Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images

Lei Ding, Dong Lin, Shaofu Lin, Jing Zhang, Xiaojie Cui, Yuebin Wang, Member, IEEE, Hao Tang, and Lorenzo Bruzzone, Fellow, IEEE

Abstract—Long-range contextual information is crucial for the semantic segmentation of high-resolution (HR) remote sensing images (RSIs). However, image cropping operations, commonly used for training neural networks, limit the perception of long-range contexts in large RSIs. To overcome this limitation, we propose a wide-context network (WiCoNet) for the semantic segmentation of HR RSIs. Apart from extracting local features with a conventional convolutional neural network (CNN), the WiCoNet has an extra context branch to aggregate information from a larger area image. Moreover, we introduce a context transformer to embed contextual information from the context branch and selectively project it onto the local features. The context transformer extends the vision transformer, an emerging kind of neural networks, to model the dual-branch semantic correlations. It overcomes the locality limitation of CNNs and enables the WiCoNet to see the bigger picture before segmenting the land-cover/land-use (LCLU) classes. Ablation studies and comparative experiments conducted on several benchmark datasets demonstrate the effectiveness of the proposed method. In addition, we present a new Beijing Land-Use (BLU) dataset. This is a large-scale HR satellite dataset with high-quality and fine-grained reference labels, which can facilitate future studies in this field.

Index Terms—Convolutional neural network, remote sensing, semantic segmentation, vision transformer (ViT).

I. INTRODUCTION

SEMANTIC segmentation of remote sensing images (RSIs) refers to their pixelwise labeling according to the ground information of interest (e.g., land-cover/land-use (LCLU) types). This is important for a variety of practical applications, such as environmental assessment, crop monitoring, natural resources management, and digital mapping. Recently, with the development of Earth observation technology and the emergence of convolutional neural networks (CNNs), it has been possible to perform automatic semantic segmentation of RSIs on easily accessible high-resolution (HR) RSIs.

Recent CNN models for visual recognition tasks are mostly based on stacked convolutional filters. A single convolution operation can extract/strengthen a certain feature, while stacked convolutions can combine and transform a variety of features. With the inclusion of numerous convolutional layers, a deep CNN can learn high-level semantic representations of the observed objects in images [1]. Since the introduction of fully convolutional network (FCN) in [2], deep CNNs have been widely used for dense classification tasks (i.e., semantic segmentation).

However, one of the limitations of CNNs is the intrinsic locality of convolution operations. The receptive field (RF) of a CNN unit is the region of input that is seen and responded to by the unit. Considering the sparse activation nature of CNNs, the valid receptive field (VRF) of a CNN unit is rather small [3]. This means that conventional CNNs model mostly the local image patterns (e.g., color and texture of objects) rather than considering the context information. Although numerous papers have proposed designs to enlarge the VRFs of CNNs [4], [5], they do not consider the long-range dependence between different image areas. The introduction of attention mechanism in CNNs [6]–[8] has allowed the network to learn biased focus under different image scenes. However, the semantic correlations between different image regions are not deeply modeled.

Recently, transformers are emerging [9] and gaining increasing research interest in the computer vision community [10], [11]. Differently from CNNs that rely on local operators to extract information, transformers employ stacked multihead attention blocks to model the global relationship between tokenized image patches. This enables them...
to exploit in-depth the long-range dependence that the data may exhibit. In recent studies, transformers are replacing CNNs in many visual recognition tasks [12]–[14]. However, training a vision transformer (ViT) requires large amount of training data to compensate its lack of inductive biases [10]. It is also more calculation-intensive compared to CNNs.

In this study, we aim to take advantage of both the CNN and transformer for the semantic segmentation of HR RSIs. The CNNs are good at preserving the spatial information, while transformer enables better modeling of the long-range dependencies. Moreover, instead of placing a plain transformer at the end of a CNN [15], we propose a dual branch context transformer to model the broader context in large RSIs. By allowing network to look at the bigger picture (i.e., seeing the wider context), it can understand better the local LCLU information. The main contributions in this study can be summarized as follows.

1) Proposing a wide-context network (WiCoNet) for the semantic segmentation of HR RSIs. The WiCoNet includes two CNNs that extract features from local and global image levels, respectively. This enables the WiCoNet to consider both local details and the wide context.

2) Proposing a context transformer to model the dual-branch semantic dependencies. The context transformer embeds the dual-branch CNN features into flattened tokens and learns contextual correlations through repetitive attention operations across the local and contextual tokens. Consequently, the projected local features are aware of the wide contextual information.

3) Presenting a benchmark dataset (i.e., the Beijing Land-Use (BLU) dataset) for the semantic segmentation of RSIs. This is a challenging HR satellite dataset annotated according to the land-use types. We believe that the release of this dataset can greatly facilitate future studies.

The remainder of this article is organized as follows. Section II introduces the literature work related to the semantic segmentation of RSIs. In Section III, we present the proposed WiCoNet. Section IV illustrates the designed experiments and introduces our BLU dataset. Section V reports the experimental results. Finally, we draw a conclusion of this study in Section VI.

II. RELATED WORK
A. Semantic Segmentation of Natural Images

In [2], deep CNNs have been first introduced for the semantic segmentation of images. CNN-based semantic segmentation can be used in many applications, such as saliency detection [16], medical segmentation [17], road scene understanding [18], and LC mapping [19]. CNN architectures for the semantic segmentation of images typically include an encoder network to aggregate the local information as well as a decoder network to retrieve the lost spatial details [17], [20]. Many network modules have been proposed to enhance the exploitation of local information, including the deformable convolution [21] and the dilated convolution [5] to enlarge the convolutional kernels and the pyramid pooling module to model multiscale context information [4]. Meanwhile, many literature works presented sophisticated CNN architectures to enhance the extraction of features, such as the multi-branch feature encoding designs in the HRNet [22] and the RefineNet [23]. In [24], the ExFuse is proposed, which is a network that includes cross-level information exchanging and multiscale feature fusion designs.

