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 CNN, the WiCoNet has an extra context branch to aggregate information from a larger image area. 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*—Remote Sensing, Semantic Segmentation, Vision Transformer, Convolutional Neural Network**

I. Introduction

Semantic segmentation of remote sensing images (RSIs) refers to their pixel-wise labelling 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 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, 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 dependency between different image areas. The introduction of attention mechanism in CNNs [6], [7], [8] has allowed the network to learn biased focus under different image scenes. However, the semantic correlations between different image regions are not deeply modelled.

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 multi-head attention blocks to model the global relationship between tokenized image patches. This enables them to exploit in-depth the long-range dependency that the data may exhibit. In recent studies transformers are replacing CNNs in many visual recognition tasks [12], [13], [14]. However, training a vision transformer 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

L. Ding is with the PLA Strategic Force Information Engineering University, ZhengZhou, China (E-mail: dinglei14@outlook.com).

D. Lin is with the Space Engineering University, No.7 Fuxue Road, Changping District, 102249 Beijing, China and also with the State Key Laboratory of Geo-Information Engineering, No.1 Yanta Road, Beilin District 710054, Xi'an, China. (E-mail: lindong_hb59@163.com).

S. Lin is with Beijing University of Technology, NO.100 Pingle Garden, Chaoyang District, 100022 Beijing, P.R. China. (E-mail: linshaofu@bjut.edu.cn).

J. Zhang and L. Bruzzone are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (E-mail: jing.zhang-1@unitn.it, lorenzo.bruzzone@unitn.it).

X. Cui is with the Beijing Institute of Remote Sensing Information, No.6 Waiguanxie Street, Chaoyang District, 100011 Beijing, China. (E-mail: cuixjgis@163.com).

Y. Wang is with the China University of Geosciences (Beijing), No.29 Xueyuan Road, Haidian District, 100084 Beijing, China. (E-mail: xxgcdxwyb@163.com).

H. Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland. (E-mail: hao.tang@vision.ee.ethz.ch).

This document is funded by State Key Laboratory of Geo-Information Engineering, No.SKLGIE2019-Z-3-3. It is also funded by the scholarship from China Scholarship Council under the grant NO.201703170123.

and transformer for the semantic segmentation of HR RSIs. The CNNs are good at preserving the spatial information, while transformer enables a better modelling 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 the release of this dataset can greatly facilitate future studies.

The remainder of this paper 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. Finally, we draw a conclusion of this study in Section V.

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 an 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 multi-scale 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 multi-scale feature fusion designs.

In recent years, the self-attention mechanism has been introduced to visual tasks in the Squeeze-and-Excitation Networks (SENet) [25]. An SE block aggregates and embed global information into features to learn biased focus in different image scenes, which is often referred 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 non-local attention [26] in a parallel manner, has been presented. In the OCRNet [27] 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 [28], building [29], and water body [30]). 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 [31] employs a DeepUNet to extract multi-scale features and adaptively construct a tree-like CNN module to fuse the features. The ResUNet [32] employs the UNet with residual convolutional blocks as the segmentation backbone and combines atrous convolution and pyramid scene parsing pooling to aggregate the context information. The MP-ResNet [33] 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 [34] and [35] the ground truth boundaries of objects are provided as a supervision to guide the network to learn edge features. In [36], the Multi-layer 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 [37], a direction supervision is introduced for the segmentation of roads. It strengthens the detection of linear features, thus the occluded and low-contrast roads are more salient to the models. In [29], the shape of object contours is modelled 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 [38], the SE design is extended to the spatial dimension to represent the patch-wise semantic focus, which bridges the semantic gap between high-level and low-level features. In [39], local and non-local attention designs are integrated in different branches of the HRNet [22], so that the local focus and long-range dependencies are captured, respectively. In [40], the channel attention and non-local attention blocks are sequentially used to augment the long-range dependencies in aerial RSIs.

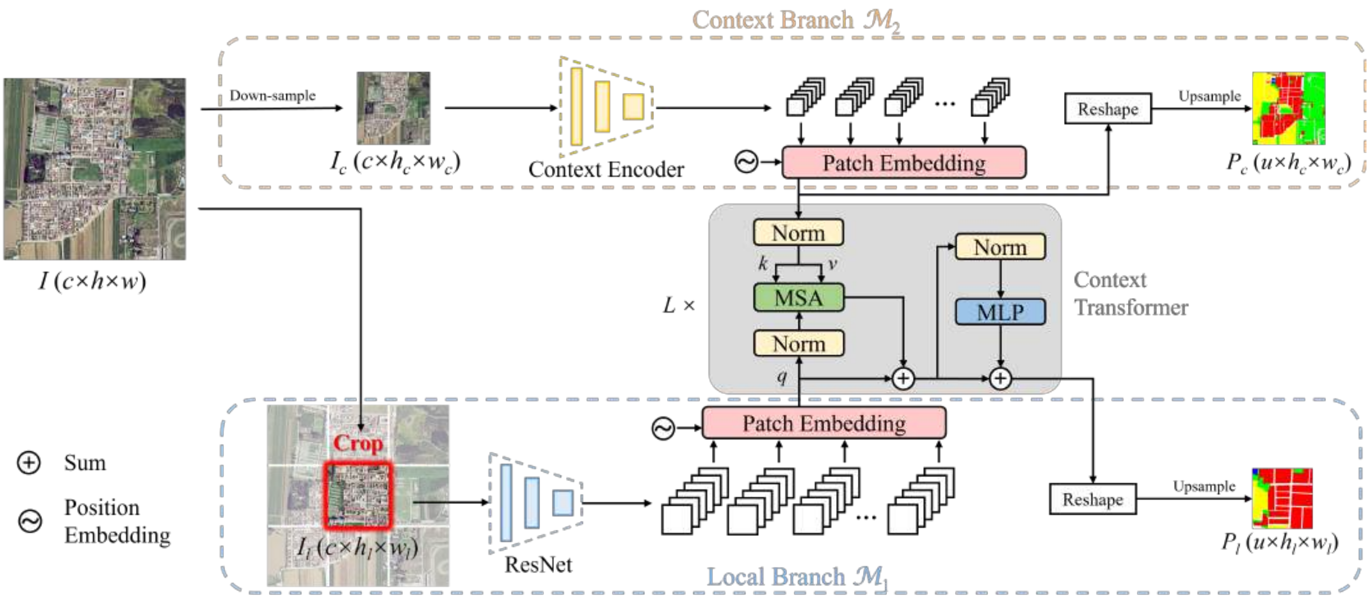

Fig. 1: The proposed **Wi**de-**Co**ntext **Net**work (WiCoNet).

