



A Benchmark High-Resolution GaoFen-3 SAR Dataset for Building Semantic Segmentation

Junshi Xia , Senior Member, IEEE, Naoto Yokoya , Member, IEEE, Bruno Adriano , Member, IEEE, Lianchong Zhang, Member, IEEE, Guoqing Li, Senior Member, IEEE, and Zhigang Wang

Abstract—Deep learning is increasingly popular in remote sensing communities and already successful in land cover classification and semantic segmentation. However, most studies are limited to the utilization of optical datasets. Despite few attempts applied to synthetic aperture radar (SAR) using deep learning, the huge potential, especially for the very high resolution (VHR) SAR, are still underexploited. Taking building segmentation as an example, the VHR SAR datasets are still missing to the best of our knowledge. A comparable baseline for SAR building segmentation does not exist, and which segmentation method is more suitable for SAR image is poorly understood. This article first provides a benchmark high-resolution (1 m) GaoFen-3 SAR datasets, which cover nine cities from seven countries, review the state-of-the-art semantic segmentation methods applied to SAR, and then summarize the potential operations to improve the performance. With these comprehensive assessments, we hope to provide the recommendation and roadmap for future SAR semantic segmentation.

Index Terms—Building segmentation, GaoFen-3, high-resolution, synthetic aperture radar (SAR).

I. INTRODUCTION

DUE to the reason that building is the main component in urban cities, building semantic segmentation attracts more attention in urban remote sensing studies. Most studies of building semantic segmentation focus on very high resolution (VHR) optical datasets and have been formed by a series of datasets, for instance, SpaceNet Challenge (1, 2, 4),¹ Inria Aerial Image Labeling Dataset, DeepGlobe Building Extraction Challenge,² 2018 Open AI Tanzania Building Footprint Segmentation Challenge³, and CrowdAI Mapping Challenge.⁴

Manuscript received January 13, 2021; revised April 23, 2021; accepted May 13, 2021. Date of publication May 31, 2021; date of current version June 23, 2021. This work was supported in part by KAKENHI under Grant 19K20309 and Grant 18K18067, in part by JSPS Bilateral Joint Research Project under Grant JPJSBP 120203211, and in part by the Open Research Fund of National Earth Observation Data Center under Grant NODAOP2020021. (Corresponding author: Junshi Xia.)

Junshi Xia and Bruno Adriano are with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), Japan (e-mail: junshi.xia@riken.jp; bruno.adriano@riken.jp).

Naoto Yokoya is with the Department of Complexity of Science and Engineering, The University of Tokyo, and with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

Lianchong Zhang and Guoqing Li are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100864, China (e-mail: zhanglc@aircas.ac.cn; ligq@radi.ac.cn).

Zhigang Wang is with the China Center for Resource Satellite Data and Application, Beijing 100094, China (e-mail: kevinwang2000@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3085122

¹<https://spacenetchallenge.github.io/>

²<http://deepglobe.org/index.html>

³<https://competitions.codalab.org/competitions/20100>

⁴<https://www.crowdai.org/challenges/mapping-challenge>

SAR building detection becomes more popular in recent years because of its imaging capability during the day or nighttime under all weather conditions for observing the land surface [1]. In [2], a global urban extent mapping using ENVISAT advanced synthetic aperture radar (ASAR) wide-swath mode datasets (resolution of 75 m) were achieved. Ban *et al.* [3] developed an effectively extractor to obtain urban extent using ENVISAT ASAR datasets (resolution of 30 m). Since the Sentinel-1 was launched on April 3, 2014 and then made the datasets freely available, many scientists have investigated the potential use of Sentinel-1 images to extract the urban areas. For instance, the combination of textural and intensity features of Sentinel-1 is applied to extract the built-up areas of Chinese cities using the region growing technique [4]. Chini *et al.* [5] automatically generated built-up areas using the intensity and interferometric coherence generated from multitemporal Sentinel-1 data. Esch *et al.* [6] investigated the potential of high-resolution X-band (i.e., TanDEM-X) data for the automatic building mapping. Then, the datasets were further used in a new unsupervised and automatic system, namely urban footprint processor (UFP), to produce the global urban footprints (resolution: 12 m) [7]–[9].

Recently, the VHR SAR satellites have been successfully launched. Since VHR SAR images at the meter- and submeter-level can provide very detailed geometric structures and radiometric features to separate the different objects, it is very beneficial and useful for accurately building segmentation. GaoFen-3, which is part of the China High-Resolution Earth Observation System (CHEOS) family intending to provide high-resolution observations and disaster monitoring [10], is the first C-band polarimetric SAR imaging satellite of China National Space Administration (CNSA), Beijing, China. GF-3 has 12 imaging models with single (HH or VV), dual (HH+HV or VH+VV), and full polarization (HH+HV+VH+VV). The spatial resolution ranges from 1 to 500 m, and a swath coverage ranges from 10 to 650 km. GF-3 datasets have been already successfully applied to the following applications, such as target detection (e.g., ship) [11], [12], crop classification [13], etc. It is also an important data source in Indonesia Tsunami, Iran Earthquake, and Laos Flood under the China GEO Disaster Data Response Mechanism (CDDR) [14]. Recently, deep learning has made rapid, significant achievements on semantic segmentation [15], [16], and many models from U-Net [17] to high-resolution net (HRNet) [18] are proposed. However, the specific problem of building semantic segmentation from SAR imagery using deep learning has received very little attention until recently. For

instance, Shahzad *et al.* [19] adopted the integration of fully convolution neural networks and conditional random field to detect buildings of TerraSAR-X SAR image. Li *et al.* [20] and Wu *et al.* [21] developed a multiscale convolution neural network and multiscale structured network for the extraction of building area using SAR datasets. Huang *et al.* [22] proposed deep SAR-Net, which considers the spatial textures and backscattering information from complex-valued SAR images.

To promote further research of building semantic segmentation using SAR datasets, the combination of VHR SAR datasets and semantic segmentation models is required to investigate. However, until now, the public VHR SAR datasets for building semantic segmentation are very limited. In the IEEE Geoscience and Remote Sensing Society (GRSS) data fusion challenge of 2012 [23], the high-resolution optical, SAR, and LiDAR datasets are constructed. However, the dataset area is limited to the downtown of San Francisco, and the datasets are no longer publicly available because of its limited license. Yao *et al.* [24] constructed the datasets from three data sources (with a resolution of 2.9 m): TerraSAR-X images, Google Earth images, and Open Street Map (OSM) data, to perform SAR and optical image semantic segmentation. The datasets cover 15 cities of North Rhine-Westphalia (NRW), Germany. Spacenet 6⁵ constructed the multisensor all-weather mapping (MSAW) dataset, including airborne full-polarized X-band SAR and spaceborne optical datasets (with a resolution of 0.5 m) over the port of Rotterdam, the Netherlands. All the datasets, as mentioned earlier, are limited to the single location of city and country.

To address the limitations mentioned above, We construct a new dataset, GaoFen-3 Building (GFB), with a resolution of 1 m. It should be emphasized that GFB datasets cover nine cities from seven countries, in which the building structures and urban design layouts are very diverse. The datasets will be open to the community. We also included the Google Earth image as optical images to thoroughly investigate the performance between different modality and their combinations using deep-learning baseline models.

These baseline models are fundamental to the community, which can help us to deeply understand the capability of state-of-the-art segmentation models for working with SAR data. The main contribution can be summarized as follows.

- 1) A high-resolution SAR dataset for building semantic segmentation is presented.
- 2) A comprehensive comparison of different segmentation methods is analyzed.
- 3) The influence and the potential solution to improve the performance is given.

II. STUDY AREA AND DATASET

A. Study Area

In total, nine cities from seven countries, including Beijing, Shanghai, Hongkong, Yokosuka, Berlin, Rennes, Barcelona,

TABLE I
PARAMETERS OF GF-3 SPOTLIGHT DATASETS

City	Date	Mode	Direction	Look/Incidence angle
Rennes	2017-04-07	VV	ASC	35.40/40.07
Berlin	2018-02-07	HH	DEC	38.64/44.05
San Diego	2019-01-09	HH	ASC	34.70/39.24
Barcelona	2018-11-19	HH	DEC	38.85/44.25
Yokosuka	2019-08-29	HH	DEC	35.17/39.82
Rio	2019-10-19	VV	ASC	28.30/31.65
	2019-02-07	HH	ASC	37.81/42.95
Beijing	2019-12-16	HH	DEC	40.98/46.90
Shanghai	2018-03-07	VV	DEC	34.94/39.53
	2019-12-26	HH	DEC	27.17/30.32
Hongkong	2018-08-25	HH	DEC	28.58/31.97

TABLE II
DATES OF GOOGLE EARTH IMAGES

City	Date	City	Date
Rennes	2016-04-19	Rio	2019-09-09
Berlin	2018-08-13	Beijing	2019-02-26
San Diego	2018-11-18	Shanghai	2019-5-13
Barcelona	2019-03-09		2018-02-08
Yokosuka	2019-08-08	Hongkong	2019-10-29
			2018-10-04

San Diego, and Rio de Janeiro (short for Rio), were selected (seen in Fig. 1).

