+	ours	ours
gray-single	Xception41	0.7601	0.7345	0.7031	0.6763	59.53	–
	Xception65	0.7688	0.7473	0.7098	0.6866	0.6097	–
	Xception71	0.7421	0.7233	0.6715	0.6586	–	–
RGB-median	Xception41	0.9035	0.8897	0.8284	0.8193	65.58	69.21
	Xception65	0.9087	0.8902	0.8301	0.8212	66.10	70.96
	Xception71	0.8956	0.8735	0.8200	0.8037	65.35	–
RGB-mean	Xception41	0.9263	0.9132	0.8576	0.8431	66.78	**72.73**
	Xception65	**0.9276**	**0.9139**	**0.8587**	**0.8446**	**67.12**	72.45
	Xception71	0.9199	0.9021	0.8468	0.8276	65.48	–

with size of 256 × 256. And the Fig. 7(b) contains the label masks. Fig. 7(c) presents the semantic segmentation results. The object-level semantic segmentation results show the clear category boundaries of the water. However, the building segmentation results illustrate inaccuracy and blurred outline. As for the road category, the main roads can be recognized, but the narrow roads are lost. The object-level semantic segmentation results demonstrate that it is difficult to precisely classify the building and road categories with the FUSAR-Map dataset. On the one hand, only the visual information of the image is considered, the characteristics of the building in the SAR image are different in the position and shape of the footprints label. The road category is closer to the background intensity information, which is difficult to accurately segment. On the other hand, the road category accounts for a relatively small proportion and there are very thin roads.

C. Experiments on 2020 Gaofen Challenge

To quantitatively evaluate the improvement of our network over the baseline method DeepLabv3+, we conduct a *K*-fold cross-validation study on the Gaofen challenge's training dataset. The training and validation sets are randomly divided from all available training data (500 image pairs) with the numeric ratio of 4:1. In this way, a total of five training and validation sets can be divided. To reduce the effects of random factors, the *K*-fold cross-validation has been conducted on all five training and validation sets. And we take the average of the indicators of *K*-fold cross-validation as the performance of the network.

Similarly, to compare different unified image preprocessing methods, we used three different methods to preprocess SAR data. Besides, we also evaluate the impact of DCNN on the segmentation results for different pretraining structures, as shown in Table V. For the 2020 Gaofen challenge test scores, in the preliminary stage, an evaluation server on the official website is provided to generate the FWIoU scores for the upload algorithms with the unified testing dataset. In the final stage, the official uses a different testing dataset from the preliminary stage which scales from 512 × 512 to 1500 × 1500, based on the offline evaluation method, considering the algorithm accuracy, and running speed.

In the data preprocessing part, we compared three different unified preprocessing methods. The first is to directly add the amplitude values of the four channels of PolSAR into a gray image with one channel, marked as gray-single in Table V. The second is to use the median of the SAR data values to cut and stretch to form an RGB pseudocolor image [42], marked as RGB-median in Table V. The third is the RGB pseudocolor image obtained by the NMT method in Section III-A, marked as RGB-mean in Table V. For the images generated by these different visualization methods, we have done the same normalization process before entering the network, that is, the standard normal process with the mean value of 0 and the standard deviation of 1. In the encoder–decoder network structure, we mainly compared the DCNN structure of different depths based on the Xception module, specifically Xception41, Xception65, and Xception71 [43]. Also, we compared the baseline model DeepLabv3+ with our adjusted network. The results are reported in Table V.

The results show that based on the unified SAR data preprocessing method, whether it is the baseline model Deeplabv3+ or our improved network structure, Xception65 as a DCNN module can achieve slightly better results than Xception41, and both are better than Xception71. This shows that in limited SAR data, too deep DCNN often leads to overfitting. While keeping the network structure unchanged, different SAR data preprocessing methods greatly affect the network segmentation effect. Whether it is the local *K*-fold cross-validation or the test results during the 2020 Gaofen Challenge, the RGB-median and RGB-mean method improve by about 6% compared to the gray-single. And the RGB-mean can increase by about 2% compared with RGB-median. This is mainly because the NMT method has the characteristics of making the SAR data of the same source and the same region look similar visually, but different sources and different regions look different, as Fig. 8. From the *K*-fold cross-validation results, based on the same DCNN structure and SAR data processing method, our adjusted network structure is better than the baseline model Deeplabv3+.

