Automated High-Resolution Earth Observation Image Interpretation: Outcome of the 2020 Gaofen Challenge

Xian Sun, *Senior Member, IEEE*, Peijin Wang, *Member, IEEE*, Zhiyuan Yan, *Member, IEEE*, Wenhui Diao, *Member, IEEE*, Xiaonan Lu, Zhujun Yang, Yidan Zhang, Deliang Xiang, Chen Yan, Jie Guo, Bo Dang, Wei Wei, Feng Xu, *Senior Member, IEEE*, Cheng Wang, *Senior Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, Martin Weinmann, *Member, IEEE*, Naoto Yokoya, *Senior Member, IEEE*, and Kun Fu, *Member, IEEE*

***Abstract*—In this article, we introduce the 2020 Gaofen Challenge and relevant scientific outcomes. The 2020 Gaofen Challenge is an international competition, which is organized by the China High-Resolution Earth Observation Conference Committee and the Aerospace Information Research Institute, Chinese Academy of Sciences and technically cosponsored by the IEEE Geoscience and Remote Sensing Society and the International Society for Photogrammetry and Remote Sensing. It aims at promoting the academic development of automated high-resolution earth observation image interpretation. Six independent tracks have been organized in this challenge, which cover the challenging problems in the field of object detection and semantic segmentation. With the development of convolutional neural networks, deep-learning-based methods have achieved good performance on image interpretation. In this article, we report the details and the best-performing methods presented so far in the scope of this challenge.**

***Index Terms*—Convolutional neural networks, Gaofen Challenge, object detection and recognition, optical images, SAR images, semantic segmentation.**

Manuscript received May 28, 2021; revised June 24, 2021; accepted August 12, 2021. Date of publication August 24, 2021; date of current version September 15, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61725105 and in part by the National Major Project on China High-resolution Earth Observation System under Grant GFZX0404120201/GFZX0404120205. *(Corresponding authors: Xian Sun; Peijin Wang.)*

Xian Sun, Peijin Wang, Zhiyuan Yan, Wenhui Diao, Xiaonan Lu, Zhujun Yang, Yidan Zhang, and Kun Fu are with the Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@aircas.ac.cn; wangpj@aircas.ac.cn; yanzy@aircas.ac.cn; dwh1031@gmail.com; luxiaonan19@mails.ucas.ac.cn; yangzhujun19@mails.ucas.ac.cn; zhangyidan19@mails.ucas.ac.cn; fukun@mail.ie.ac.cn).

Deliang Xiang is with the Beijing University of Chemical Technology, Beijing 100029, China (e-mail: xiangdeliang@gmail.com).

Chen Yan is with the Computer School, Wuhan University, Wuhan 430079, China (e-mail: yanchen1997@whu.edu.cn).

Jie Guo is with the Nanjing Research Institute of Electronics Technology, Nanjing 210039, China (e-mail: 1101860289@qq.com).

Bo Dang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: bodang@whu.edu.cn).

Wei Wei is with the Northwestern Polytechnical University, Xi'an 710060, China (e-mail: weiweinwpu@nwpu.edu.cn).

Feng Xu is with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China (e-mail: fengxu@fudan.edu.cn).

Cheng Wang is with the Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: cwang@xmu.edu.cn).

Ronny Hänsch is with the German Aerospace Center, 82234 Weßling, Germany (e-mail: rww.haensch@gmail.com).

Martin Weinmann is with the Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: martin.weinmann@kit.edu).

Naoto Yokoya is with the Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan, and also with RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: naoto.yokoya@riken.jp).

Digital Object Identifier 10.1109/JSTARS.2021.3106941

I. Introduction

With the significant progress of various earth observation missions, a large amount of high-resolution data has been widely acquired, providing a variety of earth information. Automated high-resolution earth observation image interpretation has a wide range of applications, such as flight management, urban planning, and water-body monitoring [1]–[7]. However, the automatic interpretation of high-resolution remote sensing images is still challenging due to complex background and various objects in remote sensing images [8].

The 2020 Gaofen Challenge has covered two main approaches proposed in the field of automated interpretation: first, object detection and recognition, and second, semantic segmentation. The main purpose of object detection and recognition is to obtain the categories and locations of objects in an image. In the field of interpreting remote sensing images, object detection and recognition is significant to many rigid objects, such as airplanes, ships, and bridges. Common object detection and recognition algorithms consist of two categories: anchor-based algorithms and anchor-free algorithms. Anchor-based algorithms include one-stage methods and two-stage methods. For the two-stage methods, proposals are generated using a region proposal network (RPN) first. Then, they further classify and locate objects with candidate region proposals [9]–[12]. Compared with two-stage object detection algorithms, one-stage algorithms do not need to generate region proposals and, instead, they predict the classification and localization directly. One-stage methods are more efficient than two-stage methods due to their simple structures [13]–[17]. Recently, anchor-free object detection methods have been proposed in several works. For example, CornerNet

[18] and CenterNet [19] regard an object as a pair of keypoints. Anchor-free methods have few hyperparameters and can also achieve relatively good performance [18], [20], [21].

Compared to object detection, semantic segmentation needs to obtain the categories for each pixel in an image. Common semantic segmentation methods consist of encoder–decoder models and dilated-based models. For instance, the fully convolutional network [22], UNet [23], and SegNet [24] use the encoder–decoder structure to exploit the high-level feature maps. DeepLab [25] and ENet [26] adopt atrous convolutions to enlarge the receptive field of filters and aggregate multiscale context information. The aforementioned methods have made great progress in the field of image processing. However, automated high-resolution earth observation image interpretation is challenging due to the inherent characteristics of remote sensing scenes [27]–[29]. More specifically, remote sensing images typically cover large and often complex scenes with diverse background and a wide variety of objects exhibiting large differences in size. Some object categories even reveal high intracategory and low intercategory variations, making the interpretation even more challenging [28], [30].

To promote the development of this domain, the 2020 Gaofen Challenge on automated high-resolution earth observation image interpretation serves to bring together researchers from both computer vision and earth observation domains to discuss cutting-edge technologies on image interpretation and their applications.[1] It is an international competition, which is hosted by the China High-Resolution Earth Observation Conference Committee and the Aerospace Information Research Institute, Chinese Academy of Sciences and technically cosponsored by the IEEE Geoscience and Remote Sensing Society and the International Society for Photogrammetry and Remote Sensing.

We set six tracks in the 2020 Gaofen Challenge to meet different application requirements. Tracks 1, 2, and 3 aim to promote the research of object detection and recognition in optical images and synthetic aperture radar (SAR) images. Specifically, fine-grained airplane detection, bridge detection, and ship detection tasks are set in these tracks. The other three tracks focus on semantic segmentation in optical images and SAR images with respect to object categories, such as water body, road, tree, building, vehicles, and land.

To satisfy the high-resolution earth observation system construction requirements for major national scientific and technological projects, images used in the scope of the 2020 Gaofen Challenge are collected from the Gaofen-2 satellite and Gaofen-3 satellite. Specifically, we use the Gaofen-2 optical satellite data with 0.8–4 m resolution for airplane detection, bridge detection, and water-body segmentation tasks. And the Gaofen-3 SAR data with 1–5 m resolution are used for the tasks addressing ship detection, and semantic segmentation in polarimetric SAR data. To obtain high-quality data, we invited hundreds of experts taking more than three months to prepare the dataset. Finally, a large-scale and challenging dataset with various categories and tremendous object instances has been published for the 2020 Gaofen Challenge.

[1][Online]. Available: http://en.sw.chreos.org/

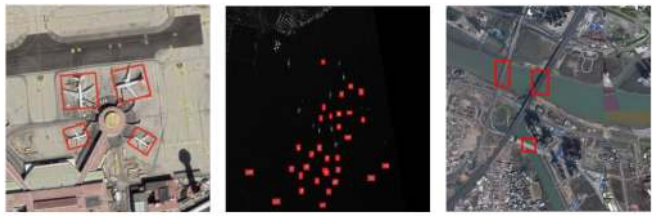

Fig. 1. Sample images and ground truths of object detection and recognition tasks. (a) Airplane detection. (b) Ship detection. (c) Bridge detection.

The rest of this article is organized as follows. We introduce the relevant details about the organization and dataset of the challenge in Section II. The overall information and results of participants in the challenge are discussed in Section III. We report the methods proposed by the winning teams of each track in Sections IV–IX. Finally, Section X, we make a conclusion to the 2020 Gaofen Challenge.

II. Data of the 2020 Gaofen Challenge

Data from Chinese Gaofen satellites are provided for all six tracks of the 2020 Gaofen Challenge. The data used in the challenge include multiscale, multiview, multiresolution optical remote sensing images and SAR images, which are all collected from Gaofen-2 and Gaofen-3 satellites with the resolution ranging from 1–4 and 1–5 m, respectively. The data containing more than 10 000 images are annotated by more than 100 experts over three months. Some images and corresponding ground truth labels of each track are shown in Figs. 1 and 2. Details of the data provided for the 2020 Gaofen Challenge Tracks 1 to 6 are presented in the following.

1) Data for Track 1 (airplane detection and recognition in optical images) are provided by the Gaofen-2 satellite. The scenes include the main civil airports in the world, such as Sydney Airport, Beijing Capital International Airport, Shanghai Pudong International Airport, Hong Kong Airport, Tokyo International Airport, and many more. The data contain 3000 satellite images with a spatial resolution of 0.8 m. Each image is of the size 1000 $\times$ 1000 pixels and contains ten categories of airplanes (i.e., Boeing 737, Airbus A321, Airbus A330, Boeing 747, Boeing 777, Boeing 787, Airbus A220, COMAC ARJ21, Airbus A350, and other) exhibiting a wide variety of orientations and scales.
2) Data for Track 2 (ship detection in SAR images) are collected from Gaofen-3 satellite. It contains 1000 SAR images with a spatial resolution ranging from 1–5 m. Each image is of the size 1000 $\times$ 1000 pixels and includes ships exhibiting a wide variety of orientations and scales. The scenes include the main civil ports in the world, such as Victoria Harbour, Port of Sanya, Incheon Port, etc.
3) Data for Track 3 (automatic bridge detection in optical satellite images) are provided by the Gaofen-2 satellite with the resolution ranging from 1–4 m. Each image contains at least one bridge. There are 3000 images with

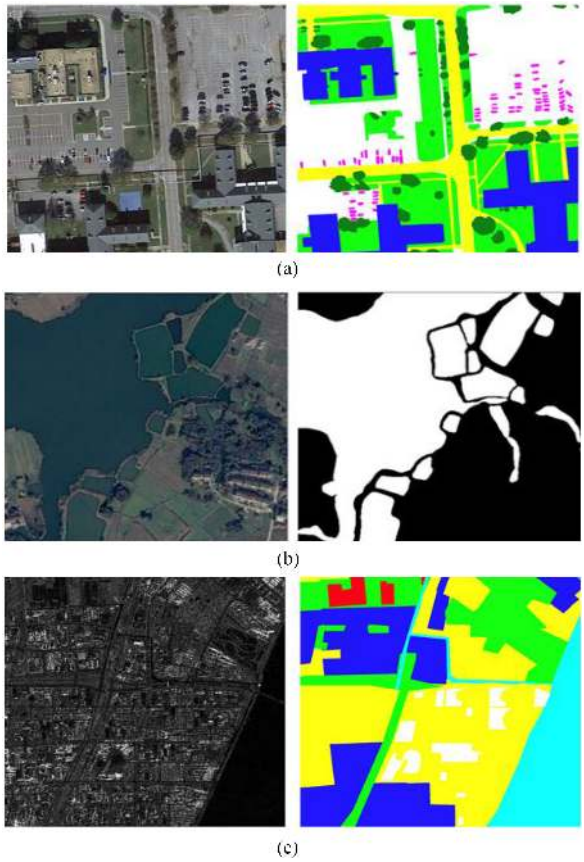

Fig. 2. Sample images and ground truths of semantic segmentation tasks. (a) Semantic segmentation in optical images. (b) Water-body segmentation in optical images. (c) Semantic segmentation in fully polarimetric SAR images.

