ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery

Rui Li [1], Shunyi Zheng [1], Ce Zhang [2, 3], Chenxi Duan [4, *], Libo Wang [1],

and Peter M. Atkinson [2,5,6]

1) School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China.

2) Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

3) UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK.

4) The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China.

5) Geography and Environmental Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

6) Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A Datun Road, Beijing 100101, China.

*Corresponding author.

Abstract—Semantic segmentation of remotely sensed imagery plays a critical role in many real-world applications, such as environmental change monitoring, precision agriculture, environmental protection, and economic assessment. Following rapid developments in sensor technologies, vast numbers of fine-resolution satellite and airborne remote sensing images are

now available, for which semantic segmentation is potentially a valuable method. However, because of the rich complexity and heterogeneity of information provided with an ever-increasing spatial resolution, state-of-the-art deep learning algorithms commonly adopt complex network structures for segmentation, which often result in significant computational demand. Particularly, the frequently-used fully convolutional network (FCN) relies heavily on fine-grained spatial detail (fine spatial resolution) and contextual information (large receptive fields), both imposing high computational costs. This impedes the practical utility of FCN for real-world applications, especially those requiring real-time data processing. In this paper, we propose a novel Attentive Bilateral Contextual Network (ABCNet), a lightweight convolutional neural network (CNN) with a spatial path and a contextual path. Extensive experiments, including a comprehensive ablation study, demonstrate that ABCNet has strong discrimination capability with competitive accuracy compared with state-of-the-art benchmark methods while achieving significantly increased computational efficiency. Specifically, the proposed ABCNet achieves a 91.3% overall accuracy (OA) on the Potsdam test dataset and outperforms all lightweight benchmark methods significantly. The code is freely available at https://github.com/lironui/ABCNet.

Index Terms—Semantic Segmentation, Attention Mechanism, Bilateral Architecture, Convolutional Neural Network, Deep Learning

1. Introduction

Driven by the rapid development of Earth observation technology, massive numbers of remotely sensed images at fine spatial resolution are commercially available for a variety of applications, such as image classification (Lyons et al., 2018; Maggiori et al., 2016), object

detection (Li et al., 2017; Xia et al., 2018) and semantic segmentation (Kemker et al., 2018; Zhang et al., 2019a). The re-visit capabilities of orbital sensors facilitate continuous monitoring of the land surface, ocean, and atmosphere (Duan and Li, 2020). Fine-resolution remotely sensed images are rich in information and contain substantial spatial detail for land cover and land use classification and segmentation. Different automatic and semi-automatic methods have been developed to identify land cover and land use categories by exploiting spectral and spectral-spatial features within remote sensing images (Gong et al., 1992; Ma et al., 2017; Tucker, 1979; Zhong et al., 2014; Zhu et al., 2017). However, these traditional approaches rely on handcrafting features and information transformation, which commonly fail to adequately capture the contextual information contained abundantly within images, and are often limited in their flexibility and general adaptability (Li et al., 2020; Tong et al., 2020). This is especially true given the detailed structural and contextual information provided at a very fine spatial resolution. Meanwhile, recent developments in deep learning, and deep convolutional neural network (CNN), in particular, have replaced feature engineering with high-level non-linear feature representations created end-to-end, hierarchically, and in an automatic fashion. This has had a transformative impact on information understanding and semantic characterization from fine-resolution remotely sensed imagery (Li et al., 2021b; Zheng et al., 2020).

Semantic segmentation, which assigns each pixel in an image to a particular category, has become one of the most important approaches for ground feature interpretation, playing a pivotal role in different application scenarios (Wang et al., 2021), such as precision agriculture (Griffiths et al., 2019; Picoli et al., 2018), environmental protection (Samie et al., 2020; Yin et al., 2018)

and economic assessment (Zhang et al., 2020; Zhang et al., 2019a). The fully convolutional network (FCN) was demonstrated to be the first effective end-to-end CNN structure for semantic segmentation (Long et al., 2015). Restricted by the oversimplified design of the decoder, the results of FCN, although encouraging in principle, are presented at a coarse resolution. Subsequently, more elaborate encoder-decoder structures, such as U-Net, have been proposed, with two symmetric paths: a contracting path for extracting features and an expanding path for achieving accurate results through precise positioning (Badrinarayanan et al., 2017; Li et al., 2021a; Ronneberger et al., 2015). The per-pixel classification is often ambiguous in the presence of only local information for semantic segmentation, while the task becomes much simpler if global contextual information, from the whole image, is available (as shown in Fig. 1). Therefore, to guarantee the accuracy of segmentation, global contextual information and multiscale semantic features were utilized comprehensively to differentiate semantic categories at different spatial scales. Through the spatial pyramid pooling module, the pyramid scene parsing network (PSPNet) aggregated contextual information across different regions (Zhao et al., 2017). The dual attention network (DANet) applied the dot-product attention mechanism to extract abundant contextual relationships (Fu et al., 2019). Subject to an enormous memory and computational demand, DANet simply attached the dot-product attention mechanism at the lowest layer without capturing the long-range dependencies from the larger feature maps in the higher layers. DeeplabV3 adopted atrous convolution to mine the multiscale features (Chen et al., 2017a) and a simple, yet useful, decoder module was added in DeepLabV3+ to further refine the segmentation results (Chen et al., 2018a).

Fig. 1. Illustration of global and local contextual information.

The extraction of global contextual information and the exploitation of large-scale feature maps are computationally expensive (Chen et al., 2017b; Diakogiannis et al., 2020b; Li et al., 2021b). Therefore, a series of lightweight networks have been developed to accelerate the computation while maintaining the trade-off between accuracy and efficiency (Hu et al., 2020; Oršić and Šegvić, 2021; Romera et al., 2017; Yu et al., 2018; Zhuang et al., 2019). For example, the asymmetric convolution used in ERFNet factorized the standard 3 × 3 convolutions into a 1 × 3 convolution and a 3 × 1 convolution, saving approximately 33% of the computational cost (Romera et al., 2017). By exploiting spatial correlations and cross-channel correlations, respectively, BiseNet achieved depth-wise separable convolution (Yu et al., 2018), which further reduced the consumption of standard convolution (Chollet, 2017). Multi-scale encoder-decoder branch pairs with skip connections were studied in ShelfNet (Zhuang et al., 2019), where a shared-

weight strategy was harnessed in the residual block to reduce the number of parameters without sacrificing accuracy. For non-local context aggregation, FANet employed the fast attention module in efficient semantic segmentation (Hu et al., 2020). SwiftNet explored the effectiveness of pyramidal fusion in compact architectures (Oršić and Šegvić, 2021). However, the CNN is designed to extract local patterns and lacks the ability to model global context in its nature. More severely, as lightweight networks normally adopted relatively shallow backbones, the capacity of those networks to extract global contextual information is further limited.

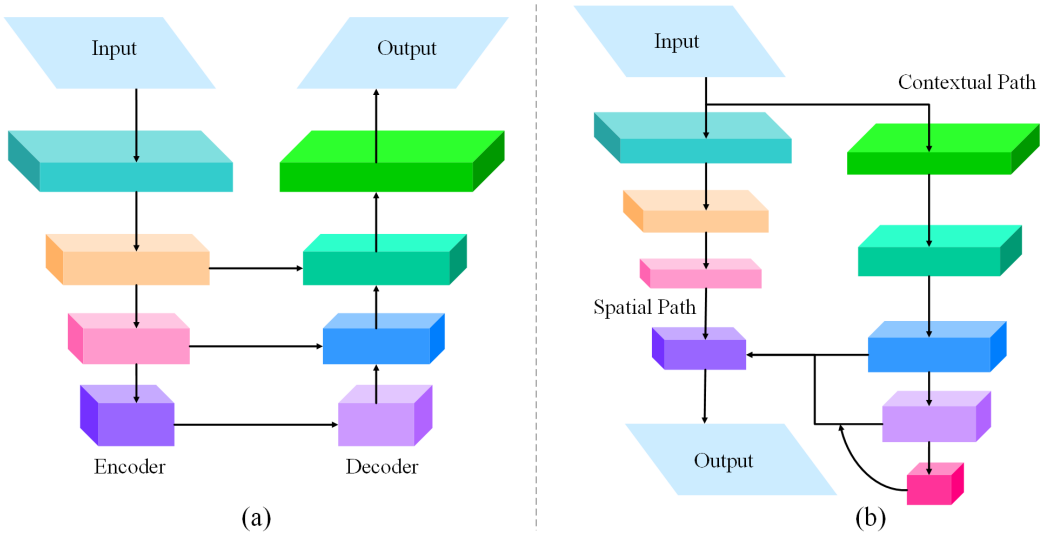

Fig. 2. Illustration of (a) the encoder-decoder structure and (b) the bilateral architecture.

Due to the limited capacity of lightweight networks to extract global contextual information, there is a huge gap in accuracy between lightweight networks and state-of-the-art deep models, which limits their applicability to fine-resolution remotely sensed images. The dot-product attention mechanism, as a powerful approach that can capture long-range dependencies, is potentially an ideal solution to address this issue (Vaswani et al., 2017). However, the memory and computational costs of the dot-product attention mechanism increase quadratically with an

increase in the spatio-temporal size of the input, which runs counter to the aim of lightweight networks. Encouragingly, previous researches on linear attention (Katharopoulos et al., 2020; Li et al., 2021b) reduce the complexity of the dot-product attention mechanism from $O(N^2)$ to $O(N)$, with a significant increase in computational speed, while maintaining high accuracy.

In this paper, we aim to further increase segmentation accuracy while ensuring the efficiency of semantic segmentation simultaneously. We address this challenge by modeling the global contextual information using the linear attention mechanism. Specifically, we propose an Attentive Bilateral Contextual Network (ABCNet) to realize efficient semantic segmentation of fine-resolution remote sensing images. Following the design philosophy of BiSeNet (Yu et al., 2018), we design the ABCNet based on a bilateral architecture: a spatial path to retain the abundant spatial detail and a contextual path to capture the global contextual information. As the features generated by the two paths are quite disparate semantically, we further design a feature aggregation module (FAM) to fuse those features. The comparison between the conventional encoder-decoder structure and the bilateral architecture used in the proposed ABCNet can be seen in Fig. 2. The main contributions are two-fold. On the one hand, we propose a novel approach for efficient semantic segmentation of fine-resolution remotely sensed imagery, i.e., ABCNet with spatial and contextual paths. On the other hand, we design two specific modules: an attention enhancement module (AEM) for exploring long-range contextual information, and a feature aggregation module (FAM) for fusing the features obtained by the two paths. A thorough benchmark comparison was undertaken against the state-of-the-art to demonstrate the effectiveness of the proposed ABCNet.

