LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images

Lei Ding, *Student Member, IEEE*, Hao Tang, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*— The trade-off between feature representation power and spatial localization accuracy is crucial for the dense classification/semantic segmentation of remote sensing images (RSIs). High-level features extracted from the late layers of a neural network are rich in semantic information, yet have blurred spatial details; low-level features extracted from the early layers of a network contain more pixel-level information but are isolated and noisy. It is therefore difficult to bridge the gap between high- and low-level features due to their difference in terms of physical information content and spatial distribution. In this article, we contribute to solve this problem by enhancing the feature representation in two ways. On the one hand, a patch attention module (PAM) is proposed to enhance the embedding of context information based on a patchwise calculation of local attention. On the other hand, an attention embedding module (AEM) is proposed to enrich the semantic information of low-level features by embedding local focus from high-level features. Both proposed modules are lightweight and can be applied to process the extracted features of convolutional neural networks (CNNs). Experiments show that, by integrating the proposed modules into a baseline fully convolutional network (FCN), the resulting local attention network (LANet) greatly improves the performance over the baseline and outperforms other attention-based methods on two RSI data sets.**

***Index Terms*— Convolutional neural network (CNN), deep learning, remote sensing, semantic segmentation.**

I. Introduction

THE dense classification of remote sensing images (RSIs), which is often referred to as semantic segmentation, is a crucial step for the automatic analysis of remote sensing data. It is widely used in a variety of applications, such as land-use and land-change mapping, urban management, environment monitoring, and so on. With the development of convolutional neural networks (CNNs) and their application to dense classification (introduced in the fully convolutional network (FCN) [1]), the accuracy of semantic segmentation on RSIs has been greatly improved [2]. A commonly used design in CNNs is based on stacked convolutions and pooling operations, which constantly reduce the spatial size of features to enhance their semantic representations [3]. Although this feature embedding design (referred to as "encoders") has the benefits of enlarging the receptive field and learning more intrinsic feature representations, it has the cost of losing detailed spatial information. Thus, the semantic segmentation results are generated by considering a large area as a whole instead of precisely classifying each pixel. As a result, small objects may be neglected and the contours of objects are ambiguous. To conquer this problem, "decoders" are introduced, which typically employ the low-level features from "encoders" to retrieve the lost spatial information [4]–[6]. However, the low-level and high-level features have significant differences in both semantic information and spatial distributions (e.g., low-level feature are more sensitive to gradient changes and distinct points, while the high-level features have stronger activation in the center of objects), thus the fusion of them does not bring significant improvements to the segmentation accuracy [7].

This trade-off between feature embedding power and spatial localization accuracy is crucial for the semantic segmentation of RSIs. On the one hand, different categories of the ground objects may share similar spectral features, thus requiring for an aggregation of context information [8]. On the other hand, many applications of the analysis of the RSIs require high precision in mapping contours of ground objects. Therefore, detailed spatial information is needed for accurately identifying both the boundary of regions and small objects.

The introduction of the attention mechanism is an effective strategy to reduce the confusion in predicted categories without losing spatial information. With the global statistics aggregated from the whole image, scene information can be embedded to highlight (or suppress) the features with strong correlations [9]. However, the spatial size of RSIs is usually much larger than that of natural images, whereas the number of object categories is often smaller. For example, each image in the ISPRS semantic labeling data set (Potsdam area) [10] has 6000×6000 pixels divided into six object categories in this data set. As a result, almost every image contains all the object categories, and no clear global scene information can be embedded at the global level. In other words, we argue that the typical attention-based techniques cannot be directly applied to the semantic segmentation of large-size RSIs.

In this article, we propose the generation of patch-level local attention to improve the semantic segmentation of RSIs. The proposed approach is based on the finding that the image-level semantic information of RSIs is not clear, whereas the local image patches have clear semantic references (an illustration example of this observation is given in Fig. 1). Therefore, we propose a novel patch attention module (PAM) to exploit patchwise local attention. This module operates on extracted

Manuscript received February 20, 2020; revised March 29, 2020; accepted May 8, 2020. This work was supported by the scholarship from the China Scholarship Council under Grant 201703170123. *(Corresponding author: Lorenzo Bruzzone.)*

The authors are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lei.ding@unitn.it; hao.tang@unitn.it; lorenzo.bruzzone@unitn.it).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2994150

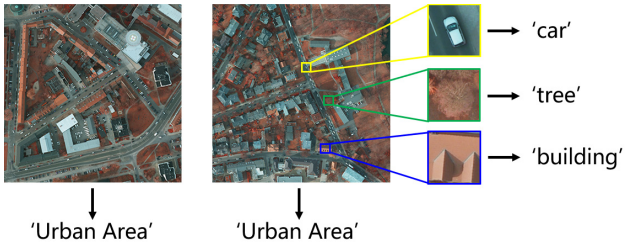


Fig. 1. Examples of the image-level information for RSIs. The information of a whole RSI cannot be deduced more specifically than just "urban area," but the information of image patches can be easily attributed to classes like "car," "tree," and "building."

feature maps and can aggregate context information from the local patch to reduce confusion. In our model, the PAM is appended after both the high-level and low-level features to enhance their representation. Moreover, to bridge the gap between high-level and low-level features, an attention embedding module (AEM) is proposed to embed semantic focus from high-level features into low-level features. This module can greatly improve the semantic representation of low-level features without losing their spatial details, thus improving the effectiveness of the fusion between high-level and low-level features. The proposed modules are lightweight and can be incorporated into existing CNN architectures to improve the segmentation accuracy. The experiments on two RSI data sets have validated the effectiveness of the proposed architecture.

To summarize, the main contributions in this article are as follows.

1) Proposing both a PAM to embed scene information from local patches and an AEM to enhance the semantic representation of low-level features by introducing attention from high-level features.
2) Proposing a local attention network (LANet) to improve the semantic segmentation of RSIs by enhancing the scene-related representation in both encoding and decoding phases.
3) Performing extensive ablation studies on two RSI data sets by incorporating the proposed modules into baseline FCN network in sequence. The resulting LANet is further compared with other networks with decoding or attention-based designs to evaluate its performance.