In recent years, the self-attention mechanism has been introduced to visual tasks in the squeeze-and-excitation networks (SENs) [6]. An SE block aggregates and embeds global information into features to learn biased focus in different image scenes, which is often referred to as channel attention in later literature. In [7], the channel attention is extended also to the spatial dimension to learn the position of focus. In [18] the DANet, which combines channel attention and nonlocal attention [25] in a parallel manner, has been presented. In the OCRNet [26], the relation between each pixel and its surrounding object regions is calculated to augment the contextual representations.

B. Semantic Segmentation of RSIs

Semantic segmentation of RSIs refers to the dense classification of either multiple LCLU classes or single interested class in RSIs (e.g., road [27], building [28], and water body [29]). Spatial accuracy is often crucial to remote sensing applications, which is a requirement for the semantic segmentation of RSIs. To improve the spatial localization accuracy, many literature works introduce U-shape networks with symmetric encoder–decoder structures. The TreeUNet [30] employs a DeepUNet to extract multiscale features and adaptively construct a tree-like CNN module to fuse the features. The ResUNet [31] employs the UNet with residual convolutional blocks as the segmentation backbone and combines atrous convolution and pyramid scene parsing (PSP) pooling to aggregate the context information. The MP-ResNet [32] includes three parallel feature embedding branches to model the context information at different scales, each of which includes a full ResNet34 (some of the residual blocks are shared). Other papers resort to strengthen the extraction of edge information. In [33] and [34], the ground-truth (GT) boundaries of objects are provided as a supervision to guide the network to learn edge features. In [35], the multilayer perceptron (MLP) is employed to rectify the uncertain areas in CNN predictions, which improves the preservation of object boundaries.

Another research focus is to model the geometric features of ground objects. In [36], direction supervision is introduced for the segmentation of roads. It strengthens the detection of linear features, and thus, the occluded and low-contrast roads are more salient to the models. In [28], the shape of object contours is modeled for the segmentation of buildings. The building contours are in-painted and sharpened through the adversarial learning of their shape information.

Recently, the attention mechanism has been widely used to augment the CNN-extracted features for the semantic segmentation of RSIs. In [37], the SE design is extended to the spatial dimension to represent the patchwise semantic focus, which
bridges the semantic gap between high- and low-level features. In [38], local and nonlocal attention designs are integrated in different branches of the HRNet [22] so that the local focus and long-range dependencies are captured, respectively. In [39], the channel attention and nonlocal attention blocks are sequentially used to augment the long-range dependencies in aerial RSIs.

C. Transformers in Vision Tasks

Transformer was first introduced for natural language processing tasks [9] where it achieved the state-of-the-art performance [40]. Recently, the use of transformer for computer vision tasks has drawn great research interests. In [10], the ViT is introduced for image classification, which shows that a pure transformer can replace CNN for image recognition tasks. In [41], transformer is first used for object detection. The resulting detection transformer (DETR) passes CNN features to a transformer, where the object class and locations are automatically generated with the encoded positional queries.

There are also literature works that use transformers for dense classification tasks. In [11], a dual-path transformer is proposed for panoptic segmentation, which includes a pixel path for segmentation and a memory path for class prediction. The transformer is used for information communication between the two paths. In [42], a two-branch architecture is proposed for the segmentation of medical images, which employs jointly a CNN and a transformer to extract features. In the swin transformer [13], cascaded transformers are constructed in an architecture similar to the ResNet. The spatial sizes of embedded patches are gradually increased to enlarge the RF.

In several recent papers, transformers have been introduced for processing RSIs. In [43], the ViT shows advantages over CNNs for scene classification in RSIs. In [44], a bitemporal transformer is introduced for the change detection of RSIs. The bitemporal semantic features are tokenized and concatenated, transformer is introduced for the change detection of RSIs. The CNNs for scene classification in RSIs. In [44], a bitemporal for processing RSIs. In [43], the ViT shows advantages over

III. PROPOSED WiCoNet

In this section, we illustrate the motivation for modeling a wide context in RSIs, followed by the architecture of the proposed network. Then, we describe the designed context transformer for communication of information between the two feature extraction branches. Finally, we report the implementation details.

A. Motivation of the Wide-Context Modeling

VRFs are known to be crucial for visual recognition tasks since they determine the maximum range of area where neural networks can gather information. In [38] and [39], the nonlocal attention blocks are introduced for the semantic segmentation of RSIs, which expands the VRFs of the networks into the whole input image. However, during training of neural networks, the input RSIs are often spatially cropped to avoid the overload of computational resources (and also to mix the samples in different image regions). Let us denote \( I \in \mathbb{R}^{c \times h \times w} \) as an RSI that consists of \( c \) spectral bands and has the spatial size of \( h \times w \). To train a standard CNN model \( M \), \( I \) is usually cropped into \( I_l \in \mathbb{R}^{c \times h_l \times w_l} \), where \( h_l \) and \( w_l \) are height and width of the cropping window, respectively. This limits the maximum possible RF of \( M \) to be \( h_l \times w_l \). Moreover, due to the locality that is inherent to CNNs [10], their VRFs are usually much smaller than \( h_l \times w_l \) [45]. Therefore, the long-range context information is insufficiently exploited in \( M \).

This issue is crucial in many LCLU mapping applications. The LCLU mapping is a complex task that requires high-level abstraction of regional information, where the context information limited in \( h_l \times w_l \) is often insufficient for recognizing some crucial samples. Moreover, for many objects that are spatially large (e.g., industrial buildings) or elongated (e.g., roads and rivers), the geometric features and semantic correlations cannot be well presented in local windows. To conquer these limitations, the context information should be modeled in a wider image range, which is the motivation of this study.