C. *Transformers in Vision Tasks*

Transformer was first introduced for natural language processing tasks [9] where it achieved the state-of-the-art performance [41]. Recently the use of transformer for computer vision tasks has drawn great research interests. In [10], the Vision Transformer (ViT) is introduced for image classification, which shows that a pure transformer can replace CNN for image recognition tasks. In [42], 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 [43], 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 [44] the vision transformer shows advantages over CNNs for scene classification in RSIs. In [45] a bi-temporal transformer is introduced for the change detection of RSIs. The bi-temporal semantic features are tokenized and concatenated, followed by the transformer to enrich the global semantic correlations.

III. Proposed Wide-Context Network

In this section, we illustrate the motivation for modelling 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 Modelling*

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 [40] and [39], the non-local attention blocks are introduced for the semantic segmentation of RSIs, which expand 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 a RSI that consists of c spectral bands and has the spatial size of $h \times w$. To train a standard CNN model $\mathcal{M}$, I is usually cropped into $I_l \in \mathbb{R}^{c\times h_l\times w_l}$ where h_l, w_l are height and width of the cropping window, respectively. This limits the maximum possible RF of $\mathcal{M}$ 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 $h_l \times w_l$ [46]. Therefore, the long-range context information is insufficiently exploited in $\mathcal{M}$.

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 modelled in a wider image range, which is the motivation of this study.

B. Network Architecture

We propose a Wide-Context Network (WiCoNet) that exploits the long-range dependencies in a larger image range in RSIs. As illustrated in Fig. 1, the proposed WiCoNet consists of two encoding branches. The local branch $\mathcal{M}_1$, 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 $\mathcal{M}_2$, which is introduced to explicitly model the wider-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 $\mathcal{M}_2$ and embedded into $\mathcal{M}_1$ through a Context Transformer $\mathcal{T}$. The final results of the WiCoNet is then produced by the context-enriched $\mathcal{M}_1$.

Formally, the training of a standard CNN model is performed on I_l:

$$P = \mathcal{M}(I_l), \tag{1}$$

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

$$P_l, P_c = \mathcal{T}[\mathcal{M}_1(I_l), \mathcal{M}_2(I_c)], \tag{2}$$

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

$$\mathcal{L}_{Seg} = \mathcal{L}_{\text{MCE}}(P_l, L_l) + \alpha \mathcal{L}_{\text{MCE}}(P_c, L_c), \tag{3}$$

where α is a weighting parameter, L_l and L_c are the ground truth (GT) maps in the local and context branches, respectively.

Since the information extracted from $\mathcal{M}_2$ is already modelled through $\mathcal{T}$, no further feature fusion operations are performed. During the testing phase, P_l 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 Vision Transformers. A typical Vision Transformer takes flattened and projected image patches as inputs. It consists of multiple layers of attention blocks, each of which has a Multi-head Self-Attention (MSA) unit and an MLP unit [9]. Normalization and residual connections are enabled in each unit. The long-range semantic correlations are learned through the stacked attention blocks. Let us consider an input 3D signal $\mathbf{x} \in \mathbb{R}^{\hat{c} \times h \times w}$ where $\hat{c}$ is the number of channels. $\mathbf{x}$ is first reshaped into a flattened 2D patch $\mathbf{x}_p \in \mathbb{R}^{N \times \hat{c}p^2}$, where $N = hw/p^2$, (p, p) is the spatial size of each flattened patch. Then, $\mathbf{x}_p$ is projected into a token vector $\mathbf{t} \in \mathbb{R}^{N \times D}$ where D is the constant latent vector size in all the layers of the Transformer. This operation that maps $\mathbf{x}$ into $\mathbf{t}$ is named *Patch Embedding*. To retain the position information, $\mathbf{t}$ is further added with trainable parameters before it is forwarded into the transformer. The operations inside a transformer block can be represented as follows:

$$\begin{aligned} \hat{\mathbf{t}} &= \text{MSA}(\text{LN}(\mathbf{t})) + \mathbf{t}, \\ \tilde{\mathbf{t}} &= \text{MLP}(\text{LN}(\hat{\mathbf{t}})) + \hat{\mathbf{t}}, \end{aligned} \tag{4}$$

where LN denotes a LayerNorm function. The calculations included in a MSA unit are:

$$\hat{\mathbf{t}} = \mathbf{A}\mathbf{v} = \text{softmax}(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{D/n}})\mathbf{v}, \tag{5}$$

where $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{N \times D/n}$ are three projections of LN($\mathbf{t}$), $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the attention matrix, n is the number of heads in the MSA.

Meanwhile, the goal of the designed Context Transformer $\mathcal{T}$ is to pass information from $\mathcal{M}_2$ to the main encoding branch $\mathcal{M}_1$. Instead of adding directly the values [47], we aim to project a biased focus to augment the features in $\mathcal{M}_1$. Specifically, for each position in the local feature, the responses from all the context windows are calculated and projected.