To more clearly reflect the difference in segmentation results between our network structure and the baseline model DeepLabv3+, we use the optimal DCNN structure Xception65 as the backbone network and obtain the segmentation results for each land-cover category through *K*-fold cross-validation. Table VI shows the results in details. The experimental comparison results prove that multiscale DCNN features and the skip-connection structure can improve the classification accuracy for the decoder structure. However, the baseline model DeepLabv3+ can obtain better PA on special categories like bare soil and vegetation based on the RGB-mean method. In the 2020 Gaofen

TABLE VI
COMPARISON RESULTS BASED ON XCEPTION65 BACKBONE

SAR Images	Encoder-Decoder Network	PA						OA	FWIoU
		industrial area	built-up area	bare soil	vegetation	water	others		
gray-single	DeepLabv3+	0.729	0.803	0.508	0.781	0.759	0.386	0.7473	0.6866
	ours	0.775	0.824	0.53	0.798	0.8	0.4	0.7688	0.7098
RGB-median	DeepLabv3+	0.888	0.908	0.609	0.844	0.865	0.506	0.8902	0.8212
	ours	0.909	0.923	0.671	0.855	0.888	0.508	0.9087	0.8301
RGB-mean	DeepLabv3+	0.925	0.934	**0.783**	**0.871**	0.905	0.544	0.9139	0.8446
	ours	**0.935**	**0.948**	0.758	0.867	**0.917**	**0.553**	**0.9276**	**0.8587**

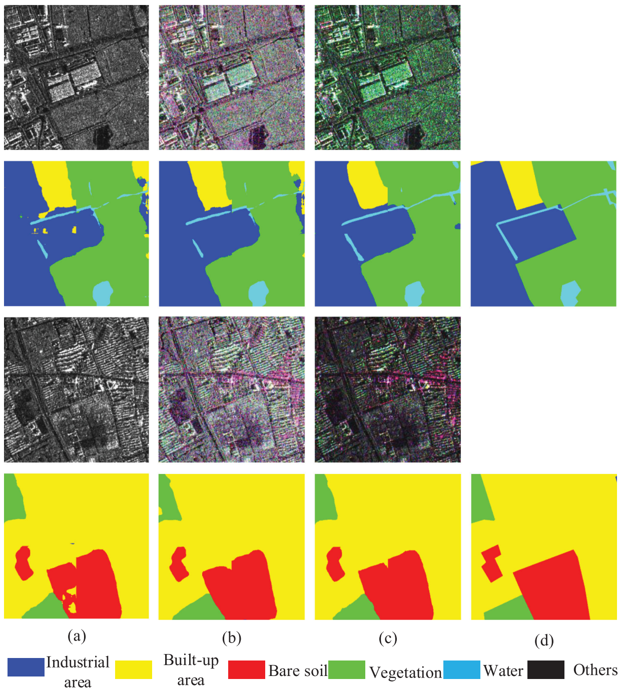


Fig. 8. Comparison of different unified SAR data preprocessing methods and segmentation results by our network on two sample areas. (a) the results of gray-single, (b) the results of RGB-median, (c) the results of rgb-mean, (d) the ground truth.

Challenge, we have won fourth place using our encoder–decoder network with unified SAR data preprocessing method NMT.

Fig. 8 shows the comparison of the SAR data preprocessing method and segmentation results on several sample dataset provided by the 2020 Gaofen Challenge. In Fig. 8, the first row and the third row are images generated by different preprocessing methods based on PolSAR data with four channels. From left to right are gray-single, RGB-mean, and RGB-mean, respectively. The second and fourth row is the corresponding segmentation results generated by our encoder–decoder network. The fourth column is the ground truth provided by the official website. Based on the unified SAR data preprocessing method, our network is capable of modeling context information from the SAR image. The segmentation results show that some confusion areas are correctly segmented and the object boundaries are continuous.

