Multi-Attention-Network for Semantic Segmentation of Fine Resolution Remote Sensing Images

Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, and P.M. Atkinson

***Abstract*—Semantic segmentation of remote sensing images plays an important role in a wide range of applications including land resource management, biosphere monitoring and urban planning. Although the accuracy of semantic segmentation in remote sensing images has been increased significantly by deep convolutional neural networks, several limitations exist in standard models. First, for encoder-decoder architectures such as U-Net, the utilization of multi-scale features causes the underuse of information, where low-level features and high-level features are concatenated directly without any refinement. Second, long-range dependencies of feature maps are insufficiently explored, resulting in sub-optimal feature representations associated with each semantic class. Third, even though the dot-product attention mechanism has been introduced and utilized in semantic segmentation to model long-range dependencies, the large time and space demands of attention impede the actual usage of attention in application scenarios with large-scale input. This paper proposed a Multi-Attention-Network (MANet) to address these issues by extracting contextual dependencies through multiple efficient attention modules. A novel attention mechanism of kernel attention with linear complexity is proposed to alleviate the large computational demand in attention. Based on kernel attention and channel attention, we integrate local feature maps extracted by ResNet-50 with their corresponding global dependencies and reweight interdependent channel maps adaptively. Numerical experiments on two large-scale fine resolution remote sensing datasets demonstrate the superior performance of the proposed MANet. Code is available at https://github.com/lironui/Multi-Attention-Network.**

***Index Terms*—fine-resolution remote sensing images, attention mechanism, semantic segmentation**

I. INTRODUCTION

Semantic segmentation of remote sensing images (i.e., the assignment of definite categories to groups of pixels in an image), plays a crucial role in a wide range of applications such as land resources management, yield estimation and economic assessment [1, 6-10].

Vegetation indices are commonly used features extracted from multispectral and hyperspectral images to characterize land surface physical properties. The normalized difference vegetation index (NDVI) [13] and soil-adjusted vegetation index (SAVI) [16] highlight vegetation over other land resources, whereas the normalized difference bareness index (NDBaI) [18] and the normalized difference bare land index (NBLI) [19] emphasize bare land. The normalized difference water index (NDWI) [20] and modified NDWI (MNDWI) [21] indicate water. These indices have been developed and applied widely in the remote sensing community. Meanwhile, different classifiers have been designed from diverse perspectives, from traditional methods such as logistic regression [22], distance measures [23] and clustering [24], to more advanced machine learning methods such as the support vector machine (SVM) [25], random forest (RF) [26] and artificial neural networks (ANN) [27] including the multi-layer perceptron (MLP) [28]. These classifiers depend critically on the quality of features that are extracted for pixel-level land cover classification. However, this high dependency on hand-crafted descriptors restricts the flexibility and adaptability of these traditional methods.

Deep Learning (DL), a powerful approach to capture nonlinear and hierarchical features automatically, has had a significant impact on various domains such as computer vision (CV) [29], natural language processing (NLP) [30] and automatic speech recognition (ASR) [31]. In the field of remote sensing, DL methods have been introduced and implemented for land cover and land use classification [32]. Compared with vegetation indices, which are based on physical and mathematical concepts and hand-coded from spectral bands only, DL methods can mine different kinds of information including temporal periods, spectra, spatial context and the interactions among different land cover categories [15].

For remotely sensed semantic segmentation, Fully Convolutional Network (FCN)-based methods [1] and encoder-

This work was supported in part by the National Natural Science Foundation of China (No. 41671452). *(Corresponding author: Rui Li.)*

R. Li and S. Zheng are with School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: lironui@whu.edu.cn; syzheng@whu.edu.cn).

C. Zhang is with the Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, U.K.; UK Centre for Ecology & Hydrology, Library Avenue, Lancaster, LA1 4AP, U.K (e-mail: c.zhang9@lancaster.ac.uk).

C. Duan is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; chenxiduan@whu.edu.cn (e-mail: chenxiduan@whu.edu.cn).

Jianlin Su is with the Shenzhen Zhuiyi Technology Co., Ltd. (e-mail: bojonesu@wezhuiyi.com).

Peter M. Atkinson is with the Lancaster Environment Center, Lancaster University, Lancaster LA1 4YQ, U.K., also with the Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, U.K., and also with the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China (e-mail: pma@lancaster.ac.uk).