2. Related Work

1) Context information extraction

Context is critically important for semantic segmentation and, thus, tremendous effort has been made to extract such information in an intelligent manner. The dilated or atrous convolution (Chen et al., 2014; Yu and Koltun, 2015) has been demonstrated to be an effective approach for enlarging receptive fields without shrinking spatial resolution. Besides, the encoder-decoder architecture (Ronneberger et al., 2015), which merges high-level and low-level features via skip connections, is an alternative for extracting spatial context. Based on the encoder-decoder framework or dilation backbone, some research has focused on exploring the use of spatial pyramid pooling (SPP) (He et al., 2015). For example, the pyramid pooling module (PPM) in PSPNet is composed of convolutions with kernels of four different sizes (Zhao et al., 2017), while DeepLab v2 (Chen et al., 2018a), equipped with the atrous spatial pyramid pooling (ASPP) module, groups parallel atrous convolution layers with varying dilation rates. However, certain limitations persist in SPP. Particularly, the SPP with the standard convolution faces a dilemma when expanding the receptive field with a large kernel size. The above operations are normally accompanied by a very large number of parameters. The SPP with small kernels (e.g. ASPP), on the other hand, lacks sufficient connection between adjacent features, and the gridding problem (Wang et al., 2018a) occurs when the field is enlarged by a dilated convolutional layer. In contrast, the dot-product attention mechanism has the powerful ability to model long-range dependencies, which enables contextual information extraction at a global scale.

2) Dot-Product attention mechanism

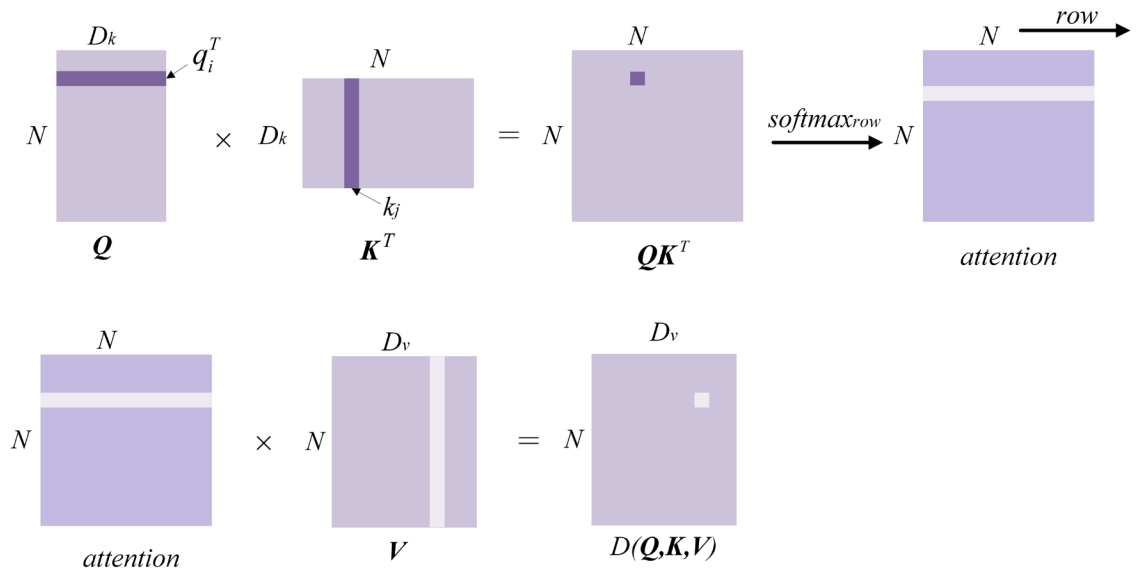

Fig. 3. Illustration of the calculation of dot-product attention mechanism.

Let *H*, *W,* and C denote the height, weight, and channels of the input, respectively. The input feature is defined as $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{N \times C}$, where $N = H \times W$. Initially, the dot-product attention mechanism utilizes three projected matrices $\boldsymbol{W}_q \in \mathbb{R}^{D_x \times D_k}$, $\boldsymbol{W}_k \in \mathbb{R}^{D_x \times D_k}$, and $\boldsymbol{W}_v \in \mathbb{R}^{D_x \times D_v}$ to generate the corresponding *query* matrix $\boldsymbol{Q}$, the *key* matrix $\boldsymbol{K}$, and the *value* matrix $\boldsymbol{V}$:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q \in \mathbb{R}^{N \times D_k}; \\ \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k \in \mathbb{R}^{N \times D_k}; \\ \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v \in \mathbb{R}^{N \times D_v}. \end{cases} \tag{1}$$

The graphical representation of the dot-product attention mechanism can be seen in Fig. 3. The dimensions of $\boldsymbol{Q}$ and $\boldsymbol{K}$ are identical, and all vectors in this section are column vectors by default. Accordingly, a normalization function ρ is employed to measure the similarity between the *i*-th *query* feature $\boldsymbol{q}_i^T \in \mathbb{R}^{D_k}$ and the *j*-th *key* feature $\boldsymbol{k}_j \in \mathbb{R}^{D_k}$ as $\rho(\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j) \in \mathbb{R}^1$. As the *query* feature and *key* feature are generated via different layers, the similarities between $\rho(\boldsymbol{q}_i^T \cdot \boldsymbol{k}_j)$ and $\rho(\boldsymbol{q}_j^T \cdot \boldsymbol{k}_i)$ are not identical. Therefore, the $N \times N$ $\boldsymbol{Q}\boldsymbol{K}^T$ matrix model the long-range

dependency between each pixel pair in the input feature maps, where the pixel at *j*-th row and *i*-th column measures the *i*-th position's impact on *j*-th position. In other words, the long-range global contextual information between every pixel of the input can be fully modeled by the $N\times N$ matrix $\boldsymbol{QK}^T$. By calculating similarities between all pairs of pixels in the input feature maps and taking the similarities as weights, the dot-product attention mechanism generates the *value* at position *i* by aggregating the *value* features from all positions using weighted summation:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \rho(\boldsymbol{QK}^T)\boldsymbol{V}. \tag{2}$$

Softmax is frequently used as the normalization function:

$$\rho(\boldsymbol{QK}^T) = softmax_{row}(\boldsymbol{QK}^T), \tag{3}$$

where $softmax_{row}$ indicates that the softmax along each row of the matrix $\boldsymbol{QK}^T$.

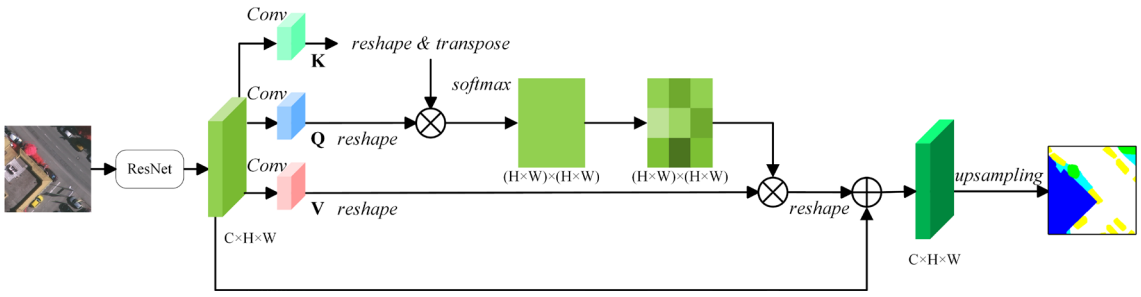

Fig. 4. Illustration of the dot-product attention mechanism utilized in computer vision.

By modeling the similarities between each pair of positions of the input, the global dependencies in the features can be extracted thoroughly by $\rho(\boldsymbol{QK}^T)$. The dot-product attention mechanism was initially designed for machine translation (Vaswani et al., 2017), while the non-local module (Wang et al., 2018b) was introduced and modified for computer vision (Fig. 4). Based on the dot-product attention mechanism, as well as its variants, different attention-based networks have been proposed to address the semantic segmentation task. Inspired by the non-

local module (Wang et al., 2018b), the double attention networks (A^2-Net) (Chen et al., 2018b), dual attention network (DANet) (Fu et al., 2019), and object context network (OCNet) (Yuan and Wang, 2018) were proposed successively for scene segmentation by exploring the long-range dependencies. Furthermore, Bello et al. (2019) augmented convolutional operators with attention mechanisms, while Zhang et al. (2019c) incorporated the attention mechanism into the generative adversarial network. Lu et al. (2019) extended the attention mechanism to CO-attention Siamese Network (COSNet) for unsupervised video object segmentation. Recently, Diakogiannis et al. (2020a) improved the attention mechanism and proposed the fractal Tanimoto attention layer for semantic change detection.

Although the introduction of attention boosts segmentation accuracy significantly, the huge resource-demand of the dot-product hinders its application to large inputs. Specifically, for $\boldsymbol{Q} \in \mathbb{R}^{N\times D_k}$ and $\boldsymbol{K}^T \in \mathbb{R}^{D_k\times N}$, the product between $\boldsymbol{Q}$ and $\boldsymbol{K}^T$ belongs to $\mathbb{R}^{N\times N}$, leading to $O(N^2)$ memory and computational complexity. Consequently, it is necessary to reduce the demand for computational resources of the dot-product attention mechanism. Substantial endeavors have been poured in aiming to alleviate the bottleneck to efficiency and push the boundaries of attention, including accelerating the generation process of the attention matrix (Huang et al., 2019a; Huang et al., 2019b; Yuan et al., 2019; Zhang et al., 2019b), pruning the structure of the attention block (Cao et al., 2019), and optimizing attention based on low-rank reconstruction (Li et al., 2019c).