B. GaoFen-3 Data and Preprocessing

GaoFen-3 SAR has 12 observing modes, including spotlight (SL), ultrafine stripmap (UFS), with different resolution/swath [25]. In this work, the SL mode with high-resolution (1 m) and wide-swath (10 km) was chosen. Due to limited wide-swath, Hongkong, Rennes, Barcelona, Berlin, and San Diego are mainly urban areas under different environmental conditions, while Shanghai, Beijing, Yokosuka, and Rio focus on the rural areas with many small villages.

Finally, 11 GF-3 images over 9 cities are chosen. Shanghai and Rio have two images. The parameters of each dataset are shown in Table I. For GF-3 data preprocessing, the Pixel Information Expert (PIE) software,⁶ designed explicitly for GF-3, was used. Following the data preprocessing modules provided by the PIE, the following steps are used for the preprocessing:

- 1) the raw images are converted into intensity;
- 2) multiple-looking with the option of 1 m resolution is applied;
- 3) the redefined Lee filter is used to reduce the speckle noise;
- 4) the DEM is used to geocode the datasets with the map projection of WGS84;
- 5) the geocode terrain correction (GTC) is applied.

Then, we applied the logarithm to the raw intensity and multiplied it by 10, making the data range in the 8-b range of 0–255. The final resolution of SAR datasets is 1 m.

⁵<https://spacenetchallenge.github.io/>

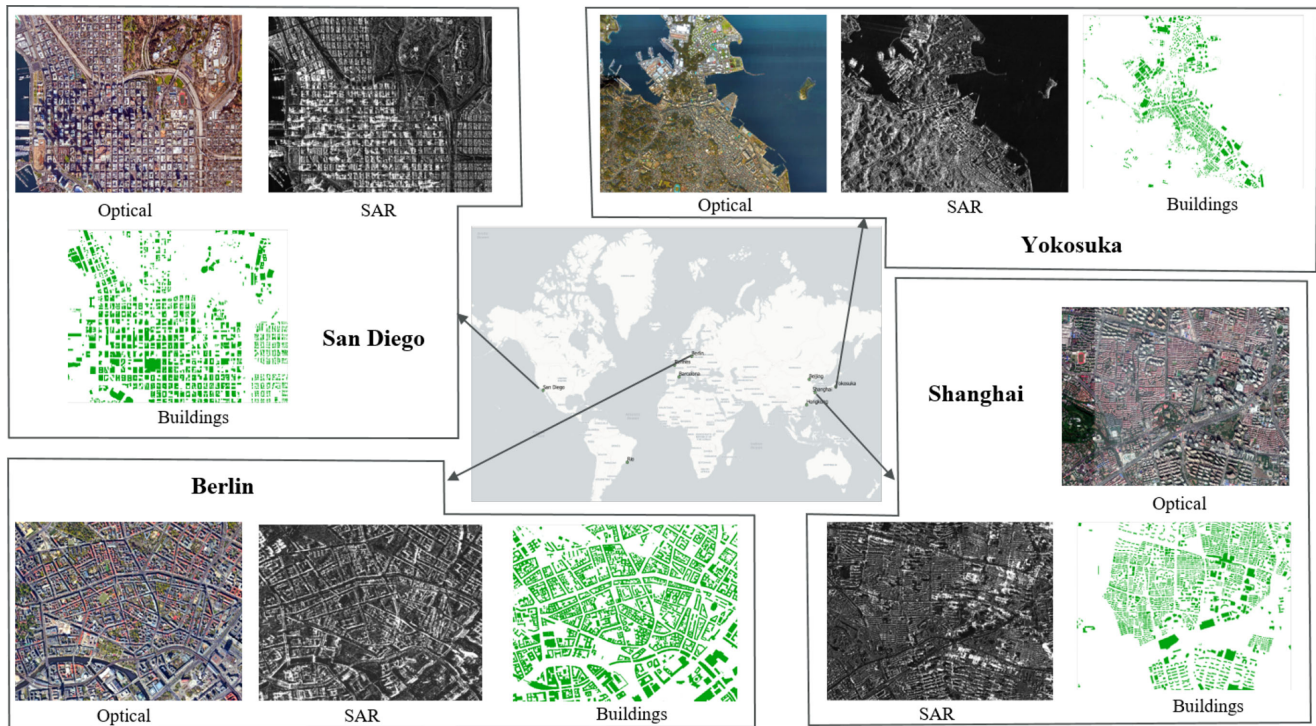


Fig. 1. Study area, and optical Google Earth, GaoFen-3 SAR, and buildings of San Diego, Yokosuka, Berlin, and Shanghai.

C. Google Earth Images

In this work, Google Earth images (seen in Table II), whose dates are close to the ones of SAR datasets, were treated as the optical RGB datasets. In this case, we ignore the building changes during the short period. Following the guideline in [26], the Google Earth images⁷ were first downloaded through Google Earth Pro software and then reconstructed to format large-scale images. To make the spatial resolution (1 m) the same as the GF-3 datasets, we set the Google Earth image's zoom level to 16. Finally, we manually select the ground control points (GCP) to fine coregister the SAR and optical datasets.

D. Content of the GFB Dataset

In this work, the open source data from OpenStreetMap⁸ are used to build the building labels. We also manually added some building labels for Beijing and Shanghai. By projecting the labels to the coregistered GF-3 and Google Earth images, the RGB and SAR patch pairs with the corresponding building labels are obtained. The datasets can be accessed in: 10.11878/db.202104.000008. The image patches' size is set to be 256 by 256 pixels (512 by 512 pixels), corresponding to a physical dimension of 256 m by 256 m (512 m by 512 m). We excluded the patches without any buildings. Finally, we obtained 8373 and 4812 pairs of patches for the sizes of 256×256 and 512×512 , respectively. Finally, we split the datasets into training, validation, and test sets with a ratio of 6:2:2. The

TABLE III
NUMBER OF PATCHES FOR EACH CITY (PATCH SIZE: 256×256)

City	No. train	No. val	No. test
Barcelona	603	201	201
Beijing	139	46	47
Berlin	903	301	302
Hongkong	530	176	178
Rennes	682	227	228
Rio	641	213	215
San Diego	542	180	182
Shanghai	585	195	196
Yokosuka	396	132	132
Total	5021	1671	1681

numbers of patches for train/validation/test from each city are shown in Tables III and IV.

III. BASELINE OF SEMANTIC SEGMENTATION

From 2015, more than a hundred semantic segmentation methods are proposed in the computer vision community. Some parts, including encoder–decoder, multiscale, and dilated convolution, are commonly used in these models. Fig. 2 has shown the timeline of popular semantic segmentation models since 2015.

The first deep learning-based semantic segmentation is fully convolutional networks (FCN) [27]. Then, the operations of encoder–decoder, multiscale, dilated convolution, skip-connection and context prior are used to construct the new models, such as SegNet [28], deep parsing network

⁶<http://www.piesat.cn/en/PIE-SAR.html>

⁷<https://developers.google.com/maps/documentation/maps-static/dev-guide>

⁸<https://planet.openstreetmap.org/>

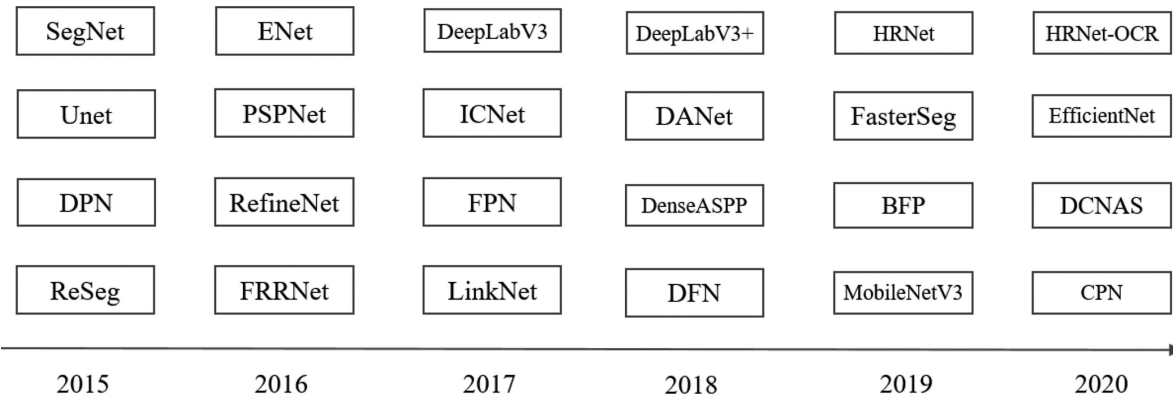


Fig. 2. Timeline of typical semantic segmentation algorithms.