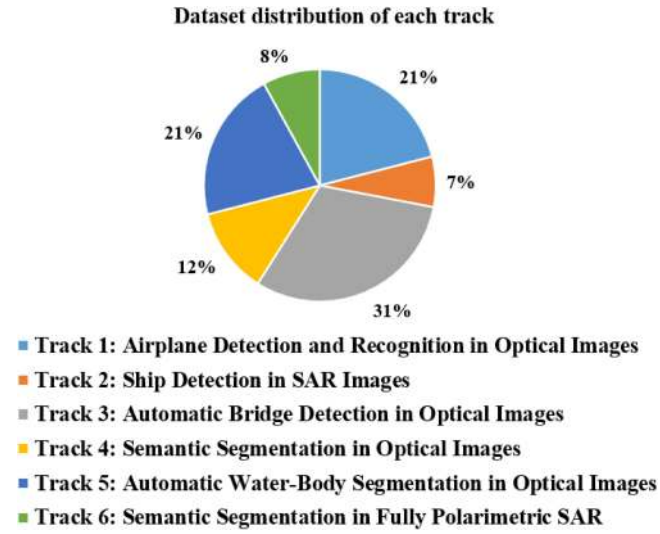

Fig. 3. Dataset distribution of 2020 Gaofen Challenge.

TABLE I
DATASET STATISTICS FOR CHALLENGE

Track	The number of images			Image resolution (pixel)	Size (GB)
	train set	preliminary test set	final test set		
1	1000	1000	1000	1000×1000	5
2	300	400	300	1000×1000	2
3	2000	1000	1500	512-12000	5.3
4	500	300	1000	512-5000	3.24
5	1000	500	1000	492-2000	1.2
6	500	300	400	512-1500	2

different sizes ranging from 667 × 667 to 1001 × 1001 pixels in the bridge dataset.

4) Data for Track 4 (semantic segmentation in optical satellite images) are provided by the Gaofen-2 satellite with 0.8 m resolution. Each image is annotated with respect to nine categories of ground objects at the pixel level, including road, building, shrub and tree, lawn, land, water body, vehicle, impervious ground, and others. There are 1800 images with the size ranging from 512 to 5000 pixels.
5) Data for Track 5 (automatic water-body segmentation in optical satellite images) are provided by the Gaofen-2 satellite with the resolution ranging from 1–4 m, covering rivers and lakes in large scope. There are 2500 images with a size ranging from 492 to 2000 pixels in the water-body dataset.
6) Data for Track 6 (semantic segmentation in fully polarimetric SAR images) are provided by the Gaofen-3 satellite with 1–3 m resolution, containing four polarization modes (i.e., HH, VV, HV, and VH). Six categories, including water body, building, industrial area, lawn, land, and others, are annotated at the pixel level for each image. There are 1200 images with the size ranging from 512 to 1500 pixels.

The aforementioned datasets are provided for the training set, preliminary test set, and final test set of the 2020 Gaofen Challenge. More information about the distribution of the data provided for the different tracks is shown in Fig. 3 and Table I.

III. ORGANIZATION, SUBMISSIONS, AND RESULTS

Six independent and distinctive tracks were organized in the 2020 Gaofen Challenge. Considering the practical application, three of the six tracks addressed the task of object detection and recognition, and the remaining three addressed the task of semantic segmentation, as described in Sections III-A–III-F. For the tracks on object detection and recognition (Tracks 1–3), the mean Average Precision (mAP) [31] with the Intersection over Union (IoU) of 0.5 is used to evaluate the results. For a given ground truth and the predicted result, TP, FP, and FN are selected according to an IoU threshold of 0.5. Then, the precision and recall are calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

According to Pascal VOC 2012, the AP of each class is calculated based on precision and recall, and then the mAP can be obtained. For the tracks on semantic segmentation (Tracks 4–6), the frequency weighted IoU (FWIoU) [32] is used as an accuracy evaluation indicator, and its calculation method is as

follows:

$$\text{FWIoU} = \frac{1}{\sum_{i=0}^{N}\sum_{j=0}^{N} s_{ij}} \sum_{i=0}^{N} \frac{\sum_{j=0}^{N} s_{ij} s_{ii}}{\sum_{j=0}^{N} s_{ij} + \sum_{j=0}^{N} s_{ji} - s_{ii}} \tag{3}$$

where N is the number of categories, and s_{ij} represents the number of pixels belonging to category i and predicted to be category j.

In addition to the accuracy of image interpretation, the inference time and the quality of technical reports are also taken into account in the final results. The final score is defined as

$$\text{score} = 70\% \cdot \text{accuracy} + 20\% \cdot \text{speed} + 10\% \cdot \text{report}. \tag{4}$$

Section III-G shows baseline solutions achieved for the 2020 Gaofen Challenge, whereas the participating and winning methods are analyzed in Sections III-H and III-I, respectively.

A. Track 1: Airplane Detection and Recognition in Optical Images

Track 1 is dedicated to the detection and recognition of airplanes in optical satellite images. For each image in the dataset, there is an XML file with the same name for describing annotation information, such as the image coordinates and object information of airplanes. Each airplane instance in the images is annotated by the corresponding category information and location with an oriented bounding box [33].

B. Track 2: Ship Detection in SAR Images

Track 2 is dedicated to the detection of ships in SAR images, where the goal is to locate the ships in SAR images. In each image, the coordinates of ships are described in a predefined format. Compared with Track 1, each XML file corresponds to one image, including the coordinates of the horizontal bounding box for each ship.

C. Track 3: Automatic Bridge Detection in Optical Satellite Images

The goal for Track 3 is to locate bridges in large-scale optical satellite images. The labeling format is similar to Track 2, and the coordinates of each bridge are given as horizontal bounding boxes.

D. Track 4: Semantic Segmentation in Optical Satellite Images

Track 4 is dedicated to semantic segmentation in optical satellite images. In this case, a pair of images are provided for each scene, as shown in Fig. 2(a). One is the original optical satellite image, and the other is an image annotated with the ground truth whose size is the same as for the previous satellite image. In ground truth images, different categories are marked with different RGB values in pixel level.

E. Track 5: Automatic Water-Body Segmentation in Optical Satellite Images

To detect the water body in remote sensing images, the 2020 Gaofen Challenge set up Track 5 whose purpose is to locate the water body in the optical satellite images with pixel level. Same as Track 4, the original optical satellite images and ground truth images are provided for water-body segmentation.

F. Track 6: Semantic Segmentation in Fully Polarimetric SAR Images

In addition to the track for semantic segmentation in optical satellite images, a semantic segmentation track for SAR images was also set up. Its goal is to classify the features in SAR satellite images with pixel level. As shown in Fig. 2(c), the dataset format is the same as for Tracks 4 and 5.

G. Baseline Solutions

Classic object detection and semantic segmentation networks are used as baseline solutions of each track separately. A two-stage object detection method in the form of a Faster RCNN [10] based on ResNet-50 is used for object detection tracks. The Faster RCNN is a detector with good performance, which generates anchors through an RPN and completes regression and classification after Region-of-Interest (RoI) pooling. For Track 1, an angle information regression is added to realize rotated boxes regression. For semantic segmentation tracks, we use DeepLab V3 [34] based on ResNet-50 as a baseline solution. DeepLab V3 improves the atrous spatial pyramid pooling (ASPP) structure and uses multiple scales to obtain better segmentation results.

H. Participation

There are 701 teams from 253 affiliations, with 2023 competitors joining in the 2020 Gaofen Challenge. The competitors come from more than 20 countries, including China, England, Germany, France, Japan, Australia, Singapore, India, Sweden, etc. The total number of track registrations is 1584 times, of which the tracks for object detection were registered 860 times with 54%, and the tracks for semantic segmentation were registered 724 times with 46%. It can be seen that the popularity of object detection tracks and semantic segmentation tracks is similar, indicating that both of them are widely studied in the field of the automated interpretation of high-resolution earth observation data. In total, there were 5719 submissions for all tracks. The specific numbers of submissions for each track are shown in Fig. 4.

I. Best-Performing Approaches and Discussion

The top six teams of each track were awarded winning places. In this article, we mainly introduce the methods of champion teams. The brief introduction of the champion teams for Tracks 1–6 is as follows.

1) *First place in Track 1*: The *Detect AI* team; Chen Yan, Wenxuan Shi, Tao Qu, Chu He, and Dingwen Wang from Wuhan University, China; with attention mechanism and deformable convolution based on Faster RCNN.
2) *First place in Track 2*: The *challenger_nriet* team; Guo Jie, Zhuang Long, Xie Cong, and Zheng Ping from the Nanjing Research Institute of Electronics Technology, China; with

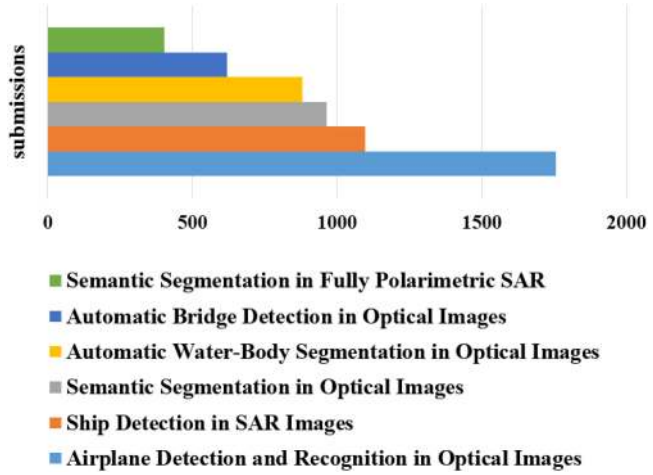

Fig. 4. Total submission of each track.

SPPNet and an ensemble of adaptively spatial feature fusion (ASFF) module with Faster RCNN.

3) *First place in Track 3*: The *MDIPL-lab* team; Yuxuan Sun, Wei Li, Wei Wei, and Lei Zhang from Northwestern Polytechnical University, China; with ResNet50 and HRNet-w32 on Faster RCNN.
4) *First place in Track 4*: The *BUCT Tu Xiang Jie Yi Xiao Fen Dui* team; Fei Ma, Jun Ni, Ruirui Li, Yingbing Liu, Feixiang Zhang, and Fan Zhang from the Beijing University of Chemical Technology, China; with an ensemble of ResNet101-V2 on DeepLab V3+ [35]–[37].
5) *First place in Track 5*: The *Wu Da Ti Shui Gao Fen Dui* team; Bo Dang, Jintao Li, Tianyi Gao, and Yansheng Li from Wuhan University, China; with multistructure deep segmentation network [38]–[41].
6) *First place in Track 6*: The *BUCT Tu Xiang Jie Yi Xiao Fen Dui* team; Fei Ma, Jun Ni, Ruirui Li, Yingbing Liu, Feixiang Zhang, and Fan Zhang from the Beijing University of Chemical Technology, China; with an ensemble of conditional random field (CRF) with DeepLab V3+ [42].