Let $\mathbf{t}_l \in \mathbb{R}^{N \times D}$ and $\mathbf{t}_c \in \mathbb{R}^{M \times D}$ (M is the number of flattened features in $\mathcal{M}_2$) denote the local and context tokens embedded from $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. In $\mathcal{T}$, a local query $\mathbf{q}_l$ is projected with $\mathbf{t}_l$, while the context key $\mathbf{k}_c$ and value $\mathbf{v}_c$ are projected with $\mathbf{t}_c$:

$$\begin{aligned} \mathbf{q}_l &= \mathbf{t}_l \mathbf{W}_q \in \mathbb{R}^{N \times D/n}, \\ \mathbf{k}_c &= \mathbf{t}_c \mathbf{W}_k \in \mathbb{R}^{M \times D/n}, \\ \mathbf{v}_c &= \mathbf{t}_c \mathbf{W}_v \in \mathbb{R}^{M \times D/n}, \end{aligned} \tag{6}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{D \times D/n}$ are the corresponding weights of the projection function.

The context attention $\mathbf{A}_c \in \mathbb{R}^{N \times M}$ is then calculated to update $\mathbf{t}_l$:

$$\hat{\mathbf{t}}_l = \mathbf{A}_c \mathbf{v}_c = \text{softmax}(\frac{\mathbf{q}_l \mathbf{k}_c^T}{\sqrt{D/n}})\mathbf{v}_c. \tag{7}$$

These operations, together with the MLP calculations, are repeated for L times, where the contextual dependencies between $\mathbf{t}_l$ and $\mathbf{t}_c$ are modelled 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 2-dimensional features.

D. Implementation Details

Here, we report detailed information of the proposed WiCoNet.

1) The feature extraction networks. We chose the ResNet50 as the feature extraction network in $\mathcal{M}_1$, which is powerful in exploiting the local features [38]. The down-sampling stride of the ResNet is ×1/8 to better preserve the spatial information. In the context branch, we employ a simple convolutional block (referred as the *Context Encoder*) to extract context features. It consists of 11 sequentially connected layers, including 8 convolutional layers and 3 max-pooling layers. Each pooling layer is placed after 2 convolutional layers following the encoder design of UNet [17].

2) Area of the context modelling. The down-sampling scale for input to $\mathcal{M}_2$ is $\times 1/4$, while the down-sampling stride of the context encoder is the same as the ResNet ($\times 1/8$). The size of context window is set to 9 times the size of local window ($w = 3w_l, h = 3h_l$). An analysis of the accuracy versus context modelling range is provided in Sec. V-A. In this study, the size of the local window is 256×256. 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 hyper-parameters in the Context Transformer include: L - number of transformer blocks, n - number of heads, p - size of the embedded parches and D - 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 Sec. V-A. Additionally, there is a weighting parameter α. 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. Beijing Land-Use Dataset

Currently there are few HR satellite benchmark datasets available for the multi-class semantic segmentation of RSIs. To facilitate future researches, we present a new benchmark dataset named Beijing Land-Use (BLU) dataset. This dataset was collected in June, 2018 in Beijing by the Beijing-2 satellite provided by the 21th Century Aerospace Technology Co.,Ltd. The collected data are RGB optical images and have a ground sampling distance (GSD) of 0.8m. We constructed fine-grained human annotations on the collected images based on 6 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 0.8m; *ii) High annotation accuracy.* The annotations were performed by an experienced annotation team dedicated to the RS applications. Fig. 4 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 consistence 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 and wet lands, etc. These features present challenges to the generalization capability of segmentation algorithms.

Fig. 2 presents an overview of the BLU dataset. The observed regions include both urban and rural scenes, covering around 150 km^2 of area in total. The dataset consists of 4 tiles of large RSIs collected in 4 sub-urban regions in Beijing, each one with a pixel size of 15680 × 15680. Each large image is further cropped into 64 images (49 for training, 7 for validation, and 8 for testing), each of which has 2048 × 2048 pixels (Fig. 3). The training, validation, and testing areas are non-overlapping, whereas the cropping windows within each area have small overlaps. The total number of images for training, validation, and testing are 196, 28, and 32, respectively. Both the original tiles and the divided sub-sets are provided. The BLU dataset will be released openly accessible to researchers [1].

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) The 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 × 6000 pixels. The provided data include true ortho photos containing 4 spectral bands (RGB and infrared) and the registered digital surface model (DSM) data. The labels are annotated with 6 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 [48].

2) The GID. This is an HR LC classification dataset collected by the Gaofen-2 (GF-2) satellite. It consists of 10 tiles of RSIs with 4 spectral bands (RGB and near infrared). Each tile has 7200 × 6800 pixels, with a GSD of 0.8m. 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 × 2048 pixels. 16 LC classes are annotated, including: industrial land (IDL), urban residential (UR), rural residential (RR), traffic land (TL), paddy field (PF), irrigated 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).