TABLE IV
NUMBER OF PATCHES FOR EACH CITY (PATCH SIZE: 512×512)

City	No. train	No. val	No. test
Barcelona	173	57	59
Beijing	68	22	24
Berlin	225	75	75
Hongkong	185	61	63
Rennes	193	64	66
Rio	245	81	83
San Diego	184	61	63
Shanghai	235	78	79
Yokosuka	147	49	50
Total	1655	548	562

(DPN) [29], U-Net [17], ReSeg [30], pyramid scene parsing network (PSPNet) [31], RefineNet [32], full-resolution residual networks (FRRN) [33], feature pyramid network (FPN) [34], LinkNet [35], DeepLabV3/V3+ [36], [37], dual attention network (DANet) [38], dense atrous spatial pyramid pooling (ASPP) [39], discriminative feature network (DFN) [40]. Among them, we selected the six typical models (seen in Fig. 3) to form the baseline algorithm, including U-Net, PSPNet, FPN, LinkNet, DeepLabV3, HRNet [41].

A. U-Net

U-Net, which was original proposed for biomedical image segmentation, is one of the most popular semantic segmentation networks and winning solutions in many challenges of the remote sensing community [42]. U-Net can be treated as an encoder network followed by a decoder network. The simple architecture of U-Net is illustrated in Fig. 3 (a), in which the downsampling blocks (yellow color) are used to extract the low-level features at multiple different levels, and the upsampling blocks (green color) are employed to concatenate the low-level features to infer the segmentation in the same resolution.

B. Linknet

Instead of concatenating in U-Net, LinkNet [shown in Fig. 3 (b)] adds the upsampled feature representation with resolution

information, which makes LinkNet as an efficient networks [35]. It combines both lower and higher layers to generate the final result. In DeepGlobe road extraction challenge [43], D-LinkNet, an improved version of LinkNet with pretrained encoder and an additional dilated convolution layers in the center part, won the first place [44].

C. Pyramid Scene Parsing Network

PSPNet utilizes a pyramid pooling module that aggregate contextual information in different regions to improve the ability to obtain global information [31].

As shown in Fig. 3 (c), the input image is fed into a pretrained model and dilated strategy to extract the feature map. The size of the extracted feature map is $1/8$ of the original size of the input image. Then, the pyramid pooling module is used to obtain context information at different spatial scales. The upsampling and concatenation operation is used to form the final feature map. Finally, a convolutional layer is used to obtain the final output.

D. Feature Pyramid Network

FPN works by creating two pyramids and combines them to generate feature-rich segmentation maps at each level [34]. As shown in Fig. 3 (d), the architecture consists of a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway contains many convolutional modules with many convolutional and pooling layers inside. It should be noted that each group of feature maps with the same size is called a stage, and the output of the last layer of each stage is the feature for the pyramid level. The top-down path includes up-sampling and depooling the last feature map while using lateral connections to enhance them at the same stage of the bottom-up path. The final step is to concatenate all the modules that have $1/4$ of the input image resolution and generate the final result.

E. Deeplab

DeepLabV1 [45] used atrous convolution to control the resolution of feature responses in CNNs. This is also known as dilated convolution and introduces another parameter, the dilation rate, to convolution layers, which spaces the convoluted

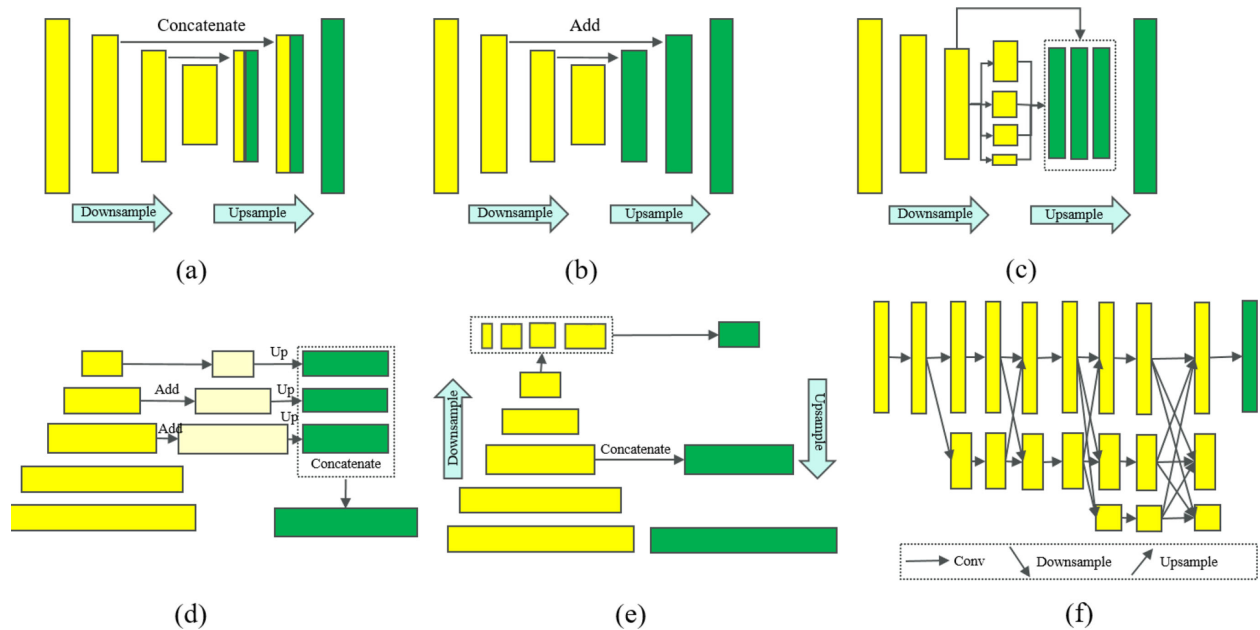


Fig. 3. Baseline segmentation models investigated in the study. (a) U-Net. (b) LinkNet. (c) PSPNet. (d) FPN. (e) DeepLabV3+. (f) HRNet.

pixels in a wider field of view while still having the same weights. DeepLabV2 [46] adopted Atrous Spatial Pyramid Pooling (ASPP) and Conditional Random Field (CRF), which helps to account for different object scales and improves accuracy. DeepLabV3 [47] add image-level features to ASPP and applying batch normalization for easier training. DeepLabV3+ [37] [see Fig. 3 (e)], as shown in extended DeepLabv3 by a decoder module to refine the segmentation results.

F. HRNet

HRNet [41] maintains high-resolution representations without losing spatial details throughout the whole process. As shown in Fig. 3 (f), HRNet starts with high-resolution subnetworks as the first stage, gradually add high-to-low resolution subnetworks, to form more stages, and connect the multiresolution subnetworks in a parallel way. In the whole process, HRNet perform multiscale repeated fusion by repeatedly exchanging information on parallel multiresolution subnetworks.

IV. EXPERIMENTAL SETTINGS

Our baselines and hyperparameters follow the publicly available code of segmentation models [47], except we add HRNet [41] and different loss functions and vary the number of training iterations. We used the F1 score ($F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$), IoU and overall accuracy (OA) to evaluate the result. The implementation is as follows.

A. Segmentation models

We investigated the attention U-Net [48], LinkNet, PSPNet, FPN, and DeepLabV3 with the encoders of Resnet [49], ResNeXt [50], Densenet [51], DPN [52], Inception [53], VGG [54], and EfficientNet [55], as well as the HRNet.

Different pretraining weights, including Imagenet, Instagram, semisupervised learning (SSL), semiweakly supervised learning (SSWL) [56], are also considered. Furthermore, we also include the ensemble of segmentation models.

B. Hyperparameters

All the segmentation networks are trained using Adam, with $\beta_1 = 0.9$ and with $\beta_2 = 0.999$, a learning rate of 0.0001, and a batch size of 16 for 100 epochs on a single NVIDIA V100 GPU with 16 GB memory. Only horizontal flipping augmentation is used for the training.

C. Loss Functions

The Dice, Jaccard, Focal [57], Lovasz [58], and their combinations are used.