The remainder of this article is organized as follows. Section II introduces the related works on semantic segmentation tasks. We then describe our LANet in Section III. In Section IV, we present a detailed experimental evaluation and discussion. Finally, we conclude this article in Section V.

II. Related Work

A. Semantic Segmentation of RSIs

Studies on CNN-based semantic segmentation of RSIs begin to thrive after the emergence of several open data sets and contests, including the ISPRS data sets,[1] the DeepGlobe contest,[2] and the SpaceNet competition.[3] One of the focuses of the studies on semantic segmentation of RSIs is the collaborative use of CNNs and statistical modeling methods to improve the accuracy [11], [12]. Another research direction is related to the multiscale feature extraction. In this context, the multiscale pyramid pooling module has been introduced to the semantic segmentation of RSIs in [13]. In [14], a two-stage design operating on seven different scales is presented to enlarge the receptive field of the network. The multiscale alignment of edges and outputs are introduced in [15] and [16], respectively. The exploitation of additional training information has also been studied in the semantic segmentation of RSIs, such as the use of Open Street Maps in [17] and [18], the explicit use of the digital surface model (DSM) in [19] and [20], and the supervision of object boundaries in [21]. However, limited attention has been paid to the special properties of RSIs, such as their large spatial size and relatively low number of categories with respect to natural images. In this article, these major differences with natural images are taken into account when designing the relevant processing modules.

B. Encoder-Decoder Designs

The encoder-decoder networks have been successfully used in many computer vision tasks such as image generation [22], [23], object/saliency detection [24], [25], crowd counting [26], and semantic segmentation [1], [27]. Usually, the encoder–decoder networks contain two subnets: 1) an encoder subnet that gradually reduces the feature maps and captures higher semantic information and 2) a decoder subnet that gradually recovers the spatial information. The encoder subnet is the focus of most existing studies. There are plenty of works on enlarging the receptive field while minimizing the number of parameters, including well-known architectures such as the PSPNet [28] and the DeepLabV3+ [27]. They both add parallel context-aggregation branches at the top of encoding networks. PSPNet employs global average pooling operations to exploit contextual information, while DeepLabV3+ employs dilated convolutions with different rates. One of the limitations of these works is that their decoder subnets are not as powerful as the encoders. Although in some studies there are cascade decoding designs that aim to exploit the features from early CNN layers [4]–[7], these features are usually concatenated or summed to the high-level features without enhancing their semantic representation. Thus, they provide a limited contribution to the accuracy. To overcome this limitation, we propose the use of the attention mechanism for enhancing the representation of low-level features during the decoding phase.

C. Attention Mechanism

Attention mechanism refers to the strategy of allocating biased computational resources to the processed signal to highlight its informative parts. In the tasks related to the understanding of image content, a typical solution for generating attention statistics is to gather information from a global scale, namely, to exploit the scene or image-level information. This is because the scene information may provide clues about the possible contents in an image. In [29], the attention of the

[1] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html
[2] http://deepglobe.org/challenge.html
[3] https://spacenetchallenge.github.io/

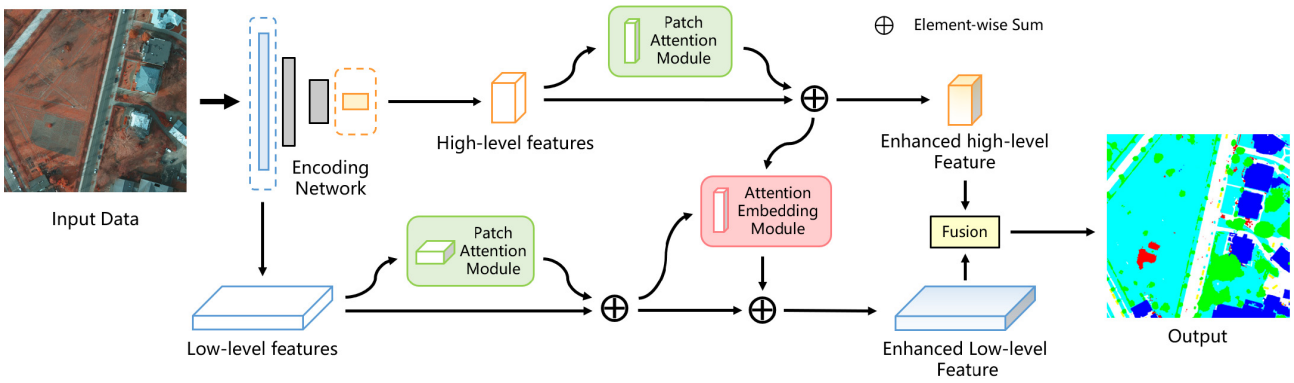


Fig. 2. Architecture of the proposed LANet. The PAM generates attention maps to highlight patchwise focus in feature maps. The AEM embeds semantic information from high-level features to low-level ones.

feature map is aggregated using an hourglass module in a residual manner. This residual attention network introduced a chunk-and-mask module, where the global attention is aggregated in the Soft Mask Branch through stacked down-sampling convolutions. In [9], a squeeze-and-excitation (SE) block is proposed, which uses global-pooling to generate channelwise attention. In this way, spatial-irrelevant information can be learned to emphasize the scene-relevant feature channels. The design of "squeezing" spatial information and the parallel connection of attention branch introduced in this article have been widely adopted in subsequent studies. In EncNet [30], a context encoding module is proposed to capture the scene-dependent global context as channelwise attention. CBAM [31] introduced a spatial attention module to highlight the informative spatial regions. The spatial attention maps are generated by using pooling operations along the channel axis. BAM [32] has a similar module to exploit spatial correlations but it is implemented by applying dilated convolutions. PSANet [33] introduced the modeling of long-range correlation for each spatial position, but the channels of its inner layers are related to the input image size and cannot be applied to the prediction of full-size RSIs. A parallel design that models both channelwise and pointwise attention is introduced in DANet [34]. A limitation of the nonlocal reasoning-based networks is that the reasoning of global spatial correlation is calculation intensive. A lightweight graph-based module for reasoning latent correlations has been presented in [35].