B. Network Architecture

We propose a WiCoNet that exploits the long-range dependencies in a larger image range in RSIs. As shown in Fig. 1, the proposed WiCoNet consists of two encoding branches. The local branch \( M_1 \), which is the main branch of the WiCoNet, employs the ResNet to extract local features. The novel design in the WiCoNet is a context branch \( M_2 \), which is introduced to explicitly model the wide-range context information in RSIs. It employs a simple CNN encoder to learn coarsely the context information (instead of gathering the spatial details). The context information is learned through \( M_2 \) and embedded into \( M_1 \) through a context transformer \( T \). The final results of the WiCoNet are then produced by the context-enriched \( M_1 \).

Formally, the training of a standard CNN model is performed on \( I_l \)

\[
P = M(I_l)
\]  

(1)

where \( P \in \mathbb{R}^{a \times h_l \times w_l} \) is the segmentation map (\( a \) is the number of classes). Differently, the WiCoNet is trained with both \( I_l \) and \( I_c \). \( I_c \in \mathbb{R}^{c \times h_l \times w_l} \) is a downsampled copy of \( I \) to provide an overview of the surrounding environment. \( I_f \) is associated with the central area of \( I_c \). Two segmentation maps \( P_f \in \mathbb{R}^{a \times h_l \times w_l} \) and \( P_c \in \mathbb{R}^{a \times h_l \times w_l} \) are produced during the training phase

\[
P_f, P_c = T[M_1(I_l), M_2(I_c)].
\]  

(2)

The training is driven by the total multiclass cross-entropy (MCE) losses of the two branches, which is calculated as

\[
\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{MCE}}(P_f, L_f) + \alpha \mathcal{L}_{\text{MCE}}(P_c, L_c)
\]  

(3)

where \( \alpha \) is a weighting parameter and \( L_f \) and \( L_c \) are the GT maps in the local and context branches, respectively.

Since the information extracted from \( M_2 \) is already modeled through \( T \), no further feature fusion operations are performed. During the testing phase, \( P_f \) is taken directly as the segmentation result.
C. Context Transformer

We introduce a context transformer to project long-range contextual information onto the local features, which is developed on top of the ViTs. A typical ViT takes flattened and projected image patches as inputs. It consists of multiple layers of attention blocks, each of which has a multihead self-attention (MSA) unit and an MLP unit [9]. Normalization and residual connections are enabled in each unit. The long-range semantic information, correlations are learned through the stacked attention blocks.

Fig. 1. Proposed WiCoNet.

In particular, for each position in the local feature, the responses from all the context windows are calculated and projected. Let \( t_l \in \mathbb{R}^{N \times D} \) and \( t_c \in \mathbb{R}^{M \times D} \) (\( M \) is the number of flattened features in \( M_2 \)) denote the local and context tokens embedded from \( M_1 \) and \( M_2 \), respectively. In \( T \), a local query \( q \) is projected with \( t_l \), while the context key \( k_c \) and value \( v_c \) are projected with \( t_c \).

\[
q_l = t_l W_q \in \mathbb{R}^{N \times D/n} \\
k_c = t_c W_k \in \mathbb{R}^{M \times D/n} \\
v_c = t_c W_v \in \mathbb{R}^{M \times D/n}
\]  
(6)

where \( W_q, W_k, W_v \in \mathbb{R}^{D \times D/n} \) are the corresponding weights of the projection function.

The context attention \( A_c \in \mathbb{R}^{N \times M} \) is then calculated to update \( t_l \):

\[
\hat{t}_l = A_c v_c = \text{softmax} \left( \frac{q_l k_c^T}{\sqrt{D/n}} \right) v_c.
\]
(7)

These operations, together with the MLP calculations, are repeated for \( L \) times, where the contextual dependencies between \( t_l \) and \( t_c \) are modeled and enforced. Consequently, the local tokens are projected with long-range dependencies from the context tokens. Finally, the local and context tokens are reshaped into the 2-D features.

D. Implementation Details

Here, we report detailed information of the proposed WiCoNet.

1) Feature Extraction Networks: We chose the ResNet50 as the feature extraction network in \( M_1 \), which is powerful in exploiting the local features [37]. The downsampling stride of the ResNet is \( 1 \times 1/8 \) to better preserve the spatial information. In the context branch, we employ a simple convolutional block
(referred to as the context encoder) to extract context features. It consists of 11 sequentially connected layers, including eight convolutional layers and three max-pooling layers. Each pooling layer is placed after two convolutional layers following the encoder design of UNet [17].

2) Area of the Context Modeling: The downsampling scale for input to $M_2$ is $1/4$, while the downsampling stride of the context encoder is the same as the ResNet ($1/8$). The size of context window is set to 9 times the size of local window ($w = 3 w_l$ and $h = 3 h_l$). An analysis of the accuracy versus context modeling range is provided in Section V-A. In this study, the size of the local window is $256 \times 256$. In cases where the local window is at the border of RSIs, empty areas in the context window are padded with reflections of the image.

3) Context Transformer: The hyperparameters in the context transformer include: $L$ is the number of transformer blocks, $n$ is the number of heads, $p$ is the size of the embedded patches, and $D$ is the dimension of the embedded tokens. $p$ is set to 1 to retain the spatial information. $D$ is set to 512, which is the number of output channels of the context encoder. $L$ and $n$ are set according to the experimental results, which are discussed in Section V-A. In addition, there is a weighting parameter $\alpha$. It is dynamically calculated at each iteration as: $(1 - iteration/all_iterations)^2$. In this way, its value declines over iterations and the WiCoNet gradually focuses on the local branch.

To find more details of the WiCoNet, readers are encouraged to visit the released codes at https://github.com/ggsDing/WiCoNet.