3) **Generalization and simplification of the dot-product attention mechanism**

If the normalization function is set as softmax, the *i*-th row of the result matrix generated by

the dot-product attention mechanism can be written as:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{N} e^{\boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j} \boldsymbol{v}_j}{\sum_{j=1}^{N} e^{\boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j}}, \tag{4}$$

where $\boldsymbol{v}_j$ is *j*-th *value* feature.

Equation (4) can be rewritten and generalized to any normalization function as:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{N} \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{N} \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)}, \quad \mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) \geq 0. \tag{5}$$

$\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)$ can be expanded as $\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$ which measures the similarity between $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$, and equation (4) can be rewritten as equation (6) and simplified as equation (7):

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\sum_{j=1}^{N} \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)\boldsymbol{v}_j}{\sum_{j=1}^{N} \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)}, \tag{6}$$

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i = \frac{\phi(\boldsymbol{q}_i)^T \sum_{j=1}^{N} \varphi(\boldsymbol{k}_j)\boldsymbol{v}_j{}^T}{\phi(\boldsymbol{q}_i)^T \sum_{j=1}^{N} \varphi(\boldsymbol{k}_j)}. \tag{7}$$

If $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) = e^{\boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j}$, equation (5) is equivalent to equation (4). The vectorized form of equation (7) is:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\phi(\boldsymbol{Q})\varphi(\boldsymbol{K})^T\boldsymbol{V}}{\phi(\boldsymbol{Q})\sum_j \varphi(\boldsymbol{K})_{i,j}^T}. \tag{8}$$

As the softmax function is substituted for $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j) = \phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$, the order of the commutative operation can be altered, thereby avoiding multiplication between the reshaped *key* matrix $\boldsymbol{K}$ and *query* matrix $\boldsymbol{Q}$. In concrete terms, we can first compute the multiplication between $\varphi(\boldsymbol{K})^T$ and $\boldsymbol{V}$, and then multiply the result with $\boldsymbol{Q}$, leading to only $O(dN)$ time complexity and $O(dN)$ space complexity. The suitable $\phi(\cdot)$ and $\varphi(\cdot)$ enable the above scheme to achieve competitive performance with finite computational complexity (Katharopoulos et al., 2020; Li et al., 2021c).

4) **Linear attention mechanism**

In our previous research (Li et al., 2021b), we proposed a linear attention mechanism to replace the softmax function with the first-order approximation of the Taylor expansion, as in equation (9):

$$e^{\boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j} \approx 1 + \boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j. \tag{9}$$

To guarantee the above approximation to be nonnegative, $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$ are normalized by the l_2 norm, thereby ensuring $\boldsymbol{q}_i{}^T \cdot \boldsymbol{k}_j \geq -1$:

$$sim(\boldsymbol{q}_i, \boldsymbol{k}_j) = 1 + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right). \tag{10}$$

Thus, equation (5) can be rewritten as equation (11) and simplified as equation (12):

$$\boldsymbol{D}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N}} \left(\boldsymbol{1} + \left(\frac{\boldsymbol{q}_{\boldsymbol{i}}}{\|\boldsymbol{q}_{\boldsymbol{i}}\|_2}\right)^{\boldsymbol{T}} \left(\frac{\boldsymbol{k}_{\boldsymbol{j}}}{\|\boldsymbol{k}_{\boldsymbol{j}}\|_2}\right)\right) \boldsymbol{v}_{\boldsymbol{j}}}{\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N}} \left(\boldsymbol{1} + \left(\frac{\boldsymbol{q}_{\boldsymbol{i}}}{\|\boldsymbol{q}_{\boldsymbol{i}}\|_2}\right)^{\boldsymbol{T}} \left(\frac{\boldsymbol{k}_{\boldsymbol{j}}}{\|\boldsymbol{k}_{\boldsymbol{j}}\|_2}\right)\right)}, \tag{11}$$

$$D(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})_i = \frac{\sum_{j=1}^{N} \boldsymbol{v}_j + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T \sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right) \boldsymbol{v}_j^T}{N + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T \sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)}. \tag{12}$$

Equation (12) can be turned into a vectorized form:

$$D(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \frac{\sum_j \boldsymbol{V}_{i,j} + \left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\left(\left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)^T \boldsymbol{V}\right)}{N + \left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right) \sum_j \left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)_{i,j}^T}. \tag{13}$$

Since $\sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right) \boldsymbol{v}_j^T$ and $\sum_{j=1}^{N} \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)$ can be calculated and reused for each *query*, the time and memory complexity of the attention based on equation (13) is $O(dN)$. For more detailed information on the proposed attention mechanism, as well as its validity and efficiency, the reader is referred to (Li et al., 2021b).

5) Scaling attention mechanism

Besides dot-product attention, there exists another genre of techniques referred to as attention mechanisms in the literature. To distinguish it from the dot-product attention mechanism, we call them scaling attention. Unlike dot-product attention which models global dependencies from feature maps, scaling attention reinforces informative features and whittles information-lacking features. For example, Wang et al. (2017) proposed a residual attention network (RAN) which introduces the scaling attention mechanism inserted into deep residual networks. As a high-capacity structure, the residual attention is mainly built on max-pooling layers, convolutional layers, and residual units. In contrast, Hu et al. (2018) presented the squeeze-and-excitation (SE) module, a lightweight gating mechanism constructed on the global average pooling layer and linear layers, to calculate a scaling factor for each channel, thereby weighting the channels accordingly. The convolutional block attention module (CBAM) (Woo et al., 2018), selective kernel unit (SK unit) (Li et al., 2019b) and efficient channel attention module (ECA) (Wang et al., 2020) further boost the SE block's performance. Despite both names containing attention, the principles and purposes of dot-product attention and scaling attention are entirely divergent.

6) Efficient semantic segmentation

For many practical applications, efficiency is critical, and this is especially pertinent for real-time ($\geqslant$30FPS) scenarios such as autonomous driving. Therefore, huge efforts have been made to accelerate models for efficient semantic segmentation, by employing lightweight operations or down-sampling the input size. The utilization of lightweight convolutions (e.g., asymmetric

convolution and depth-wise separable convolution) is a common strategy for designing lightweight networks (Romera et al., 2017; Yu et al., 2018). The down-sampling of the input size is a trivial solution to speed up semantic segmentation by reducing the resolution of the input images, which inevitably results in the loss of information. To extract spatial details at the original resolution, some of the latest methods include a further shallow branch, forming a two-path architecture (Yu et al., 2020; Yu et al., 2018).

3. Attentive Bilateral Contextual Network

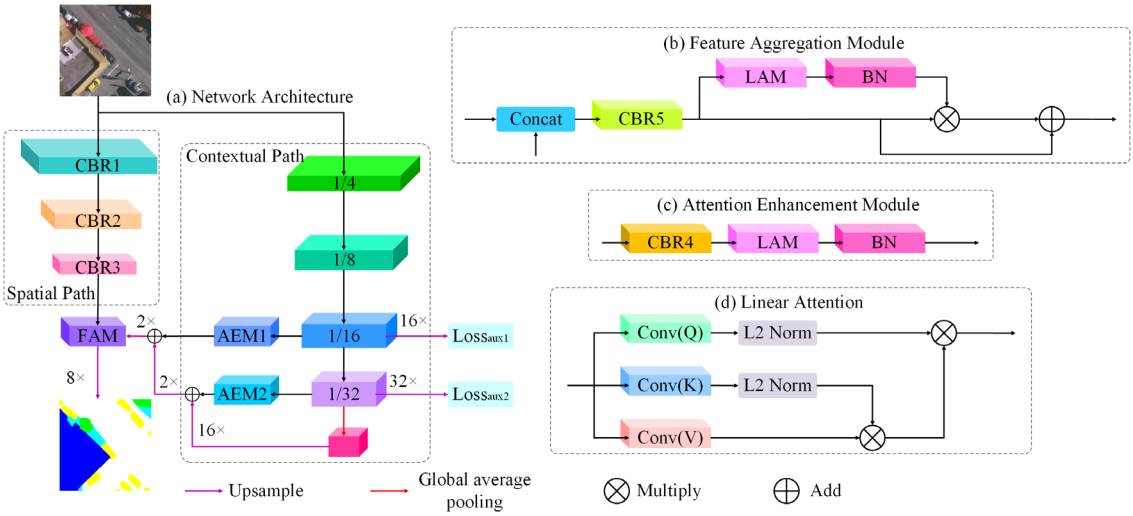

Fig. 5. An overview of the Attentive Bilateral Contextual Network. (a) network architecture. (b) the Feature Aggregation Module (FAM). (c) the Attention Enhancement Module (AEM). (d) the Linear Attention Mechanism. Note that CBR means Convolution+BatchNorm+ReLU, LAM denotes Linear Attention Mechanism, Conv signifies Convolution layer, Concat represents Concatenate operation, BN illustrates BatchNorm layer, and Mul is Multiplication operation.

The proposed Attentive Bilateral Contextual Network (ABCNet), as well as the components, are demonstrated in Fig. 5.

1) Spatial path

It is very challenging to reconcile the requirement for spatial detail with a large receptive field simultaneously. However, both of them are crucial to achieving high segmentation accuracy. Especially, for efficient semantic segmentation, mainstream solutions focus on down-sampling of the input image or speeding up the network by channel pruning. The former loses the majority of the spatial detail, whereas the latter can change its character deleteriously. By contrast, in the proposed ABCNet, we adopt a bilateral architecture (Yu et al., 2018), which is equipped with a spatial path to capture spatial details and generate low-level feature maps. Therefore, a rich channel capacity is essential for this path to encode sufficient spatial detailed information. Meanwhile, since the spatial path focuses merely on low-level details, a shallow structure with a small stride is sufficient for this branch. Specifically, the spatial path is comprised of three layers as shown in Fig. 5(a). The kernel size, channel number, stride and padding for each layer is [7, 64, 2, 3], [3, 64, 2, 1], and [3, 64, 2, 1], respectively. Each layer is followed by batch normalization (Ioffe and Szegedy, 2015) and ReLU (Glorot et al., 2011). Therefore, the output feature maps of this path are 1/8 of the original image, which encodes abundant spatial details resulting from the large spatial size.