Some works use the attention mechanism for the segmentation of RSIs. In [36], a channel attention block is designed to enhance the decoding branch of the CNN. In [37], the attention mechanism is used to match the caption nouns with the objects in RSIs. The global attention upsampling module [38] is introduced in [39] to provide global guidance from high-level features to low-level ones. In [8], the attention-based reasoning of both positional and channelwise relations and their integration in serial and parallel manners have been studied. In [40], a multiscale design has been introduced to aggregate context information through different branches.

Building on top of these studies, we propose a simple yet effective approach that extends the use of the attention mechanism to the spatial dimension without significantly increasing the computational load.

III. Proposed Approach

In this section, we present the proposed LANet devised for improving semantic segmentation of RSIs. First, an overview of the network is given to introduce the general motivation and architecture. After this, the proposed modules are described in detail. Finally, further explanation is given on the integration of the proposed modules into two backbone networks (ResNet and HRNet).

A. Overview of the Proposed LANet

Contextual information is known to be crucial for the semantic segmentation of RSIs. Global pooling is an effective operation to aggregate contextual information since it utilizes the scene information to learn biased focus on object categories. However, this approach is less effective on RSIs, since the image-level information is not clear, as discussed in Section I. To address this problem, we propose the LANet to utilize patch-based scene information on RSIs.

The motivation of this article is twofold: 1) employing patch-based attention to enhance the embedding of contextual information and 2) enriching the semantic representation of low-level features to better utilize the spatial information. To achieve this goal, two separate modules are introduced in LANet: 1) a PAM to enhance the embedding of local context information and 2) an AEM to improve the use of spatial information. Specifically, we designed two parallel branches to process features from different layers. As shown in Fig. 2, in the upper branch, high-level features (produced by late layers of a CNN) go through a PAM to enhance their feature representation; in the lower branch, low-level features (produced by early layers of a CNN) are first enhanced by PAM, then embedded with semantic information from high-level through AEM. The final segmentation results are produced by the fusion of the features from both branches.

B. PAM

Semantic segmentation of RSIs suffers greatly from the problem of intraclass inconsistence since the discrimination of object categories is a comprehensive task affected by both the surface type and the context of an image. To alleviate this

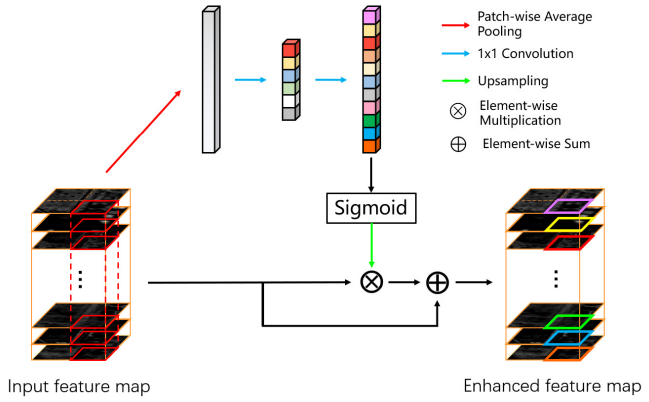


Fig. 3. Detailed design of the PAM. Descriptors are calculated patchwisely to aggregate local context information.

problem, we propose a PAM to enhance the aggregation of context information in the extracted features.

Fig. 3 shows the design of the PAM. This article is inspired by the design of the SE-block [9]. The original SE-block introduced global average pooling to generate one single descriptor for each feature channel. However, as discussed in Section I, this cannot be applied to the processing of large-size RSIs. In our approach, we limit the generation of descriptors to local patches, so that each descriptor contains meaningful information of the local context. Let us first consider a single patch. The descriptor z_c for the cth channel of a generic patch is calculated as

$$z_c = \frac{1}{h_p w_p} \sum_{i=1}^{h_p} \sum_{j=1}^{w_p} x_c(i, j) \tag{1}$$

where h_p and w_p denote the horizontal and vertical spatial size of the pooling window, respectively, and x_c denotes a pixel at cth channel. In this way, a c-channel vector $\mathbf{z}_p$ can be generated, which contains the statistics describing the patch p. After this, we follow the bottleneck gating design in [9] to learn an attention vector $\mathbf{a}_p \in \mathbb{R}^{c \times h_p \times w_p}$ for the patch p. Instead of using fully connected layers, we employ convolutional operations so that they can be applied to process other patches without assigning extra weights. The gating operation to generate attention maps can be symbolized as

$$\mathbf{a}_p = F_U\{\sigma[H_i\delta(H_r\mathbf{z}_p)]\} \tag{2}$$

where σ and δ denote sigmoid and ReLU functions [41], respectively; H_r represents the 1×1 dimension-reduction convolution with the reduction ratio r; H_i denotes the 1×1 dimension-increasing convolution that recovers the feature dimension back to c; and F_U is the upsampling operation.

Let us now extend the case of a single local patch to the global level. Given a feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, maps of descriptors $\mathbf{Z} \in \mathbb{R}^{C \times H' \times W'}$ can be generated. Here, H' and W' are determined by the size of each patch (pooling window) as

$$H' = \frac{H}{h_p}, \quad W' = \frac{W}{w_p} \tag{3}$$

where h_p and w_p are set according to the spatial reduction ratio of the corresponding encoding layer to ensure a remarkable enlargement of the receptive field. An alternative is to use a sliding window for generating the descriptors, so that the descriptor maps have the same size of input images. However, this option will tremendously increase the calculation; thus, it is not adopted in our implementation. After the convolutional layers, attention maps $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$ can be produced. Finally, the original input features $\mathbf{X}$ are multiplied elementwise with $\mathbf{A}$ to enhance their representation. A residual design is adopted to ensure the stable backpropagation of gradients.

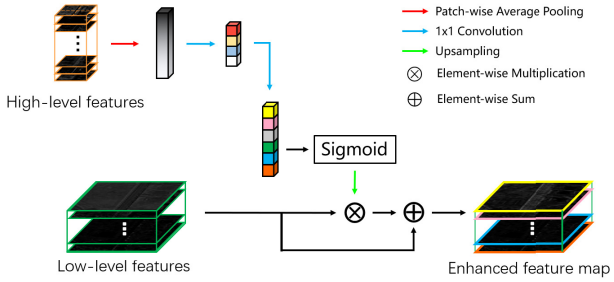


Fig. 4. Detailed design of the AEM. Low-level features are semantically enriched by embedding local focus from high-level features.