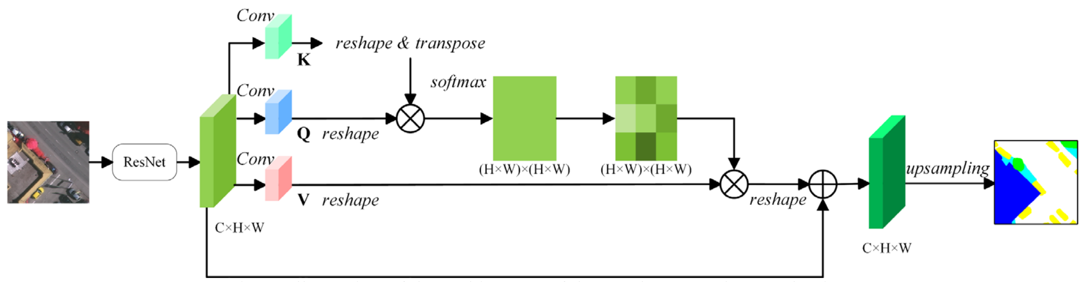


Fig. 1. Illustration of the architecture of dot-product attention mechanism.

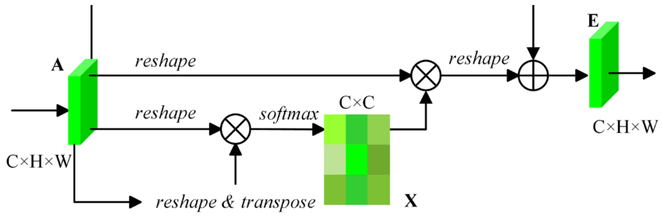


Fig. 2. Details of the channel attention mechanism.

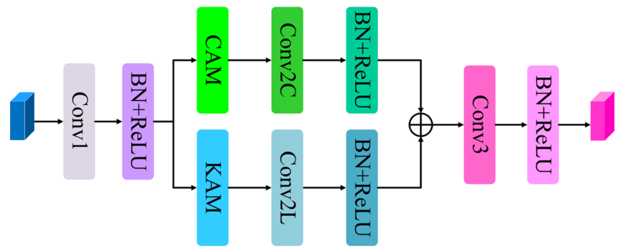


Fig. 3. Illustration of the attention block.

decoder architectures such as SegNet [33] and U-Net [34] have been adopted widely. Generally, the FCN-based architectures comprise a contracting path that extracts information from the input image and generates high-level feature maps, and an expanding path, where high-level feature maps are utilized to reconstruct the mask for pixel-wise segmentation by the single [1] or multi-level [34, 35] up-sampling procedures. Despite their powerful representation capability, however, information flow bottlenecks limit the potential of these multi-scale approaches [36]. For example, the low-level and fine-grained detailed feature maps generated by the encoder are concatenated with high-level and coarse-grained semantic information generated by the decoder without any further refinement, leading to inadequate exploitation and deficient discrimination of features. Besides, the discriminative ability of the feature representations might be insufficient for challenging tasks such as semantic segmentation of fine spatial resolution remote sensing images.

The utilization of context fusion at multiple scales is a feasible solution [12, 14, 37-41], increasing the discriminative power of feature representations. The multi-scale context information can be aggregated using techniques such as atrous spatial pyramid pooling [14, 37], pyramid pooling module [12], or context encoding module [39]. Although context captured by the above strategies is beneficial to characterizing objects at different scales, the contextual dependencies for whole input regions are homogeneous and non-adaptive, without considering the disparity between contextual dependencies and local representation of different categories. Further, these multi-scale context fusion strategies are designed manually, with limited flexibility in modeling multi-context representations. The long-range dependencies of feature maps are insufficiently leveraged in these approaches, which may be of paramount importance for remotely sensed semantic segmentation.

With strong capabilities to capture long-range dependencies, dot-product attention mechanisms have been applied in vision and natural language processing tasks. The dot-product-attention-based Transformer has demonstrated state-of-the-art performance in a majority of tasks in natural language processing [30, 42-44]. The non-local module [45], a dot-product-based attention modified for computer vision, has shown great potential in image classification [46], object detection [47], semantic segmentation [5] and panoptic segmentation [48].