V. RESULTS AND ANALYSIS

A. Investigation on a Single Model

First, we investigated the performance using different single models. Tables V and VI presented the accuracies for SAR, RGB, and SAR+RGB of the patch sizes of 256×256 and 512×512 , respectively. Here, the encoders of U-Net, LinkNet, PSPnet, FPN, and DeepLabV3 is Resnet18. Generally, RGB got higher accuracies, following by SAR+RGB and SAR. Simply stacking RGB and SAR decreased the accuracy because of the difference between the view angle of RGB and the incidence angle of SAR. HRNet obtained the best performance for RGB and SAR datasets in terms of IoU and F1 scores, followed by U-Net. For the patch size of 256×256 , Unet achieved the best OAs for RGB and SAR. For the patch size of 512×512 , Unet and PSPnet yielded the best OAs for RGB and SAR, respectively. Moreover, HRNet

TABLE V
RESULTS WITH DIFFERENT MODELS USING RGB, SAR, AND RGB+SAR (PATCH SIZE: 256×256): THE COLUMN OF GFLOPS REPRESENTS INFERENCE COMPUTATIONAL COST

Model	Datasets	Params(M)	Size(M)	GFLOPs	F1	IoU	OA
U-Net	RGB	14.4	56.6	5.4	0.5872	0.4157	0.6513
	SAR	14.4	56.5	5.3	0.4431	0.2846	0.5429
	RGB+SAR	14.5	56.6	5.5	0.5636	0.3923	0.6312
LinkNet	RGB	11.7	45.7	3.1	0.5729	0.4013	0.6463
	SAR	11.7	45.7	2.9	0.4198	0.2657	0.4800
	RGB+SAR	11.7	45.7	3.1	0.5588	0.3878	0.6410
PSPnet	RGB	11.3	44.4	1.5	0.5667	0.3954	0.6203
	SAR	11.3	44.4	1.3	0.4409	0.2828	0.5103
	RGB+SAR	11.3	44.4	1.5	0.5423	0.3721	0.6443
FPN	RGB	13.1	51.1	4.5	0.5760	0.4044	0.5842
	SAR	13.0	51.1	4.4	0.4395	0.2816	0.5030
	RGB+SAR	13.1	51.1	4.5	0.5471	0.3766	0.5738
DeepLabV3	RGB	15.9	62.2	16.8	0.5804	0.4088	0.6436
	SAR	15.9	62.2	16.7	0.4405	0.2824	0.5029
	RGB+SAR	15.9	62.2	16.8	0.5471	0.3766	0.5029
HRNet	RGB	9.9	37.2	4.5	0.6178	0.4516	0.6411
	SAR	9.9	37.2	4.5	0.4567	0.2963	0.5314
	RGB+SAR	9.9	37.2	4.5	0.5783	0.4010	0.6098

The bold entities mean the best performance.

TABLE VI
RESULTS WITH DIFFERENT MODELS USING RGB, SAR, AND RGB+SAR (PATCH SIZE: 512×512): THE COLUMN OF GFLOPS REPRESENTS INFERENCE COMPUTATIONAL COST

Model	Datasets	Params(M)	Size(M)	GFLOPs	F1	IoU	OA
U-Net	RGB	14.4	56.6	21.7	0.5715	0.4001	0.6155
	SAR	14.4	56.5	21.3	0.4426	0.2842	0.5423
	RGB+SAR	14.5	56.6	21.9	0.5416	0.3713	0.6584
LinkNet	RGB	11.7	45.7	12.2	0.5414	0.3712	0.6542
	SAR	11.7	45.7	11.8	0.4143	0.2612	0.5398
	RGB+SAR	11.7	45.7	12.4	0.5248	0.3557	0.6124
PSPnet	RGB	11.3	44.4	5.8	0.5314	0.3619	0.6267
	SAR	11.3	44.4	5.4	0.4255	0.2702	0.5787
	RGB+SAR	11.3	44.4	6.0	0.5137	0.3456	0.5934
FPN	RGB	13.1	51.1	17.8	0.5508	0.3800	0.5368
	SAR	13.0	51.1	17.4	0.4407	0.2826	0.5350
	RGB+SAR	13.1	51.1	18.0	0.5350	0.3651	0.5702
DeepLabV3	RGB	15.9	62.2	68.8	0.5712	0.4097	0.6018
	SAR	15.9	62.2	66.8	0.4307	0.2811	0.5278
	RGB+SAR	15.9	62.2	69.2	0.5413	0.3675	0.6001
HRNet	RGB	9.9	37.2	17.6	0.5878	0.4316	0.6187
	SAR	9.9	37.2	17.3	0.4501	0.2887	0.5328
	RGB+SAR	9.9	37.2	18.0	0.5678	0.3817	0.6088

The bold entities mean the best performance.

has fewer parameters and reducing FLOPs when compared to U-Net.

In order to deeply analyze the results, we provide the accuracies (F1 score) of each city (patch size: 256×256) in Fig. 4. Barcelona, Rennes, and Berlin yielded better results than other cities. One reason could be that the number of patches from the three cities is higher than that of other cities. The other reason is that most of the buildings in these three cities are moderate size and height. As point out in [59], the size and height of buildings influence the performance. Smaller buildings could not be detected, which exist in the datasets of Shanghai, Beijing, Rio, and Yokosuka. Performance gradually increases and then decreases as buildings become taller. Geometric distortions (lay-over or foreshortening) become more extreme as building height increases. In this case, many tall buildings in San Diego and Hongkong cause lower accuracies. The performance of different

models varied. For instance, U-Net, DeepLabV3, and HRNet achieved the best segmentation results of SAR for Rennes, Barcelona, and Berlin, respectively. The similar observations can be found in the patch size of 512×512 .

Second, we investigated the performance with U-Net using different encoders, including Resnet18, Resnet50, ResNeXt50_32 \times 4 d, SE-Resnet50, SE_ResNeXt50_32 \times 4 d, Densenet201, Dpn68, Inception-v4, VGG11 and newly EfficientNet-b5/b7 (Tables VII and VIII). For RGB results, SE_ResNeXt50_32 \times 4 d and ResNeXt50_32 \times 4 d achieved the best results with the F1 (IoU) of 0.6238 (0.4533) and 0.6056 (0.4343) for the patch sizes of 256×256 and 512×512 , respectively. SE_Resnet50 and EfficientNet-b7 achieve the best performance in terms of OAs for the patch sizes of 256×256 and 512×512 . For the results of SAR, the newly EfficientNet-b5 generated better results than others. Compared to VGG, DPN,

TABLE VII
RESULTS WITH DIFFERENT ENCODERS USING RGB, SAR, AND RGB+SAR (PATCH SIZE: 256×256): THE COLUMN OF GFLOPS REPRESENTS INFERENCE COMPUTATIONAL COST

Model	Datasets	Params(M)	Size(M)	GFLOPs	F1	IoU	OA
Resnet18	RGB	14.4	56.6	5.4	0.5872	0.4157	0.6513
	SAR	14.4	56.5	5.3	0.4431	0.2846	0.5429
	RGB+SAR	14.5	56.6	5.5	0.5636	0.3923	0.6312
Resnet50	RGB	33.8	132.5	10.7	0.6197	0.4489	0.6308
	SAR	33.8	132.4	10.6	0.4563	0.2956	0.5766
	RGB+SAR	33.8	132.5	10.8	0.5781	0.4066	0.6048
ResNeXt50_32×4d	RGB	33.3	130.5	10.9	0.6222	0.4516	0.6289
	SAR	33.3	130.4	10.8	0.4560	0.2954	0.5265
	RGB+SAR	33.3	130.5	11.0	0.5841	0.4126	0.5932
SE_Resnet50	RGB	36.4	142.4	10.4	0.5974	0.4259	0.7142
	SAR	36.3	142.3	10.3	0.4567	0.2959	0.5107
	RGB+SAR	36.4	142.4	10.5	0.5746	0.4031	0.6164
SE_ResNeXt50_32×4d	RGB	35.8	140.4	10.9	0.6238	0.4533	0.6434
	SAR	35.8	140.3	10.8	0.4628	0.3010	0.5535
	RGB+SAR	35.8	140.4	11.0	0.5943	0.4228	0.6461
Densenet201	RGB	30.4	120.1	11.4	0.6093	0.4381	0.6313
	SAR	30.4	120.1	11.3	0.4563	0.2956	0.5143
	RGB+SAR	30.4	120.1	11.5	0.5764	0.4048	0.5989
Dpn68	RGB	17.3	68.1	6.8	0.6027	0.4313	0.6908
	SAR	17.3	68.1	6.8	0.4462	0.2871	0.6533
	RGB+SAR	17.3	68.1	6.8	0.5764	0.4142	0.6927
Inception-v4	RGB	49.7	194.7	15.4	0.6066	0.4353	0.6578
	SAR	49.7	194.7	15.4	0.4620	0.3005	0.5622
	RGB+SAR	49.7	194.7	15.4	0.5676	0.3963	0.6068
VGG11	RGB	18.5	72.4	14.5	0.6040	0.4327	0.6697
	SAR	18.5	72.4	14.5	0.4330	0.2763	0.5123
	RGB+SAR	18.5	72.4	14.6	0.5744	0.4029	0.6660
EfficientNet-b5	RGB	31.3	123.3	2.9	0.5999	0.4285	0.7097
	SAR	31.3	123.2	2.9	0.4664	0.3041	0.5673
	RGB+SAR	31.3	123.3	2.9	0.5866	0.4150	0.6626
EfficientNet-b7	RGB	67.2	264.2	3.1	0.6139	0.4430	0.6886
	SAR	67.2	264.2	3.1	0.4168	0.2632	0.6741
	RGB+SAR	67.2	264.2	3.1	0.5620	0.3909	0.5620

The bold entities mean the best performance.