C. AEM

An effective exploitation of low-level features is difficult due to their difference with high-level features in terms of spatial distribution and physical meaning. The most frequently used way of employing low-level features is to concatenate them with high-level features, which brings only slight improvement in performance (refer to discussion in Section IV). To make the best use of low-level features, we propose an AEM to enrich their semantic meaning. This operation bridges the gap between high-level and low-level features without sacrificing the spatial details of the latter.

Fig. 4 shows the design of the proposed AEM. The intuition of this approach is to embed local attention from high-level features into the low-level features. In this way, low-level features are embedded with context information that goes beyond the limitation of their receptive fields, while their spatial details are kept. First, we generate descriptors from high-level features through the same calculation as in (1). Let us denote these maps of descriptors as $\mathbf{Z}_h \in \mathbb{R}^{C_h \times H' \times W'}$, and the low-level features as $\mathbf{X}_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. We generate attention maps for the low-level features $\mathbf{A}_l$ by transforming $\mathbf{Z}_h$ through bottleneck convolutions as

$$\mathbf{A}_l = F_U\{\sigma[H_l\delta(H_r\mathbf{Z}_h)]\} \tag{4}$$

where H_r is a dimension reduction convolution and H_l changes the number of channels to be the same as $\mathbf{X}_l$. To avoid excessive interference of high-level features, we add a residual design to emphasize the importance of low-level features. The enhanced low-level features are calculated as

$$\mathbf{X}_l = \mathbf{X}_l + \mathbf{X}_l\mathbf{A}_l. \tag{5}$$

D. Feature Fusion Between Different Layers

After being processed by AEM, low-level features are semantically enriched and can potentially give a higher

contribution to the prediction of the pixel class. Both the high-level and low-level features keep their dimensions after the processing of PAM and AEM. Accordingly, classic feature fusion operations (e.g., concatenation) can be applied to the outputs of the two branches. Since the specific feature fusion operation is not the focus of this article, also considering the interest in validating the output from each branch, we simply train two separate classifiers for each branch and perform an elementwise sum to generate the final results.

IV. Data Set Description and Design of Experiments

To assess the effectiveness of the proposed method, experiments have been conducted on two RSI data sets, i.e., the Potsdam data set and the Vaihingen data set. In this section, we provide a short description of both data sets and then present the design of experiments providing implementation details.

A. Descriptions of Data Sets

We employ two publicly available data sets to evaluate the proposed methods. The first data set is the Potsdam data set [10], which consists of 38 true orthophoto (TOP) tiles and the corresponding DSMs collected from a historic city with large building blocks; 24 imageries are used for training and the remaining 14 for testing. There are four spectral bands in each TOP image (red, green, blue, and near-infrared) and one band in each DSM. All data files have the same spatial size, equal to 6000 $\times$ 6000 pixels. The ground sampling distance (GSD) of this data set is 5 cm. The reference data are labeled according to six land-cover types: impervious surfaces, building, low vegetation, tree, car, and clutter/background.

The second data set is the Vaihingen data set [10], which contains 33 TOP tiles and the corresponding DSMs collected from a small village; 16 images are used for training and the remaining 17 ones for testing. Different from the Potsdam data set, each TOP in the Vaihingen data set contains three spectral bands (near-infrared, red, and green bands) and one DSM band. The spatial size of the images varies from 1996 $\times$ 1995 pixels to 3816 $\times$ 2550 pixels. The GSD of this data set is 9 cm. The reference data are divided into the same six categories as the Potsdam data set.

B. Design of Experiments

Following the evaluation method provided by the data publisher [10] and used in literature [13], [21], [42], three evaluation metrics are used to evaluate the performance of methods, i.e., overall accuracy (OA), per-class F1 score and average F1 score. OA is calculated by dividing the correctly classified number of pixels with the total number of pixels. The F1 score for a certain class is defined as the harmonic mean of precision and recall

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \tag{6}$$

The same preprocessing, data augmentation, and weight initialization settings have been used in all the experiments.

TABLE I

Results of the Ablation Study on the Potsdam Data Set. (*) Low-Feat Indicates the Use of Low-Level Features

Method	low-feat*	PAM	AEM	mean F1	OA
FCN				88.66	89.42
FCN+PAM		$\surd$		89.03	89.61
FCN	$\surd$			91.23	89.58
FCN+PAM	$\surd$	$\surd$		91.76	90.65
FCN+AEM	$\surd$		$\surd$	91.78	90.60
LANet	$\surd$	$\surd$	$\surd$	**91.95**	**90.84**

The DSMs are concatenated with TOPs as input data, so that we obtain five channels for the Potsdam data set and four channels for the Vaihingen data set. Due to the limitation of computational resources, the input data are cropped using a 512 $\times$ 512 window during the training phase. However, the prediction for the test set is performed whole-imagewise to obtain an accurate evaluation of the compared methods. Random-flipping and random-cropping operations are conducted during each iteration of the training phase as an augmentation approach. We use ResNet50 as the backbones for all compared networks with the pretrained weight for Pascal VOC data set loaded from the PyTorch library. Following the design of DeepLabv3+ [27], we choose the output features of the first convolutional block of ResNet50 as the low-level features in the implementation. This has been done considering as empirical criterion a spatial scaling rate of the features equal to 1/4. Considering the different GSD of the two data sets, the downsampling stride for the Potsdam data set is set to 32, while for the Vaihingen data set it is set to 16. The networks are implemented with PyTorch, and the experiments are conducted on a server with NVIDIA Quadro P6000 23GB GPU.

V. Experimental Results

In this section, we present the tests of the proposed modules through an ablation study. Then, we compare the proposed LANet with state-of-the-art methods and conclude our experimental validation.

A. Ablation Study

In order to verify the effectiveness of the proposed modules, ablation studies have been conducted on the two data sets. FCN (ResNet-50) is used as the baseline network for comparison. Since the proposed LANet uses low-level features, the effect of considering low-level features has also been measured.