D. DCNN Kernel Visualization

To explain why the convolution kernel with ImageNet pretraining weights can be used for feature extraction of SAR images through fine-tuning, we use the gradient ascent method [44] to visualize the convolution kernel in Xception65. Fig. 9 shows the comparison of the several convolution kernel visualization in the same DCNN layer for three different datasets. It has a relatively clear texture structure in the convolution kernel visualization. Through the comparison of convolution kernel visualization, it shows that there is wavy and angular textures structure of DCNN pretrained by ImageNet, which are similar to fine-tuning on the FUSAR-Map dataset. For SAR images, the convolution kernel can still extract wavy and angular texture structures. But the texture information highlighted by the convolution kernel is different, like color or specific texture features. Fig. 9 also shows that the same DCNN structure has similarities in high-dimensional features and structural differences for different datasets. The visualization of the convolution kernel can explain to a certain extent that the use of pretraining weights to reasonably improve the starting point of network optimization, therefore the network takes less training time and retains some similar texture structures for convolution kernel.

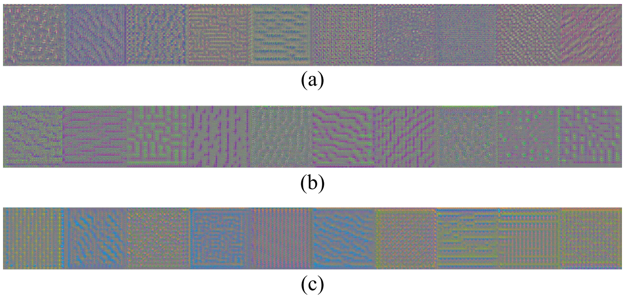

(a)
(b)
(c)

Fig. 9. Comparison of DCNN kernel visualization. (a) based on ImageNet dataset, (b) based on PolSAR dataset, and (c) based on FUSAR-Map dataset.

V. CONCLUSION

In this article, in the field of object-level semantic segmentation with high-resolution single-pol SAR images, we employ an encoder–decoder model based on transfer learning for fine-tuning to obtain the semantic segmentation results. To solve the lack of well pixel-labeled land cover datasets with single-pol SAR images, we develop a novel architecture of a semiautomated process to label high-resolution single-pol SAR images. Besides, we constructed a well pixel-labeled object-level semantic segmentation dataset FUSAR-Map. This new dataset contains 610 high-resolution GF-3 single-pol SAR images, with detailed information on latitude and longitude. Moreover, FUSAR-Map has an accurate outline of the water, road, and building categories, which provides a strong basis for object-level semantic segmentation task. Furthermore, based on our encoder–decoder network, our algorithm with a unified SAR data preprocessing method can achieve high performance of PolSAR semantic segmentation on the 2020 Gaofen Challenge.

The experimental results demonstrate an acceptable object-level semantic segmentation performance using a deep learning algorithm on FUSAR-Map. On the one hand, object-level semantic segmentation with single-pol SAR images is difficult to research work. On the other hand, at present, the well-annotated datasets and excellent algorithms accumulated in this field are rare. Also, compared with raw encoder–decoder networks, the networks pretrained for fine-tuning can often improve the semantics segmentation accuracy. Meanwhile, when training on FUSAR-Map, they always have the advantages of less time consumption and a more stable learning gradient.

In our future research, we hope to construct larger and category-rich datasets for object-level semantic segmentation of high-resolution GF-3 single-pol SAR images. Furthermore, we consider adding prior knowledge to the encoder–decoder network to improve the object-level semantic segmentation performance of single-pol SAR images.

Acknowledgment

The authors would like to thanks to Q. Song, X. Hou, Q. Guo, S. Li, Y. Qian, Z. Zhou, Y. Shi, X. Zhang, and S. Liu for helping to complete the work of image registration.

References

[1] G. Li, D. Lu, E. Moran, L. Dutra, and M. Batistella, “A comparative analysis of ALOS PALSAR L-band and RADARSAT-2 C-band data for land-cover classification in a tropical moist region,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 70, pp. 26–38, 2012.

[2] C. Rossi and E. Erten, “Paddy-rice monitoring using tandem-x,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 900–910, Feb. 2015.