Utilization of the dot-product attention mechanism often comes with significant memory and computational costs, which increase quadratically with the size of the input over space and time. It remains an intractable problem to model global dependency on large-scale inputs, such as video, long sequences and fine-resolution images. To alleviate the substantial computational requirement, Child et al. [49] designed a sparse factorization of the attention matrix and reduced the complexity from $O(N^2)$ to $O(N\sqrt{N})$. Using locality sensitive hashing, Kitaev et al. [50] reduced the complexity to $O(N \log N)$. Katharopoulos et al. [11] represented self-attention as a linear dot-product of kernel feature maps to further reduce the complexity to $O(N)$, and Shen et al. [3] modified the position of the softmax functions.

In this paper, by comparison, we not only dramatically decrease the complexity, but also amply exploit the potential of the attention mechanism by designing a multilevel framework. Specifically, we reduce the complexity of the dot-product attention mechanism to $O(N)$ by treating attention as a kernel function. As the complexity of attention is reduced dramatically by kernel attention, we propose a Multi-Attention-Network (MANet) with a ResNet-50 backbone which explores the complex combinations between attention mechanisms and deep networks for the task of semantic segmentation using fine-resolution remote sensing images. The performance of the proposed algorithm is compared comprehensively with various benchmarks. The major contributions of this research are two-fold: 1) a novel attention mechanism involving kernel attention with linear complexity is proposed to alleviate the huge computational demand from attention module; . 2) we propose a novel Multi-Attention-Network (MANet) with a multi-scale