Table I shows the results of the ablation study on the Potsdam data set. Three groups of observations can be done from the results. When no low-level features are involved in the decoding stage, the use of only one PAM (added on top of the FCN) increases the OA of 0.19%. With the inclusion of low-level features (concatenated with high-level features), the OA of the baseline FCN increases by only 0.16%. However, when two PAMs are added to process the high-level and low-level features separately, the OA increases by

TABLE II
RESULTS OF THE ABLATION STUDY ON THE VAIHINGEN DATA SET. (*) LOW-FEAT INDICATES THE USE OF LOW-LEVEL FEATURES

Method	low-feat*	PAM	AEM	mean F1	OA
FCN				86.14	88.66
FCN+PAM		√		86.42	88.68
FCN	√			86.52	88.84
FCN+PAM	√	√		87.49	89.36
FCN+AEM	√		√	86.80	89.05
LANet	√	√	√	**88.09**	**89.83**

another 1.07%. When the proposed AEM is used instead to enhance low-level features, the OA increases by 1.02%. With the use of both PAM and AEM, the proposed LANet increases the OA and average F1 compared with the baseline FCN (with the use of low-level features) of 1.26% and 0.72%, respectively.

The results of the ablation study on the Vaihingen data set are presented in Table II. The inclusion of low-level features improves the OA of the baseline FCN of 0.18%. However, the use of both low-level features and PAM brings an increase of 0.7% on OA and 1.35% on average F1. The use of low-level features and AEM brings an increase of 0.39% on OA and 0.63% on average F1. Under the condition that low-level features are considered, the proposed LANet improves the average F1 score and OA by about 1.57% and 0.99%, respectively.

B. Qualitative Analysis of Features

To visually confirm the effectiveness of the proposed modules, we present comparisons of the segmented features generated independently before and after the use of the proposed modules. Fig. 5 shows the effect of applying the PAM module on high-level features. Since high-level layers already have relatively large receptive field before using the PAM, the enhancement is not significant. However, one can still observe that some of the meaningless small segments are removed, and the segmentation of easily confused areas is improved.

Fig. 6 shows the changes of the segmented low-level features before and after the sequenced use of PAM and AEM. In the original low-level feature maps, pixels are only related to their neighborhoods due to the limitation of the small receptive field. This leads to fragmented results and confusion of object classes. However, after the enhancement obtained with the proposed modules, the semantic representation of low-level features is significantly improved. The pixels are classified based on not only the surface type of objects but also the context information. Moreover, one can verify from the clearly segmented boundaries the spatial details of low-level features.

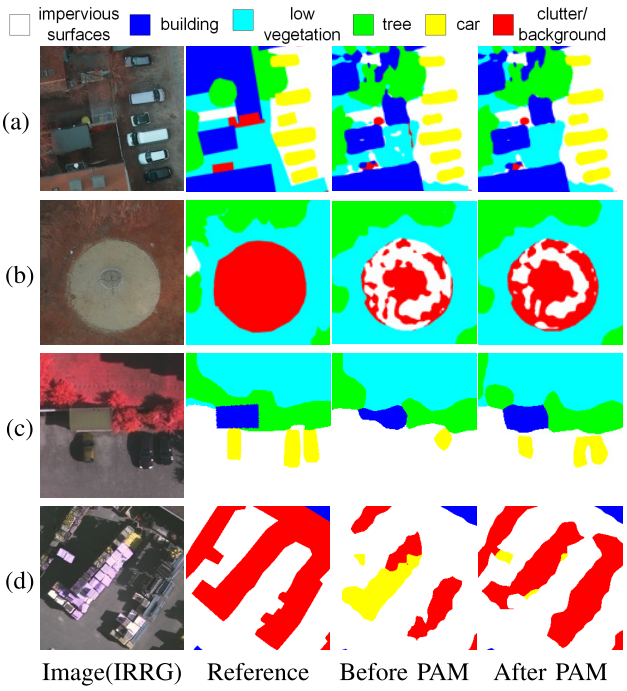


Fig. 5. Comparison of segmented high-level features before and after the use of PAM. (a) and (b) are selected from the Potsdam data set. (c) and (d) are selected from the Vaihingen data set.

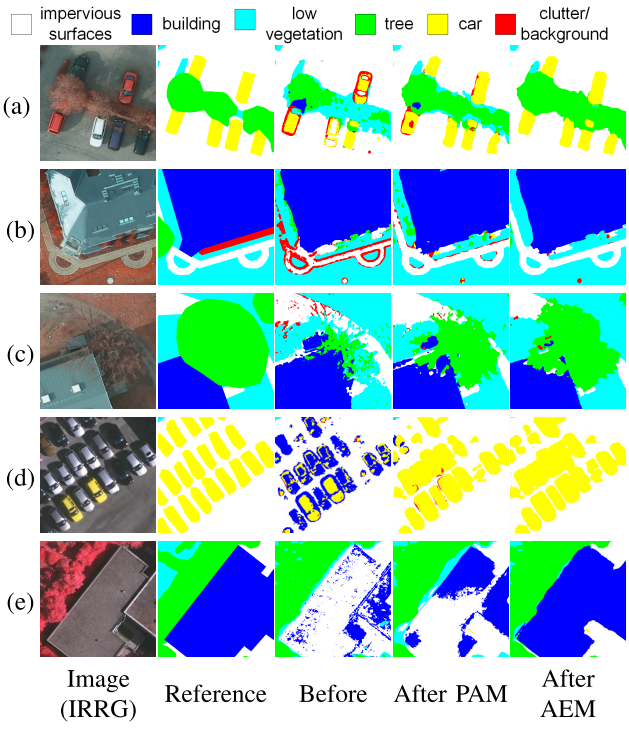


Fig. 6. Comparison of segmented low-level features before and after the use of PAM and AEM. (a)–(c) are selected from the Potsdam data set. (d) and (e) are selected from the Vaihingen data set.

C. Quantitative Comparison With State-of-the-Art Methods

Comparisons are made between the proposed LANet and approaches presented in the literature. All the tested approaches use the same backbone network (resnet50) and conduct the prediction on full-size test data. The experiments consider several recent works that have used the attention mechanism, including the SE block [9], the BAM [32], the CBAM [31], the GloRe [35], and the DANet [34]. The PSPNet [43] and DeepLabv3+ [27] with receptive-field-enlarging designs are also included in the comparisons.