2) Contextual path

In parallel to the spatial path, the contextual path is designed to provide a sufficient receptive field, thereby extracting global high-level contextual information. For segmentation, as the receptive field determines the richness of context, several recent approaches attempt to address

the issue by taking advantage of the spatial pyramid pooling. However, huge computational demand and memory consumption will be brought when expanding the receptive field by a large kernel size. Instead, we develop the contextual path with the linear attention mechanism (Li et al., 2021b), which considers the long-range contextual information and efficient computation simultaneously.

In the contextual path as shown in Fig. 5(a), we harness the lightweight backbone (i.e., ResNet-18) (He et al., 2016) to down-sample the feature map and encode the high-level semantic information. We deploy two attention enhancement modules (AEM) on the last two layers of the backbone to fully extract the global contextual information. Besides, a global average pooling operation is attached to the tail of the contextual path to extract the contextual information, while the obtained features are added with the enhanced features generated by AEM2. Thereafter, the acquired features are upsampled by scale=2 to restore the shape. Finally, the features obtained by the AEM1 and AEM2 are added and then fed into the feature aggregation module (FAM).

3) **Feature aggregation module**

The feature representations of the spatial path and the contextual path are complementary, but provided in different domains (i.e., the spatial path generates the low-level and detailed features, while the contextual path provides the high-level and semantic features). Specifically, the output feature captured by the spatial path encodes mainly rich detail information, while the information generated by the contextual path mostly encodes contextual information. Thus, even though summation and concatenation can merge those features (Poudel et al., 2019), these simple fusion schemes are less effective to fuse information in diverse domains (Yang et al., 2021). Here, we

design a feature aggregation module (FAM) to merge both types of feature representation in consideration of the need for high accuracy and efficiency.

As shown in Fig. 5(b), with two domains of features, we first concatenate the output of the spatial and contextual paths. Thereafter, a convolutional layer with batch normalization (Ioffe and Szegedy, 2015) and ReLU (Glorot et al., 2011) is attached to balance the scales of the features. Then, we capture the long-range dependencies of the generated features using the linear attention mechanism, thereby weighing the features selectively. Finally, the weighted features are multiplied and added with the balanced features. As both the scales and contributions of features are readjusted adaptively, the outputs of spatial and contextual paths can be fused effectively.

4) **Loss function**

As shown in Fig. 5(a), besides the principal loss function used to supervise the output of the entire network, we utilize two auxiliary loss functions along the contextual path to accelerate the convergence velocity. We select the cross-entropy loss as the principal loss:

$$\mathrm{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)}, \tag{14}$$

where N and K are the number of samples and number of classes, respectively. $y^{(n)}$ and $\hat{y}^{(n)}$ with $\mathrm{n} \in [1, \cdots, N]$ are one-hot vectors of the true labels and the corresponding softmax output from the network. Essentially, $\hat{y}_k^{(n)}$ depicts the network's confidence of sample n being classified as k. The auxiliary loss functions are chosen as the focal loss:

$$\mathrm{L}_{Focal} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \left(1-\hat{y}_k^{(n)}\right)^{\gamma} y_k^{(n)} \log \hat{y}_k^{(n)}, \tag{15}$$

where γ is the focusing parameter, which controls the down-weighting of the easily classified examples, parameterized as 2 in the experiments. Hence, the overall loss of the network is:

$$\mathrm{L} = \mathrm{L}_{CE} + \mathrm{L}_{Focal}^{aux1} + \mathrm{L}_{Focal}^{aux2}. \quad (16)$$

5) Network variants

There are four main parts in our proposed ABCNet, i.e., the contextual path, the spatial path, the attention enhancement module (AEM), and the feature aggregation module (FAM). Hence, there are mainly five variants of our ABCNet.

Baseline: The baseline (denoted as Cp) can be constructed based on the contextual path without AEM and FAM, while the backbone is set as ResNet-18. The baseline can be utilized as the benchmark to evaluate the effectiveness of components in the network.

Cp + AEM: In the contextual path, the attention enhancement module is designed to capture global contextual information. Hence, a simple variant is a contextual path with attention enhancement modules. The performance of Cp + AEM compared with the baseline will illustrate the effectiveness of the attention enhancement module.

Cp + Sp + AEM (Sum) and **Cp + Sp + AEM (Cat)**: As abundant spatial information is crucial for semantic segmentation, the spatial path is designed to provide a relatively large spatial size and extract spatial information. Two simple fusion schemes including summation (Sum) and concatenation (Cat) can be utilized to merge features. The effectiveness of the spatial path can be validated by merging the spatial information into the network.

Cp + Sp + AEM + FAM: Given that the features obtained by the spatial and contextual paths are in different domains, neither summation nor concatenation provides the optimal fusion scheme. The full version of the proposed ABCNet is fusing the contextual information and spatial information by the feature aggregation module. By comparing the accuracy with Cp + Sp + AEM

(Sum) and Cp + Sp + AEM (Cat), the superiority of the feature aggregation module will be demonstrated.

4. Eeperimental Results and Discussion

1) Experimental settings

a) Datasets

The effectiveness of the proposed ABCNet was tested using the ISPRS Vaihingen dataset and the ISPRS Potsdam dataset (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html). There are two types of ground truth provided in the ISPRS datasets: with and without eroded boundaries. We conducted all experiments on the ground truth with eroded boundaries.

Vaihingen: The Vaihingen dataset contains 33 images with an average size of 2494 × 2064 pixels and a ground sampling distance (GSD) of 9 cm. The near-infrared, red, and green channels together with corresponding digital surface models (DSMs) and normalized DSMs (NDSMs) are provided in the dataset. We utilized ID: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38 for testing, ID: 30 for validation, and the remaining 15 images for training. The DSMs were not used in the experiments. The reference data are labeled according to six land-cover types: impervious surfaces, building, low vegetation, tree, car, and clutter/background.

Potsdam: There exist 38 fine-resolution images of size 6000 × 6000 pixels with a GSD of 5 cm in the Potsdam dataset. The dataset provides the near-infrared, red, green, and blue channels as well as DSMs and NDSMs. We utilized ID: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13 for testing, ID: 2_10 for validation, and the remaining 22

images, except for image named 7_10 with error annotations, for training. We employed only the red, green, and blue channels in the experiments. The reference data are divided into the same six categories as the Vaihingen data set.

b) Training and testing setting

All the training processes were implemented with PyTorch on a single Tesla V100 with 32 batch size, and the optimizer was set as AdamW with a learning rate of 0.0003 and a weight decay value of 0.0025. For the learning rate scheduler, we adopted available ReduceLROnPlateau in PyTorch with the patience of 5 and the learning rate decrease factor as 0.5. If OA on the validation set does not increase for more than 10 epochs, the training procedure will be stopped, while the maximum iteration period is 1000 epochs. For training, we cropped the raw images as well as corresponding labels into 512 × 512 patches and augmented them via rotating on a random angle (90°, 180°, or 270°), resizing by a random scale (from 0.5 to 2.0), flipping by the horizontal axis, flipping by the vertical axis, and adding stochastic Gaussian noise. The probabilities to conduct those augmentation strategies for a patch were set as 0.15, 0.15, 0.25, 0.25, and 0.1, respectively. The comparative benchmark methods selected included the contextual information aggregation methods designed initially for natural images, such as pyramid scene parsing network (PSPNet) (Zhao et al., 2017) and dual attention network (DANet) (Fu et al., 2019), the multi-scale feature aggregation models proposed for remote sensing images, including multi-stage attention ResU-Net (MAResU-Net) (Li et al., 2021b) and edge-aware neural network (EaNet) (Zheng et al., 2020), as well as lightweight networks developed for efficient semantic segmentation, including depth-wise asymmetric bottleneck network (DABNet) (Li et al., 2019a), efficient residual factorized

convNet (ERFNet) (Romera et al., 2017), bilateral segmentation network V1 (BiSeNetV1) (Yu et al., 2018) and V2 (BiSeNetV2) (Yu et al., 2020), fast attention network (FANet) (Hu et al., 2020), ShelfNet (Zhuang et al., 2019) and SwiftNet (Oršić and Šegvić, 2021). In the inference stage, we also utilized the data augmentation operation including random rotation and horizontal as well as vertical flipping which is also known as test-time augmentation (TTA).

c) Evaluation metrics

The performance of ABCNet was evaluated using the overall accuracy (OA), mean Intersection over Union (mIoU), and F1 score (F1). Based on the accumulated confusion matrix, the OA, mIoU, and F1 are computed as:

$$OA = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FP_k + TN_k + FN_k}, \tag{17}$$

$$mIoU = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{18}$$

$$precision_k = \frac{TP_k}{TP_k + FP_k}, \tag{19}$$

$$recall_k = \frac{TP_k}{TP_k + FN_k}, \tag{20}$$

$$F1_k = 2 \times \frac{precision_k \times recall_k}{precision_k + recall_k}, \tag{21}$$

TABLE 1. Ablation study of each component in the proposed ABCNet.

Dataset	Method	Mean F1	OA (%)	mIoU (%)
Vaihingen	Cp	83.9	88.1	73.9
	Cp + AEM	85.8	88.8	75.6
	Cp + Sp + AEM(Sum)	86.6	89.8	77.4
	Cp + Sp + AEM(Cat)	87.1	89.7	77.8
	Cp + Sp + AEM + FAM	89.5	90.7	81.3
Potsdam	Cp	89.7	87.9	81.6
	Cp + AEM	90.6	89.3	83.0
	Cp + Sp + AEM(Sum)	91.0	89.4	83.4
	Cp + Sp + AEM(Cat)	91.2	89.8	84.1
	Cp + Sp + AEM + FAM	92.7	91.3	86.5

where TP_k, FP_k, TN_k, and FN_k represent the true positive, false positive, true negative, and false negatives, respectively, for a particular object indexed as class k. The OA was computed for all categories including the background class.