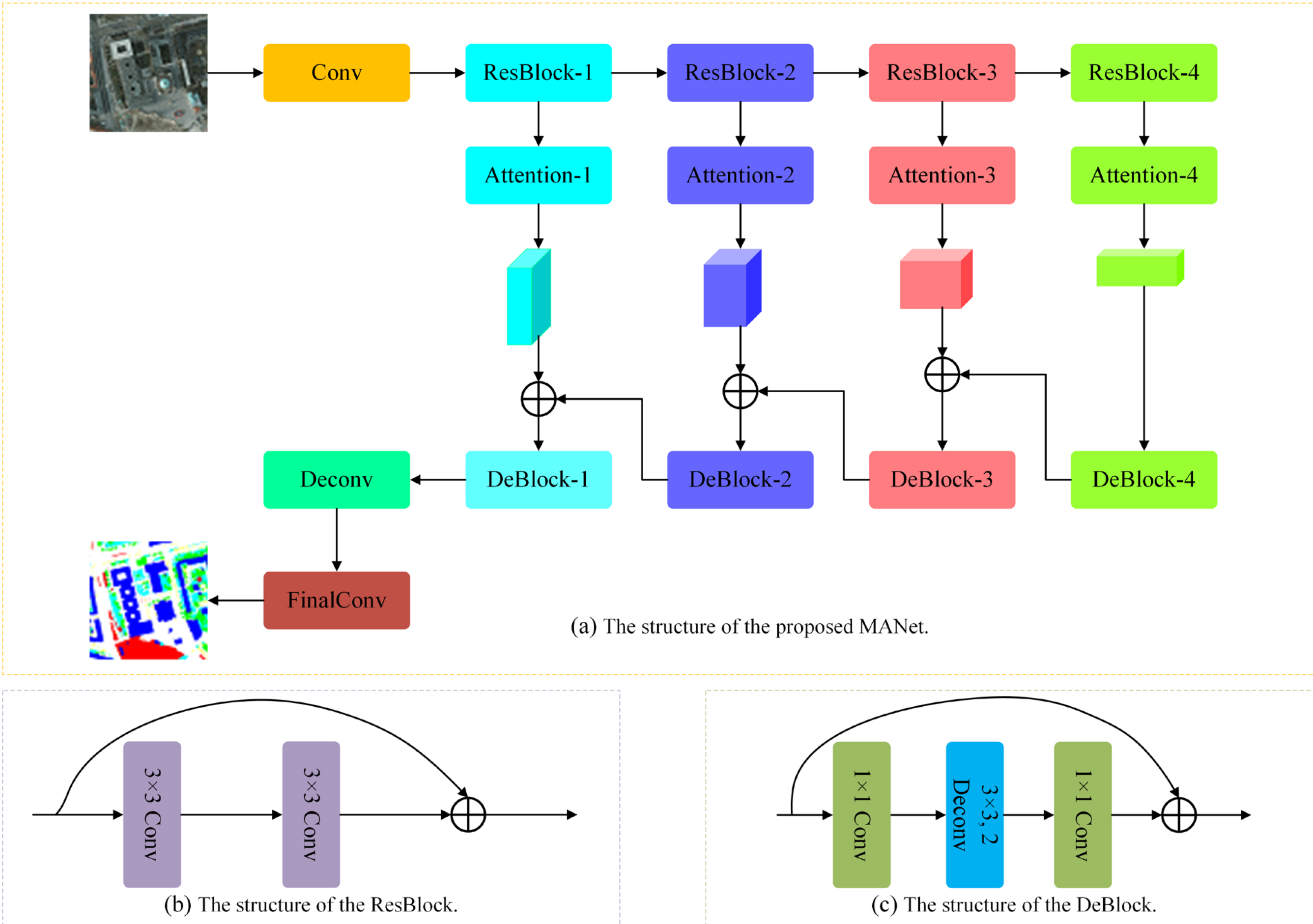


Fig. 4. The structure of (a) the proposed MANet, (b) the ResBlock, and (c) the DeBlock.

strategy to aggregate relevant contextual features hierarchically. The MANet extracts global contextual dependencies using multi-kernel attention.

II. Related Work

A. Attention Inspired by Human Perception

Due to the overwhelming computational requirement for perceiving surrounding scenes with detail equivalent to foveal vision, the selective visual attention endows humans with the ability to orientate rapidly towards salient objects in a sophisticated visual scene [51] and choose a subset of the available perceptual information before further processing. Inspired by the human attention mechanism, substantial algorithms have been developed over the last few decades [52-54].

Recently, a very large number of domains has been influenced significantly by the wave of DL, which emphasizes end-to-end hierarchical feature extraction in an automatic fashion. Integration of DL with the attention mechanism has great potential to transform the paradigm in this field. Attention in DL could be regarded as a weighted combination of the input feature maps, where the weights are hinged on the similarities between elements of the input [55]. Given that kernel learning [56] processes all inputs simultaneously and order-independently by computing the similarity between the inputs, attention could be interpreted as a kernel smoother [57] applied over the inputs in a sequence, where the kernel evaluates the similarity between different inputs. The formulae and mathematical proofs can be found in [55].

B. Dot-Product Attention

To enhance word alignment in machine translation, Bahdanau et al. [58] proposed the initial formulation of the dot-product attention mechanism. Subsequently, recurrences are entirely replaced by attention in the Transformer [44]. State-of-the-art records in most natural language processing tasks demonstrate the superiority of attention mechanisms amongst others. Wang et al. [45] modified dot-product attention for computer vision and proposed the non-local module. This method has been developed and applied to many tasks of computer vision, including image classification [46], object detection [47], semantic segmentation [5] and panoptic segmentation [48]. These successful applications demonstrated further the effectiveness and general utility of attention mechanisms.

C. Scaling Attention

Besides dot-product attention, there exists another set of techniques for scaling attention (or simply attention) in the literature. Unlike dot-product attention which models global dependency, scaling attention reinforces informative features and whittles information-lacking features. In the squeeze-and-excitation (SE) module [2], a global average pooling layer and a linear layer are harnessed to calculate a scaling factor for each channel, and then the channels are weighted accordingly. The convolutional block attention module (CBAM) [4], selective kernel unit (SK unit) [59] and efficient channel attention module (ECA) [60] further boost the SE block's performance. The principles and purposes of dot-product attention and scaling attention are entirely divergent. This paper focuses on dot-product attention due to its superiority in many computer vision and pattern recognition tasks.

D. Semantic Segmentation

FCN-based methods have brought tremendous progress and evolution in semantic segmentation. DilatedFCN and EncoderDecoder are two prominent directions followed by FCN. In DilatedFCNs [12, 14, 37-40, 61], dilate or atrous convolutions are harnessed to retain the receptive field-of-view, and a multi-scale context module is utilized to cope with high-level feature maps. Alternatively, EncoderDecoders [34, 35, 62-68] utilize an encoder to capture multi-level feature maps, which are then incorporated into the final prediction using a decoder.