TABLE III
RESULTS IN TERMS OF PER-CLASS F1 SCORE, AVERAGE F1 SCORE AND OA (POTSDAM DATA SET)

Method	Per-class F1 Score (%)					Average F1 (%)	OA (%)
	Impervious Surface	Building	low vegetation	Tree	Car		
FCN	91.46	96.63	85.99	86.94	82.28	88.66	89.42
FCN+SE [9]	91.47	96.57	86.21	87.51	81.07	88.56	89.55
FCN+BAM [32]	90.43	94.97	85.84	87.47	85.63	88.87	88.83
FCN+CBAM [31]	91.37	96.49	86.00	87.40	83.22	88.89	89.46
FCN+GloRe [35]	91.55	96.54	86.17	87.42	82.69	88.87	89.57
DANet [34]	91.61	96.44	86.11	**88.04**	83.54	89.14	89.72
PSPNet [28]	91.61	96.30	86.41	86.84	91.38	90.51	89.45
DeepLabv3+ [27]	92.35	96.77	85.22	86.79	93.58	90.94	89.74
Proposed LANet	**93.05**	**97.19**	**87.30**	**88.04**	**94.19**	**91.95**	**90.84**

TABLE IV
RESULTS IN TERMS OF PER-CLASS F1 SCORE, AVERAGE F1 SCORE AND OA (VAIHINGEN DATA SET)

Method	Per-class F1 Score (%)					Average F1 (%)	OA (%)
	Impervious Surface	Building	low vegetation	Tree	Car		
FCN	90.98	94.10	81.25	87.58	76.80	86.14	88.66
FCN+SE [9]	90.43	93.95	81.33	87.50	63.33	83.31	88.27
FCN+BAM [32]	90.77	94.01	81.54	87.78	71.76	85.17	88.62
FCN+CBAM [31]	90.86	94.03	81.16	87.63	76.26	85.99	88.61
FCN+GloRe [35]	90.57	93.99	81.28	87.49	70.09	84.68	88.41
DANet [34]	90.78	94.11	81.40	87.42	75.85	85.91	88.59
PSPNet [28]	91.44	94.38	81.52	87.91	78.02	86.65	88.99
DeepLabv3+ [27]	91.35	94.34	81.32	87.84	78.14	86.60	88.91
Proposed LANet	**92.41**	**94.90**	**82.89**	**88.92**	**81.31**	**88.09**	**89.83**

TABLE V
COMPARISON OF MODEL SIZE AND CALCULATIONS EXPRESSED IN TERMS OF PARAMS (MB) AND FLOPS (GBPS), RESPECTIVELY

Method	FCN	FCN+SE	FCN+BAM	FCN+CBAM	FCN+GloRe	DANet	PSPNet	DeepLabv3+	Proposed LANet
Params (Mb)	23.79	23.80	24.15	23.97	23.81	47.73	46.94	39.73	23.80
FLOPS (Gbps)	21.95	21.95	22.38	21.95	21.95	28.01	31.67	30.72	21.98

Tables III and IV report the quantitative results on the Potsdam data set and the Vaihingen data set, respectively. Compared with the baseline FCN, the use of most attention-based modules, such as SE, BAM, and CBAM, does not involve noticeable performance improvement. The use of the SE-block even causes decreases in terms of F1 scores, especially for the car class. This is because the channelwise descriptors are calculated on the whole feature map, and the classes that account for a small portion of total pixels are suppressed. This proves our assumption that the global-level calculation of attention descriptors is not suitable for processing large-size RSIs. The DANet with a spatial dependence modeling design improves the OA of 0.3% on the Potsdam data set, but there is a decrease of OA on the Vaihingen data set. DeepLabv3+, which uses both low-level features and dilated convolutions, has good performance in F1 scores. The proposed LANet, with the use of both context aggregation and attention embedding strategies, shows significant advantages over the compared methods. It shows the best performance in terms of both average F1 score and OA, and obtains better F1 scores in all the predicted categories.

To evaluate the required amount of calculation resources of the compared models, Table V represents the values of two metrics, i.e., the size of parameters and the floating point operations per second (FLOPS) (for processing each batch of data). The calculations are based on the input channels and pooling stride of processing the Potsdam data set. Overall, the attention-based methods (SE, BAM, CBAM, and GloRe) are lightweight, whereas the context-aggregation-based methods (PSPNet and DeepLabv3+) require more calculations. The proposed LANet does not significantly increase the calculations compared to the baseline FCN.

D. Qualitative Analysis of the Semantic Segmentation Results

Examples of the predicted patches on the two data sets are shown in Fig. 7. The segmentation maps provided by FCN are ambiguous (especially at the contours of objects) due to the