### IV. EXPERIMENTAL DATASETS AND SETTINGs

In this section, the experimental datasets and settings are reported. First, the experimented datasets are introduced, including the novel Beijing LU dataset and two open datasets. Then, the experimental settings and evaluation metrics are reported.

#### A. BLU Dataset

Currently, there are few HR satellite benchmark datasets available for the multiclass semantic segmentation of RSIs. To facilitate future research, we present a new benchmark dataset named BLU dataset. This dataset was collected in June 2018 in Beijing by the Beijing-2 satellite provided by the 21st Century Aerospace Technology Company Ltd. The collected data are RGB optical images and have a ground sampling distance (GSD) of 0.8 m. We constructed fine-grained human annotations on the collected images based on six LU classes: background/barren, built-up, vegetation, water, agricultural land, and road. These are the most interesting and frequently investigated land-use classes in both research studies and real-world applications (e.g., environment monitoring, traffic analysis, and urban and rural management). The detailed statistics of the class distributions are shown in Table I.

Compared to the existing datasets, the BLU dataset shows several remarkable features: i) high spatial resolution—as a satellite dataset, it has a high GSD of 0.8 m and ii) high annotation accuracy—the annotations were performed by an experienced annotation team dedicated to the RS applications. Fig. 2 shows some sample image patches selected from this dataset. One can observe that the LU classes in this dataset are easy to be discriminated due to the high GSD of RSIs. Moreover, the annotations are up to the pixel-level and the ground objects have been precisely annotated and geometrically optimized (to ensure both local consistency and topological correctness). Meanwhile, the observed areas include a variety of scenes, including farmland, residential areas, highways, airport, wet land, and others. This ensures that each LU class contains diverse samples. For example, the “built-up” class includes residential buildings, industrial buildings, and villages; the “water” class includes rivers, ponds, wet lands, and so on. These features present challenges to the generalization capability of segmentation algorithms.

Fig. 3 presents an overview of the BLU dataset. The observed regions include both urban and rural scenes, covering around 150 km² of area in total. The dataset consists of four tiles of large RSIs collected in four suburban regions.
in Beijing, each one with a pixel size of $15,680 \times 15,680$. Each large image is further cropped into 64 images (49 for training, 7 for validation, and 8 for testing), each of which has $2048 \times 2048$ pixels (Fig. 4). The training, validation, and testing areas are nonoverlapping, whereas the cropping windows within each area have small overlaps. The total numbers of images for training, validation, and testing are 196, 28, and 32, respectively. Both the original tiles and the divided subsets are provided. The BLU dataset will be released openly accessible to researchers.\footnote{https://rslab.disi.unitn.it/dataset/BLU/}

| Dataset | Method       | Components | Transformer | OA(%) | mean F1(%) | mIoU(%) |
|---------|--------------|------------|-------------|-------|------------|---------|
| BLU     | FCN [2]      | ✓          | ✓           | ✓     | 86.51      | 70.09   |
|         | FCN+Transformer WiCoNet (Ours) | ✓ | ✓ | ✓ | 86.74      | 70.92   |
|         | WiCoNet (Ours) | ✓ | ✓ | ✓ | 87.35      | 71.50   |
| GID     | FCN [2]      | ✓          | ✓           | ✓     | 74.71      | 49.02   |
|         | FCN+Transformer WiCoNet (Ours) | ✓ | ✓ | ✓ | 75.82      | 51.36   |
|         | WiCoNet (Ours) | ✓ | ✓ | ✓ | 77.14      | 53.07   |
| Potsdam | FCN [2]      | ✓          | ✓           | ✓     | 88.96      | 83.24   |
|         | FCN+Transformer WiCoNet (Ours) | ✓ | ✓ | ✓ | 88.69      | 82.66   |
|         | WiCoNet (Ours) | ✓ | ✓ | ✓ | 90.24      | 84.93   |

| Dataset | Metrics | Size of context windows |
|---------|---------|-------------------------|
|         |         | 512×512  768×768  1024×1024 |
| BLU     | OA      | 86.91      87.35      87.20   |
|         | mean F1 | 82.11      82.77      82.35   |
|         | mIoU    | 70.41      70.58      70.81   |
| GID     | OA      | 77.06      77.14      77.28   |
|         | mean F1 | 66.03      66.26      66.55   |
|         | mIoU    | 53.04      53.07      53.38   |
| Potsdam | OA      | 90.16      90.24      90.34   |
|         | mean F1 | 91.59      91.71      91.76   |
|         | mIoU    | 84.72      84.93      85.03   |
B. Standard Benchmark Datasets

To make a comprehensive analysis on the performance of the proposed WiCoNet, we conducted experiments on two additional open benchmark datasets, i.e., the ISPRS Potsdam dataset and the Gaofen Image dataset (GID).

1) Potsdam Dataset: This is an area dataset collected in urban scenes. It consists of 38 tiles of very high resolution (VHR) RSIs, each having $6000 \times 6000$ pixels. The provided data include true ortho photographs containing four spectral bands (RGB and infrared) and the registered digital surface model (DSM) data. The labels are annotated with six LC categories: impervious surfaces, building, low vegetation, tree, car, and clutter/background. We use 18 tiles of images for training, 6 for validation, and the remaining 14 ones for testing. The division of training and validation tiles follows the practice in [49].