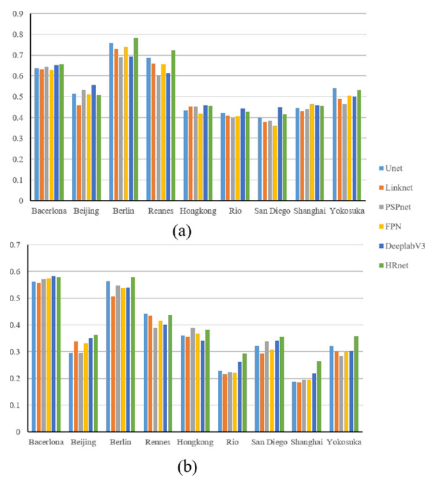


Fig. 4. Segmentation results (F1 score) of individual city using different single models (patch size: 256×256). (a) RGB. (b) SAR.

Densenet201, Resnet18, EfficientNet-b5 improves F1 and IoU scores by 1.01–3 percentage points for SAR, and the GFLOPs is

much lower and less than one third, while the number of parameters of EfficientNet-b5 is larger. Compared to Inception-v4, SE_ResNeXt50_32×4 d and EfficientNet-b7, EfficientNet-b5 achieves 0.2–5 percentage points SAR in terms of F1 and IoU scores with smaller model size and complexity.

From the SAR results' accuracies (patch size: 256×256) of each city in Fig. 5, EfficientNet-b5 obtained the best results for six cities, including Barcelona, Beijing, Hongkong, Rio, San Diego, and Yokosuka. SE_ResNeXt50_32×4 d, Inception-v4, and Resnet50 produced the best results for Berlin, Rennes, and Shanghai, respectively. For the patch size of 512×512 , EfficientNet-b5 obtained the best results for five cities. Fig 6 has shown the building segmentation results (patch size: 256×256) for RGB and SAR image patches using U-Net with SE_ResNeXt50_32×4 d and EfficientNet-b5. SAR results got the strong backscatter locations, which show as bright spots areas in the datasets. Due to different imaging conditions between RGB and SAR, the boundaries of SAR are hard to be determined. Thus, segmentation networks cannot model the strong backscatter locations together with their shadows and boundaries, which leads to more distortion of SAR results.

TABLE VIII
RESULTS WITH DIFFERENT ENCODERS USING RGB, SAR, AND RGB+SAR (PATCH SIZE: 512×512): THE COLUMN OF GFLOPS REPRESENTS INFERENCE COMPUTATIONAL COST

Model	Datasets	Params(M)	Size(M)	GFLOPs	F1	IoU	OA
Resnet18	RGB	14.4	56.6	21.7	0.5715	0.4001	0.6155
	SAR	14.4	56.5	21.3	0.4426	0.2842	0.5423
	RGB+SAR	14.5	56.6	21.9	0.5416	0.3713	0.6584
Resnet50	RGB	33.8	132.5	42.8	0.5985	0.4270	0.6549
	SAR	33.8	132.4	42.4	0.4535	0.2932	0.5985
	RGB+SAR	33.8	132.5	43.0	0.5613	0.3901	0.5613
ResNeXt50_32×4d	RGB	33.3	130.5	43.6	0.6056	0.4343	0.6361
	SAR	33.3	130.4	43.2	0.4576	0.2967	0.5723
	RGB+SAR	33.3	130.5	43.8	0.5670	0.3957	0.5670
SE-Resnet50	RGB	36.4	142.4	41.6	0.6011	0.4298	0.6813
	SAR	36.3	142.3	41.2	0.4569	0.2961	0.5955
	RGB+SAR	36.4	142.4	41.8	0.5792	0.4077	0.6483
SE_ResNeXt50_32×4d	RGB	31.3	123.3	11.4	0.5563	0.3854	0.6812
	SAR	31.3	123.2	11.4	0.4153	0.2621	0.4689
	RGB+SAR	31.3	123.3	11.4	0.5143	0.3462	0.6904
Densenet201	RGB	30.4	120.1	45.6	0.5946	0.4231	0.6654
	SAR	30.4	120.1	45.2	0.4572	0.2963	0.5682
	RGB+SAR	30.4	120.1	45.8	0.5586	0.3875	0.6939
Dpn68	RGB	17.3	68.1	27.0	0.5726	0.4011	0.6917
	SAR	17.3	68.1	27.0	0.4370	0.2796	0.5924
	RGB+SAR	17.3	68.1	27.0	0.5692	0.3978	0.6792
Inception-v4	RGB	49.7	194.7	61.6	0.5616	0.3904	0.7044
	SAR	49.7	194.7	61.5	0.4580	0.2970	0.5093
	RGB+SAR	49.7	194.7	61.6	0.5593	0.3883	0.6209
VGG11	RGB	18.5	72.4	58.1	0.5755	0.4040	0.6632
	SAR	18.5	72.4	57.8	0.4416	0.2834	0.5755
	RGB+SAR	18.5	72.4	58.2	0.5352	0.3654	0.6757
EfficientNet-b5	RGB	35.8	140.4	43.6	0.6016	0.4302	0.6208
	SAR	35.8	140.3	43.2	0.4596	0.2984	0.5567
	RGB+SAR	35.8	140.4	43.8	0.5516	0.3809	0.6297
EfficientNet-b7	RGB	67.2	264.2	12.5	0.5312	0.3617	0.7124
	SAR	67.2	264.2	12.5	0.4351	0.2780	0.5215
	RGB+SAR	67.2	264.2	12.5	0.5600	0.3884	0.6592

The bold entities mean the best performance.

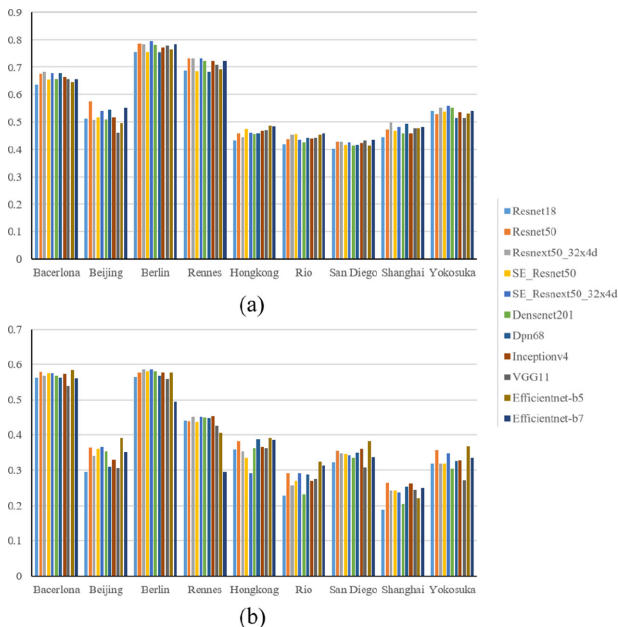


Fig. 5. Segmentation results (F1 score) of individual city using different encoders (patch size: 256×256). (a) RGB. (b) SAR.

Third, we investigated the performance with different loss functions using U-Net with Resnet18. We compared the performance using Dice, Jaccard, Focal, Lovasz, and their combinations in this work. The results are listed in Tables IX and X. The combination of Jaccard and Focal gained the best results in both cases, followed by combining all loss functions.

Fourth, we investigated the performance with different pretraining weights, including Imagenet, Instagram, SSL on Imagenet, SWSL on Imagenet, from the encoder of ResNeXt101_32×8 d. We also consider the transfer learning approach [59]. In this case, the model is first trained on RGB datasets, and then the generated weights are used as the initial weights for training on SAR. Since SAR datasets only have one channel, the process is simplified by averaging three RGB channels to make it one channel.