[1] https://rslab.disi.unitn.it/dataset/BLU/

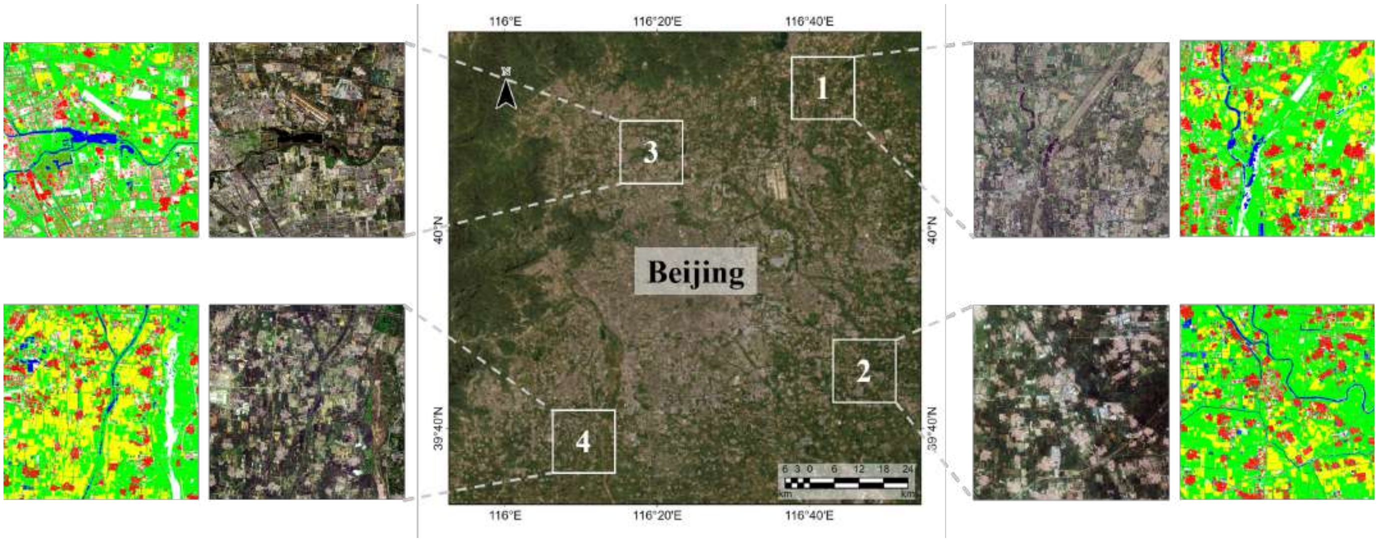

Fig. 2: Overview of the BLU dataset.

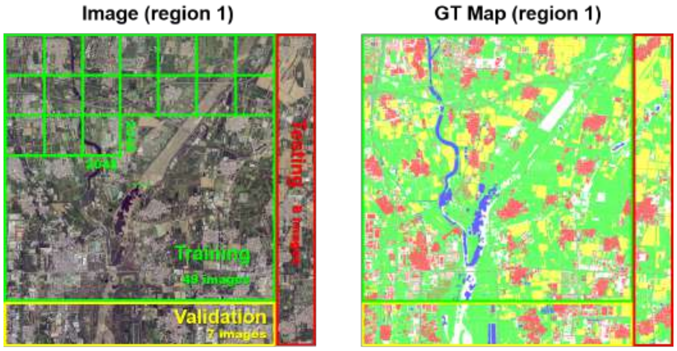

Fig. 3: Split of the training, validation and testing sets.

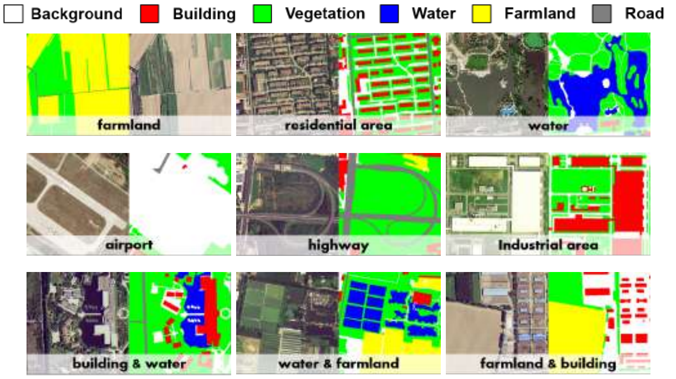

Fig. 4: Sample images taken from different scenes in the BLU dataset.

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 lr is dynamically calculated at each iteration as: $0.1 * (1 - iterations/total\_iterations)^{1.5}$. 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 OA (evaluated on the validation set) is saved.

TABLE I: Class distribution in the BLU dataset.

	Class Name	Number of pixels	Proportions (%)
	Background	156,190,234	15.88
🟥	Built-up	125,695,683	12.78
🟩	Vegetation	478,668,644	48.67
🟦	Water	28,364,259	2.88
🟨	Agricultural	159,386,020	16.20
⬛	Road	35,144,760	3.57
	Total	983,449,600	-

In this study we adopt the most frequently used metrics [35], [40] to evaluate the tested methods, including: i) Overall Accuracy (OA), which is the numeric ratio of correctly classified pixels versus all the pixels in RSIs, ii) F_1 score of each class, which is the harmonic mean of the $Precision$ and $Recall$, and iii) 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:

$$\begin{aligned} OA &= (TP + TN)/(TP + TN + FP + TN), \\ Precision &= TP/(TP + FP), Recall = TP/(TP + FN), \\ F_1 &= 2 \times \frac{Precision \times Recall}{Precision + Recall}, \\ IoU &= TP/(TP + FP + FN). \end{aligned} \tag{8}$$

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 modelling range is analyzed. Finally, the proposed WiCoNet is compared with several CNN models with context-aggregation designs in recent studies.

A. Ablation Study

Choice of Hyper-Parameters As introduced in Sec.III-D, L and n are two adjustable hyper-parameters 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 III. One can observe that the best OA on the BLU and GID datasets is obtained when $L = 4, n = 4$. Meanwhile, the optimal hyper-parameter values for the Potsdam dataset are $L = 2, n = 4$. The OA is lower when L is set to 8. We assume that this is caused by over-fitting, since the long-range context information in RSIs is relatively simple, thus too many Transformer layers may be redundant. The tested optimal parameters for different datasets are fixed in the following experiments.

Quantitative Results An ablation study is conducted to test the effectiveness of context modelling. 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 II.

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 modelling with the WiCoNet, significant improvements are obtained. The improvements over the baseline FCN are 0.84%, 1.01%, and 1.41% in OA and 1.41%, 4.05, and 1.69% in mIoU, respectively, on the BLU dataset, GID, and Potsdam dataset. These results show that the wide context modelling in the WiCoNet stably improves the LCLU segmentation accuracy of HR RSIs.