Looking at the overall trend, the methods used by the winning teams were all improved and extended on the basis of the well-established models. The methods used by the champion teams of each track are described in detail in Sections IV–IX.

IV. First Place in the Airplane Detection and Recognition in Optical Images: Detect AI

In this section, we introduce the winning method proposed for airplane detection and recognition in optical images. Airplane detection is one of the most common detection applications in rotation detection. The similarity of airplanes increases the difficulty of fine-grained detection regarding different types of airplanes. To solve this problem, Detect AI team proposes a rotation detection method based on an attention mechanism. First, they use the attention mechanism to extract the texture features of the aircraft in the feature representation stage for classification and add deformable convolutional network (DCN) to extract the irregular structure features of the aircraft. Finally, Detect AI team used many common techniques in the training process without spending extra time.

Detect AI team first selected R2CNN [43], RRPN [44], RoI transformer [45], S^2A-Net [46], and other algorithms in the competition. After basic training and verification of these algorithms, S^2 A-Net has achieved the highest detection performance, so Detect AI team uses S^2 A-Net as their detection benchmark.

An S^2 A-Net-based airplane detection method is proposed and optimized in the feature representation stage and the object regression stage. The overall framework of the method is shown in Fig. 5. The optimization of each part will be introduced below.

A. Deformable Convolutional Network

It is challenging to acquire the structural features and information of the airplanes by common convolution because of their irregular shapes. The common convolutional neural network mainly uses regular square grid points to sample the fixed position, which cannot learn the structural characteristics of the airplanes.

To solve the aforementioned problems, this section introduces deformable convolution by adding two-dimensional offset values and pooling operations to achieve the freedom of convolutional kernel and pooling to learn the irregular shape of the airplanes [47]. Specifically, the bias value of the convolutional kernel and pooling layer are obtained through an additional convolutional layer and the feature map with the RoI together, respectively. Since the biased models are all simple layers, the number of parameters and calculations required for this process are relatively small, and end-to-end training can be achieved through the gradient backpropagation algorithm.

B. Orientation-Sensitive Regression

Detect AI team first adopts active rotating filters (ARFs) to learn the orientation information. The ARF filter can rotate several times during convolution to generate orientation features. Using ARF in the deep learning network can obtain orientation-invariant features with encoded orientation information. Object classification tasks benefit from orientation-invariant features, whereas bounding box regression tasks require sensitive features. Then, Detect AI team conducts the pooling layer to the orientation-invariant feature and obtains the orientation-sensitive features for the bounding box (bbox) regression.

C. Experiment

There are 1000 images with ground-truth labels in the training data, and the size of each image is 1024×1024 pixels. The data contain ten types of airplane samples. Detect AI team first divides the training set into two parts, 800 images are used for training, and 200 images are used for validation. They randomly rotate the training set and expand the training set to five times. They made an automatic contrast argumentation based on the dataset and applied mixup [48] to the dataset, which greatly expands the training samples. At the same time, they collect airplane images from the public remote sensing dataset as training

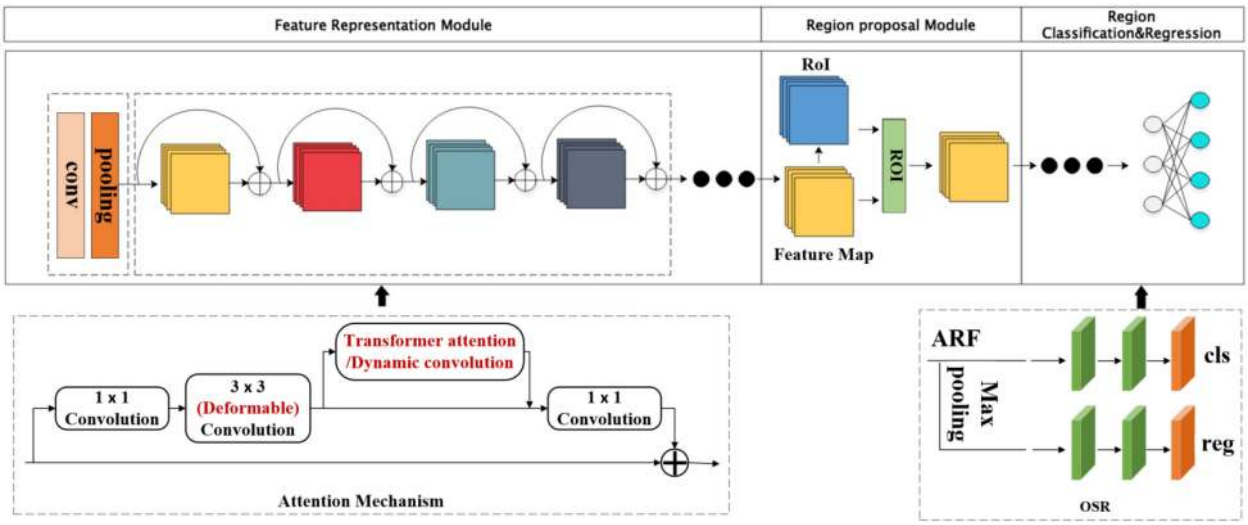

Fig. 5. Network of airplane detection method.

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS ON AIRPLANE DETECTION

Model	mAP(%)
M1: S^2A-Net+pretrained model	80.32
M2: Attention+DCN+OSR	84.24
Our: mixup+multi scales	**85.09**

Bold entities mean the best performance model in the table.

data for pretraining. The data source is mainly from DOTA [59], UCAS-AOD [49], NWPU VHR-10 [50], and RSOD-Dataset [51]. A total of 7449 images containing airplanes are collected for pretraining, and the model is tested on the competition data.

This article verifies the proposed method on the test set, and compares the performance of the S^2 A-Net. The object detection results are provided in Table II. The visualization of airplane detection results are shown in Fig. 6.

D. Discussion

Airplane detection and recognition play an important role in both military and civilian fields. Detect AI team analyzes the characteristics of optical airplane remote sensing images, and carry out research on its object characteristics. According to its existing problems and challenges, they improve and optimize the existing detection framework. On the one hand, they use attention and DCN to learn the texture features and irregular shape features of the airplanes. On the other hand, they propose a new orientation-sensitive bbox regression method, with which the bbox of the object is regressed more accurately.

V. FIRST PLACE IN THE SHIP DETECTION IN SAR IMAGES: CHALLENGER_NRIET

In this section, we introduce the winning method proposed for ship detection in SAR images. There are a few particular challenges for SAR ship detection, as analyzed as follows.

1) A large number of small objects. Compared with natural scenes, there are many objects in small size in remote sensing imagery. The SAR images provided by the official website are acquired from the Gaofen-3 satellite with a spatial resolution ranging from 1–5 m. This means that for a 20-m ship, it will be only 4–20 pixels in the provided SAR images.
2) Rotation invariance. Objects in satellite imagery may have any orientation. For example, a ship can sail at any angle on the sea.
3) Insufficient training data. Compared with optical images, it is more difficult to obtain SAR images [52]. Therefore, the number of available SAR images is less than that of optical images.
4) Wide range of aspect ratios. Ships may have a relatively large aspect ratio in satellite images compared with most other objects. Therefore, anchor-based CNN methods have difficult setting anchors covering ships with different aspect ratios [53], [54].

Challenger_nriet team uses a bag of tricks to alleviate these problems, which are described in the following sections.

A. Baseline Model

In the face of these challenges, they adopt YOLOv3 [55] as the baseline model. Ships have a large range of aspect ratios in SAR images compared with general objects in optical images. Thus, the nine anchors in the YOLOv3 model cannot cover scales and aspect ratios of ships in SAR images very well. Therefore, they use guided anchors to adjust the shape of the anchor to fit the desired shape.

As shown in Fig. 7, a spatial pyramid pooling (SPP) layer added in the YOLOv3 model can combine local and global features, making features contain richer information and have stronger representation power.

Furthermore, they use the ASFF [56] model to filter conflictive information to control the inconsistency between different scales

outputted by the feature pyramid network (FPN) of YOLOv3. An extra IoU loss function [57] is added to the original smooth L1 loss for more accurate bounding box regression.

The baseline model is trained with the 300 training images downloaded from the official website for 100 epochs. The proposed model is trained using stochastic gradient descent (SGD) [58] algorithms with the cosine learning rate schedule from 0.001 to 0.00001. The values of weight decay and momentum are 0.0005 and 0.9, respectively.

B. Bells and Whistles

In this part, we introduce some bells and whistles to improve the model's ability in their method.

1) Data Augmentation: They add SAR-Ship-Dataset [59] to train their model. The image size of SAR-Ship-Dataset is 256 × 256 pixels. Therefore, they randomly select 2 × 2/3 × 3/4 × 4 images and stitch them together and rescale the stitched images to a size of 1000 × 1000 pixels. The Fig. 8 shows the results of data augmentation. They also involve mirroring, cropping, distorting, and random-affine transformations for data augmentation. Moreover, Challenger_nriet team adds the HRSID dataset [60] to the training set to train the model.

2) Finer-Grained Features and Denser Grid: Many ships in the SAR images are relatively small compared with objects in natural scenes. As a result, they remove stage 5, which has a stride of 32 in the YOLOv3 backbone network. Instead, they output stage 2, stage 3, and stage 4 to detect ships in different scales. To keep the depth of output features consistent, they add more convolutional layers with shortcut connections in stage 2. Finally, they get finer-grained features while still keeping enough semantic information.

3) Multiscale Training: As shown in Fig. 9, challenger_nriet team adopts multiscale training with the random crop. First, they randomly crop image patches from images in the dataset, and the scale of cropped patches is randomly sampled from 384, 416, 448, 480, 512, 544, 576, 608, and 640, then the cropped patches are rescaled to a fixed size of 512 × 512 pixels for training. They only keep those patches with ships.

4) Scale-Aware Loss Function: To focus more on small ships, Challenger_nriet team set different weights on the loss function according to the size of ships. The weight of L1 loss and IoU loss of small ships are larger than for large ships. The weights are calculated as

$$\text{weight} = \begin{cases} 1, & \text{if } \frac{(w \cdot h)}{(W \cdot H)} > 0.01 \\ 3 - \frac{(200 \cdot w \cdot h)}{(W \cdot H)}, & \text{otherwise} \end{cases} \tag{5}$$

where w and h represent the width and height of the ship, respectively. W and H represent the width and height of the image, respectively.

5) Multiple Weights Fusion: Challenger_nriet team trains the model for 100 epochs, then averages the weights from epochs of 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, and 100 to achieve the final network weights for more robust testing.

6) Deeper Network and More Training Epochs: They add more convolutional layers in stage 2 and train the models for more epochs (200 epochs). This strategy can keep finer-grained features while still catching enough semantic information.

7) Multiscale Testing: They adopt three scales for testing, 800 × 800, 1056 × 1056, and 1248 × 1248 pixels. They first obtain the network outputs of each scale. Then, they concatenate them together and perform nonmaximum suppression (NMS) to get the final detection results. Besides, they change the NMS threshold from 0.65 to 0.55.