[3] D. Cantorna, C. Dafonte, A. Iglesias, and B. Arcay, “Oil spill segmentation in SAR images using convolutional neural networks. A comparative analysis with clustering and logistic regression algorithms,” *Appl. Soft Comput.*, vol. 84, 2019, Art. no. 105716.

[4] M. Tao, F. Zhou, Y. Liu, and Z. Zhang, “Tensorial independent component analysis-based feature extraction for polarimetric SAR data classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2481–2495, May 2015.

[5] Q. Chen, G. Kuang, J. Li, L. Sui, and D. Li, “Unsupervised land cover/land use classification using PolSAR imagery based on scattering similarity,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1817–1825, Mar. 2013.

[6] Y. Li and Y. Q. Jin, “Target decomposition and recognition from wide-angle SAR imaging based on a Gaussian amplitude-phase model,” *Sci. China Inf. Sci.*, vol. 60, no. 6, 2017, Art. no. 062305.

[7] B. Zou, X. Xu, and L. Zhang, “Object-based classification of PolSAR images based on spatial and semantic features,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 609–619, 2020.

[8] Z. Zhang, H. Wang, F. Xu, and Y. Jin, “Complex-valued convolutional neural network and its application in polarimetric SAR image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.

[9] Y. Zhou, H. Wang, F. Xu, and Y. Jin, “Polarimetric SAR image classification using deep convolutional neural networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016.

[10] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, “A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 151,pp. 223 – 236, 2019.

[11] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, “Buildings detection in VHR SAR images using fully convolution neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.

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

[13] X. Hou, W. Ao, Q. Song, J. Lai, H. Wang, and F. Xu, “Fusar-ship: Building a high-resolution SAR-ais matchup dataset of Gaofen-3 for ship detection and recognition,” *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140303.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

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

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[17] 2020 Gaofen challenge on automated high-resolution earth observation image interpretation. [Online]. Available: http://en.sw.chreos.org

[18] M. Salehi, M. R. Sahebi, and Y. Maghsoudi,“Improving the accuracy of urban land cover classification using Radarsat-2 PolSAR data,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1394–1401, Apr. 2014.

[19] M. Jafari, Y. Maghsoudi, and M. J. Valadan Zoej, “A new method for land cover characterization and classification of polarimetric SAR data using polarimetric signatures,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3595–3607, Jul. 2015.

[20] R. G. Negri, L.V. Dutra, C. C. daFreitas, and D. Lu, “Exploring the capability of ALOS PALSAR L-band fully polarimetric data for land cover classification in tropical environments,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5369–5384, Dec. 2016.

[21] S. Uhlmann and S. Kiranyaz, “Integrating color features in polarimetric SAR image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2197–2216, Apr. 2014.

[22] A. Jain and D. Singh, “An optimal selection of probability distribution functions for unsupervised land cover classification of palsar-2 data,” *Adv. Space Res.*, vol. 63, no. 2, pp. 813 – 825, 2019.

[23] S. Arisoy and K. Kayabol, “Mixture-based superpixel segmentation and classification of SAR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1721–1725, Nov. 2016.

[24] L. Jiao, X. Tang, B. Hou, and S. Wang, “SAR images retrieval based on semantic classification and region-based similarity measure for earth observation,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3876–3891, Aug. 2015.

[25] A. Zhang, X. Yang, S. Fang, and J. Ai, “Region level SAR image classification using deep features and spatial constraints,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 36 – 48, 2020.

[26] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, “Deep metric learning based on scalable neighborhood components for remote sensing scene characterization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

[27] F. Xu, C. Hu, J. Li, A. Plaza, and M. Datcu, “Special focus on deep learning in remote sensing image processing,” *Sci. China Inf. Sci.*, vol. 63, pp. 1–2, 2020.

[28] C. O. Dumitru, S. Cui, G. Schwarz, and M. Datcu, “Information content of very-high-resolution SAR images: Semantics, geospatial context, and ontologies,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1635–1650, Apr. 2015.

[29] N. He, L. Fang, and A. Plaza, “Hybrid first and second order attention UNET for building segmentation in remote sensing images,” *Sci. China Inf. Sci.*, vol. 63, 2020, Art. no. 140305. [Online]. Available: https://doi.org/10.1007/s11432-019-2791-7

[30] 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, Jun. 2019.