DilatedFCN The dilated or atrous convolution [38, 61] has been demonstrated to be an effective technology for dense prediction and has achieved high accuracy in semantic segmentation. In DeepLab [14, 37], the atrous spatial pyramid pooling (ASPP), comprised of parallel dilated convolutions with diverse dilated rates, is able to embed context information, while the pyramid pooling module (PPM) enables PSPNet [12] to incorporate the contextual prior among different scales. Alternatively, EncNet [39] utilizes a context encoding module to exploit global context information. FastFCN [40] further replaces the dilated convolutions with a joint pyramid upsampling (JPU) module to reduce computational complexity. To extract abundant contextual relationships, a dot-product attention mechanism is attached to the DANet [5]. For further differentiating the same-object-class contextual pixels from the different-object-class contextual pixels, the object-contextual representation (OCR) module is elaborated by the OCRNet [69].

EncoderDecoder Skip connections are employed to integrate the high-level features generated by the decoder and the low-level features generated by the corresponding encoder, which are the essential structure of U-Net [34]. In the recent literature [62-64], the plain skip connections in U-Net are substituted by more subtle and elaborate skip connections which reduce the semantic gap between the encoder and decoder. Meanwhile, the structural development based on residual connections is also a promising direction [35, 65-68]. Taking DeepLab V3 as the encoder, DeepLab V3+ [37] combined the merits of DilatedFCN and EncoderDecoder in a single framework.

E. Attention-based Networks for Semantic Segmentation

Based on dot-product attention as well as its variants, various attention-based networks have been proposed to cope with the semantic segmentation task. Inspired by the non-local module [45], the Double Attention Networks (A^2-Net) [70], Dual Attention Network (DANet) [5], Point-wise Spatial Attention Network (PSANet) [71], Object Context Network (OCNet) [72], and Co-occurrent Feature Network (CFNet) [73] were proposed for scene segmentation by exploring the long-range dependency.

The computing resource required by dot-product attention modules is normally huge, which severely limits the application of attention mechanisms. Therefore, substantial researches have been implemented which aim to alleviate the bottleneck to efficiency and push the boundaries of attention, including accelerating the generation process of the attention matrix [69, 74-76], pruning the structure of the attention block [77], and optimizing attention based on low-rank reconstruction [78].

Meanwhile, another burgeoning research area for semantic segmentation is how to embed the dot-product attention into a Graph Convolutional Network (GCN) and optimize the complexity of the attention [79-83].

III. METHODOLOGY

A. Definition of Dot-Product Attention

Supposing N and C denote the length of input sequences and the number of input channels, respectively, where $N = H \times W$, and H and W denote the height and width of the input, given a feature $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in P^{\mathrm{N}\times\mathrm{C}}$, dot-product attention utilizes three projected matrices $\boldsymbol{W}_q \in P^{D_x \times D_k}$, $\boldsymbol{W}_k \in P^{D_x \times D_k}$, and $\boldsymbol{W}_v \in P^{D_x \times D_v}$ to generate the corresponding query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$ and value matrix $\boldsymbol{V}$ as:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{X}\boldsymbol{W}_q \in P^{N\times D_k}, \\ \boldsymbol{K} &= \boldsymbol{X}\boldsymbol{W}_k \in P^{N\times D_k}, \\ \boldsymbol{V} &= \boldsymbol{X}\boldsymbol{W}_v \in P^{\boldsymbol{N}\times \boldsymbol{D}_v}. \end{aligned} \tag{1}$$

where $D_{(\cdot)}$ means the dimension of $(\cdot)$. Please note that the shapes of $\boldsymbol{Q}$ and $\boldsymbol{K}$ are supposed to be identical. Therefore, we use the same symbol to represent their shapes.