Qualitative Results To qualitatively assess the effects of context modelling, Fig. 5 and Fig. 6 show 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: *i) Discriminating the critical areas.* By modelling 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., Fig. 5(b), Fig. 6(b)(f)); *ii) 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), the rivers in Fig. 5(c) and Fig. 6(a)). The WiCoNet better preserves their long-range consistency; *iii) 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)).

Effects of the Context Modelling 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 modelling range may cause loss of focus on the local content. To find the best context modelling 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., $w \times h = 2w_l \times 2h_l$, $w \times h = 3w_l \times 3h_l$, and $w \times h = 4w_l \times 4h_l$). 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 pyramid scene parsing (PSP) module, the DANet [18] with channel attention and non-local attention, the SCAttNet [49] with spatial and channel attention, the MSCA [39] with multi-scale context aggregation designs, and the LANet [38] with local attention.

We implement all the tested methods with the experimental settings described in Sec. IV-C and report the results in Tables V, VI and VII. The reported values are the average of the metrics derived in 3 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.

TABLE II: Quantitative results of the ablation study on the considered data sets.

Dataset	Method	Components			OA(%)	mean F1(%)	mIoU(%)
		local branch	context branch	Transformer			
BLU	FCN [2]	√			86.51	81.88	70.09
	FCN+Transformer	√		√	86.74	82.48	70.92
	WiCoNet (Ours)	√	√	√	**87.35**	**82.89**	**71.50**
GID	FCN [2]	√			74.71	63.13	49.02
	FCN+Transformer	√		√	75.82	65.20	51.36
	WiCoNet (Ours)	√	√	√	**77.14**	**66.26**	**53.07**
Potsdam	FCN [2]	√			88.96	90.72	83.24
	FCN+Transformer	√		√	88.69	90.39	82.66
	WiCoNet (Ours)	√	√	√	**90.24**	**91.70**	**84.93**

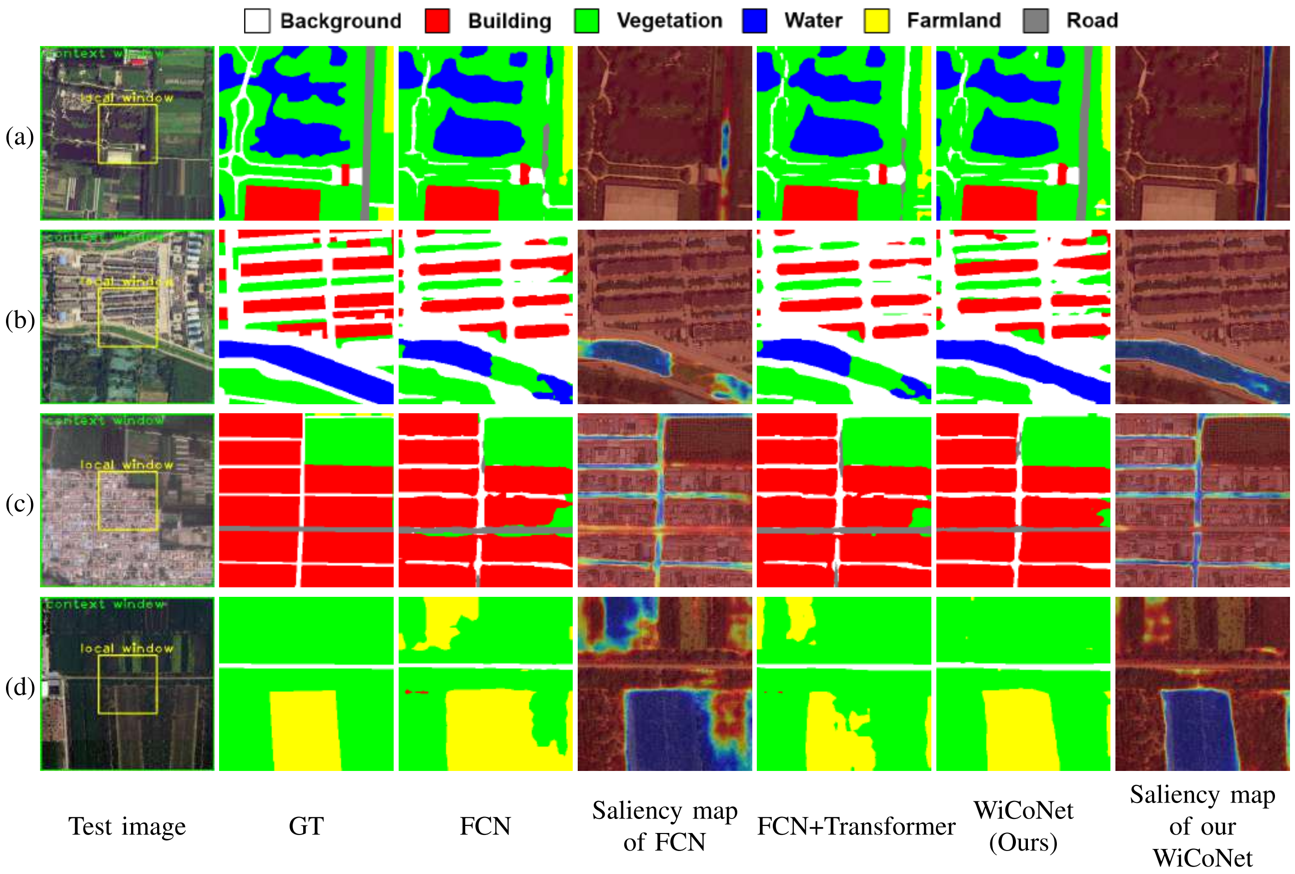

Fig. 5: Qualitative results of the ablation study on the BLU datasets. The saliency maps of the critical classes are presented. The selected challenging scenes include: (a) occluded road, (b) green algae-covered river, (c) streets in a residential area, and (d) farmland surrounded by vegetation.

TABLE III: The OA obtained by the WiCoNet with different hyper-parameters.