[31] [Online]. Available: http://www.piesat.com.cn/product/pie-sar/index.html

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[33] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[34] François Chollet, *Xception: Deep learning with depthwise separable convolutions*, in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1251–1258.

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

[36] S. Lazebnik, C. Schmid, and J. Ponce,"Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Proc.*IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 2169–2178.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

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

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

[40] Martín Abadi, A. Agarwal, and P. Barham, "Tensorflow: Large-Scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[41] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

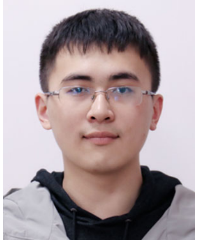
[42] Y. Ren, J. Yang, L. Zhao, P. Li, and L. Shi, "SIRV-based high-resolution PolSAR image speckle suppression via dual-domain filtering. *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5923–5938, Aug. 2019.

[43] C. Liu *et al.*, "Auto-Deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 82–92.

[44] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

Xianzheng Shi (Student Member, IEEE) was born in Shangrao, Jiangxi, China, in 1997. He received the B.S. degree in communication science and engineering from Fudan University, Shanghai, China, in 2019. He is currently working toward the master's degree with the Key Laboratory of Information Science of Electromagnetic Waves(MoE), Fudan University.

His research interests include SAR image processing, image semantic segmentation, and deep learning.



Shilei Fu (Student Member, IEEE) received the B.E. degree in communication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2017, and is currently working toward the Ph.D. degree in electromagnetic field and microwave technology from Fudan University, Shanghai, China.

His research interests include reciprocal translation between SAR and optical images, and 3D structural reconstruction from SAR images.

Jin Chen received the B.S. degree in electronic science and engineering from Nanjing University, Nanjing, China, in 2004, the M.S. and Ph.D. degree in electronic science and engineering from the National University of Defense Technology, Changsha, China, in 2006 and 2010, respectively.

From 2008 to 2009, he was a Visiting Scholar with the University of Calgary, Calgary, AB, Canada. He is currently a Research Associate with Beijing Institute of Remote Sensing Information, Beijing, China. His research interests include image classification and information fusion.

Feng Wang (Member, IEEE) received the B.S. and M.S. degrees from Information Engineering University, Zhengzhou, China, in 2006 and 2010, respectively, and the Ph.D. degree from Fudan University, Shanghai, China, in 2017.

He was with the Shanghai Institute of Technical Physics of The Chinese Academy of Sciences, Shanghai, as an Assistant Research Fellow. He is currently an Assistant Professor with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University. His research interests include radar imaging, hyperspectral remote sensing, and target recognition.

Feng Xu (Senior Member, IEEE) received the B.E. (Hons.) degree in information engineering from Southeast University, Nanjing, China, in 2003, and the Ph.D. (Hons.) degree in electronic engineering from Fudan University, Shanghai, China, in 2008.

From 2008 to 2010, he was a Postdoctoral Fellow with the NOAA Center for Satellite Application and Research, Camp Springs, MD, USA. From 2010 to 2013, he worked with Intelligent Automation Inc., Rockville, MD, USA, and with the NASA Goddard Space Flight Center, Greenbelt, MD, USA, as a Research Scientist. In 2012, he was selected for China's Global Experts Recruitment Program and subsequently returned to Fudan University, Shanghai, China, in 2013, where he is currently a Professor and the Vice Dean at the School of Information Science and Technology and the Vice Director of the MoE Key Laboratory for Information Science of Electromagnetic Waves. He has authored more than 40 papers in peer-reviewed journals, coauthored three books, and holds two patents, as well as publishing many conference papers. His research interests include electromagnetic scattering modeling, SAR information retrieval, and radar system development.

Dr. Xu was a recipient of the Second-Class National Nature Science Award of the IEEE Geoscience and Remote Sensing Society and the 2014 SUMMA Graduate Fellowship in the advanced electromagnetics area. He currently serves as an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and IEEE GRSS AdCom member. He is the Founding Chair of the IEEE GRSS Shanghai Chapter.