TABLE 2. The complexity and speed of the proposed ABCNet and lightweight methods. 'G' indicates Gillion (i.e., units for the number of floating point operations) and 'M' signifies Million (i.e., units for the number of parameters). For an extensive comparison, we chose 256×256, 512×512, 1024×1024, 2048×2048, and 4096×4096 pixels as the sizes of the input image and report the inference speed measured in frames per second (FPS) on a midrange notebook graphics card 1660Ti. mIoU is measured using patches of 512×512 pixels, where the first number is the mIoU on the Vaihingen dataset and the second one is on the Potsdam dataset.

Method	Backbone	Complexity(G)	Parameters(M)	256×256	512×512	1024×1024	2048×2048	4096×4096	mIoU
DABNet (Li et al., 2019a)	-	5.22	0.75	90.67	87.74	27.41	7.44	*	70.2/79.6
ERFNet (Romera et al., 2017)	-	14.75	2.06	90.51	59.04	17.59	4.87	1.25	69.1/76.2
BiSeNetV1 (Yu et al., 2018)	ResNet18	15.25	13.61	143.50	87.63	25.89	7.23	1.84	75.8/81.7
PSPNet (Zhao et al., 2017)	ResNet18	12.55	24.03	151.12	105.03	34.83	10.16	2.66	68.6/75.9
BiSeNetV2 (Yu et al., 2020)	-	13.91	12.30	124.49	82.84	25.64	7.07	*	75.5/82.3
DANet (Fu et al., 2019)	ResNet18	9.90	12.68	181.66	124.18	40.80	11.42	*	69.4/80.3
FANet (Hu et al., 2020)	ResNet18	21.66	13.81	112.59	67.97	20.41	5.57	*	75.6/84.2
ShelfNet (Zhuang et al., 2019)	ResNet18	12.36	14.58	123.59	90.41	30.93	9.06	2.40	78.7/84.4
SwiftNet (Oršić and Šegvić, 2021)	ResNet18	13.08	11.80	157.63	97.62	30.79	8.65	*	78.3/83.8
MAResU-Net (Li et al., 2021b)	ResNet18	25.43	16.17	70.12	37.55	13.35	3.51	*	78.6/83.9
EaNet (Zheng et al., 2020)	ResNet18	18.75	34.23	73.98	55.95	17.94	5.53	1.54	79.6/83.4
ABCNet	ResNet18	18.72	14.06	113.09	72.13	22.73	6.23	1.60	81.3/86.5

2) **Experimental results**

a) Ablation Study

To evaluate the effectiveness of the components in the proposed ABCNet, we conducted extensive ablation experiments; the setting details and quantitative results are listed in Table 1.

Baseline: The baseline was constructed based on the contextual path, while the generated feature maps were up-sampled directly to the same shape as the original input image.

Ablation for attention enhancement module: To capture the global contextual information, we designed an attention enhancement module (AEM) in the contextual path. As presented in Table 1, for two datasets, the utilization of AEM (indicated as Cp + AEM) produced an increase of greater than 1.4% in the mIoU.

Ablation for the spatial path: Table 1 demonstrates that even simple fusion schemes for merging spatial information such as summation (represented as Cp + Sp + AEM(Sum)) and concatenation (represented as Cp + Sp + AEM(Cat)) boosted the performance of the mIoU by about 1.8% on Vaihingen dataset, and 0.4% on Potsdam dataset.

Ablation for feature aggregation module: As shown in Table 1, the significant gap in performance (more than 2.4% in the mIoU) demonstrates the validity of the feature aggregation module (signified as Cp + Sp + AEM + FAM).

b) The complexity and speed of the network

Complexity and speed are important criteria for measuring the merit of an algorithm, and this is especially true for practical applications. We first compared the computation and memory requirements between the linear attention mechanism and dot-product attention mechanism which can be found in Fig. 6.

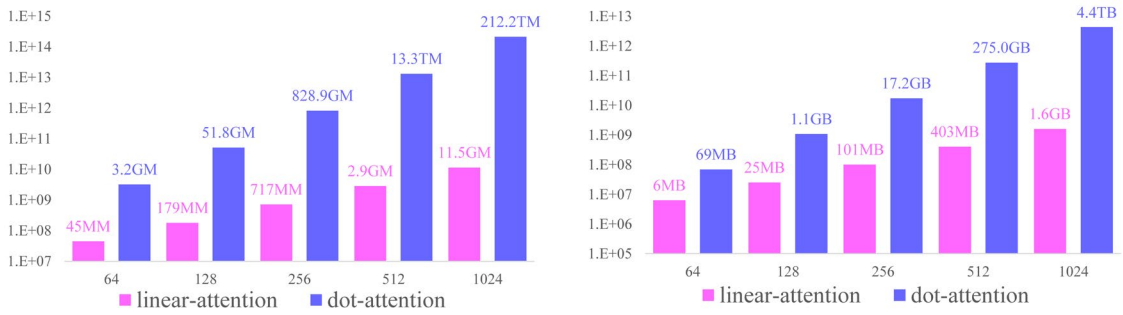

Fig. 6. Comparison between the (a) computational and (b) memory requirements of the linear attention mechanism and dot-product attention mechanism under different input sizes. The calculation assumes $C = D_v = 2D_k = 64$. MM denotes 1 Mega multiply-accumulation (MACC), where 1 MACC means 1 multiplication and 1 addition operation. GM means 1 Giga MACC, while TM signifies 1 Tera MACC. Similarly, MB, GB, and TB represent 1 MegaByte, 1 GigaByte, and 1 TeraByte, respectively. Note the figure is shown on the log scale.

For a comprehensive comparison, we further implemented the experiments under different settings. A comparison between the parameters and computational complexity of the different networks is reported in Table 2. The proposed ABCNet maintained both high speed and high accuracy simultaneously. As listed in the last column of Table 2, the mIoU on the Potsdam dataset achieved by the ABCNet is at least 2.0% higher than the benchmark methods. Meanwhile, the ABCNet was able to achieve a 72.13 FPS speed for a 512×512 input. The remarkable performance of the speed and occupation of memory not only derives from the linear attention mechanism but also results from that we only utilized the AEM in deeper layers with small spatial dimensionality. Besides, the elaborate design enabled the ABCNet to handle the massive input (4096×4096), while more than half of the benchmark methods ran out of memory for a such large input.

TABLE 3. Quantitative comparison results on the Vaihingen test set with the lightweight networks.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
DABNet (Li et al., 2019a)	-	87.8	88.8	74.3	84.9	60.2	79.2	84.3	70.2
ERFNet (Romera et al., 2017)	-	88.5	90.2	76.4	85.8	53.6	78.9	85.8	69.1
BiSeNetV1 (Yu et al., 2018)	ResNet18	89.1	91.3	80.9	86.9	73.1	84.3	87.1	75.8
PSPNet (Zhao et al., 2017)	ResNet18	89.0	93.2	81.5	87.7	43.9	79.0	87.7	68.6
BiSeNetV2 (Yu et al., 2020)	-	89.9	91.9	82.0	88.3	71.4	84.7	88.0	75.5
DANet (Fu et al., 2019)	ResNet18	90.0	93.9	82.2	87.3	44.5	79.6	88.2	69.4
FANet (Hu et al., 2020)	ResNet18	90.7	93.8	82.6	88.6	71.6	85.4	88.9	75.6
EaNet (Zheng et al., 2020)	ResNet18	91.7	94.5	83.1	89.2	80.0	87.7	89.7	78.7
ShelfNet (Zhuang et al., 2019)	ResNet18	91.8	94.6	83.8	89.3	77.9	87.5	89.8	78.3
MAResU-Net (Li et al., 2021b)	ResNet18	92.0	95.0	83.7	89.3	78.3	87.7	90.1	78.6
SwiftNet (Oršić and Šegvić, 2021)	ResNet18	92.2	94.8	84.1	89.3	81.2	88.3	90.2	79.6
ABCNet	ResNet18	**92.7**	**95.2**	**84.5**	**89.7**	**85.3**	**89.5**	**90.7**	**81.3**

c) Results on the ISPRS Vaihingen and Potsdam datasets

The ISPRS Vaihingen is a relatively small dataset. All images represent the same city, such that the statistical characters of the training and test datasets are similar (Ghassemi et al., 2019). Therefore, high accuracy can be achieved relatively easily by specifically designed networks, especially for those that fuse orthophoto (TOP) images with auxiliary DSMs or NDSMs. In this section, we demonstrate that the proposed ABCNet model using only TOP images with an