A normalization function ρ evaluates the similarity between the *i*-th query feature $\boldsymbol{q}_i^T \in P^{D_k}$ and the *j*-th key feature $\boldsymbol{k}_j \in P^{D_k}$ by $\rho(\boldsymbol{q}_i^T \boldsymbol{k}_j) \in P^1$. Please note that the vectors in this paper default to column vectors. Generally, as the query feature and key feature are generated by diverse layers, the similarities between $\rho(\boldsymbol{q}_i^T \boldsymbol{k}_j)$ and $\rho(\boldsymbol{q}_j^T \boldsymbol{k}_i)$ are not symmetric. By calculating the similarities between all pairs of positions and taking the similarities as weights, the dot-product attention module computes the value at position i by aggregating the value features from all positions based on weighted summation:

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

The softmax is a standard normalization function as:

$$\rho(\boldsymbol{Q}^T\boldsymbol{K}) = \mathrm{softmax}_{\mathrm{row}}(\boldsymbol{Q}\boldsymbol{K}^T), \tag{3}$$

where $\mathrm{softmax}_{\mathrm{row}}$ indicates the application of the softmax function along each row of the matrix $\boldsymbol{Q}\boldsymbol{K}^T$.

The $\rho(\boldsymbol{Q}\boldsymbol{K}^T)$ models the similarities between all pairs of positions. However, as $\boldsymbol{Q} \in P^{N\times D_k}$ and $\boldsymbol{K}^T \in P^{D_k \times N}$, the product between $\boldsymbol{Q}$ and $\boldsymbol{K}^T$ belongs to $P^{N\times N}$, leading to $O(N^2)$ memory complexity and $O(N^2)$ computational complexity. As a consequence, the high resource-demand of the dot-product critically limits its application to large-scale inputs. One way to solve this problem is to modify the softmax [3], and another is to rethink the attention via the lens of the kernel. An illustration of the architecture for the dot-product attention mechanism is shown in Fig. 1.

B. Generalization of Dot-Product Attention Based on Kernel

Under the condition of the softmax normalization function, the *i*-th row of the result matrix generated by the dot-product attention module (equation 2) can be written as:

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

Then, equation (4) can be generalized for any normalization function and rewritten as:

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

where $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)$ indicates the function calculating the similarity between $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$. If $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=e^{\boldsymbol{q}_i{}^T\boldsymbol{k}_j}$, equation (5) is equivalent to equation (4). And $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)$ can be further expanded as $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=\phi(\boldsymbol{q}_i)^T\varphi(\boldsymbol{k}_j)$, where $\phi(\cdot)$ and $\varphi(\cdot)$ can be considered as kernel smoothers [55] if $\phi=\varphi$. Accordingly, the corresponding inner product space can be defined as $\langle\phi(\boldsymbol{q}_i),\varphi(\boldsymbol{k}_j)\rangle$.

Equation (4) can then be further rewritten as:

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

which can be simplified as:

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

As $\boldsymbol{K}\in P^{D_k\times N}$ and $\boldsymbol{V}^T\in P^{N\times D_v}$, the product between $\boldsymbol{K}$ and $\boldsymbol{V}^T$ belongs to $P^{D_k\times D_v}$, which reduces the complexity of the dot-product attention mechanism considerably.

C. Kernel Attention

We take $\phi(\cdot)=\varphi(\cdot)=\mathrm{softplus}(\cdot)$, where

$$\mathrm{softplus}(x)=\log(1+e^x).\tag{8}$$

The reason why we select $\mathrm{softplus}(\cdot)$ instead of $\mathrm{ReLU}(\cdot)$ is that the nonzero property of the softplus enables the attention to avoid zero gradients when the input is negative. Then, the similarity function can be embodied as:

$$\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{k}_j)=\mathrm{softplus}(\boldsymbol{q}_i)^T\mathrm{softplus}(\boldsymbol{k}_j),\tag{9}$$

thereby rewriting the equation (5) as:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i=\frac{\mathrm{softplus}(\boldsymbol{q}_i)^T\sum_{j=1}^{N}\mathrm{softplus}(\boldsymbol{k}_j)\boldsymbol{v}_j^T}{\mathrm{softplus}(\boldsymbol{q}_i)^T\sum_{j=1}^{N}\mathrm{softplus}(\boldsymbol{k}_j)},\tag{10}$$

which can be further written in a vectorized form as:

$$D(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\frac{\mathrm{softplus}(\boldsymbol{Q})\mathrm{softplus}(\boldsymbol{K})^T\boldsymbol{V}}{\mathrm{softplus}(\boldsymbol{Q})\sum_j\mathrm{softplus}(\boldsymbol{K})_{i,j}^T}.\tag{11}$$

As $\sum_{j=1}^{N}\mathrm{softplus}(\boldsymbol{k}_j)\boldsymbol{v}_j^T$ and $\sum_{j=1}^{N}\mathrm{softplus}(\boldsymbol{k}_j)$ can be calculated and reused for each query, the time and memory complexity of the proposed linear attention mechanism based on equation (11) is $O(N)$ only.