In Table XI (patch size: 256×256), the model F1 scores of RGB and SAR dropped to 0.6017 and 0.4181 in the absence of pretraining. Then, we replaced Imagenet with Instagram, SWSL, and RGB, the scores of SAR can be slightly increased. The score can be increased to 0.4695 by using the pretraining weights of SSL. For the patch size of 512×512 (seen in Table XII), the IoU and F1 scores decreased without any pretraining. When the

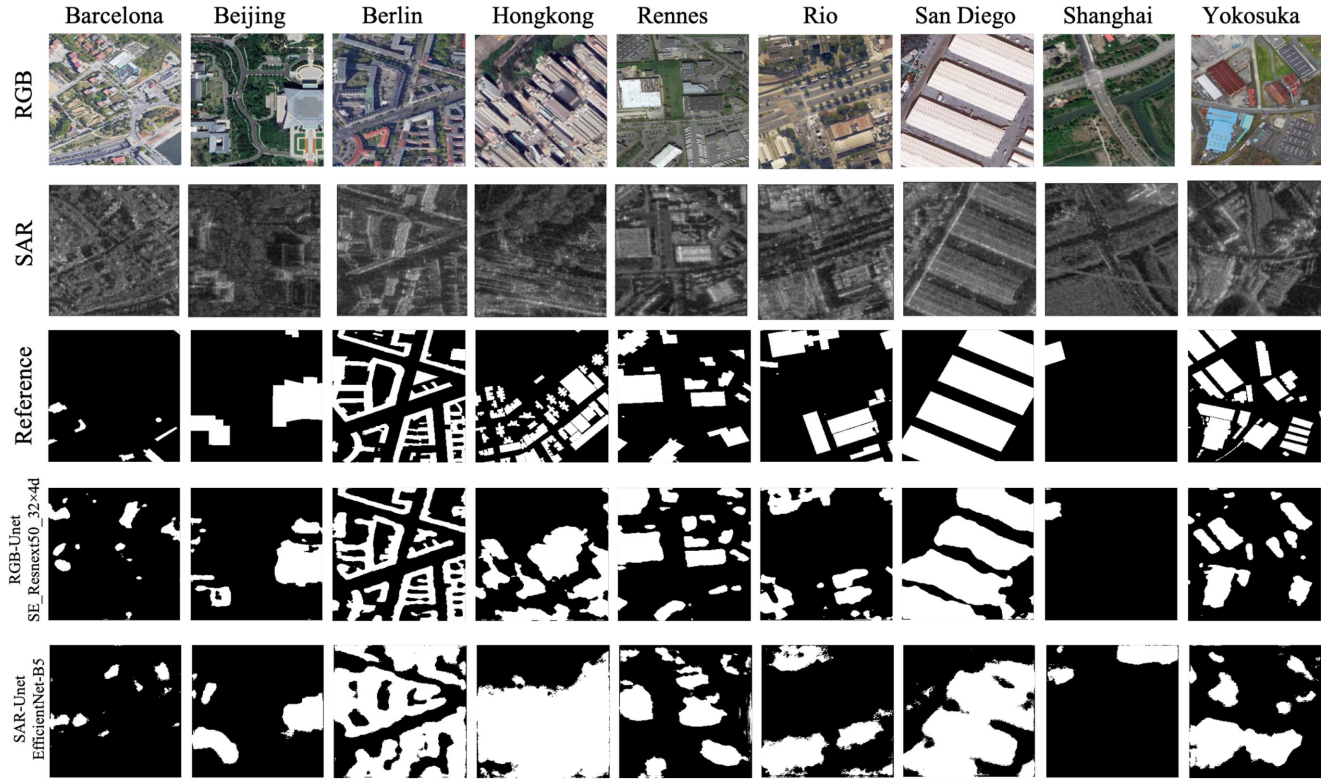


Fig. 6. Segmentation results of RGB and SAR datasets (patch size: 256×256) from individual city. For RGB and SAR datasets, U-Net with SE_ResNeXt50_32x4d and EfficientNet-B5 are respectively used.

TABLE IX
RESULTS WITH DIFFERENT LOSS FUNCTIONS (PATCH SIZE: 256×256)

Dice	Jaccard	Focal	Lovasz	RGB			SAR		
				F1	IoU	OA	F1	IoU	OA
✓	✓	✓	✓	0.6008	0.4192	0.7810	0.3808	0.2796	0.7267
				0.5997	0.4283	0.7611	0.4402	0.2821	0.7410
✓	✓	✓	✓	0.5998	0.4294	0.5659	0.4245	0.2351	0.3040
				0.4765	0.1508	0.3630	0.3456	0.1418	0.8764
✓	✓	✓	✓	0.6197	0.4489	0.6308	0.4563	0.2956	0.5766
				0.6002	0.4317	0.6224	0.4462	0.2873	0.5713
✓	✓	✓	✓	0.6102	0.4411	0.6302	0.4481	0.2914	0.5742

The bold entities mean the best performance.

TABLE X
RESULTS WITH DIFFERENT LOSS FUNCTIONS (PATCH SIZE: 512×512)

Dice	Jaccard	Focal	Lovasz	RGB			SAR		
				F1	IoU	OA	F1	IoU	OA
✓	✓	✓	✓	0.5518	0.3816	0.6434	0.3914	0.2612	0.6015
				0.5614	0.3814	0.6537	0.4026	0.2624	0.6126
✓	✓	✓	✓	0.5618	0.3902	0.6512	0.4217	0.2683	0.6678
				0.5124	0.2346	0.4874	0.3812	0.1982	0.7026
✓	✓	✓	✓	0.5715	0.4001	0.6155	0.4426	0.2842	0.5423
				0.5711	0.3982	0.6812	0.4378	0.2801	0.5306
✓	✓	✓	✓	0.5703	0.3912	0.6928	0.4414	0.2813	0.5614

The bold entities mean the best performance.

TABLE XI
RESULTS WITH DIFFERENT PRETRAINING WEIGHTS (PATCH SIZE: 256×256)

re-training	RGB			SAR		
	F1	IoU	OA	F1	IoU	OA
None	0.6017	0.4234	0.5017	0.4181	0.2645	0.4658
Imagenet	0.6417	0.4725	0.5325	0.4589	0.2936	0.4850
Instagram	0.6577	0.4900	0.6812	0.4659	0.3037	0.5644
SSL	0.6539	0.4857	0.6348	0.4695	0.3068	0.4599
SWSL	0.6586	0.4909	0.7095	0.4628	0.3011	0.5714
RGB	-	-	-	0.4621	0.2916	0.5693

The bold entities mean the best performance.

TABLE XII
RESULTS WITH DIFFERENT PRETRAINING WEIGHTS (PATCH SIZE: 512×512)

Pre-training	RGB			SAR		
	F1	IoU	OA	F1	IoU	OA
None	0.5984	0.4212	0.5867	0.4316	0.2816	0.5414
Imagenet	0.6104	0.4393	0.6894	0.4644	0.3024	0.5728
Instagram	0.6060	0.4347	0.6790	0.4677	0.3052	0.6220
SSL	0.6107	0.4396	0.7093	0.4584	0.2974	0.5410
SWSL	0.6062	0.4350	0.6899	0.4700	0.3072	0.6097
RGB	-	-	-	0.4641	0.2952	0.5730

The bold entities mean the best performance.

TABLE XIII
RESULTS WITH DIFFERENT ENSEMBLE SCHEMES (PATCH SIZE: 256×256)

No	Model	Pre-training	F1	IoU	OA
1	U-Net, LinkNet, PSPnet, FPN, DeepLabV3, HRNet	Imagenet	0.4682	0.3092	0.5802
2	Resnet50, ResNeXt50_32×4d, SE_ResNeXt50_32×4d, Inception-v4, EfficientNet-B5	Imagenet	0.4727	0.3122	0.5824
3	EfficientNet-B5	Imagenet	0.4738	0.3176	0.5826
4	ResNeXt101_32×8d	Instagram SSL SWSL RGB	0.4718	0.3142	0.5801

The bold entities mean the best performance.

pretraining weights of Instagram, SSL, SWSL, and RGB are included, the performance was increased. SWSL and Instagram obtained the best results of F1 (IoU) and OA, respectively.

B. Investigation of Multiple Models

From the above subsection (Figs. 4 and 5), we can find that different models shows different advantages in different cities. In this case, combining different models may improve performance. In ensemble learning, the accuracy of a single model and the diversity among different models should be considered [60]. The following ensemble models are adopted by considering different methods to promote diversity.

- 1) Model 1: ensemble of U-Net, LinkNet, PSPnet, FPN, DeepLabV3, and HRNet.

- 2) Model 2: ensemble of U-Net with different encoders of Resnet50, ResNeXt50_32×4 d, SE_ResNeXt50_32×4 d, Inception-v4, and EfficientNet-B5.
- 3) Model 3: ensemble of U-Net with the encoder of EfficientNet-B5 with four different combination of loss functions.
- 4) Model 4: ensemble of U-Net with encoder of ResNeXt101_32×8 d with different pretraining weights.

Tables XIII and XIV has shown the results obtained from different ensemble schemes. All used ensembles of neural networks achieved better performance than a single neural network. EfficientNet has achieved state-of-the-art performance on ImageNet while being markedly smaller and faster than other networks. Specifically, EfficientNet-B5 obtained the best results in SAR datasets. Thus, the combination of diverse

TABLE XIV
RESULTS WITH DIFFERENT ENSEMBLE SCHEMES (PATCH SIZE: 512×512)

No	Model	Pre-training	F1	IoU	OA
1	U-Net, LinkNet, PSPnet, FPN, DeepLabV3, HRNet	Imagenet	0.4573	0.2801	0.5404
2	Resnet50, ResNeXt50_32×4d, SE_ResNeXt50_32×4d, Inception-v4, EfficientNet-B5	Imagenet	0.4612	0.2984	0.5523
3	EfficientNet-B5	Imagenet	0.4678	0.3088	0.5618
4	ResNeXt101_32×8d	Instagram SSL SWSL RGB	0.4712	0.3104	0.5698

The bold entities mean the best performance.