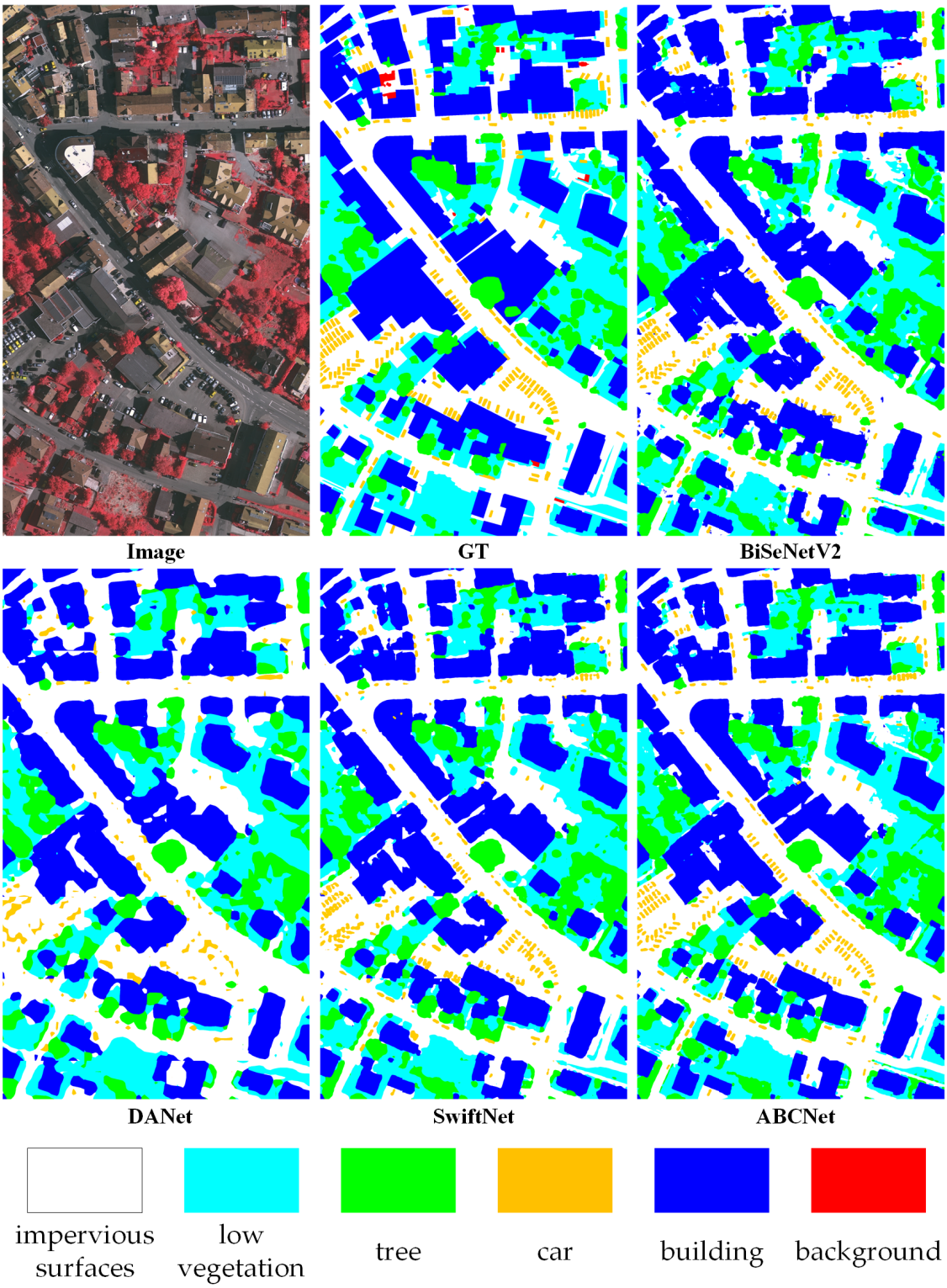

Fig. 7. Mapping results for test images of Vaihingen tile-27.

efficient architecture can not only transcend lightweight networks (Table 3) but also achieve

highly competitive accuracy compared to specially designed models (Table 4).

As shown in Table 3, the numeric scores for the ISPRS Vaihingen test dataset demonstrated that ABCNet delivers high accuracy, exceeding other lightweight networks in the mean F1, OA, and mIoU by a significant margin. Particularly, the ‘‘car’’ class in the Vaihingen dataset is difficult to handle as it is a relatively small object. Nonetheless, ABCNet produced an 85.3% F1 score for this class, which is at least 4.1% higher than for the benchmark methods. In addition, we visualize area 27 in Fig. 7 to qualitatively demonstrate the effectiveness of ABCNet, while the enlarged

TABLE 4. Quantitative comparison results on the Vaihingen test set with the state-of-the-art networks.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)	Speed
DeepLabV3+ (Chen et al., 2018a)	ResNet101	92.4	95.2	84.3	89.5	86.5	89.6	90.6	81.5	13.27
PSPNet (Zhao et al., 2017)	ResNet101	92.8	95.5	84.5	89.9	**88.6**	90.3	90.9	**82.6**	22.03
DANet (Fu et al., 2019)	ResNet101	91.6	95.0	83.3	88.9	87.2	89.2	90.4	81.3	21.97
EaNet (Zheng et al., 2020)	ResNet101	**93.4**	**96.2**	85.6	90.5	88.3	**90.8**	91.2	-	9.97
DDCM-Net (Liu et al., 2020)	ResNet50	92.7	95.3	83.3	89.4	88.3	89.8	90.4	-	37.28
HUSTW5 (Sun et al., 2019)	ResegNets	93.3	96.1	**86.4**	**90.8**	74.6	88.2	**91.6**	-	-
CASIA2 (Liu et al., 2018)	ResNet101	93.2	96.0	84.7	89.9	86.7	90.1	91.1	-	-
V-FuseNet# (Audebert et al., 2018)	FuseNet	91.0	94.4	84.5	89.9	86.3	89.2	90.0	-	-
DLR_9# (Marmanis et al., 2018)	-	92.4	95.2	83.9	89.9	81.2	88.5	90.3	-	-
ABCNet	ResNet18	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3	**72.13**

- means the results are not reported in the original paper.

means the DSM or NDSM are used in the network.

results are shown in Fig. 9 (top).

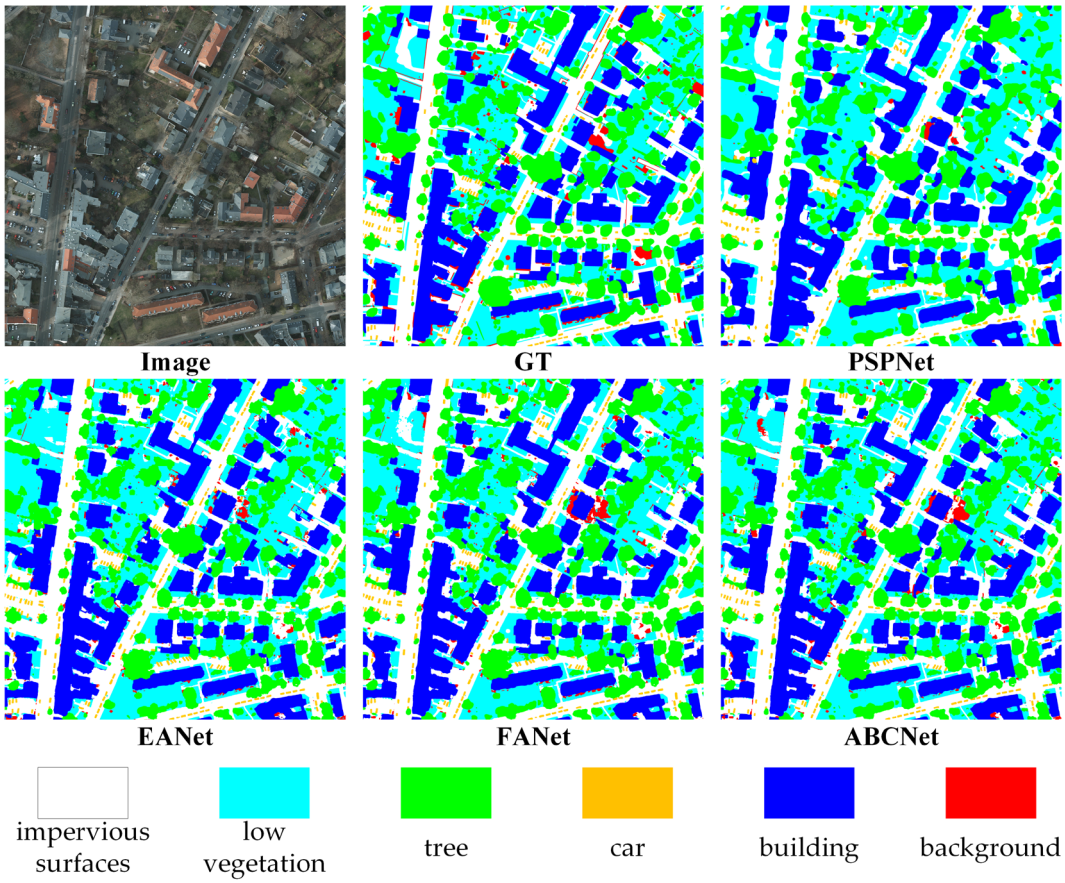

Fig. 8. Mapping results for the test images of Potsdam tile-3_13.

For a comprehensive evaluation, ABCNet was also compared with other state-of-the-art methods. As can be seen in Table 4, as a lightweight network, the proposed ABCNet achieved a competitive performance even compared with those models designed with complex structures. It is worth noting that the speed of ABCNet is two-to-seven times faster than those methods.

Furthermore, we undertook experiments on the ISPRS Potsdam dataset to further evaluate the performance of ABCNet. Compared with the encoder-decoder structure, the bilateral architecture can retain more spatial information without reducing the speed of the model (Yu et al., 2018). The spatial path stacks only three convolution layers to generate 1/8 feature maps, while the contextual

TABLE 5 Quantitative comparison results on the Potsdam test set with the lightweight networks.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
ERFNet (Romera et al., 2017)	-	88.7	93.0	81.1	75.8	90.5	85.8	84.5	76.2
DABNet (Li et al., 2019a)	-	89.9	93.2	83.6	82.3	92.6	88.3	86.7	79.6
PSPNet (Zhao et al., 2017)	ResNet18	89.1	94.5	84.0	85.8	76.6	86.0	87.2	75.9
BiSeNetV1 (Yu et al., 2018)	ResNet18	90.2	94.6	85.5	86.2	92.7	89.8	88.2	81.7
BiSeNetV2 (Yu et al., 2020)	-	91.3	94.3	85.0	85.2	94.1	90.0	88.2	82.3
EaNet (Zheng et al., 2020)	ResNet18	92.0	95.7	84.3	85.7	95.1	90.6	88.7	83.4
MAResU-Net (Li et al., 2021b)	ResNet18	91.4	95.6	85.8	86.6	93.3	90.5	89.0	83.9
DANet (Fu et al., 2019)	ResNet18	91.0	95.6	86.1	87.6	84.3	88.9	89.1	80.3
SwiftNet (Oršić and Šegvić, 2021)	ResNet18	91.8	95.9	85.7	86.8	94.5	91.0	89.3	83.8
FANet (Hu et al., 2020)	ResNet18	92.0	96.1	86.0	87.8	94.5	91.3	89.8	84.2
ShelfNet (Zhuang et al., 2019)	ResNet18	92.5	95.8	86.6	87.1	94.6	91.3	89.9	84.4
ABCNet	ResNet18	**93.5**	**96.9**	**87.9**	**89.1**	**95.8**	**92.7**	**91.3**	**86.5**

path includes two attention enhancement modules (AEM) to refine the features and capture contextual information. Numerical comparisons with other lightweight methods are shown in Table 5. Remarkably, ABCNet achieved 91.3% overall accuracy and 86.5% in mIoU. Visualization of area 3_13 is displayed in Fig. 8, and the enlarged results are exhibited in Fig. 9 (bottom). As there are sufficient images in the Potsdam dataset to train the network, the performance of ABCNet can be equivalent to the state-of-the-art methods with a much faster

speed. The comparison results are listed in Table 6.