2) Gaofen Image Dataset: This is an HR LC classification dataset collected by the Gaofen-2 (GF-2) satellite. It consists of ten tiles of RSIs with four spectral bands (RGB and near infrared). Each tile has $7200 \times 6800$ pixels, with a GSD of 0.8 m. Since the division of training and testing sets is not provided, we further crop and divide the tiles into 90 training images, 30 validation images, and 40 testing images (each one with $2048 \times 2048$ pixels; 16 LC classes are annotated, including industrial land (IDL), urban residential (UR), rural residential (RR), traffic land (TL), paddy field (PF), irrigated

### Table V

| Method     | Background | Built-up | Vegetation | Water | Agricultural | Road |
|------------|------------|----------|------------|-------|--------------|------|
| FCN [2]    | 72.92      | 87.56    | 90.41      | 85.15 | 86.42        | 68.88|
| PSPNet [49]| 72.66      | 87.40    | 90.41      | 86.30 | 86.71        | 68.84|
| DeepLabv3+ [15] | 73.99       | 87.93    | 90.76      | 86.46 | 87.32        | 68.55|
| DANet [18]  | 73.06      | 87.73    | 90.55      | 85.45 | 86.77        | 69.07|
| SCAttNet [48]| 73.21      | 87.62    | 90.54      | 86.26 | 86.87        | 69.32|
| MSCA [38]   | 73.71      | 88.34    | 90.74      | 85.92 | 86.86        | 70.31|
| LANet [37]  | 73.81      | 87.48    | 90.60      | 85.99 | 87.02        | 68.49|
| WiCoNet (Ours) | 74.43      | 88.55    | 90.94      | 86.01 | 87.23        | 70.21|

**OA (%)**, mean $F_1$ (%) and mIoU (%)

| Method     | Background | Built-up | Vegetation | Water | Agricultural | Road |
|------------|------------|----------|------------|-------|--------------|------|
| FCN [2]    | 86.51      | 80.67    | 85.18      | 84.91 | 88.30        | 70.07|
| PSPNet [49]| 86.59      | 82.05    | 85.18      | 86.71 | 86.84        | 70.84|
| DeepLabv3+ [15] | 87.08      | 82.55    | 85.18      | 86.71 | 86.84        | 71.07|
| DANet [18]  | 86.76      | 82.10    | 85.18      | 86.71 | 86.84        | 71.07|
| SCAttNet [48]| 86.77      | 82.10    | 85.18      | 86.71 | 86.84        | 71.07|
| MSCA [38]   | 87.17      | 82.64    | 85.18      | 86.86 | 87.02        | 70.21|
| LANet [37]  | 86.89      | 82.28    | 85.18      | 86.86 | 87.02        | 70.21|
| WiCoNet (Ours) | 87.35      | 82.89    | 85.18      | 87.23 | 87.23        | 71.50|

### Table VI

| Method     | Background | Built-up | Vegetation | Water | Agricultural | Road |
|------------|------------|----------|------------|-------|--------------|------|
| FCN [2]    | 72.92      | 87.56    | 90.41      | 85.15 | 86.42        | 68.88|
| PSPNet [49]| 72.66      | 87.40    | 90.41      | 86.30 | 86.71        | 68.84|
| DeepLabv3+ [15] | 73.99       | 87.93    | 90.76      | 86.46 | 87.32        | 68.55|
| DANet [18]  | 73.06      | 87.73    | 90.55      | 85.45 | 86.77        | 69.07|
| SCAttNet [48]| 73.21      | 87.62    | 90.54      | 86.26 | 86.87        | 69.32|
| MSCA [38]   | 73.71      | 88.34    | 90.74      | 85.92 | 86.86        | 70.31|
| LANet [37]  | 73.81      | 87.48    | 90.60      | 85.99 | 87.02        | 68.49|
| WiCoNet (Ours) | 74.43      | 88.55    | 90.94      | 86.01 | 87.23        | 70.21|

**OA (%)**, mean $F_1$ (%) and mIoU (%)

### Table VII

| Method     | Background | Built-up | Vegetation | Tree | Car |
|------------|------------|----------|------------|------|-----|
| FCN [2]    | 91.08      | 95.21    | 86.17      | 86.51 | 94.63|
| PSPNet [49]| 88.85      | 93.20    | 83.89      | 82.69 | 91.62|
| DeepLabv3+ [15] | 91.79      | 96.46    | 86.17      | 86.39 | 94.34|
| DANet [18]  | 91.94      | 96.05    | 86.74      | 87.11 | 94.42|
| SCAttNet [48]| 91.66      | 95.57    | 86.44      | 86.79 | 94.13|
| MSCA [38]   | 92.31      | 96.74    | 86.59      | 87.01 | 95.11|
| LANet [37]  | 91.63      | 95.83    | 85.96      | 86.35 | 93.98|
| WiCoNet (Ours) | 92.50      | 96.53    | 87.03      | 87.31 | 95.13|

**OA (%)**, mean $F_1$ (%) and mIoU (%)

### Table VIII

| Methods      | FCN        | PSPNet    | DeepLabv3+ | DANet | SCAttNet | MSCA    | LANet    | WiCoNet (proposed) |
|--------------|------------|-----------|------------|-------|----------|---------|---------|-------------------|
| Params (Mb)  | 23.78      | 44.37     | 39.47      | 48.22 | 24.62    | 66.06   | 23.79   | 38.24             |
| FLOPs (Gbps) | 25.27      | 46.58     | 41.10      | 50.29 | 26.09    | 21.78   | 8.28    | 41.74             |
Fig. 5. Qualitative results of the ablation study on the BLU datasets. The saliency maps of the critical classes are presented. The selected challenging scenes include: (a) occluded road, (b) green algae-covered river, (c) streets in a residential area, and (d) farmland surrounded by vegetation.

land (IL), dry cropland (DC), garden plot (GP), arbor woodland (AW), shrub land (SL), natural grassland (NG), artificial grassland (AG), river (RV), lake (LK), and pond (PN).

C. Experimental Settings

The proposed WiCoNet and the compared methods are implemented with PyTorch. The hardware environment of this study is a server equipped with a GTX3090 GPU. For each dataset, we fix the training epochs to 50, the batch size to 32, and the initial learning rate to 0.1. The learning rate $\eta$ is dynamically calculated at each iteration as: $0.1 \times (1 - \text{iterations/total\_iterations})^{0.5}$. The optimization algorithm is the stochastic gradient descent with the momentum of 0.9. Random flipping and random cropping operations are adopted to augment the data. They are performed at each iteration of the training process. At the end of training, the model file with the best overall accuracy (OA) (evaluated on the validation set) is saved.