D. Multi-Attention-Network

For the spatial dimension, as the computational complexity of the dot-product attention mechanism exhibits a quadratic relationship with the size of the input ($N=H\times W$), we design an attention mechanism based on kernel attention, named KAM. For the channel dimension, the number of input channels C is normally far less than the number of pixels contained in the feature maps (i.e., $C\ll N$). Therefore, the complexity of the softmax function for channels (i.e., $O(C^2)$), is not large according to equation (3). Thus, we utilize the channel attention mechanism based on the dot-product [5], named CAM (Fig. 2). Using the kernel attention mechanism (KAM) and channel attention mechanism (CAM) which model the long-range dependencies of positions and channels, respectively, we design an attention block to enhance the discriminative ability of feature maps extracted by each layer (Fig. 3).

The structure of the proposed Multi-Attention-Network is illustrated in Fig. 4. We harness ResNet-50 pre-trained on ImageNet to extract feature maps. Specifically, five feature maps at different scales acquired from the outputs of [Conv, ResBlock-1, ResBlock-2, ResBlock-3, ResBlock-4] are adopted. The lowest level feature Res-4 is up-sampled directly by the DeBlock-4 which is comprised of a 3 × 3 deconvolution layer with stride=2 and two 1 × 1 convolution layers before and after the deconvolution layer. The feature maps generated by ResBlocks are then refined by corresponding attention blocks and added with the up-sampled lower feature maps. Subsequently, the fused features are up-sampled by the DeBlocks correspondingly. Finally, the output of the last DeBlock is up-sampled to the identical spatial resolution of the input by employing a deconvolution operation and fed into the final convolution layer to obtain the predicted segmentation map.

IV. Dataset and Experimental Setting

A. Datasets

The effectiveness of the linear attention mechanism is tested using the ISPRS Potsdam dataset and the ISPRS Vaihingen dataset(http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html). All the results for the ISPRS dataset are tested using ground reference data without eroded boundaries, so the evaluation indices are not as high as reported in certain elements of the literature.

Vaihingen: The Vaihingen semantic labeling dataset is composed of 33 images with an average size of 2494 × 2064 pixels and a GSD of 5 cm. The near-infrared, red and green channels together with DSM are provided in the dataset. There are 16 images in the training set and 17 images in the test set. We exploited 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. We did not use DSM in our experiments to reduce computation. Note that we use only the red, green and blue channels in our experiments. For training, we crop the raw images into 512 × 512 patches and augmented them by rotating, resizing, horizontal axis flipping, vertical axis flipping, and adding random noise.

Potsdam: The Potsdam dataset contains 38 fine-resolution images of size 6000 × 6000 pixels with a ground sampling distance (GSD) of 5 cm. The dataset provides near-infrared, red, green and blue channels as well as DSM and normalized DSM (NDSM). There are 24 images in the training set and 16 images in the test set. Specifically, we utilize 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 image named 7_10 with error annotations, for training. The process of the training dataset is identical to that for Vaihingen.

B. Evaluation Metrics

The performance of MANet on the three datasets is evaluated using the overall accuracy (OA), the mean Intersection over Union (mIoU), and the F1 score (F1), which are computed on the accumulated confusion matrix:

$$OA = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k}, \tag{12}$$

$$mIoU = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{13}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{14}$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negatives, respectively, for object indexed as class k. OA is calculated for all categories including the background.