TABLE 6. Quantitative comparison results on the Potsdam test set with state-of-the-art networks.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)	Speed
DeepLabV3+ (Chen et al., 2018a)	ResNet101	93.0	95.9	87.6	88.2	96.0	92.1	90.9	84.3	13.27
PSPNet (Zhao et al., 2017)	ResNet101	93.4	97.0	87.8	88.5	95.4	92.4	91.1	84.9	22.03
DDCM-Net (Liu et al., 2020)	ResNet50	92.9	96.9	87.7	**89.4**	94.9	92.3	90.8	-	37.28
CCNet (Huang et al., 2020)	ResNet101	93.6	96.8	86.9	88.6	**96.2**	92.4	91.5	85.7	5.56
AMA_1	-	93.4	96.8	87.7	88.8	96.0	92.5	91.2	-	-
SWJ_2	ResNet101	**94.4**	97.4	87.8	87.6	94.7	92.4	**91.7**	-	-
HUSTW4 (Sun et al., 2019)	ResegNets	93.6	**97.6**	**88.5**	88.8	94.6	92.6	91.6	-	-
V-FuseNet# (Audebert et al., 2018)	FuseNet	92.7	96.3	87.3	88.5	95.4	92.0	90.6	-	-
DST_5# (Sherrah, 2016)	FCN	92.5	96.4	86.7	88.0	94.7	91.7	90.3	-	-
ABCNet	ResNet18	93.5	96.9	87.9	89.1	95.8	**92.7**	91.3	**86.5**	**72.13**

- means the results are not reported in the original paper.

means the DSM or NDSM are used in the network.

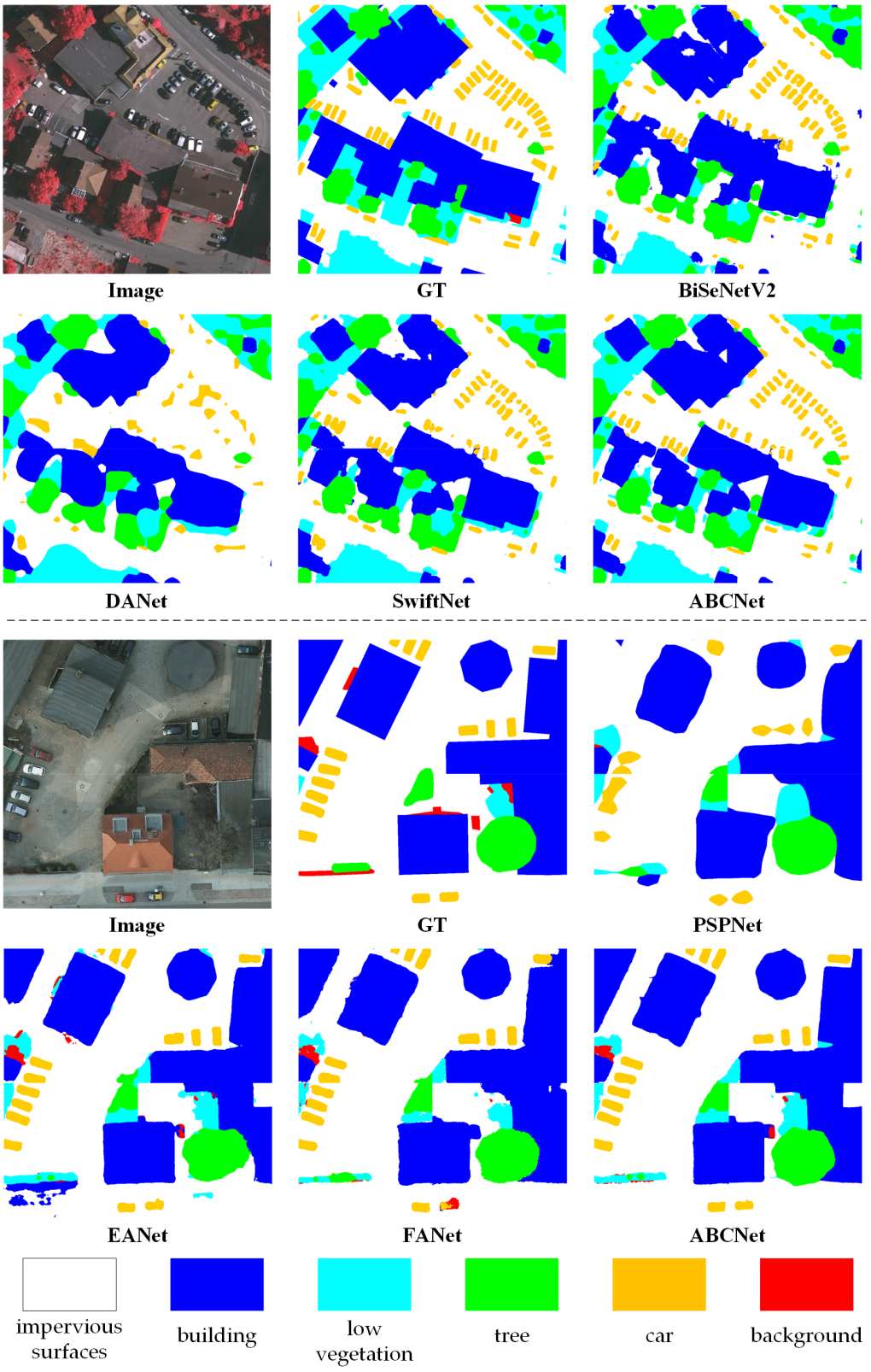

Fig. 9. Enlarged visualization of results on (top) the Vaihingen and (bottom) Potsdam datasets.

3) **Discussion**

The comprehensive experiments undertaken demonstrate the superiority of ABCNet, not only for segmentation accuracy but also efficiency. There are three important factors that guarantee accuracy without drastically increasing computational consumption. First, the bilateral architecture resolves the contradiction between sufficient contextual information and fine-grained spatial detail. The channel pruning or input cropping operations are commonly used in the encoder-decoder structure to boost inference speed, leading to the loss of low-level and spatial details which cannot be recovered easily. In contrast, the proposed ABCNet adopts a bilateral architecture, where a spatial path extracts low-level features and a contextual path exploits high-level features. To demonstrate the difference between the contextual path (Cp) and spatial path (Sp) visually, we visualize the feature maps generated by the Cp and Sp in Fig. 10. Please note that the features maps of are upsampled to restore the shape. As can be seen in the figure, the information provided by the contextual path and spatial has indeed differences. Specifically, in feature maps of the contextual path, objects have a more consistent character with those pixels in the same class. By contrast, more detailed information is preserved in the spatial path. Meanwhile, the relatively efficient design of the spatial path (three stacked identical layers) and contextual path (the ResNet-18 backbone) avoids large computational requirements. Second, the attention enhancement module balances the trade-off between global contextual information and huge calculation complexity. Conventionally, the dot-product attention mechanism employed to capture long-range dependencies is accompanied by quadratic increases in time and memory consumption with input size. Instead, we harness the linear attention mechanism, developed in

our previous research, to provide a calculation-friendly scheme for global contextual information extraction. Third, the feature aggregation module merges the spatial features and contextual features in an appropriate fashion. The spatial features generated by the spatial path are low-level and detailed, while the contextual features generated by the contextual path are high-level and semantically rich. In other words, the features have entirely different semantic meanings. Hence, although a degree of improvement in accuracy can be brought, the simple summation or concatenation operations are not the optimal feature fusion scheme. The elaborate feature aggregation module developed here ensures reasonable fusion and full utilization of both sets of features.

5. Conclusion

In this paper, we propose a novel lightweight framework for efficient semantic segmentation in the field of remote sensing, namely the Attentive Bilateral Contextual Network (ABCNet). As both sufficient contextual information and fine-grained spatial detail are crucial for the accuracy of segmentation, we design the ABCNet based on the bilateral architecture which captures simultaneously and adaptively the abundant spatial details in fine-resolution remotely sensed imagery via a spatial path and the global contextual information via a contextual path. Extensive experiments on the ISPRS Vaihingen and Potsdam datasets demonstrate the effectiveness and efficiency of the proposed ABCNet, with huge potential for practical real-time applications. Although achieving a relatively fine balance between effectiveness and efficiency, the speed of the proposed ABCNet has a certain room for improvement, especially when compared with those single-branch lightweight networks. As the contextual path occupies the majority of parameters

and complexities, our future work will focus on further optimizing the contextual path of the ABCNet, especially to design an efficient Transformer backbone using our linear attention mechanism, thereby replacing the original ResNet backbone with this novel structure.