Dataset	L			n	
	2	4	8	4	8
BLU	87.03	87.35	87.02	87.35	87.13
GID	77.04	77.14	76.96	77.14	77.05
Potsdam	90.24	90.21	89.95	90.24	90.22

The parameter size and computational cost of each model are reported in Table VIII. The number of floating point operations per second (FLOPS) is calculated based on the experimental settings for the BLU dataset (including input & output size and hyper-parameters), except for the batch size which 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 FLOPS are very close to those of the DeepLabv3+.

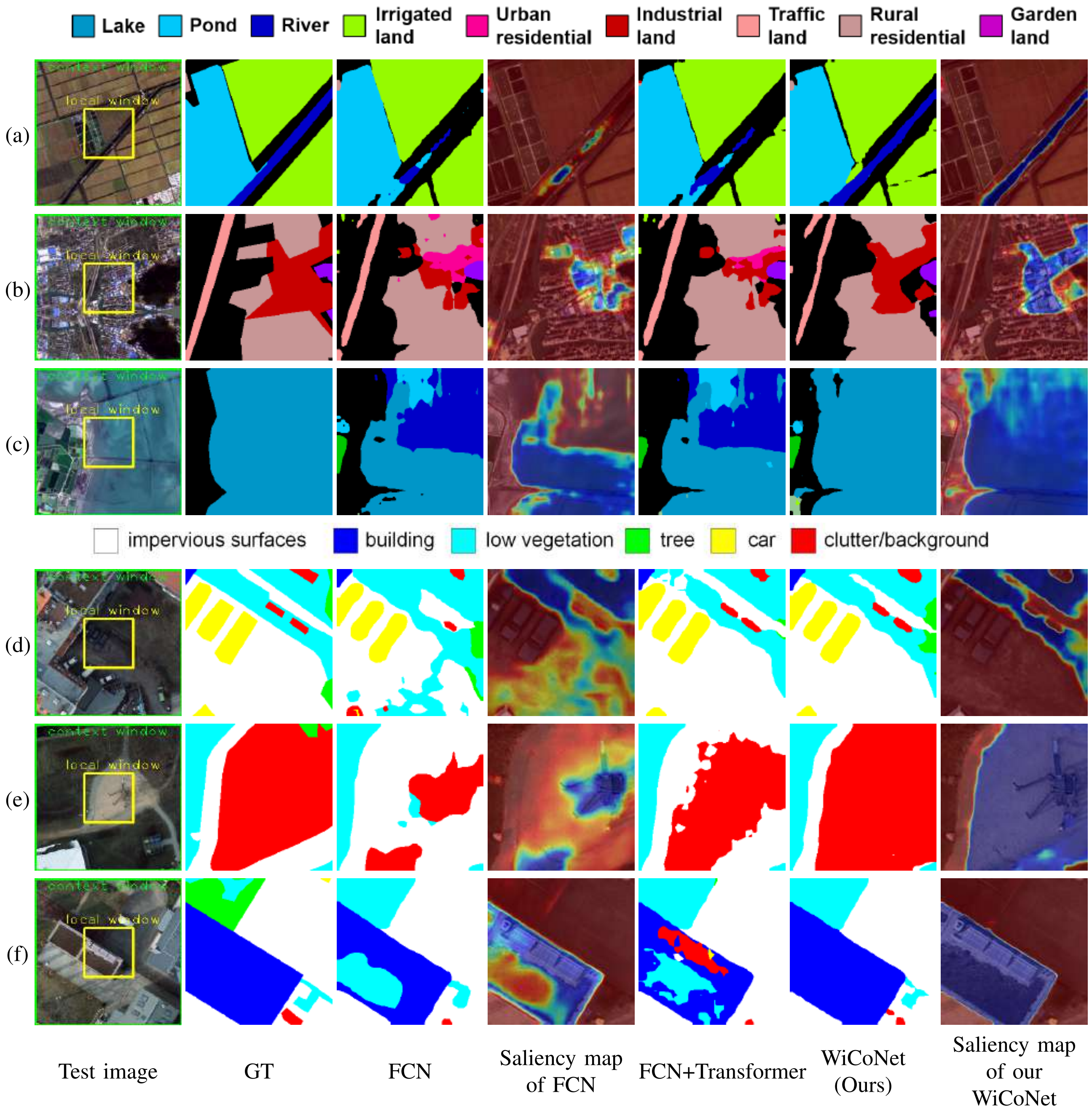

Fig. 6: Qualitative results of the ablation study on the additional datasets. The saliency maps of the critical classes are presented. (a)˜(c) Results selected from the GID, (d)˜(f) Results selected from the Potsdam dataset.

VI. Conclusions

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 Wide-Context Network (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 researches, we also release a high-quality and large-scale benchmark dataset for the semantic segmentation on HR RSIs, i.e., the Beijing Land-Use (BLU) dataset. Through experiments on the BLU dataset and two additional datasets, we i) verified the effectiveness of the long-range context modelling, ii) analyzed the accuracy of different context modelling sizes, and iii) compared the WiCoNet with several literature works that models context information in RSIs. Experimental results

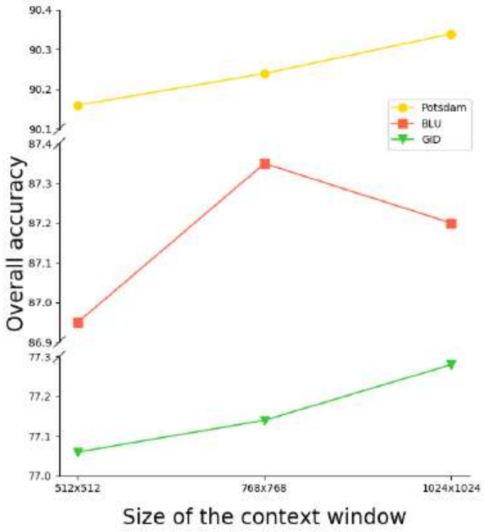

Fig. 7: The OA of results versus different size of context windows.