TABLE III
RESULTS ACHIEVED ON THE TEST SET DATASET

Methods	mAP (%)
Baseline	22.95
+ Data augmentation	39.51
+ More training data	47.29
+ Finer grained features and denser grid	50.72
+ Multi-scale training and Scale-aware loss function	51.60
+ Multiple weights fusion	51.92
+ Deeper network and more training epochs	53.72
+ Multi-scale testing	**55.14**

Bold entities mean the best performance model in the table.

C. Results and Discussion

Challenger_nriet team reports the detection results on the preliminary test set downloaded from the official website of the contest. Details are demonstrated in Table III. The detection results of ships are shown in Fig. 10.

The model achieves 59.95% mAP for the test dataset provided in phase 3 of the contest. Finally, to ensure wider testing scales, they instead adopt three scales of 736 × 736, 1056 × 1056, and 1344 × 1344 pixels for testing. And finally get 60.58% mAP for the test dataset provided in phase 3 of the contest.

VI. First Place in the Automatic Bridge Detection in Optical Satellite Images: MDIPL-lab

Bridge detection aims at automatically detecting and locating bridges in remote sensing images. As a branch of the object detection task, many detection methods for natural scenes can be also used for bridge detection. For example, Faster-RCNN [61], a representative two-stage detector, can have stable performance under different tasks. Therefore, the proposed method is modified based on Faster-RCNN. To solve the problem of the small dataset and single scene, ResNet50+DCN and HRNet-W32 are adopted as the backbone network in this method. In view of the characteristics of remote sensing images with large variation in object orientation and complicated illumination conditions, we adopt horizontal flip and random 90° rotation in data argumentation. In addition, FPN and multiscale training are adopted to deal with the variance of object sizes.

A. Model Structure

Faster RCNN+FPN and the random horizontal flip with probability $P = 0.5$ are used as the benchmark methods. Considering that some of the bridges are located on the diagonal of the target box, and most of these target boxes are rivers, we used DCNv2

TABLE IV
RESULTS OF DIFFERENT METHODS OR STRATEGIES WITH BASELINE

Methods or Strategies	mAP(%)
Faster RCNN with R50-FPN + Filp	59.11
+ DCNv2	59.72
+ Multi-Scale Train	70.40
+ RandomRotate 90°	74.10
+ Multi-Scale Test	**78.00**

Bold entities mean the best performance model in the table.

to extract features more effectively and focus on effective information.

It was observed that the captured scenes in the bridge dataset are relatively monotonous while the size of the bridges varied greatly and there are many small bridges. Therefore, we chose the backbone network HRNet-W32, which is more advantageous in integrating multiscale features compared with ResNet50 and ResNet101. At the same time, due to the relatively monotonous scene, deeper networks, such as ResNet101, are not significantly improved over ResNet50.

Integration of multiple different models has proven to be a relatively effective way to improve accuracy. For the experiments based on test dataset, they tried NMS, SoftNMS [62], VOTE, and NMS using IOF instead of IOU, and finally chose SoftNMS as the integration method.

B. Data and Training Strategy

Reasonable data argumentation can artificially control the prior rules of scene distribution and increase the amount of data, which is another strategy to improve the performance of the model. To simulate the change of camera rotation angle during data acquisition and enhance the diversity of data, we added random rotations of 0°, 90°, 180°, and 270° to the random horizontal flip in the method.

A multiscale method is introduced in the training and testing process to solve the large object size difference problem. At the same time, considering that the images in the dataset have two resolution sizes of 1001×1001 and 668×668 pixels, the size of image is randomly scaled between 600 and 1200 pixels during the training.

C. Experiment

All experiments are conducted on the object detection framework MMDetection [63]. There are a total of 2000 images in the dataset. Since some consecutive images are taken with the same scene, the first 667 images are selected as the validation set to avoid data duplication. Twelve epochs are trained in each experiment. The initial learning rate is 0.00125, the batch size is usually 4 or 8, and the learning rate decreased by 1/10 in the 8th and 11th epoch. SGD with a momentum of 0.9 and weight attenuation of 0.0001 is used as the optimizer. The probability of both a horizontal flip and a subsequent rotation of 90° is 0.5.

On the baseline model, the results using different methods and strategies are shown in Table IV. The team adds the DCNv2, multiscale training, and data enhancement strategy. The detection results of bridges are shown in Fig. 11.

TABLE V
RESULTS OF DIFFERENT MODELS

Models	mAP(%)
R50-FPN-DCNv2 + Aug	78.00
HRW32-FPN + Aug	82.31
Ensemble of the above two using softNMS	83.60

Fig. 6. Airplane detection results of the proposed method.

The different models and their integration effects are shown in Table V. HRW32 represents HRNet-W32, and the Aug represents the argumentation strategy. Finally, the integration of the two models using SoftNMS yielded slightly better results than either vote or NMS.

D. Discussion

This method is aimed at the automatic bridge detection in remote sensing imagery. On the basis of Faster-RCNN, it adjusts the backbone network selection and data argumentation strategy according to the characteristics of single scenes in the dataset, and finally selects two models to integration and obtain 83.6% AP.

VII. FIRST PLACE IN THE SEMANTIC SEGMENTATION IN OPTICAL IMAGES: BUCT

In this section, we introduce the winning method proposed for the semantic segmentation in optical images. The method proposed by the team is a deep semantic segmentation network combined with multiscale spatial features. The purpose is to obtain features of different scales and use the regional features of superpixels to combine global information to improve the performance of segmentation. This method first uses ResNet101-V2 as the backbone network of Deeplab V3+ [37] to extract image features and then uses two subnetworks of "pixel-level semantic segmentation" and "superpixel-level semantic segmentation based on boundary feature enhancement" for semantic segmentation. The framework is shown in Fig. 12.

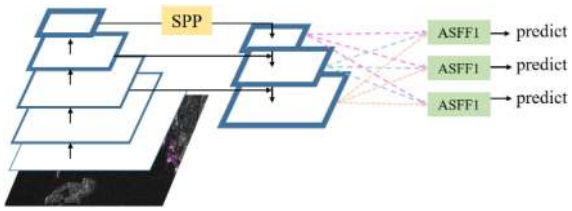

Fig. 7. YOLOv3 architecture with SPP module and ASFF module.

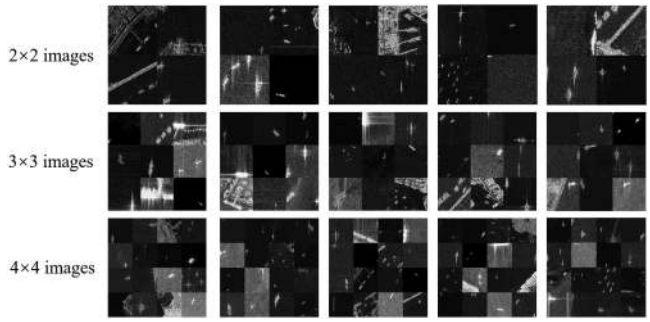

Fig. 8. Randomly selecting 2 × 2/3 × 3/4 × 4 images and stitching them together.

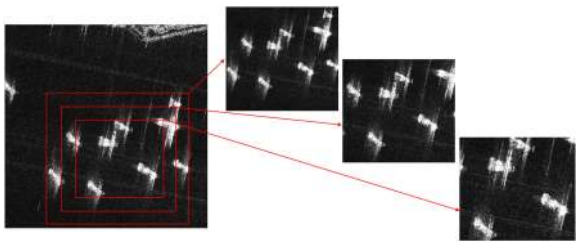

Fig. 9. Adopting multiscale training with random crop.

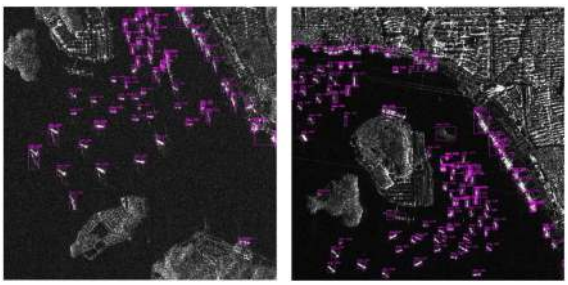

Fig. 10. Detection results of ships in SAR images.

A. Data Preprocessing

Due to the complexity of the categories of the objects in the dataset, the BUCT team augments the existing data. BUCT team used the following methods to augment the data.

1) Overlap cropping: The training images are cropped into a fixed size with overlap.
2) Spatial transformation: It includes horizontal and vertical flipping, random rotation at any angle with the image center as the origin, scaling outward or inward with a certain proportion, random cropping, and shifting in the X or Y direction (or both).
3) Random noise addition: Gaussian noise is randomly added to the data to prevent the CNN from learning useless high-frequency features, thereby reducing the probability of overfitting.

Fig. 11. Visualization of the bridge detection results.

B. Deep Semantic Segmentation Network Combined With Multiscale Spatial Features

1) Feature Extraction Based on ResNet101: Using the activation function on the residual branch, the information propagation speed of ResNet101 will be faster in the back propagation and forward propagation. It allows the network to get better results and avoids the problem of vanishing gradients.

2) Pixel-Level Semantic Segmentation Based on Deeplab V3+: The pixel-level feature classification subnetwork uses the DeepLab V3+ network to segment objects at the pixel level. The feature maps embedded in the first four convolutional blocks of ResNet101 are sent to the ASPP module to represent different local and global information proportions. Then, the feature extraction result and the low-resolution information in the encoder are cascaded up-sampling, and finally the pixel loss is obtained. DeepLab is a method that combines deep convolutional neural networks (DCNNs) and probabilistic graphical models (Dense CRFs). DCNNs use atrous convolution to expand the receptive field to solve resolution reduction caused by down-sampling or pooling in DCNNs. Dense CRFs can consider the mutual influence between adjacent pixels.

3) Superpixel-Level Semantic Segmentation Branch Based on Edge Feature Enhancement: The segmentation results of DeepLab V3+ at the edge are not very good, and there is strong segmentation noise and fuzzy edge. As a result, the BUCT team uses a high-precision end-to-end superpixel generation