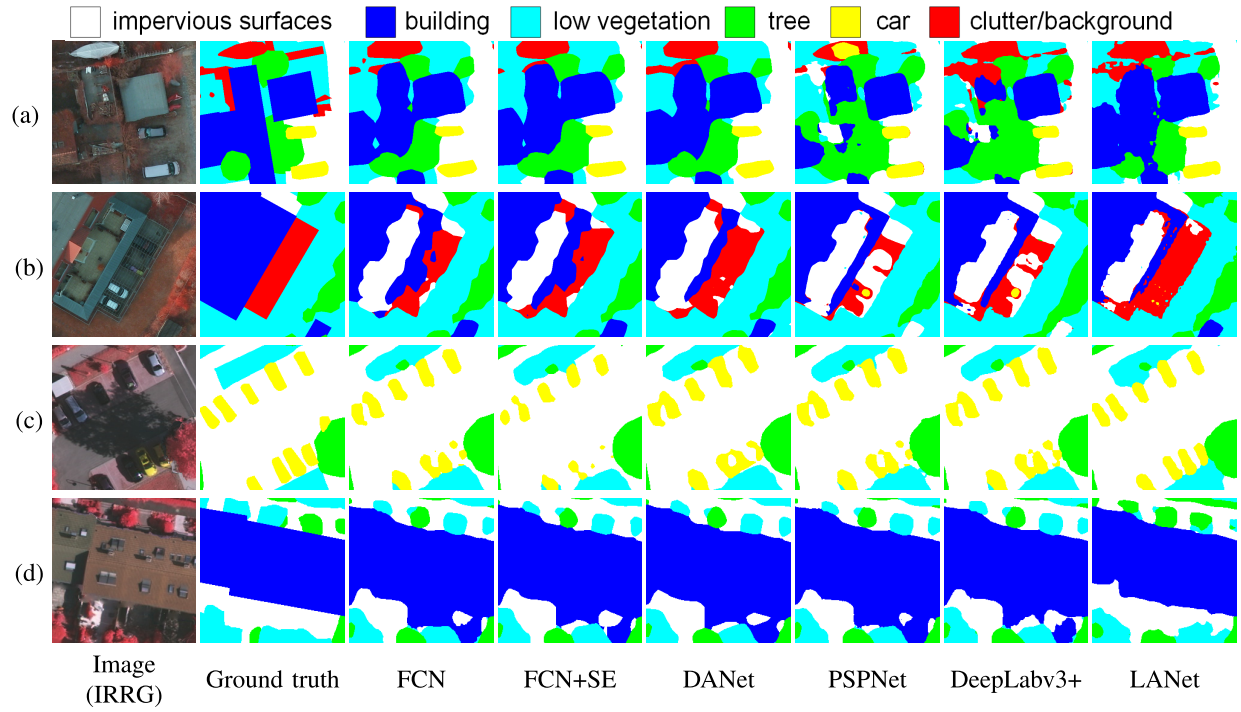


Fig. 7. Examples of semantic segmentation results. (a) and (b) are selected from the Potsdam data set. (c) and (d) are selected from the Vaihingen data set.

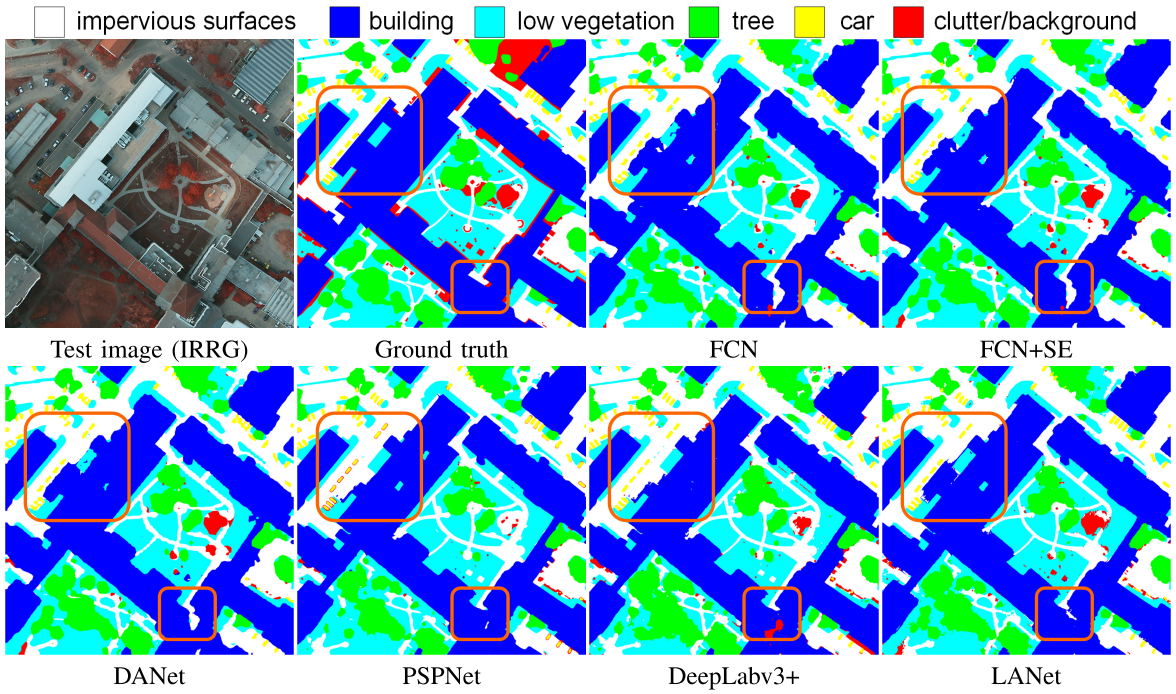


Fig. 8. Example of large-size semantic segmentation results (Potsdam data set). Major differences are marked with orange squares (zoomed-in view for more details).

loss of spatial information. The direct use of attention-based methods (e.g., SE and DANet) brings limited improvements. The context-aggregation based approaches (e.g., PSPNet and Deeplabv3+) not only show improvements in segmenting confusing areas but also produce many fragmented segments. With the aggregation of local contextual information, the proposed LANet not only significantly reduces the errors but also better preserves the spatial details. Specifically, the discrimination between cars and impervious surfaces, as well as between buildings and clutters has been greatly improved. There are also noticeable improvements in preserving the boundaries of objects. Figs. 8 and 9 show the large-size predictions on the Potsdam data set and the Vaihingen data set, respectively. Observing from a larger scale, the results of DANet are more reliable compared to FCN, but still suffer from low spatial accuracy; the results of PSPNet and Deeplabv3+ are more fragmented. As a comparison, in the predicted maps of the proposed LANet, there are less false alarms in the surrounding

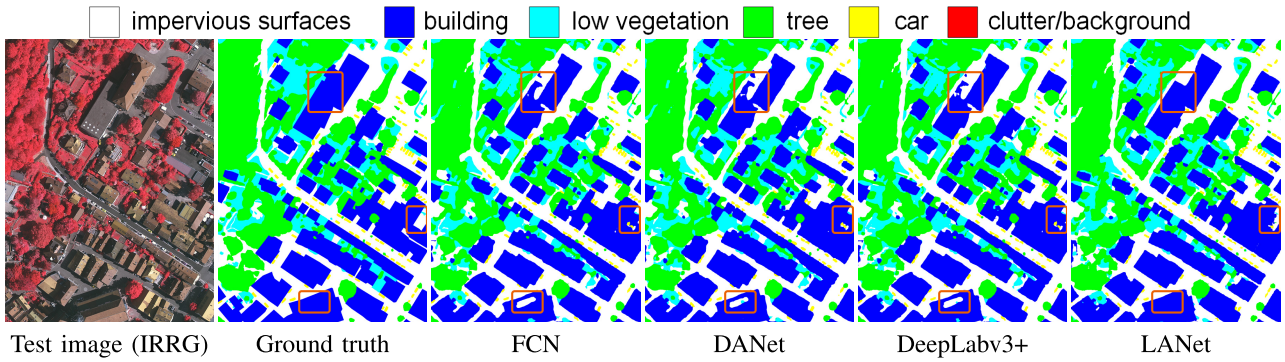


Fig. 9. Example of large-size semantic segmentation results (Vaihingen data set). Major differences are marked with orange squares (zoomed-in view for more details).

areas of buildings, which can be attributed to the embedding of contextual information. Meanwhile, the segmentation of small objects (e.g., cars, paths, and small clutters) is more accurate, which is due to the incorporation of enhanced low-level features. This points out that the proposed method improves both the discrimination of critical categories and the preservation of spatial details.