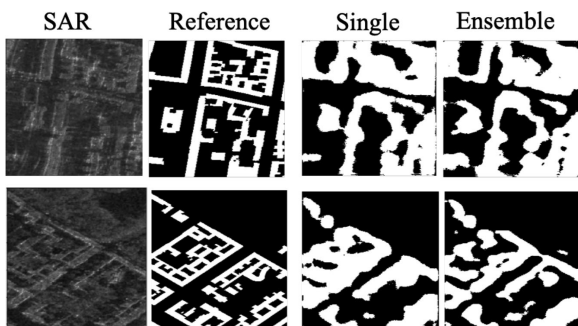


Fig. 7. Segmentation results of SAR datasets for single and ensemble models.

EfficientNet-B5 models yielded the best accuracy with F1 and IoU of 0.4738 and 0.3176 for the patch size of 256×256 (seen in Fig. 7). The pretraining weights of Instagram, SSL, SWSL, and RGB can boost the performance of a single model. Then, the performance of combining Instagram, SSL, SWSL, and RGB pretraining weights is better than other schemes for the patch size of 512×512 . It should be emphasized that different ensemble schemes obtain the best results with different patch sizes. To improve the performance of the ensemble, the diversity among the single results should be considered.

C. Other Influences

In VHR remote sensing semantic segmentation, the main factor is the boundaries. In general, the building segmentation boundaries are not perfect because the spatial information may be often lost in the training process. In this case, the following strategies are often adopted.

- 1) Multichannel mask: In this case, building labels are split into three channels: interiors, edges, and contacts between buildings.
- 2) Postprocessing step is used to refine the segmentation results. The typical one is the conditional random field (CRF).

In this work, we consider including multichannel masks, postprocessing, and both of them (seen results in Tables XV

TABLE XV
RESULTS (F1 SCORE) WITH MULTICHANNEL MASK AND POSTPROCESSING (PATCH SIZE: 256×256)

Efficientnet-B5	+Multi-channel	+CRF	both
0.4738	0.4751	0.4758	0.4763

TABLE XVI
RESULTS (F1 SCORE) WITH MULTICHANNEL MASK AND POSTPROCESSING (PATCH SIZE: 512×512)

ResNeXt101_32×8d	+Multi-channel	+CRF	both
0.4712	0.4728	0.4733	0.4756

and XVI). Both multichannel mask and CRF can slightly improve the performance.

VI. DISCUSSIONS

We summarized the main observations from our experiments as follows:

- 1) segmentation results using SAR are worse than the ones using RGB;
- 2) simple stacking RGB and SAR cannot improve accuracy when compared to RGB;
- 3) U-Net is efficient to obtain accurate segmentation results;
- 4) efficient achieved the SOTA results of SAR segmentation results;
- 5) ensemble, multichannel mask, and postprocessing can enhance the result.

Based on these observations, we provide some suggestions that may help people to choose the model for SAR segmentation results:

A. Is U-Net Enough?

Yes. When we use the same encoder, U-Net outperformed LinkNet, PSPnet, FPN, and DeepLabV3. HRNet performed better than U-Net. However, when we adopt a more advanced encoder (e.g., SE-Resnet50), U-Net is better than HRNet. It is sufficient to directly train with the U-Net with an advanced encoder to get accurate results.

B. Which Encoder of U-Net Do You Recommend?

EfficientNet. In our work, EfficientNet-B5 shows very accurate results with fewer parameters and reducing GFLOPs. Four of the five winners of SpaceNet 6 used slight variants of the newly introduced EfficientNet (B5, B7, B8).

C. Does Advanced Pretraining Required?

No. If only one single model is adopted, using advanced pretraining weights (i.e., SSL and SWSL) or training on RGB first and then fine-tuning on SAR is useful. If the ensemble scheme is selected, using Imagenet pretraining weights is sufficient.

D. Is Ensemble Helpful?

Yes. Since different models can make a difference in the results, combining multiple models can increase accuracy. However, it is trivial to select which models and how many models for the ensembles.

E. Do We Need a Multichannel Mask and Postprocessing?

Maybe, but recommended. In our work, multichannel mask and postprocessing slightly improve the performance of SAR datasets with 1 m resolution. The winners of SpaceNet demonstrated that such tricks are beneficial for the datasets of 0.5 m resolution. We recommend using such tricks for VHR (≤ 1 m) SAR datasets.

VII. CONCLUSION

This article provides a representative benchmark of high-resolution SAR datasets for building segmentation and reviews the current state-of-the-art segmentation methods. To investigate the segmentation performance of SAR, a comprehensive assessment with different models, encoders, pretraining weights, ensemble schemes are performed. Based on the evaluation, we give some suggestions to improve the segmentation results of SAR datasets. Particular attention should be given to U-Net with the encoder of EfficientNet-B5 with potential improvements of applying ensemble, multichannel mask, and postprocessing.

Since the next generation spaceborne SAR (such as Capella Space) will be launched and can provide submeter resolution datasets with global coverage, it will bring both opportunities and challenges. Specifically to SAR building segmentation, the following directions should be considered in future studies:

- 1) Large-scale submeter benchmark datasets: As we pointed out in this work, SAR benchmark datasets for building segmentation are missing. Our datasets cover nine cities but are limited to one band. SpaceNet 6 datasets have four channels but are limited to one city. In the futures, the datasets covered several cities with full-polarization mode are particularly needy in the community.
- 2) Weakly supervised learning: The ground truth or reference of buildings is mainly from OSM. As we all know, the primary source is collected from the volunteers that use GPS trackers *in situ* or manually digitize on VHR aerial or satellite images, which is not specific to SAR datasets.

There are lots of missing, incorrect, and misalign building labels for SAR datasets in this case. It is thus meaningful to use weakly supervised learning to correct the inaccurate labels to improve performance.

- 3) Unsupervised learning: Unsupervised learning can directly learn the data itself, which can bypass the influence of incorrect labels. In this case, we can fully exploit the characteristics of SAR by using feature learning, clustering, representation learning. Two typical potential techniques are Noise2Noise [61] or contrastive learning [62], which can despeckle the datasets and classify the objects without any clean data and ground truth.

ACKNOWLEDGMENT

The authors would like to thank the China National Space Administration (CNSA) for providing the GF-3 datasets, respectively. The authors would also like to thank Piesat Information Technology Co., Ltd. to provide SAR software.

REFERENCES

- [1] T. Kobayashi, M. Satake, H. Masuko, T. Manabe, and M. Shimada, "CRL/NASDA airborne dual-frequency polarimetric interferometric SAR system," *Proc. SPIE*, vol. 3497, pp. 2–12, 1998.
- [2] P. Gamba and G. Lisini, "Fast and efficient urban extent extraction using ASAR wide swath mode data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 5, pp. 2184–2195, Oct. 2013.
- [3] Y. Ban, A. Jacob, and P. Gamba, "Spaceborne SAR data for global urban mapping at 30 m resolution using a robust urban extractor," *ISPRS J. Photogrammetry Remote Sens.*, vol. 103, pp. 28–37, 2015.
- [4] H. Cao, H. Zhang, C. Wang, and B. Zhang, "Operational built-up areas extraction for cities in china using sentinel-1 SAR data," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 874.
- [5] M. Chini, R. Pelich, R. Hostache, P. Matgen, and C. Lopez-Martinez, "Towards a 20 m global building map from sentinel-1 SAR data," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1833.
- [6] T. Esch *et al.*, "Tandem-x mission-new perspectives for the inventory and monitoring of global settlement patterns," *J. Appl. Remote Sens.*, vol. 6, no. 1, Oct. 2012, Art. no. 061702.
- [7] T. Esch *et al.*, "Urban footprint processor—fully automated processing chain generating settlement masks from global data of the tandem-x mission," *IEEE Geosci. Remote. Sens. Lett.*, vol. 10, no. 6, pp. 1617–1621, Nov. 2013.
- [8] U. Gessner *et al.*, "Multi-sensor mapping of west African land cover using MODIS, ASAR and TanDEM-X/TerraSAR-X data," *Remote Sens. Environ.*, vol. 164, pp. 282–297, 2015.
- [9] M. Klotz, T. Kemper, C. Geiß, T. Esch, and H. Taubenböck, "How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe," *Remote Sens. Environ.*, vol. 178, pp. 191–212, 2016.
- [10] L. Xu, H. Zhang, C. Wang, and Q. Fu, "Classification of chinese Gaofen-3 fully-polarimetric SAR images: Initial results," in *Proc. Prog. Electromagn. Res. Symp. - Fall*, 2017, pp. 700–705.
- [11] M. Ma, J. Chen, W. Liu, and W. Yang, "Ship classification and detection based on CNN using GF-3 SAR images," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2043.