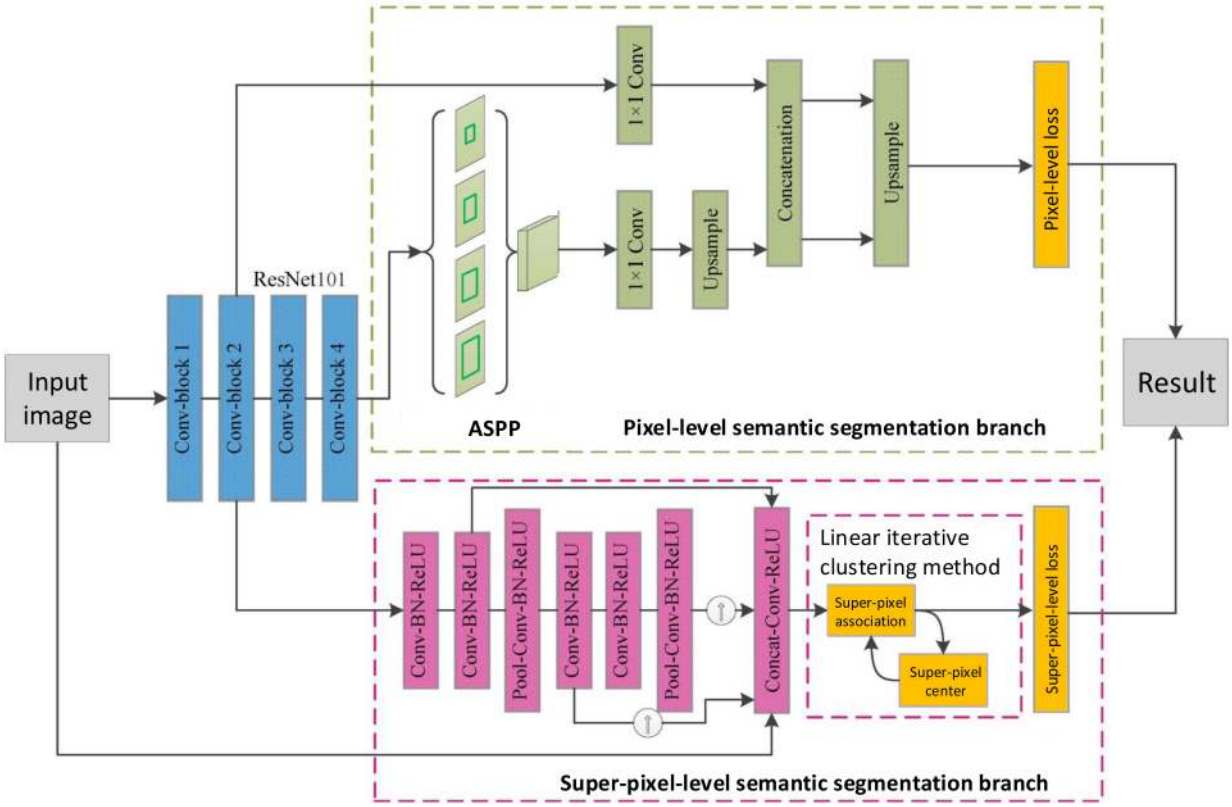

Fig. 12. Framework of the proposed semantic segmentation method.

method. This method can be implemented using a deep convolutional network, which is trained together with the semantic segmentation network, so the segmentation accuracy is greatly improved. Then, with the help of the pixel and each superpixel correlation matrix and ground truth, the superpixel-level loss function is calculated. The goal of the loss function is to ensure that the labels of pixels belonging to the same superpixel are as consistent as possible. In addition, when calculating the loss function, the BUCT team uses the ground truth to give more weight to the pixels near the edge of the objects.

At the superpixel generation stage, traditional simple linear iterative clustering method has nondifferentiable step. Therefore, this method cannot be introduced into convolutional neural networks. As a result, the BUCT team proposes a differentiable linear iterative clustering method. This method models the association between pixels and superpixels $Q \in R^{n\times m}$. For the pixel p and superpixel i in the tth step, the association is denoted as

$$Q_{pi}^{t} = e^{-D(I_p, S_i^{t-1})} = e^{-\|I_p - S_i^{t-1}\|^2} \tag{6}$$

where n and m denote the number of pixels and superpixels, respectively. I_p and S_i are the features of pixels and superpixels, respectively. The cluster center of the superpixels is defined as the weighted sum of pixel features

$$S_i^t = \frac{1}{Z_i^t} \sum_{p=1}^{n} Q_{pi}^t I_p. \tag{7}$$

$Z_i^t = \sum_p Q_{pi}^t$ is the normalized constant. Considering the calculation to obtain Q_{pi}, m is set to be 9 in the training stage. Since the superpixel is an oversegmentation of the image, the segmentation label of the image can be used as the supervision information of the superpixel segmentation. Associated matrix $Q_{(p,sp)}^t$ represents the relationship between pixels and superpixels. The annotation results of pixels can be mapped to the superpixels by applying the column normalization to $Q_{(p,sp)}^t$. Similarly, the annotation results of superpixels can be mapped to the pixels by applying the row normalization to $Q_{(p,sp)}^t$. If the annotation of pixels is defined as G, that of superpixels can be denoted as

$$G^* = Q_{\text{row}}\, Q_{\text{col}}^{\top}\, G. \tag{8}$$

C. Loss Function

In the training process, the pixel-level segmentation branch outputs the predicted pixel label matrix $P \in \mathcal{R}^{n\times 1}$. The superpixel segmentation branch outputs the pixel-superpixel correlation matrix $Q \in \mathcal{R}^{n\times m}$, where n and m are the number of pixels and superpixels, respectively, and the true label matrix is denoted as $G \in \mathcal{R}^{n\times 1}$.

1) Pixel-Level Loss Function: The pixel-level loss can be described as the cross-entropy loss between the predicted label and the ground truth, which is defined as

$$\mathcal{L}_{\text{pixel}} = \mathcal{L}(G, P). \tag{9}$$

TABLE VI
COMPARISON OF SEMANTIC SEGMENTATION RESULTS IN OPTICAL IMAGES

Models	Augmentation	ResNet50		ResNet101	
		Accuracy	Time(ms)	Accuracy	Time(ms)
UNet	No	58.71	143.11	59.52	167.25
	Yes	59.83	144.32	61.15	166.43
D-LinkNet	No	59.42	157.28	60.37	183.27
	Yes	60.64	156.79	61.29	183.42
DeepLab-V3	No	60.33	204.31	61.45	235.62
	Yes	61.27	204.77	62.38	235.17
DeepLab-V3+	No	61.55	235.38	62.57	264.33
	Yes	63.04	237.59	63.81	263.87
Parallel Forms of DeepLab-V3+	No	62.72	932.66	63.44	1011.57
	Yes	63.82	933.31	65.03	1012.89
Series with D-LinkNet of DeepLab-V3+	No	62.55	415.27	63.37	464.39
	Yes	63.57	417.63	64.11	465.80

2) Superpixel-Level Loss Function: We calculate the area loss between the superpixel reconstruction result and the ground truth, which can be defined as

$$\begin{aligned}\mathcal{L}_{\text{region}} &= W_{\text{over}}\mathcal{L}(G, G^{*}) \\ &= W_{\text{over}}\mathcal{L}(G, Q_{\text{row}}Q_{\text{col}}^{\top}G). \end{aligned} \tag{10}$$

Among them, W_{over} represents the excessive subdivision matrix. For the pixels on the edge of the ground truth, let $\omega = 1+\gamma_i$, otherwise $\omega = 1$. The overall loss function can be denoted as the sum of pixel loss and area loss

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{region}} + \mathcal{L}_{\text{pixel}} \\ &= W_{\text{over}}\mathcal{L}(G, Q_{\text{row}}Q_{\text{col}}^{\top}G) + \mathcal{L}(G, P). \end{aligned} \tag{11}$$

D. Implementation Details

For the DeepLab V3+, the number of convolutional kernels in each convolutional layer is set to 256, and the stride of the atrous convolutional is set to [6, 12, 18]. The initial learning rate and batch size are set to 0.007 and 5, respectively.

E. Results and Discussion

The experiment uses multiple models to analyze the performance, including U-Net [23], D-LinkNet [64], DeepLab-V3, DeepLab-V3+, and various forms of DeepLab-V3+. At the same time, it is analyzed whether to use data expansion and augmentation. The models use ResNet50 and ResNet101 to be baseline models to conduct experiments. All experiments used 400 images as the validation set and other images as the training set. The experimental results are shown in Table VI.

As a single network, DeepLab-V3+ has better feature extraction capabilities than U-Net, D-LinkNet, and DeepLab-V3. As a result, the obtained segmentation accuracy using DeepLab-V3+ is the highest. As shown in Table VI, for each model, the segmentation accuracy after the data augmentation has been improved. For the backbone network, the segmentation accuracy of the models using ResNet101 network is higher than that of the ResNet50 network. In addition, the accuracy of multinetwork segmentation model is higher than the single-network segmentation model, but it has longer inference time [65].

Based on the comprehensive results of inference time and segmentation accuracy, using DeepLab-V3+ network and ResNet101 backbone network can achieve the best performance of semantic segmentation.

VIII. FIRST PLACE IN THE AUTOMATIC WATER-BODY SEGMENTATION IN OPTICAL IMAGES: WHU

In this section, we introduce the winning method designed for the automatic water-body segmentation in optical images. The WHU team proposes water-body extraction method based on spatial consistency boundary optimization and rotation consistency constraint in multistructure segmentation network. The method integrates three network architectures with different characteristics, including large receptive field, high-resolution representation, and reduction of information loss caused by pooling. Thus, the noise and missing points caused by accidental errors can be reduced. The fully connected CRF is used for postprocessing of the predicted results. Then, weighted fusion of the postprocessing results and the original network prediction results are performed. In the testing stage, the original image and the image after rotation of 90°, 180°, and 270° are comprehensively predicted with multiangle rotation consistency. Weighted fusion of the results and CRF postprocessing results are performed to obtain the multichannel water-body prediction results, and then the multichannel prediction results are voted. Finally, the automatic extraction results of optical images are obtained.

A. Multistructure Deep Segmentation Network

This method trains three deep segmentation networks with different characteristics, as shown in Fig. 13. The training set is processed using data enhancement methods, such as rotation and stretching. The focal loss function [16] is used for water-body segmentation.

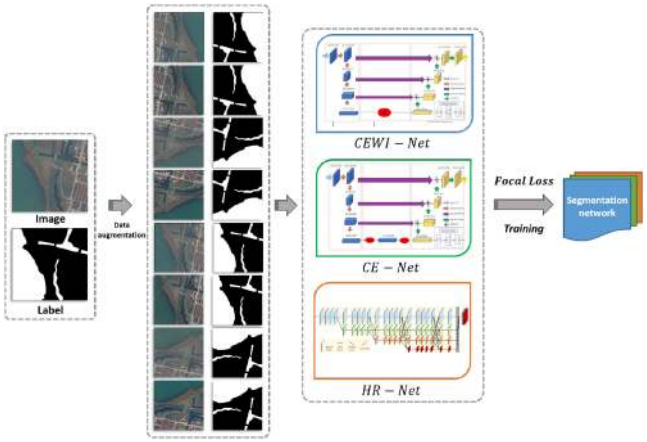

Fig. 13. Structure of multistructure deep segmentation network optimization.

1) Context Encoder Network (CE-Net): The CE-Net is first used in 2-D medical image segmentation [66]. A context extraction module is added into the traditional encoder–decoder structure to capture higher level features and obtain the spatial information for semantic segmentation, thus reducing the loss of information caused by pooling and convolution. Its network structure mainly includes a feature encoder module, a context information extraction module, and a decoder module. Among them, the ResNet-152 is used as the fixed feature extractor. The context information extractor module consists of a dense atrous convolutional (DAC) module and a residual multikernel pooling (RMP) module, whereas the decoder uses convolutional layers and transposed convolutional layers. At the same time, the weight of pretraining on the ImageNet dataset is used to accelerate the network convergence. The DAC module aims to enlarge the receptive field, and the parallel structure reduces the conflict between the segmentation and image details. The RMP module uses different scales of the pooling kernel to segment water body of various sizes.

2) Deep Segmentation Network With Dense Convolutional Pooling (CEWI-Net): Inspired by the CE-Net and Inception V1 [67], the WHU team proposes a deep segmentation network based on dense convolutional pooling (CEWI-Net), which adds a dense convolutional pooling block (DCP Block) to the encoder–decoder structure. This module is composed of convolutional layers with three convolutional kernel scales (1×1, 3×3, 5×5) and a maximum pooling layer. Each layer in the module can learn the characteristics of "sparse" and "not sparse," which has the advantage of multiscale. At the same time, they use 1×1 convolutional layer to reduce the dimension of channels so as to reduce the number of network parameters and accelerate the convergence while ensuring accuracy.