TABLE IV: The effects of context modeling range on the segmentation accuracy.

Dataset	Metrics	Size of context windows		
		512×512	768×768	1024×1024
BLU	OA	86.91	**87.35**	87.20
	mean F_1	82.11	**82.77**	82.35
	mIoU	70.41	70.58	**70.81**
GID	OA	77.06	77.14	**77.28**
	mean F_1	66.03	66.26	**66.55**
	mIoU	53.04	53.07	**53.38**
Potsdam	OA	90.16	90.24	**90.34**
	mean F_1	91.59	91.71	**91.76**
	mIoU	84.72	84.93	**85.03**

show that the WiCoNet enables a better understanding and modeling of both the local scene information and the global class distribution, thus brings significant accuracy improvements. However, there are still global inconsistency and some local fragmentation errors remain, indicating that there is still margin to improve the modelling of long-range context information in large RSIs. This is left for future works, where adversarial learning strategies [29] 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 us improve the quality of this paper.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[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 Transactions on Geoscience and Remote Sensing*, 2020.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 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 *ECCV*, 2018.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[7] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[8] H. Tang, S. Bai, and N. Sebe, "Dual attention gans for semantic image synthesis," in *ACM MM*, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," *arXiv preprint arXiv:2012.00759*, 2020.

[12] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[14] G. Yang, H. Tang, Z. Zhong, M. Ding, L. Shao, N. Sebe, and E. Ricci, "Transformer-based source-free domain adaptation," *arXiv preprint arXiv:2105.14138*, 2021.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[16] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *CVPR*, 2018, pp. 3127–3135.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[18] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *CVPR*, 2017.

[19] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *CVPR*, 2019, pp. 9308–9316.

[22] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[23] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2017.

[24] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *ECCV*, 2018.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[27] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, vol. 12351. Springer, 2020, pp. 173–190.

[28] L. Ding, Q. Yang, J. Lu, J. Xu, and J. Yu, "Road extraction based on direction consistency segmentation," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 131–144.

TABLE V: Comparison of segmentation accuracy provided by different methods on the BLU dataset.

Method	Per-class F_1 (%)						OA (%)	mean F_1 (%)	mIoU (%)
	Background	Built-up	Vegetation	Water	Agricultural	Road			
FCN [2]	72.92	87.56	90.41	85.15	86.42	68.88	86.51±0.06	81.88±0.15	70.09±0.21
PSPNet [50]	72.66	87.40	90.41	86.30	86.71	68.84	86.59±0.05	82.05±0.25	70.35±0.36
DeepLabv3+ [5]	73.99	87.93	90.76	**86.46**	**87.32**	68.85	87.08±0.12	82.55±0.20	71.07±0.25
DANet [18]	73.06	87.73	90.55	85.45	86.77	69.07	86.76±0.07	82.10±0.40	70.40±0.57
SCAttNet [49]	73.21	87.62	90.54	86.26	86.87	69.32	86.77±0.10	82.30±0.17	70.68±0.25
MSCA [39]	73.71	88.34	90.74	85.92	86.86	**70.31**	87.17±0.02	82.64±0.02	71.21±0.05
LANet [38]	73.81	87.48	90.60	85.99	87.02	68.49	86.89±0.14	82.28±0.09	70.60±0.17
WiCoNet (Ours)	**74.43**	**88.55**	**90.94**	86.01	87.23	70.21	**87.35**±0.18	**82.89**±0.22	**71.50**±0.30

TABLE VI: Comparison of segmentation accuracy provided by different methods on the GID dataset. The LC classes include: industrial land (IDL), urban residential (UR), rural residential (RR), traffic land (TL), paddy field (PF), irrigated 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).

Method	Per-class F_1 (%)															OA	mean F_1	mIoU
	IDL	UR	RR	TL	PF	IL	DC	GP	AW	SL	NG	AG	RV	LK	PN	(%)	(%)	(%)
FCN [2]	59.75	75.57	57.52	68.08	74.815	81.88	36.23	28.55	84.91	8.97	70.07	58.33	81.41	74.11	75.56	74.71±0.04	63.13±0.18	49.02±0.23
PSPNet [50]	59.84	76.29	58.50	67.70	75.25	82.45	**39.23**	31.69	85.34	7.58	73.37	62.79	83.11	76.70	75.94	75.44±0.06	64.41±0.05	50.44±0.03
DeepLabv3+ [5]	60.44	76.67	58.49	67.67	75.65	82.5	38.62	33.03	84.39	7.13	71.12	64.83	83.17	74.60	74.93	75.38±0.35	64.27±0.74	50.21±0.76
DANet [18]	62.53	76.50	56.73	68.08	75.29	82.76	38.03	26.72	85.75	12.62	73.99	62.95	83.45	77.68	77.25	75.68±0.19	64.7±0.11	50.81±0.27
SCAttNet [49]	61.87	77.32	**59.19**	68.75	74.66	82.29	35.75	**33.32**	86.31	5.66	71.53	**74.26**	81.72	80.96	80.67	76.05±0.28	65.59±0.37	52.01±0.42
MSCA [39]	62.06	77.27	56.51	68.69	74.36	82.46	35.99	24.51	87.08	**16.00**	72.75	70.65	83.78	78.61	79.09	76.10±0.03	65.33±0.59	51.60±0.69
LANet [38]	**63.65**	**77.67**	58.77	**69.13**	**76.80**	82.71	37.01	25.68	86.14	7.71	72.42	73.58	84.55	83.53	**82.02**	76.75±0.26	66.06±0.06	52.83±0.43
WiCoNet (Ours)	63.41	77.21	57.62	68.54	76.37	**83.38**	40.67	32.75	**87.57**	4.9	73.08	62.44	**87.76**	**86.86**	81.8	**77.14**±0.13	**66.26**±0.57	**53.07**±0.20

TABLE VII: Comparison of segmentation accuracy provided by different methods on the Potsdam dataset.