VI. Conclusion

The attention mechanism is a commonly used strategy in CNNs for aggregating context information in images. However, RSIs have a large spatial size and a relatively small number of classes with respect to natural images and do not express clear image-level scene information, which limits the use of the attention mechanism. In this article, we present a LANet that employs patch-level scene information to improve the semantic segmentation of RSIs. Specifically, two modules are proposed for enhancing the representation of features based on the exploitation of local attention: 1) the PAM enhances the encoding of context information based on the patchwise calculation of local descriptors and 2) the AEM embeds attention from high-level layers into low-level ones to enrich their semantic information.

Experimental results on two benchmark RSI data sets (Potsdam and Vaihingen data sets) show that the proposed approach greatly improves the representation of extracted features. The aggregation of local attention (using the PAM) is beneficial for classifying the easily confused areas, while the embedding of attentions from high-level features to low-level ones improves the preservation of spatial details. Comparative results show that the proposed LANet outperforms other global-attention- and receptive-field-enlarging-based approaches. However, one of the remaining problems in the semantic segmentation of RSIs is that the objects in segmented maps are still more-or-less fragmented, especially at the boundaries. To conquer this limitation, as a further development of this article, we plan to study feature encoding strategies to improve the embedding of high-level features in the network.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[2] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

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

[5] 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*. Springer, 2015, pp. 234–241.

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

[7] 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.

[8] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. CVPR*, 2019, pp. 12416–12425.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[10] ISPRS. *2D Semantic Labeling Contest—Potsdam*. Accessed: Sep. 4, 2018. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[11] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 36–43.

[12] 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 Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4507–4521, Aug. 2018.

[13] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.

[14] 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.*, to be published.

[15] 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, 2018.

[16] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, 2019.

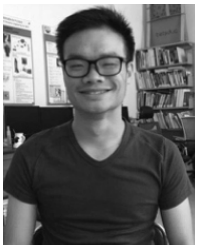
[17] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.

[18] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint learning from Earth observation and OpenStreetMap data to get faster better semantic maps," in *Proc. CVPR*, 2017, pp. 67–75.

[19] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.
[20] Z. Cao *et al.*, "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019.
[21] 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.
[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 1125–1134.
[23] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation," in *Proc. CVPR*, 2019, pp. 2417–2426.
[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 2117–2125.
[25] K. Min and J. J. Corso, "TASED-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proc. CVPR*, 2019, pp. 2394–2403.
[26] X. Jiang *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. CVPR*, 2019, pp. 6133–6142.
[27] 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.
[28] 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.
[29] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. CVPR*, 2017, pp. 3156–3164.
[30] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. CVPR*, 2018, pp. 7151–7160.
[31] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.
[32] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*. [Online]. Available: http://arxiv.org/abs/1807.06514
[33] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. ECCV*, 2018, pp. 267–283.
[34] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. CVPR*, 2017, pp. 299–307.
[35] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. CVPR*, 2019, pp. 433–442.
[36] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, p. 83, 2019.
[37] W. Cui *et al.*, "Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model," *Remote Sens.*, vol. 11, no. 9, p. 1044, 2019.
[38] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180
[39] Y. Su, Y. Wu, M. Wang, F. Wang, and J. Cheng, "Semantic segmentation of high resolution remote sensing image based on batch-attention mechanism," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3856–3859.
[40] 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, 2020.
[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.
[42] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-ShapeNetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, p. 522, 2017.
[43] 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, Jan. 2015.

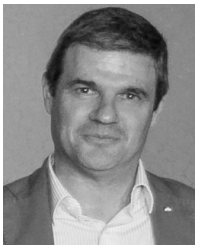
Lei Ding (Student Member, IEEE) received the B.S. degree in measurement and control engineering and the M.S. degree in photogrammetry and remote sensing from the University of Information Engineering, Zhengzhou, China, in 2013 and 2016, respectively. He is pursuing the Ph.D. degree with RSLab, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.

His research interests are related to remote sensing image processing and machine learning.



Hao Tang received the master's degree in computer application technology from the School of Electronics and Computer Engineering, Peking University, Beijing, China, in 2016. He is pursuing the Ph.D. degree with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.

He is a member of the Multimedia and Human Understanding Group (MHUG) led by Prof. Nicu Sebe, University of Trento. His research interests are machine learning, (deep) representation 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 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, inlcuding 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 (ESA), and the Science Lead for the High Resolution Land Cover project in the framework of the Climate Change Initiative of ESA. He is the author (or coauthor) of 259 scientific publications in referred international journals (193 in IEEE journals), more than 330 articles in conference proceedings, and 22 book chapters. He is editor/co-editor of 18 books/conference proceedings and one scientific book. He was invited as keynote speaker in more than 40 international conferences and workshops. Since 2009 he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS), where since 2019 he has been Vice-President for Professional Activities.

Dr. Bruzzone ranked First Place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, in July 1998. Since 1998, he has been a recipient of many international and national honors and awards, including the recent the IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, and the 2019 WHISPER Outstanding Paper Award. He 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 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 is the Founder of the *IEEE Geoscience and Remote Sensing Magazine* for which he has been the Editor-in-Chief from 2013 to 2017. He is an Associate Editor of the IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING. He was a Distinguished Speaker of the IEEE GRSS from 2012 to 2016. His articles are highly cited, as proven from the total number of citations (more than 31600) and the value of the H-index (83; source: Google Scholar).