In this study, we adopt the most frequently used metrics [34], [39] to evaluate the tested methods, including: 1) OA, which is the numeric ratio of correctly classified pixels versus all the pixels in RSi; 2) $F_1$ score of each class, which is the harmonic mean of the precision and recall; and 3) mean intersection over union (mIoU). The metrics can be calculated with the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) pixels as follows:

$$
\text{OA} = \frac{TP + TN}{TP + TN + FP + FN}
$$

$$
\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}
$$

$$
F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
$$

$$
\text{IoU} = \frac{TP}{TP + FP + FN}.
$$

V. EXPERIMENTAL RESULTS

This section reports the results of the conducted experiments. First, an ablation study is developed to verify the accuracy improvements. Then, the effect of context modeling range is analyzed. Finally, the proposed WiCoNet is compared with several CNN models with context-aggregation designs in recent studies.

A. Ablation Study

1) Choice of Hyperparameters: As introduced in Section III-D, $L$ and $n$ are two adjustable hyperparameters in the context transformer. First, we conduct a group of experiments to set their values. The initial values of $L$ and $n$ are set to 2 and 4, respectively. We change the values of $L$ and $n$ by sequence and report the OA obtained by the WiCoNet in Table II. One can observe that the best OA on the BLU and GID datasets is obtained when $L = 4$ and
Fig. 6. Qualitative results of the ablation study on the additional datasets. The saliency maps of the critical classes are presented. (a)–(c) Results selected from the GID. (d)–(f) Results selected from the Potsdam dataset.

$n = 4$. Meanwhile, the optimal hyperparameter values for the Potsdam dataset are $L = 2$ and $n = 4$. The OA is lower when $L$ is set to 8. We assume that this is caused by overfitting since the long-range context information in RSIs is relatively simple, and thus, too many transformer layers may be redundant. The tested optimal parameters for different datasets are fixed in the following experiments.

2) Quantitative Results: An ablation study is conducted to test the effectiveness of context modeling. The novel designs in the WiCoNet include an extra context branch and the context transformer. First, we compare the results of the proposed WiCoNet and the FCN [2]. To exclude the improvements brought by the transformer, we also constructed a variant of the FCN where a transformer is placed at the end of its encoder, denoted as FCN + Transformer. The experimental results are reported in Table III.

Compared to FCN, the improvements brought by adding the transformer as an encoder head (FCN + Transformer) are limited. This can be attributed to the limited long-range context information in local patches. However, after performing the wide context modeling with the WiCoNet, significant improvements are obtained. The improvements over the baseline FCN are 0.84%, 3.57%, and 1.38% in OA and 1.41%, 4.05, and 1.69% in mIoU, respectively, on the BLU, GID, and...
Potsdam datasets. These results show that the wide context modeling in the WiCoNet stably improves the LCLU segmentation accuracy of HR RSIs.

3) Qualitative Results: To qualitatively assess the effects of context modeling, Figs. 5 and 6 show the comparisons of the results in some sample areas on the BLU and the additional datasets, respectively. In the sample images, both the context window and the local window of the WiCoNet are presented. The salience maps of the FCN and the WiCoNet are also shown to highlight their perception of the critical classes. One can observe that there are many fragmentation errors and inconsistency in the segmentation results of the FCN. In many cases, learning only the local bias is not sufficient to overcome these shortcomings, as shown in the results of the FCN + Transformer.

The proposed WiCoNet shows advantages in the following.

a) Discriminating the Critical Areas: By modeling contextual dependencies on similar samples in the context window, the discrimination of certain critical or minority classes in the local window is improved [e.g., Figs. 5(b) and 6(b) and (f)].

b) Improving the Connectivity of Segmented Objects: The spatial layout of certain objects is clearer in a wider image context [e.g., the road in Fig. 5(a) and the rivers in Figs. 5(c) and 6(a)]. The WiCoNet better preserves their long-range consistency.

c) Reducing Fragmentation Errors: By looking into the context window, the WiCoNet understands better the local scenes, thus eliminating some false predictions [e.g., the lake in Fig. 6(c) and an empty field in Fig. 6(e)].

d) Effects of the Context Modeling Range: The size of the context window \((w \times h)\) determines up to which range the context information is modeled, which is critical for the WiCoNet. To allow enough coverage of the surrounding regions, the size of the context window should be several times bigger than the size of the local window \((w_l \times h_l)\). Meanwhile, since transformer is based on self-attention mechanism, too large context modeling range may cause loss of focus on the local content. To find the best context modeling range, we further conduct experiments by varying the size of context windows.

The results are reported in Table IV. The tested context windows have \(\times 4\), \(\times 9\), and \(\times 16\) times the area of local windows (i.e., \(w \times h = 2 w_l \times 2 h_l\), \(w \times h = 3 w_l \times 3 h_l\), and \(w \times h = 4 w_l \times 4 h_l\)). One can observe that the \(\times 16\) context window results in the best accuracy on the GID and the Potsdam dataset, whereas the \(\times 9\) context window leads to better accuracy on the BLU dataset. The relationship between OA and the size of context window is presented in Fig. 7. Overall, the increase in OA from \(\times 4\) to \(\times 9\) windows is noticeable, whereas that from \(\times 9\) to \(\times 16\) windows is not significant.

B. Comparative Study

We further compare the proposed WiCoNet with several recent works on context-aggregation designs. The compared models include the baseline FCN, the DeepLabv3+ [5] with dilated convolutions, the PSPNet [4] with the PSP module, the DANet [18] with channel attention and nonlocal attention, the SCAttNet [47] with spatial and channel attention, the MSCA [38] with multiscale context aggregation designs, and the LANet [37] with local attention.