Method	Per-class F_1 (%)					OA (%)	mean F_1 (%)	mIoU (%)
	Impervious Surface	Building	Low Vegetation	Tree	Car			
FCN [2]	91.08	95.21	86.17	86.51	94.63	88.96±0.30	90.72±0.20	83.24±0.33
PSPNet [50]	88.85	93.20	83.89	82.69	91.62	86.47±0.78	88.05±0.66	78.91±1.07
DeepLabv3+ [5]	91.79	96.46	86.17	86.39	94.34	89.47±0.34	91.03±0.23	83.81±0.39
DANet [18]	91.94	96.05	86.74	87.11	94.42	89.74±0.13	91.25±0.12	84.14±0.20
SCAttNet [49]	91.66	95.57	86.44	86.79	94.13	89.41±0.31	90.92±0.27	83.56±0.46
MSCA [39]	92.31	**96.74**	86.59	87.01	95.11	90.00±0.07	91.55±0.07	84.69±0.12
LANet [38]	91.63	95.83	85.96	86.35	93.98	89.91±0.10	91.45±0.11	84.47±0.19
WiCoNet (Ours)	**92.50**	96.53	**87.03**	**87.31**	**95.13**	**90.24**±0.09	**91.70**±0.04	**84.93**±0.07

TABLE VIII: Comparison of model size and computational cost expressed in terms of number of parameters and FLOPS, respectively.

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

[29] 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 Transactions on Image Processing*, vol. 31, pp. 678–690, 2022.

[30] L. Duan and X. Hu, "Multiscale refinement network for water-body segmentation in high-resolution satellite imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 4, pp. 686–690, 2019.

[31] 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 Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1–13, 2019.

[32] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[33] L. Ding, K. Zheng, D. Lin, Y. Chen, B. Liu, J. Li, and L. Bruzzone, "Mp-resnet: Multi-path residual network for the semantic segmentation of high-resolution polsar images," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[34] 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 Sensing*, vol. 10, no. 9, p. 1339, 2018.

[35] 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 Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[36] C. Zhang, I. Sargent, X. Pan, A. Gardiner, J. Hare, and P. M. Atkinson, "Vprs-based regional decision fusion of cnn and mrf classifications for very fine resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4507–4521, 2018.

[37] L. Ding and L. Bruzzone, "Diresnet: Direction-aware residual network for road extraction in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[38] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

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

[40] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *CVPR*, 2019.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[43] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *arXiv preprint arXiv:2102.08005*, 2021.

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

[45] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[46] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[47] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in *CVPR*, 2019, pp. 8885–8894.

[48] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2016.

[49] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 905–909, 2020.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *TPAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.

Lei Ding received the MS's degree in Photogrammetry and Remote Sensing from the Information Engineering University (Zhengzhou, China), and the PhD (cum laude) in Communication and Information Technologies from the University of Trento (Trento, Italy). He is now a lecturer at the PLA Strategic Force Information Engineering University. His research interests are related to semantic segmentation, change detection and domain adaptation with Deep Learning techniques. He is a referee for many international journals, including IEEE TIP, IEEE TNNLS and IEEE TGRS.

Dong Lin received the MS's degree in Photogrammetry and Remote Sensing from the Information Engineering University (Zhengzhou, China), and the PhD in Photogrammetry and Remote Sensing from the Technische Universität Dresden (Dresden, Germany). He is now a lecturer at the Space Engineering University. His research interests include deep learning, change detection and thermal image processing.

Shaofu Lin received the Ph.D. degree in mapping GIS from the Institute of Remote Sensing and Geographic Information System, Peking University in 2002. He worked on the research, construction and management of informatization and e-government in Hainan Information Center, Beijing Information Resource Management Center and Beijing Municipal Office of Informatization from 1990 to 2009. He was engaged in the promotion of smart city, e-government and industrial technology innovation in Beijing Economic and Information Commission from 2009 to 2014, successively serving as the director of E-government Department, and Science Technology Standards Department. He has been a professor of Software College at Beijing University of Technology since 2014, ever serving as the director of Information Department, the executive director of Beijing Institute of Smart City, and the executive director of Beijing Advanced Innovation Center for Future Internet Technology. His research interests include spatial-temporal big data, data fusion and intelligence, and block chain. He has senior memberships of China Computer Federation (CCF) and Chinese Institute of Electronics (CIE), and has memberships of the Blockchain Commission of CCF, Expert Committee of China Big Data Industry Ecological Alliance, and Network Information Technology Expert Committee of China Artificial Intelligence Industry Alliance. He is a board member of Beijing Institute of Big Data.

Jing Zhang received a master's degree in software engineering from Beijing University of Technology. She is currently a Ph.D student at the department of Information Engineering and Computer Science, University of Trento, Italy. Her current research interests are related to the change detection and semantic segmentation of remote sensing image.

Xiaojie Cui received the MS's degree and PhD degree in Cartography and Geographic Information System from the Information Engineering University (Zhengzhou, China). She is now an engineer at the Beijing Institute of Remote Sensing Information. Her research interests include remote sensing image processing and big data analysis.

Yuebin Wang 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 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 is currently a Postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received the master's degree from the School of Electronics and Computer Engineering, Peking University, China and the Ph.D. degree from the Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.

Lorenzo Bruzzone (S'95-M'98-SM'03-F'10) received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and 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 coauthor) of 215 scientific publications in referred international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. He was invited as keynote speaker in more than 30 international conferences and workshops. Since 2009 he is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

Dr. Bruzzone was a Guest Co-Editor 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 between 2013-2017. Currently he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing. He has been Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012-2016. His papers are highly cited, as proven form the total number of citations (more than 27000) and the value of the h-index (78) (source: Google Scholar).