3) Deep Segmentation Network of Multiscale Object Context (HR-Net): In general, existing methods encode the input image as a low-resolution representation by a module and then recovering the high-resolution representation. Instead, HR-Net [68] takes a high-resolution subnet and adds four stages from high-resolution to low-resolution subnet one by one. Four kinds of resolution subnets are connected in parallel. The information in the parallel multiresolution subnet is exchanged in the whole network to complete the repeated multiscale fusion. Finally, bilinear up-sampling of the low-resolution output in the network is carried out to obtain the high-resolution output.

Given the complex types of water body and the relationship between ground objects in high-resolution remote sensing images, we introduce the object context representation (OCR) based on high-resolution representation [69]. It is difficult to segment water body according to a single-pixel point. OCR can effectively extract context information. OCR combines the category information of water body and nonwater body to weigh each pixel and connects with the original feature to obtain the feature representation of each pixel.

4) Optimization Loss Function: Water bodies in remote sensing images mainly include rivers, lakes, and ponds with different scales and shapes, which bring different difficulties to the deep semantic segmentation network. Focal loss [70] is used as the loss function of network optimization to address the problem of an unbalanced number of difficult and easy samples in the training images. The calculation method is as follows:

$$L_{\text{Focal}} = \frac{1}{N}\sum_{i=1}^{N} -\alpha y_i'(1-y_i)^{\gamma}\log(y_i) - (1-\alpha)(1-y_i')\,y_i^{\gamma}\log(1-y_i) \tag{12}$$

where y_i' is the ground-truth and y_i denotes the predicted result. Focal loss uses two parameters α and γ to make the network pay more attention to difficult images. To prevent the loss of simple samples from being too small, both of them adjust together to achieve balance.

B. Spatial Consistency Boundary Optimization and Rotation Consistency Constraints Based on Multistructure Segmentation Network

The testing phase includes the comprehensive prediction of rotation consistency from multiple angles, spatial consistency boundary optimization, and the voting of the multistructure segmentation network, as shown in Fig. 14.

1) Comprehensive Prediction of Rotation Consistency From Multiple Angles: In remote sensing images, water body is characterized by diverse types, various scales, and complex spatial relations, which restricts the consistency of regional prediction and the integrity of extraction results. The method is to improve the accuracy of different water-body extraction results and reduce misclassification and hole phenomena by synthesizing the prediction results of the original image and the image rotated from three angles.

The concrete method structure is shown in Fig. 15. First, the original image and the image rotated by 90°, 180°, and 270° are sent into the segmentation network in turn for prediction, and the probability matrix for prediction of P_0, P_{90}, P_{180}, and P_{270} is obtained. It is then rotated to correspond to the pixels of the original image. The prediction probability matrix of water body is then obtained by averaging the prediction probability values of four water bodies. The calculation formula is as follows:

$$P_D = (P_0 + P_{90} + P_{180} + P_{270})/4. \tag{13}$$

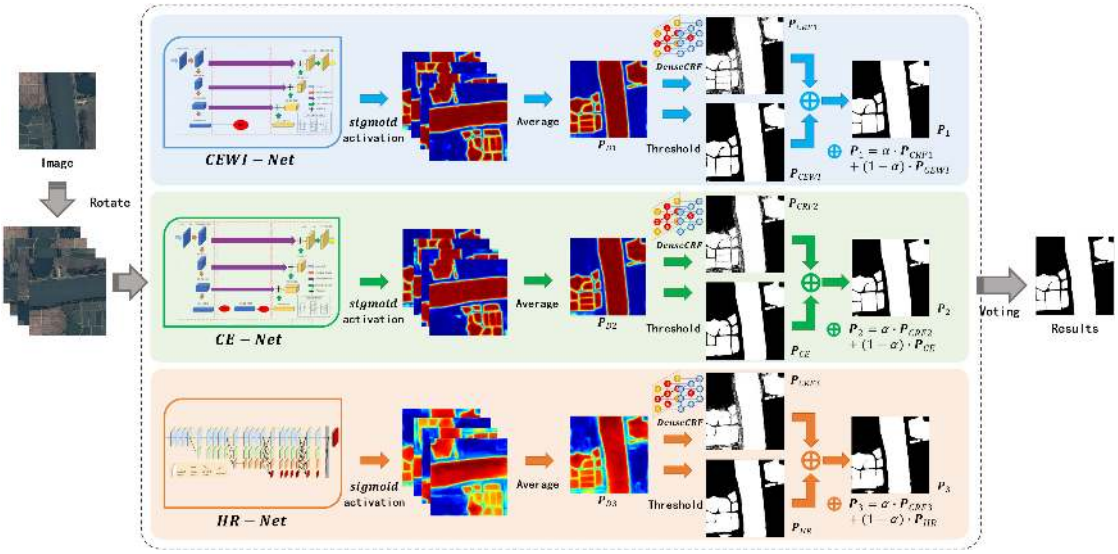

Fig. 14. Structure diagram of multistructure deep segmentation network of water-body extraction based on spatial consistency boundary optimization and rotation consistency constraints.

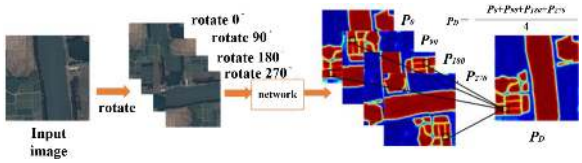

Fig. 15. Comprehensive prediction of consistent rotation (red in the prediction maps indicates the water, blue indicates the background, and colors between red and blue represent the confidence score.)

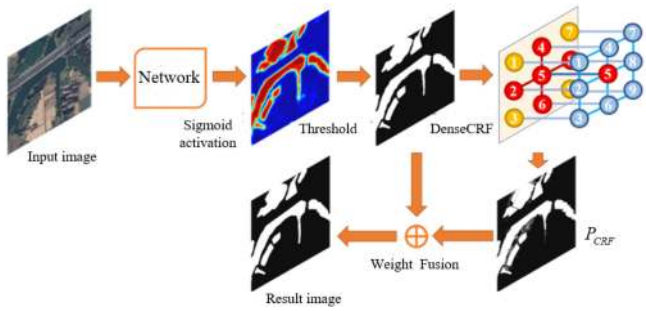

Fig. 16. Dense CRF weighted fusion model.

Here, P_D is the water-body prediction probability matrix, which is activated by the sigmoid function. i and j are the rows and column numbers of pixel points, respectively, and the final probability matrix of the water body P is the weighted fusion of P_D and P_{CRF}, and the formula is as follows:

$$P_{D_{ij}} = \delta(w_{ij}) \tag{14}$$

$$P_{ij} = \beta \cdot P_{D_{ij}} + (1-\beta) \cdot P_{CRF_{ij}} \tag{15}$$

where $\delta(\cdot)$ is the sigmoid activation function, and β is an adjustable weight parameter.

2) Spatially Consistent Boundary Optimization: It uses the fully connected CRF to postprocess the segmentation results of the network, and the weighted fusion of the processed results and the original network prediction results is carried out to recover the boundary details of the predicted results. The structure of the algorithm is shown in Fig. 16.

3) Multistructure Deep Segmentation Network Voting: It integrates three network architectures with different characteristics, and the prediction results are voted pixel by pixel to obtain the final water-body automatic extraction results. It uses CE-Net to reduce pooling information loss, CEWI-Net with multiscale characteristics, and HR-Net with high-resolution representation and spatial context relationships.

C. Implementation Details

In the experiment, the mean and standard deviation of optical images are used to normalize the images. The sigmoid activation function limits the output value within the range of [0, 1], indicating the probability of water-body prediction. The team selects the Adam [71] to be optimization method, and sets learning rate and batch size are 0.0001 and 4, respectively.

D. Results and Discussion

To validate the performance of this method, the WHU team compared their method with (1) U-Net [23]; (2) CE-Net; (3) CEWI-Net; (4) HR-Net; (5) the network only using multistructure voting mechanism without spatial consistency boundary optimization; (6) the network without the comprehensive prediction of rotation consistency from multiple angles.

Fig. 17 is an example of the automatic extraction results of water body on the test set of the Gaofen-2 high-resolution optical images. From the results, a single segmentation network will frequently have the missed and misclassified situations. The use of spatial consistency, the fully connected CRF weighted

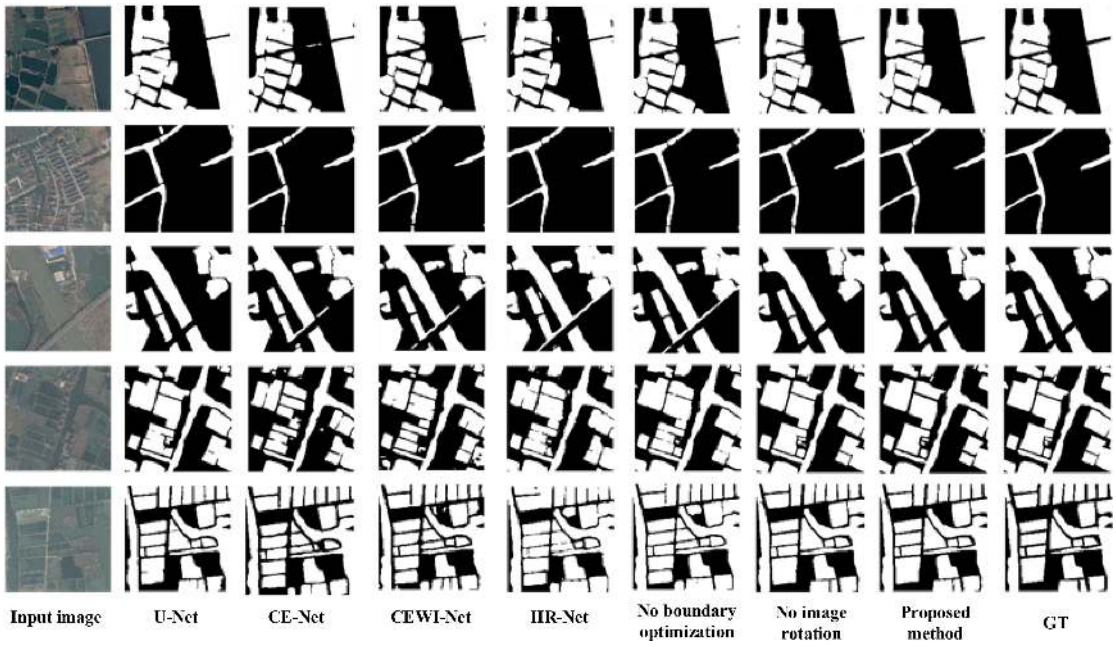

Fig. 17. Visualization of water-body extraction results.

TABLE VII
ACCURACY COMPARISON OF SEVEN METHODS ON WATER-BODY DATASET

Metrics	U-Net	CE-Net	CEWI-Net	HR-Net	w/o BO	w/o RC	ours
FWIoU	0.8552	0.8835	0.8810	0.8775	0.8872	0.8875	0.8903
System Score	/	0.8956	/	/	0.9026	0.9035	0.9052

fusion, will optimize the predicted image boundary. As shown in Table VII, the multistructure deep segmentation network can integrate the characteristics of the three networks to improve the extraction accuracy of different types of water body. Comprehensive prediction of rotation consistency can synthesize diverse spatial information from multiple angles, thereby improving the reliability of water-body prediction.