We implement all the tested methods with the experimental settings described in Section IV-C and report the results in Tables V–VII. The reported values are the average of the metrics derived in three trials. One can observe that DeepLabv3+, a well-known network in the computer vision community, shows stable improvements over FCN on the three datasets. The recent attention-based approaches (DANet, LANet, and SCAttNet) obtain good results on the BLU and Potsdam datasets. In particular, the LANet obtains the second best OA on the BLU dataset and the GID. The MSCA that integrates attention designs into the HRNet architecture achieves the second best results on the Potsdam dataset. By extending attention into wider image areas through transformers, the proposed WiCoNet obtains the best accuracy metrics (in both OA, mean \(F_1\) and mIoU) on the three datasets. Its improvements are particularly noticeable on the GID where context information is crucial to determine the LC classes.

The parameter size and computational cost of each model are reported in Table VIII. The number of floating-point operations per second (FLOOPS) is calculated based on the experimental settings for the BLU dataset (including input and output size and hyperparameters), except for the batch size that is set to 1 for clarity. The overall consumption of the WiCoNet is higher than that of the FCN, the SCAttNet, and the LANet, but it is lower than that of the PSPNet and the DANet. Its parameter size and FLOOPS are very close to those of the DeepLabv3+.

VI. CONCLUSION

While long-range context information is crucial for the semantic segmentation of VHR RSIs, most existing studies
only focus on modeling the local context information within cropped image patches. To overcome this limitation, we propose a WiCoNet. The WiCoNet employs an extra context branch to aggregate the context information in bigger image areas (i.e., context windows), which greatly broadens the possible RFs of the models. Moreover, instead of using simple feature fusion designs, we introduce a context transformer to communicate the information between its dual branches. The context information is calculated and projected into the local query tokens, which overcomes the locality limitations of CNNs.

To support this study and to facilitate future research, we also release a high-quality and large-scale benchmark dataset for the semantic segmentation on HR RSIs, i.e., the BLU dataset. Through experiments on the BLU dataset and two additional datasets, we verified the effectiveness of the long-range context modeling, analyzed the accuracy of different context modeling sizes, and compared the WiCoNet with several literature works that model context information in RSIs. Experimental results show that the WiCoNet enables a better understanding and modeling of both the local scene information and the global class distribution, thus bringing significant accuracy improvements. However, there is still global inconsistency and some local fragmentation errors remain, indicating that there is still margin to improve the modeling of long-range context information in large RSIs. This is left for future works, where adversarial learning strategies [28] can be employed to model the semantic correlations.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions, which have helped them improve the quality of this article.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[2] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[3] L. Ding, J. Zhang, and L. Bruzzone, “Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2881–2890.

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

[6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[7] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “CBAM: Convolutional block attention module,” in Proc. ECCV, 2018, pp. 3–19.

[8] H. Tang, S. Bai, and N. Sebe, “Dual attention GANs for semantic image synthesis,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1994–2002.

[9] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., Annua. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[10] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

[11] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, “MaX-DeepLab: End-to-end panoptic segmentation with mask transformers,” 2020, arXiv:2012.00759.

[12] S. Khan, M. Naseer, M. Hayat, S. Waqs Zamir, F. Shahbaz Khan, and M. Shah, “Transformers in vision: A survey,” 2021, arXiv:2101.01169.

[13] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted Windows,” 2021, arXiv:2103.14030.

[14] G. Yang et al., “Transformer-based source-free domain adaptation,” 2021, arXiv:2105.14138.

[15] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, arXiv:2102.04306.

[16] T. Wang et al., “Detect globally, refine locally: A novel approach to saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3127–3135.

[17] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Munich, Germany: Springer, 2015, pp. 234–241.

[18] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 299–307.

[19] X.-Y. Tong et al., “Land-cover classification with high-resolution remote sensing images using transferable deep models,” Remote Sens. Environ., vol. 237, Feb. 2020, Art. no. 111322.

[20] 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, Oct. 2016.

[21] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets V2: More deformable, better results,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9308–9316.

[22] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[23] 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. (CVPR), Jul. 2017, pp. 1925–1934.

[24] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, “ExFuse: Enhancing feature fusion for semantic segmentation,” in Proc. ECCV, 2018, pp. 269–284.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. CVPR, Jun. 2018, pp. 7794–7803.

[26] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” in Proc. 16th Eur. Conf. Comput. Vis. (Lecture Notes in Computer Science), vol. 12351. Glasgow, U.K.: Springer, 2020, pp. 173–190.

[27] L. Ding, Q. Yang, J. Lu, J. Xu, and J. Yu, “Road extraction based on direction consistency segmentation,” in Proc. Chin. Conf. Pattern Recognit. Munich, Germany: Springer, 2016, pp. 131–144.

[28] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone, “Adversarial shape learning for building extraction in VHR remote sensing images,” IEEE Trans. Image Process., vol. 31, pp. 678–690, 2022.

[29] L. Duan and X. Hu, “Multiscale refinement network for water-body segmentation in high-resolution satellite imagery,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 4, pp. 686–690, Apr. 2020.

[30] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, “TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation,” ISPRS J. Photogramm. Remote Sens., vol. 156, pp. 1–13, Oct. 2019.

[31] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, “ResUNet–A: A deep learning framework for semantic segmentation of remotely sensed data,” ISPRS J. Photogramm. Remote Sens., vol. 162, pp. 94–114, Apr. 2020.

[32] L. Ding et al., “MP-ResNet: Multipath residual network for the semantic segmentation of high-resolution PolSAR images,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[33] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, “ERN: Edge loss reinforced semantic segmentation network for remote sensing images,” Remote Sens., vol. 10, no. 9, p. 1339, Aug. 2018.
[34] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, “Classification with an edge: Improving semantic image segmentation with boundary detection,” *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[35] C. Zhang, I. Sargent, X. Pan, A. Gardiner, J. Hare, and P. M. Atkinson, “VPFR-based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4507–4521, Aug. 2018.