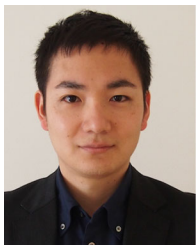
- [12] X. Hou, W. Ao, Q. Song, J. Lai, H. Wang, and F. Xu, "FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition," *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140303.
- [13] H. Gao *et al.*, "A crop classification method integrating GF-3 PolSAR and Sentinel-2A optical data in the Dongting Lake Basin," *Sensors*, vol. 18, no. 9, 2018, Art. no. 3139.
- [14] L. Zhang, G. Li, C. Zhang, H. Yue, and X. Liao, "Approach and practice: Integrating earth observation resources for data sharing in China GEOSS," *Int. J. Digit. Earth*, vol. 12, no. 12, pp. 1441–1456, 2019.
- [15] H. Yu *et al.*, "Methods and datasets on semantic segmentation: A review," *Neurocomputing*, vol. 304, pp. 82–103, 2018.
- [16] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.
- [17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
- [18] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, 2020.
- [19] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in VHR SAR images using fully convolution neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.
- [20] J. Li, R. Zhang, and Y. Li, "Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images." in *Proc. Int. Geosci. Remote Sens. Symp.*, 2016, pp. 910–913.
- [21] R. Wu, Y. Zhang, and Y. Li, "The Detection of Built-Up Areas in High-Resolution Sar Images Based on Deep Neural Networks," in *Proc. Int. Conf. Image Graph.*, 2017, pp. 646–655.
- [22] Z. Huang, M. Datcu, Z. Pan, and B. Lei, "Deep SAR-Net: Learning objects from signals," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 179–193, 2020.
- [23] C. Berger *et al.*, "Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 3, pp. 1324–1340, Jun. 2013.
- [24] W. Yao, D. Marmanis, and M. Datcu, "Semantic segmentation using deep neural networks for SAR and optical image pairs," in *Proc. Big Data Space*, 2017, pp. 2–5.
- [25] B. Han *et al.*, "The GF-3 SAR data processor," *Sensors*, vol. 18, no. 3, 2018, Art. no. 835.
- [26] X. Cao, Y. Liu, Q. Liu, X. Cui, X. Chen, and J. Chen, "Estimating the age and population structure of encroaching shrubs in arid/semiarid grasslands using high spatial resolution remote sensing imagery," *Remote Sens. Environ.*, vol. 216, pp. 572–585, 2018.
- [27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
- [28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [29] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1377–1385.
- [30] F. Visin *et al.*, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 426–433.
- [31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," 2016. [Online]. Available: <http://arxiv.org/abs/1612.01105>
- [32] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.
- [33] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3309–3318.
- [34] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
- [35] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.
- [36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv: 1706.05587*, 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. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [38] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.
- [39] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.
- [40] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.
- [41] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv: 1904.04514*.
- [42] V. Iglovikov and A. Shvets, "Ternausnet: U-Net with VGG11 encoder pre-trained on Imagenet for image segmentation," 2018, *arXiv: 1801.05746*.
- [43] I. Demir *et al.*, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 172–181.
- [44] 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*, Jun. 2018, pp. 182–186.
- [45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2016, *arXiv: 1412.7062*.
- [46] 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," 2017, *arXiv: 1606.00915*.
- [47] P. Yakubovskiy, "Segmentation Models Pytorch," 2020. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
- [48] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze excitation in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2018, pp. 421–429.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [50] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
- [51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
- [52] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Advances in neural information processing systems," vol. 30, pp. 4467–4475, 2017.
- [53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, pp. 2818–2826, 2016.
- [54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 2818–2826.
- [55] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1–14.
- [56] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019, *arXiv: 1905.00546*.
- [57] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
- [58] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421.
- [59] J. Shermeyer *et al.*, "Spacenet 6: Multi-sensor all weather mapping dataset," 2020, *arXiv: 2004.06500*.
- [60] G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, G. I. Webb and C. Sammut, Eds. New York, NY, USA: Springer, 2010.
- [61] J. Lehtinen *et al.*, "Noise2noise: Learning image restoration without clean data," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2965–2974.
- [62] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 9726–9735.



Junshi Xia (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2008 and 2013, respectively, and the Ph.D. degree in image processing from the Grenoble Images Speech Signals and Automatics Laboratory, Grenoble Institute of Technology, Grenoble, France, in 2014.

From 2014 to 2015, he was a Visiting Scientist with the Department of Geographic Information Sciences, Nanjing University, Nanjing, China. From 2015 to 2016, he was a Postdoctoral Research Fellow with the University of Bordeaux, Bordeaux, France. From 2016 to 2018, he was the Japan Society for the Promotion of Science (JSPS) Postdoctoral Overseas Research Fellow with the University of Tokyo, Tokyo, Japan. Since 2018, he has been a Research Scientist with RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan. His research interests include multiple classifier systems in remote sensing, hyperspectral remote sensing image processing, and deep learning in remote sensing applications.

Dr. Xia was the recipient of the first place prize in the IEEE Geoscience and Remote Sensing Society Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee in 2017. Since 2019, he has been an Associate Editor for the *IEEE Geoscience and Remote Sensing Letters* (GRSL), and Guest Editor for Remote Sensing and the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS).



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

He is currently a Lecturer with the University of Tokyo and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he leads the Geoinformatics Unit. He was an Assistant Professor with the University of Tokyo from 2013 to 2017. During 2015–2017, he was an Alexander von Humboldt Fellow, working with the German Aerospace Center (DLR), Oberpfaffenhofen, and Technical University of Munich (TUM), Munich, Germany. His research interests include the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management.

Dr. Yokoya won the first place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was the Chair during 2019–2021 and was a Co-Chair during 2017–2019 of IEEE GRSS IADF TC and also the secretary of the IEEE GRSS All Japan Joint Chapter since 2018. He is an Associate Editor for the *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (JSTARS) since 2018. He is/was a Guest Editor for the IEEE JSTARS during 2015–2021, for *Remote Sensing* during 2016–2021, and for the *IEEE Geoscience and Remote Sensing Letters* (GRSL) during 2018–2019.



Bruno Adriano (Member, IEEE) received the M.Eng. degree in disaster management from the National Graduate Institute for Policy Studies, Tokyo, Japan, and the International Institute of Seismology and Earthquake Engineering, Building Research Institute, Japan, in 2010, and the Ph.D. degree in civil and environmental engineering from the Department of Civil and Environmental Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan, in 2016.

From 2016 to 2018, he was a Fellow Researcher with the Japan Society for the Promotion of Science, the International Research Institute of Disaster Science, Tohoku University. Since 2018, he is a Postdoctoral Researcher with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. His research interests include integrating machine learning and high-performance computing for disaster management using remote sensing technologies.



Lianchong Zhang (Member, IEEE) received his Master degree in natural geography from Hebei Normal University, Shijiazhuang, China, in 2013, and the Ph.D. degree in signal and information processing from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2019. He is a Postdoctoral Fellow with the Aerospace Information Research Institute (AIR), Chinese Academy of Sciences, Guangzhou, China. His research interests include earth observation data integrating, management, and sharing.



Guoqing Li (Senior Member, IEEE) received Bachelor of Space Physics from Peking University in 1990, and then got Master of Science and Ph.D. of Science from Chinese Academy of Sciences in 1999 and 2005 respectively, majored in Cartography and Geographic Information System. He is a Professor and the Director of National Earth Observation Data Center (NODA). His research interests include on high-performance geo-computation, big earth data management, and spaceborne disaster mitigation.



Zhigang Wang received Bachelor of Engineering from Northwestern Polytechnical University (NPU), majoring in Electronic Engineering in Xi'an, China, 1987 and got Master of Science and Ph.D of Science in Peking University in 2002 and 2006 respectively, majoring in Cartography and Geographic Information System. He is currently the Executive Secretary with Space and Major Disasters International Charter on behalf of China National Space Administration (CNSA), Beijing, China, and engaged in several international organizations such as Earth Observation

Organization (GEO) and Commission on Earth Observation Satellites (CEOS). He has solid theoretical ground and rich experiences in remote sensing application, disaster management, GIS information system, and system engineering. He has worked with the China Resource Satellite Data and Application Center since 2006 and now is the Professor. Over the year's responsibility for international cooperation and communication work, he has participated in many international cooperation projects to promote Chinese satellite data and applications to international society such as CBERS-02B, 04 data reception and distribution in Africa, as well as CBERS-04 data sharing and servicing in ASEAN countries. He has authored or coauthored more than 20 papers.

Dr. Wang was the recipient of two international rewards.