IX. FIRST PLACE IN THE SEMANTIC SEGMENTATION IN FULLY POLARIMETRIC SAR: BUCT

In this section, we introduce the winning method proposed for the semantic segmentation in fully polarimetric SAR. The method proposed by the team consists of a set of fully polarized SAR image preprocessing methods and a multiscale deep network collaboration with superpixel constraints. This method uses Deeplab V3+ for pixel-level classification and simultaneously extracts local gradient ratio patterns (LGRPs) from the original fully polarimetric SAR image, then performs weighted K-means [72] clustering to generate superpixels. Under the constraints of superpixels, the classification loss function is further optimized to improve the segmentation performance. The framework of the method is shown in Fig. 18.

A. Data Preprocessing

The dataset used in this method is divided into two parts, one is the Gaofen-3 fully polarimetric SAR training dataset provided by the organizers, and the other part is the fully polarimetric SAR data collected by team. BUCT team has augmented the existing data, including the following.

1) Overlap cropping: They crop the original image to a fixed size with overlap.
2) Spatial transformation: It includes horizontal and vertical flipping, random rotation of the image at any angle with the center as the origin, scaling of the image at a certain ratio, random cropping, and shifting.
3) Adding noise: Gamma noise fitting and noise addition of different visual numbers is performed on the image, thereby enriching the training samples.
4) Polarization simulation: They perform polarization simulation for the specific objects, and then obtain the HH, HV, and VH channels of the simulation data.

B. Pixel-Level Semantic Segmentation: DeepLab V3+

BUCT team uses DeepLab V3+ for semantic segmentation of the fully polarimetric SAR image. In the DeepLab V3+ network, feature extraction is performed on the input image through the backbone network to obtain low-level feature and high-level feature. In the encoding stage, the advanced features go through the FPN, including a 1×1 convolution, three atrous convolutional layers with different atrous rates (6, 12, 18), a global average pooling, and an up-sampling layer. Then, the outputs of the five layers are cascaded, and the number of

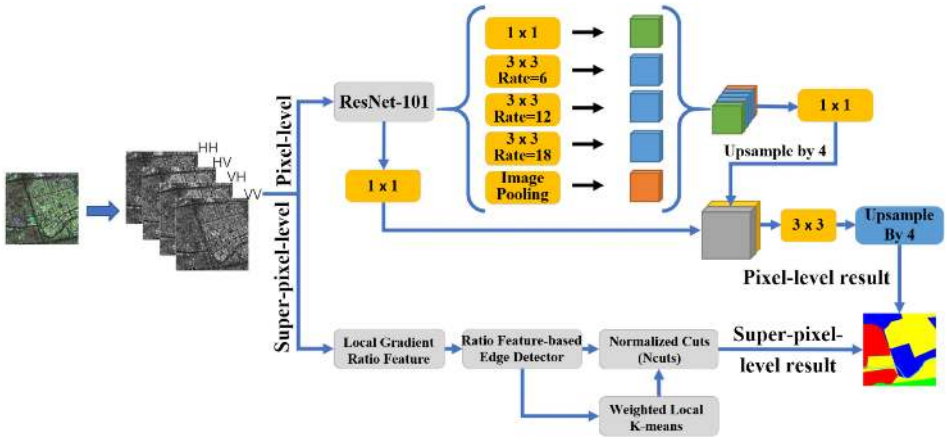

Fig. 18. Framework of the proposed method for SAR image semantic segmentation.

channels is changed through 1×1 convolution. In the decoding stage, the low-level features are dimensionally adjusted by 1×1 convolution (output stride $= 4$), and the encoder output is up-sampled 4 times (output stride changes from 16 to 4). Then, we concatenate the features and perform 3×3 convolution, then up-sample 4 times to get dense prediction. All up-sampling layers in the decoder use bilinear interpolation.

C. Superpixel Segmentation Technology for Fully Polarimetric SAR Image

Due to geometric distortion and speckle noise in fully polarized SAR images, it is difficult to adopt a effective method to generate superpixels with high boundary fitting, compactness, and low computational cost. This method adopts an superpixel generation algorithm with linear feature clustering and edge constraint for SAR images [35]. There are three stages. First, BUCT team extracts the LGRP of each pixel. This feature has strong robustness to coherent speckle noise. LGRP characteristics can be defined as

$$\mathrm{LGRP}_{P,R}(g_c) = \sum_{P-1}^{p=0} s\left(G_{\mathrm{ratio}}(g_p) - \overline{G_{\mathrm{ratio}}(g_c)}\right) 2^p \quad (16)$$

where $G_{\mathrm{ratio}}(g_p)$ and $G_{\mathrm{ratio}}(g_c)$ are the gradient ratio characteristics of neighboring pixels and center pixels, respectively. Second, for the edge detector, the traditional rectangular edge detector uses a series of windows with various directions to calculate the edge strength map (ESM). The windows are divided into a pair of parallel subwindows. BUCT team uses the ratio-feature-based edge detector of Gaussian windows instead of the traditional rectangular windows. The horizontal Gaussian window is defined as

$$\mathrm{GW}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sqrt{2\pi}\sigma_y} \exp\left(-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right)\right). \quad (17)$$

In addition to the ESM, the Gaussian window can obtain edge direction map and edge map of the SAR image. Finally, an improved superpixel generation strategy based on normalized cuts (Ncuts) is adopted, which uses distance metrics and also considers spatial proximity and feature similarity. In this strategy, the BUCT team approximates the similarity using a positive semidefinite kernel function instead of traditional feature-based algorithms. The best point can be obtained by weighted K-means and Ncuts function, thereby effectively reducing the computational cost. The weighted local K-means clustering function is denoted as

$$\Phi_{\mathrm{K-means}} = \sum_{k=1}^{K} \sum_{u\in\omega(k)} w(u)\|\Psi(u) - m_k\|^2. \quad (18)$$

The Ncuts function is defined as

$$\Phi_{\mathrm{Ncuts}} = \frac{1}{K}\sum\nolimits_{k=1}^{K} \frac{\sum_{u\in\omega(k)}\sum_{v\in\omega(k)} W(u,v)}{\sum_{u\in\omega(k)}\sum_{v\in V} W(u,v)}. \quad (19)$$

Among them, for each pixel p, $\Psi(u)$ is an eight-dimensional feature vector composed of LGRP features. Given two pixels $u = (l_{4u}, l_{8u}, x_u, y_u)$ and $v = (l_{4v}, l_{8v}, x_v, y_v)$, the similarity measure between them is denoted as

$$\hat{W}(u,v) = \hat{W}_f(u,v) + \beta_{\mathrm{adp}} \cdot \hat{W}_s(u,v). \quad (20)$$

The variation coefficient is used to learn the tradeoff factor between spatial proximity and feature similarity during linear feature clustering, which helps to adaptively adjust the shape and scale of superpixels according to image uniformity. The coefficient of variation is calculated as follows:

$$\beta_{\mathrm{adp}} = 1 - \frac{1}{2}\left[\mathrm{CoV}(x_u, y_u) + \mathrm{CoV}(x_v, y_v)\right]. \quad (21)$$

The superpixel generation method used in this method has some characteristics, which are as follows.

1) The structure of the image can be maintained well because of edge information and Ncuts strategy.
2) The method is not sensitive to the coherent speckle noise.
3) The method has higher computational efficiency.
4) The shape and compactness of super pixels can be adaptively changed according to the complexity of the image.

TABLE VIII
COMPARISON OF EXPERIMENTAL RESULTS IN FULLY POLARIMETRIC SAR IMAGES

Models	Augmentation	ResNet50		ResNet101	
		Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)
U-Net	No	67.9964	109.56	68.6675	120.55
	Yes	68.6436	114.78	69.5546	123.66
U-Net+CRF	Yes	67.6357	114.34	68.8866	121.77
D-LinkNet	No	69.8965	132.93	70.4575	148.33
	Yes	72.0065	140.36	72.6654	149.88
D-LinkNet+CRF	Yes	70.9658	136.93	71.2238	149.67
DeepLab-V3+	No	73.1343	157.22	73.6674	161.33
	Yes	74.9964	159.89	75.2939	166.37
D-LinkNet+CRF	Yes	74.0364	160.16	74.1985	165.25
Parallel Forms of DeepLab-V3+ (HH+HV+VV and HH+VH+VV)	No	73.1325	280.35	74.2235	288.66
	Yes	73.5613	290.87	74.8085	301.22

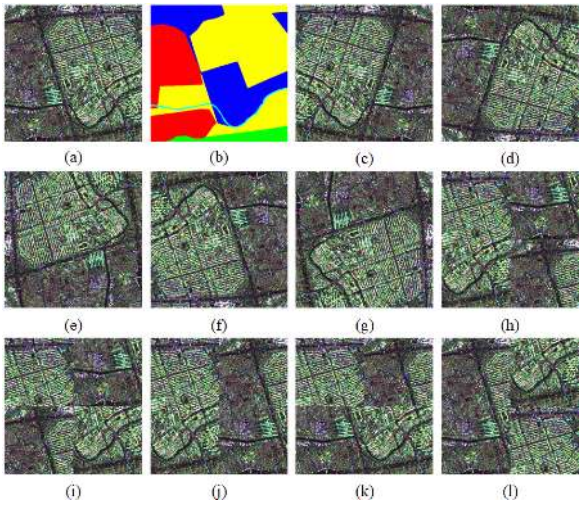

Fig. 19. Image enhancement results of proposed method. (a) ground-truth image. (b) label image. (c)-(g) the images obtained by rotating the ground-truth image at different angles. (h)-(l) the images obtained by cutting and stitching the ground-truth image.

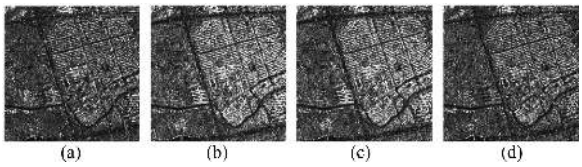

Fig. 20. Gray image of polarization modes. (a) HH. (b) HV. (c) VH. (d) VV.

D. Results and Discussion

In this competition, the BUCT team uses U-Net, D-LinkNet, and DeepLab V3+ for pixel-level semantic segmentation and use a CRF as postprocessing after U-Net and D-LinkNet network, defined as U-Net+CRF and D-LinkNet+CRF, respectively.

In terms of data augmentation, the BUCT team has performed methods, such as rotation, cropping, and stitching on the original image, enhancing the sensitivity of the model to image edges. The specific transformation is shown in Fig. 19. Fig. 19(a) is the image A-10 in the training dataset, and Fig. 19(b) is the corresponding colored label map. Fig. 19(c)–(g) shows the images obtained by rotating A-10 at different angles; Fig. 19(h)–(l) shows the images formed after cropping and then stitching. Performing the same operation on all the original images can get the augmented dataset. Fig. 20 shows a grayscale image of the four polarization channels of image A-10. Adding these images to the training can enhance the model's sensitivity to edges and improve the overall accuracy. However, through training, it is found that the model performs not well enough on the edges of rivers and small objects.

The results of different methods are shown in Table VIII. From the table, for a single network, DeepLab V3+ has good performance on feature extraction. BUCT team attempted to use a CRF as postprocessing, but the accuracy has not improved because the CRF overlooked some small objects. It is obvious that the performance of the model after data augmentation has improved, reflecting the importance of the amount of data. To get higher accuracy, the BUCT team try to parallelize the dual networks in DeepLab V3+. However, the accuracy is still slightly lower than using DeepLab V3+, and the inference time is also longer.