[36] L. Ding and L. Bruzzone, “DiResNet: Direction-aware residual network for road extraction in VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10243–10254, Dec. 2021.

[37] L. Ding, H. Tang, and L. Bruzzone, “LANet: Local attention embedding to improve the semantic segmentation of remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[38] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, “Multi-scale context aggregation for semantic segmentation of remote sensing images,” *Remote Sens.*, vol. 12, no. 4, p. 701, Feb. 2020.

[39] L. Mou, Y. Hua, and X. X. Zhu, “A relation-augmented fully convolutional network for semantic segmentation in aerial scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[40] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT)*, vol. 1, Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. ECCV*, Munich, Germany: Springer, 2020, pp. 213–229.

[42] Y. Zhang, H. Liu, and Q. Hu, “TransFuse: Fusing transformers and CNNs for medical image segmentation,” 2021, arXiv:2102.08005.

[43] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, “Vision transformers for remote sensing image classification,” *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.

[44] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Object detectors emerge in deep scene CNNs,” in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, MAY 2015.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jul. 2015.

[47] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, “SCANet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[48] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Semantic correlation promoted shape-variant context for segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8885–8894.

[49] M. Volpi and D. Tuia, “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2016.

Lei Ding received the M.S. degree in photogrammetry and remote sensing from Information Engineering University, Zhengzhou, China, in 2016, and the Ph.D. degree (cum laude) in communication and information technologies from the University of Trento, Trento, Italy, in 2021. He is currently a Lecturer with PLA Strategic Force Information Engineering University, Zhengzhou, China. His research interests are related to semantic segmentation, change detection, and domain adaptation with deep learning techniques.

Dr. Ding is a Referee for many international journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Lei Ding

Dong Lin received the M.S. degree in photogrammetry and remote sensing from Information Engineering University, Zhengzhou, China, in 2015, and the Ph.D. degree in photogrammetry and remote sensing from the Technische Universität Dresden, Dresden, Germany, in 2019. He is currently a Lecturer with Space Engineering University, Beijing, China. His research interests include deep learning, change detection, and thermal image processing.

Dong Lin

Shaofu Lin received the Ph.D. degree in mapping and geographic information system from the Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China, in 2002. He worked on the research, construction, and management of informatization and e-government at the Hainan Information Center, Beijing Information Resource Management Center, and Beijing Municipal Office of Informatization, Beijing, from 1990 to 2009. He was engaged in the promotion of smart city, e-government, and industrial technology innovation in the Beijing Economic and Information Commission from 2009 to 2014, successively working as the Director of the E-government Department and the Science and Technology Standards Department. He has been a Professor with the Software College, Beijing University of Technology, Beijing, since 2014, ever serving as the Director of the Information Department, the Executive Director of the Beijing Institute of Smart City, and the Executive Director of the Beijing Advanced Innovation Center for Future Internet Technology. His research interests include spatial-temporal big data, data fusion and intelligence, and blockchain.

Dr. Lin is a Senior Member of the China Computer Federation (CCF) and the Chinese Institute of Electronics (CIE) and a member of the Block-Chain Commission of CCF, the Expert Committee of China Big Data Industry Ecological Alliance, and the Network and Information Technology Expert Committee of China Artificial Intelligence Industry Alliance. He is a Board Member of the Beijing Institute of Big Data.

Shaofu Lin

Jing Zhang received the master’s degree in software engineering from the Beijing University of Technology, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. Her research interests are related to the change detection and semantic segmentation of remote sensing images.

Jing Zhang

Xiaojie Cui received the M.S. and Ph.D. degrees in cartography and geographic information system from Information Engineering University, Zhengzhou, China, in 2016 and 2019, respectively. She is currently an Engineer with the Beijing Institute of Remote Sensing Information, Beijing, China. Her research interests include remote sensing image processing and big data analysis.

Xiaojie Cui
Yuebin Wang (Member, IEEE) received the Ph.D. degree from the School of Geography, Beijing Normal University, Beijing, China, in 2016. He was a Post-Doctoral Researcher with the School of Mathematical Sciences, Beijing Normal University, Beijing. He is currently an Associate Professor with the School of Land Science and Technology, China University of Geosciences (Beijing), Beijing. His research interests include remote sensing imagery processing and 3-D urban modeling.

Hao Tang received the master’s degree from the School of Electronics and Computer Engineering, Peking University, Beijing, China, in 2016, and the Ph.D. degree from the Multimedia and Human Understanding Group, University of Trento, Trento, Italy, in 2021. He was a Visiting Scholar with the Department of Engineering Science, University of Oxford, Oxford, U.K. He currently holds a post-doctoral position with the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland. His research interests are deep learning, machine learning, and their applications to computer vision.

Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree (summa cum laude) in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively. He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. His research interests are in the areas of remote sensing, radar and synthetic aperture radar (SAR), signal processing, machine learning and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the principal investigator of many research projects. Among the others, he is the Principal Investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the Jupiter ICy moons Explorer (JUICE) mission of the European Space Agency. He is the author (or co-author) of 215 scientific publications in refereed international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is the editor/coeditor of 18 books/conference proceedings and one scientific book. He was invited as a keynote speaker at more than 30 international conferences and workshops.

Dr. Bruzzone has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) since 2009. He was a guest coeditor of many special issues of international journals. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the Founder of the IEEE Geoscience and Remote Sensing Magazine, for which he has been Editor-in-Chief from 2013 to 2017. He is also an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has been a Distinguished Speaker at the IEEE Geoscience and Remote Sensing Society from 2012 to 2016. His papers are highly cited, as proven by the total number of citations (more than 27000) and the value of the H-index (78) (source: Google Scholar).