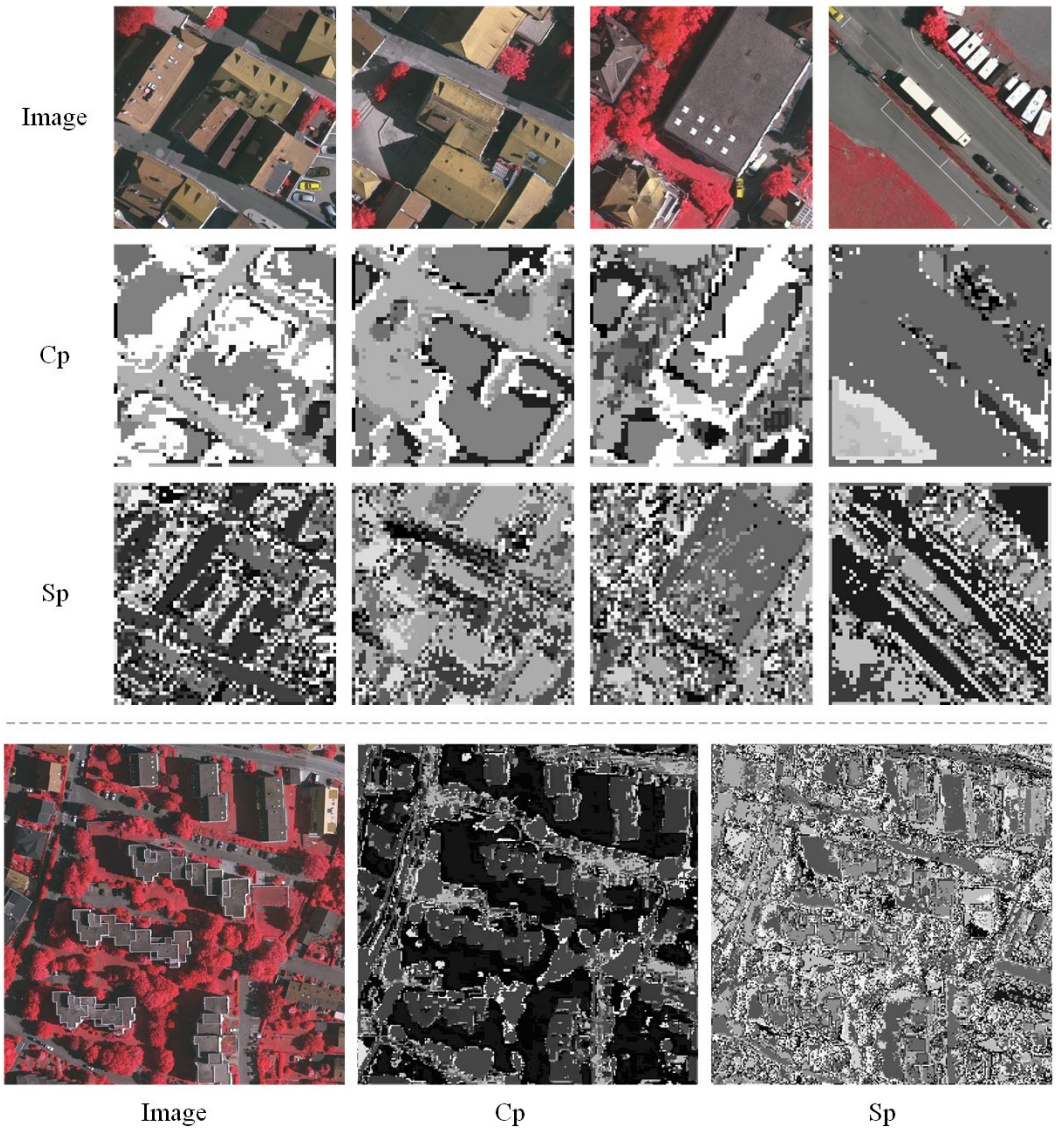

Fig. 10. Illustration of feature mapsgenerated by Cp and Sp, where the input size of the image in the top part is 512×512 and 2048×2048 in the bottom.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal

relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, Grant No. 41671452.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140, 20-32.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 2481-2495.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V., 2019. Attention augmented convolutional networks, Proceedings of the IEEE/CVF international conference on computer vision, pp. 3286-3295.

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond, Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0-0.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017a. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation, Proceedings of the European conference on computer vision (ECCV), pp. 801-818.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S., 2017b. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5659-5667.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018b. A^ 2-nets: Double attention networks, Advances in neural information processing systems, pp. 352-361.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258.

Diakogiannis, F.I., Waldner, F., Caccetta, P., 2020a. Looking for change? Roll the Dice and demand Attention. arXiv preprint arXiv:2009.02062.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020b. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing 162, 94-114.

Duan, C., Li, R., 2020. Multi-Head Linear Attention Generative Adversarial Network for Thin Cloud Removal. arXiv preprint arXiv:2012.10898.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation,

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146-3154.
Ghassemi, S., Fiandrotti, A., Francini, G., Magli, E., 2019. Learning and adapting robust features for satellite image segmentation on heterogeneous data sets. IEEE Transactions on Geoscience and Remote Sensing 57, 6517-6529.
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, pp. 315-323.
Gong, P., Marceau, D.J., Howarth, P.J., 1992. A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. Remote sensing of environment 40, 137-151.
Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote sensing of environment 220, 135-151.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence 37, 1904-1916.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141.
Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S., 2020. Real-time semantic segmentation with fast attention. IEEE Robotics and Automation Letters 6, 263-270.
Huang, L., Yuan, Y., Guo, J., Zhang, C., Chen, X., Wang, J., 2019a. Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273.
Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019b. Ccnet: Criss-cross attention for semantic segmentation, Proceedings of the IEEE International Conference on Computer Vision, pp. 603-612.
Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.S., 2020. CCNet: Criss-Cross Attention for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, International conference on machine learning. PMLR, pp. 448-456.
Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F., 2020. Transformers are rnns: Fast autoregressive transformers with linear attention, International Conference on Machine Learning. PMLR, pp. 5156-5165.
Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS journal of photogrammetry and remote sensing 145, 60-77.
Li, G., Yun, I., Kim, J., Kim, J., 2019a. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357.
Li, K., Cheng, G., Bu, S., You, X., 2017. Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 56, 2337-2348.
Li, R., Duan, C., Zheng, S., Zhang, C., Atkinson, P.M., 2021a. MACU-Net for semantic segmentation of fine-resolution remotely sensed images. IEEE Geoscience and Remote Sensing Letters.
Li, R., Zheng, S., Duan, C., Su, J., Zhang, C., 2021b. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters.
Li, R., Zheng, S., Duan, C., Yang, Y., Wang, X., 2020. Classification of Hyperspectral Image Based on Double-Branch Dual-Attention Mechanism Network. Remote Sensing 12, 582.
Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2021c. Multiattention network for

semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing.

Li, X., Wang, W., Hu, X., Yang, J., 2019b. Selective kernel networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 510-519.

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H., 2019c. Expectation-maximization attention networks for semantic segmentation, Proceedings of the IEEE International Conference on Computer Vision, pp. 9167-9176.

Liu, Q., Kampffmeyer, M., Jenssen, R., Salberg, A.-B., 2020. Dense dilated convolutions' merging network for land cover classification. IEEE Transactions on Geoscience and Remote Sensing 58, 6309-6320.

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C., 2018. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS journal of photogrammetry and remote sensing 145, 78-95.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F., 2019. See more, know more: Unsupervised video object segmentation with co-attention siamese networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3623-3632.

Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., Elith, J., 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sensing of Environment 208, 145-153.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS Journal of Photogrammetry and Remote Sensing 130, 277-293.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 55, 645-657.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing 135, 158-172.

Oršić, M., Šegvić, S., 2021. Efficient semantic segmentation with pyramidal fusion. Pattern Recognition 110, 107611.

Picoli, M.C.A., Camara, G., Sanches, I., Simões, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R.A., 2018. Big earth observation time series analysis for monitoring Brazilian agriculture. ISPRS journal of photogrammetry and remote sensing 145, 328-339.

Poudel, R.P., Liwicki, S., Cipolla, R., 2019. Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502.

Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R., 2017. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems 19, 263-272.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234-241.

Samie, A., Abbas, A., Azeem, M.M., Hamid, S., Iqbal, M.A., Hasan, S.S., Deng, X., 2020. Examining the impacts of future land use/land cover changes on climate in Punjab province, Pakistan: implications for environmental sustainability and economic growth. Environmental Science and Pollution Research 27, 25415-25433.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial

imagery. arXiv preprint arXiv:1606.02585.
Sun, Y., Tian, Y., Xu, Y., 2019. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning. Neurocomputing 330, 297-304.
Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment 237, 111322.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote sensing of Environment 8, 127-150.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, Advances in neural information processing systems, pp. 5998-6008.
Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164.
Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X., 2021. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. Remote Sensing 13, 3065.
Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018a. Understanding convolution for semantic segmentation, 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp. 1451-1460.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. ECA-net: Efficient channel attention for deep convolutional neural networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534-11542.
Wang, X., Girshick, R., Gupta, A., He, K., 2018b. Non-local neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803.
Woo, S., Park, J., Lee, J.-Y., So Kweon, I., 2018. Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), pp. 3-19.
Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. DOTA: A large-scale dataset for object detection in aerial images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974-3983.
Yang, M.Y., Kumaar, S., Lyu, Y., Nex, F., 2021. Real-time Semantic Segmentation with Context Aggregation Network. ISPRS Journal of Photogrammetry and Remote Sensing 178, 124-134.
Yin, H., Pflugmacher, D., Li, A., Li, Z., Hostert, P., 2018. Land use and land cover change in Inner Mongolia-understanding the effects of China's re-vegetation programs. Remote Sensing of Environment 204, 918-930.
Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N., 2020. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. arXiv preprint arXiv:2004.02147.
Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation, Proceedings of the European conference on computer vision (ECCV), pp. 325-341.
Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
Yuan, Y., Chen, X., Wang, J., 2019. Object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065.

Yuan, Y., Wang, J., 2018. Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916.

Zhang, C., Harrison, P.A., Pan, X., Li, H., Sargent, I., Atkinson, P.M., 2020. Scale Sequence Joint Deep Learning (SS-JDL) for land use and land cover classification. Remote Sensing of Environment 237, 111593.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019a. Joint Deep Learning for land cover and land use classification. Remote sensing of environment 221, 173-187.

Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E., 2019b. Acfnet: Attentional class feature network for semantic segmentation, Proceedings of the IEEE International Conference on Computer Vision, pp. 6798-6807.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019c. Self-attention generative adversarial networks, International conference on machine learning. PMLR, pp. 7354-7363.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881-2890.

Zheng, X., Huan, L., Xia, G.-S., Gong, J., 2020. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. ISPRS Journal of Photogrammetry and Remote Sensing 170, 15-28.

Zhong, Y., Zhao, J., Zhang, L., 2014. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 52, 7023-7037.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5, 8-36.

Zhuang, J., Yang, J., Gu, L., Dvornek, N., 2019. Shelfnet for fast semantic segmentation, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0-0.