X. CONCLUSION

The development of earth observation programs and accessible high-resolution data can provide abundant information about the earth and promote various applications. Due to the insufficient amount of annotated data and the complex background, it is of great challenge to apply the automated interpretation for such data. Therefore, it is significant that highly advanced techniques need to be proposed.

To enhance the academic development in this field, the 2020 Gaofen Challenge focuses on the automated high-resolution earth observation image interpretation for optical and SAR images. More than 10 000 images from Gaofen-2 and Gaofen-3 satellites are annotated for this challenge. Complex background, various scales, and fine-grained types make the 2020 Gaofen Challenge more difficult.

The 2020 Gaofen Challenge is arranged in six tracks according to different application requirements. Tracks 1–3 aim to promote the development of object detection and recognition in optical and SAR images. Tracks 4–6 focus on semantic segmentation in optical and SAR images.

The 2020 Gaofen Challenge has attracted 701 teams from 253 affiliations with 2023 competitors to participate in. The competitors come from more than 20 countries, including China, England, Germany, France, Japan, Australia, Singapore, India, Sweden, etc. All winners use deep-learning-based methods for image interpretation.

Although many excellent algorithms have emerged in the challenge, the exploration of earth observation technology cannot be stopped. After the challenge, the datasets are still accessible for further research.

In the future, we will also continue to promote this event and hope it can help the earth observation community to develop deep-learning-based methods. We will dedicate to improve the professional level of the Gaofen Challenge. For the data, we will continue to build larger scale high-resolution multisource datasets and enhance the quality of annotations. After the challenge, we will provide a repository to share datasets and codes for competitors. For the tracks in the challenge, we will set more tracks that are combined with practical applications in the field of remote sensing. For the competitors, we will encourage more foreign scholars to participate in the competition to make it more international. Moreover, we will improve the evaluation system to obtain more authoritative and fair results. With the improvement of Gaofen Challenge, we hope that more and more scholars from all over the world will participate in the challenge.

ACKNOWLEDGMENT

The authors would like to thank the IEEE Geoscience and Remote Sensing Society for the support, especially to Prof. P. Gamba, Prof. Jun Li, and the Image Analysis and Data Fusion Technical Committee for their valuable comments. They would also like to thank the International Society for Photogrammetry and Remote Sensing, especially to Prof. C. Toth and Prof. S. Hinz for their great support.

REFERENCES

[1] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[2] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 50–65, 2021.

[3] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[4] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[5] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 131.

[6] G. S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[7] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, Jul. 2021.

[8] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, 2016.

[9] Y. Ren, C. Zhu, and S. Xiao, "Deformable faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1470.

[10] S. Ren, K. He, R. Girshick, and S. Jian, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst. 28: Ann. Conf.Neural Inf. Process. Syst.*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Montreal, Quebec, Canada, Dec. 7-12, 2015, pp. 91.-99.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[12] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Comput. Soci. IEEE Conf. Comput. Vision Pattern Recognition*, Las Vegas, NV, USA, Jun. 27-30, 2016, pp. 779–788.

[14] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," *CoRR*, vol. abs/1701.06659, 2017.

[15] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[16] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[17] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[18] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[19] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[20] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *IEEE/CVF Int. Conf. Comput. Vision*, Seoul, Korea (South), Oct. 27–Nov. 2, 2019, pp. 9656–9665.

[21] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

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

[23] 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.

[24] 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.

[25] 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, Apr. 2018.

[26] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016.

[27] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS ^{4} Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5398–5413, Sep. 2020, doi: 10.1109/JSTARS.2020.3021098.

[28] 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.

[29] K. Fu, Z. Chang, Y. Zhang, and X. Sun, "Point-based estimator for arbitrary-oriented object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4370–4387, May 2021.

[30] L. Jun, L. Yunfei, H. Lin, C. Jin, and P. Antonio, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140301.

[31] L. Liu *et al.*, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[32] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *CoRR*, vol. abs/1704.06857, 2017.

[33] X. Sun *et al.*, "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *CoRR*, vol. abs/2103.05569, 2021.
[34] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.
[35] D. Xiang, T. Tang, S. Quan, D. Guan, and Y. Su, "Adaptive superpixel generation for SAR images with linear feature clustering and edge constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3873–3889, Jun. 2019.
[36] F. Gao, F. Ma, J. Wang, J. Sun, E. Yang, and H. Zhou, "Visual saliency modeling for river detection in high-resolution SAR imagery," *IEEE Access*, vol. 6, pp. 1000–1014, 2017.
[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[38] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112045.
[39] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 182–196, 2018.
[40] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2021.
[41] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.
[42] F. Ma, F. Gao, J. Wang, A. Hussain, and H. Zhou, "A novel biologically-inspired target detection method based on saliency analysis for synthetic aperture radar (SAR) imagery," *Neurocomputing*, vol. 402, pp. 66–79, 2020.
[43] Y. Jiang *et al.*, "R2CNN: rotational region CNN for orientation robust scene text detection," *CoRR*, vol. abs/1706.09579, 2017.
[44] R. Nabati and H. Qi, "RRPN: Radar region proposal network for object detection in autonomous vehicles," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3093–3097.
[45] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
[46] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *CoRR*, vol. abs/2008.09397, 2020.
[47] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*,Venice, Italy, 2017, pp. 764–773.
[48] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018.
[49] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.
[50] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[51] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, "Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618–644, 2015.
[52] Q. Song and F. Xu, "Zero-shot learning of SAR target feature space with deep generative neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2245–2249, Dec. 2017.
[53] 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, pp. 1–19, 2020.
[54] Q. Song, H. Chen, F. Xu, and T. J. Cui, "EM simulation-aided zero-shot learning for SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1092–1096, Jun. 2020.
[55] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.
[56] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *CoRR*, vol. abs/1911.09516, 2019.
[57] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. S. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Conf. Multimedia*, Amsterdam, The Netherlands, 2016, pp. 516–520.
[58] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, Apr. 24-26, 2017.
[59] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "SAR target detection based on SSD with data augmentation and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 150–154, Jan. 2019.
[60] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120 234–120254, 2020.
[61] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, vol. 97, pp. 6105–6114.
[62] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 5562–5570.
[63] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," *CoRR*, vol. abs/1906.07155, 2019.
[64] L. Zhou, C. Zhang, and M. Wu, "D-Linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 182–186.
[65] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Sci. Rep.*, vol. 10, no. 1, pp. 1–7, 2020.
[66] Z. Gu *et al.*, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.
[67] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.
[68] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[69] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. 16th Eur. Conf.*, Glasgow, U.K., 2020, vol. 12351, pp. 173–190.
[70] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollér, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 2999–3007.
[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7-9, 2015.
[72] T. Gadhiya and A. K. Roy, "Superpixel-driven optimized Wishart network for fast PolSAR image classification using global k-means algorithm," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 1, pp. 97–109, Jan. 2020.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009, all in electronic information engineering.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Peijin Wang (Member, IEEE) received the B.Sc. degree in automation from Tianjin University, Tianjin, China, in 2017, and the M.Sc. degree in signal and information processing from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2020.

She is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include computer vision, deep learning, and remote sensing.

Zhiyuan Yan (Member, IEEE) received the B.Sc. degree from Xiamen University, Xiamen, China, in 2016, and the M.Sc. degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2019, all in electronic information engineering.

She is currently an Assistant Engineer with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include computer vision and remote sensing image analysis.

Wenhui Diao (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016, all in electronic information engineering.

He is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Xiaonan Lu received the B.Sc. degree in communication engineering from Xidian University, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree in signal and information processing with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, and the University of Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision, pattern recognition, and remote sensing image processing, especially on image scene classification and object detection.

Zhujun Yang received the B.Sc. degree in communication engineering from Chongqing University, Chongqing, China, in 2019. She is currently working toward the Ph.D. degree in signal and information processing with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision, edge intelligence, and remote sensing processing, especially on semantic segmentation.

Yidan Zhang received the B.Sc. degree in communication engineering from Tianjin University, Tianjin, China, in 2019. She is currently working toward the master's degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision and deep learning, especially on remote sensing object detection and model compression.

Deliang Xiang received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2010, the M.S. degree in photogrammetry and remote sensing from the National University of Defense Technology, Changsha, China, in 2012, and the Ph.D. degree in geoinformatics from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2016.

Since 2020, he has been a Full Professor with the Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing, China. His research interests include urban remote sensing, synthetic aperture radar (SAR)/polarimetric SAR image processing, artificial intelligence, and pattern recognition.

Chen Yan photograph and biography not available at the time of publication.

Jie Guo received the B.E. degree in optoelectronic engineering and the Ph.D. degree in optical engineering from the Beijing Institute of Technology, Beijing, China, in 2013 and 2019, respectively.

He is currently with the Nanjing Research Institute of Electronic Technology, Nanjing, China. His current research interests include object detection, visual tracking, and related computer vision problems.

Bo Dang is currently working toward the Undergraduate degree in remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include deep learning and remote sensing processing, especially on semantic segmentation.

Wei Wei (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2012.

He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. He has been authored more than 40 articles, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, *Patten Recognition*, CVPR, ICCV, ECCV, AAAI, and IJCAI.

Dr. Wei has served as an Associate Editor for the *Journal of Selected Topics in Applied Earth Observations* and *Remote Sensing*.

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

He is currently a Professor with the School of Information Science and Technology and the Vice Director of the MoE Key Laboratory for Information Science of Electromagnetic Waves, Fudan University, Shanghai, China. His research interests include electromagnetic scattering modeling, synthetic aperture radar information retrieval, and radar system development.

Cheng Wang (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2002.

He is currently a Professor and an Associate Dean of the School of Informatics, and an Executive Director of the Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China. He has coauthored more than 150 articles. His research interests include image processing, mobile LiDAR data analysis, and multisensor fusion.

Dr. Wang is the Chair of the ISPRS Working Group I/6 on Multisensor Integration and Fusion from 2016 to 2020, and a council member of the China Society of Image and Graphics.

Ronny Hänsch (Senior Member, IEEE) received the Undergraduate degree in computer science and the Ph.D. degree in computer vision from the Technische Universität Berlin, Berlin, Germany, in 2007 and 2014, respectively.

He is currently with the German Aerospace Center, Weßling, Germany. His current research interests include ensemble methods for image analysis.

Dr. Hänsch is the Co-Chair of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society from 2017 to 2021 and the Co-Chair of the International Society for Photogrammetry and Remote Sensing Working Group II/1 (Image Orientation).

Martin Weinmann (Member, IEEE) received the Diploma degree in electrical engineering and information technology from the Technical University of Karlsruhe, Karlsruhe, Germany, in 2009, and the Ph.D. degree in computer vision from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2015.

He is currently a Postdoctoral Researcher with the Institute of Photogrammetry and Remote Sensing, KIT. His research interests include computer vision, pattern recognition, machine learning, photogrammetry, and remote sensing, where he published respective work in a diversity of reputable journals and conference proceedings.

Dr. Weinmann served for several years as a Reviewer for the International Society for Photogrammetry and Remote Sensing, the Institute of Electrical and Electronics Engineers, and the German Society for Photogrammetry, Remote Sensing and Geoinformation.

Naoto Yokoya (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees in aerospace engineering from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He is currently a Lecturer with The University of Tokyo, Tokyo, Japan, and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he leads the Geoinformatics Unit.

Dr. Yokoya is the Chair of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society from 2019 to 2021.

Kun Fu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic information engineering from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.