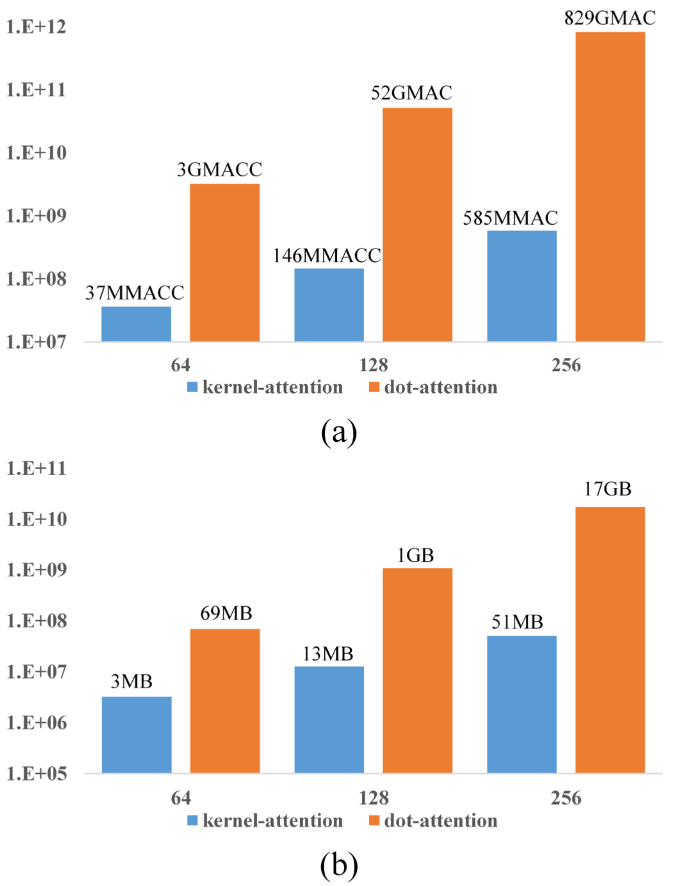


Fig. 5. Computation (a) and memory (b) requirements under different input sizes. The blue and orange bars depict the resource requirements of the kernel attention and dot-attention, respectively. The calculation assumes $D = D_v = 2D_k = 64$. The figure is in log scale.

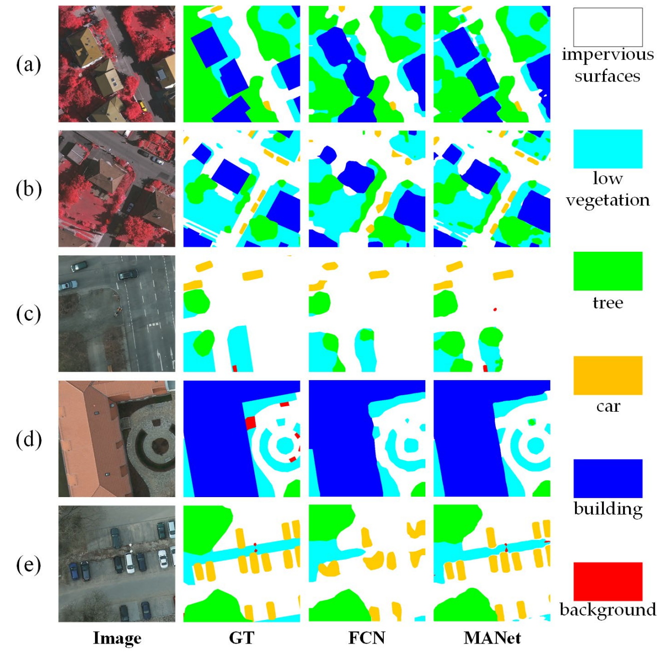


Fig. 6. Comparison of segmentation maps generated by FCN and our MANet, where (a) and (b) are from the Vaihingen dataset while (c)-(e) are from the Potsdam dataset

C. Experimental Setting

We select ResNet-50 pre-trained on ImageNet as the backbone for all comparative methods which are implemented with PyTorch. The optimizer is set as the Adam with a 0.0003 learning rate and 4 batch sizes. All the experiments are implemented on a single NVIDIA Tesla V100 GPU with 16 GB RAM. The cross-entropy loss function is used as a quantitative evaluation coupled with backpropagation to measure the disparity between the achieved segmentation maps and the ground reference:

$$loss(p, y) = -y\log(p) - (1 - y)\log(1 - p), \tag{15}$$

where p is the prediction generated by the network and y is the ground reference.

V. Experimental Results and Analysis

A. The Complexity of Kernel Attention

We analyze the efficiency merit of kernel attention over dot-product attention in both memory and computation in this section. Given a feature $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in P^{N\times C}$, both the dot-attention and kernel attention will generate the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$.

For the dot-attention, to compute the similarity using softmax function, we have to generate the $N \times N$ matrix by multiplying the transposed key matrix $\boldsymbol{K}$ and the value matrix $\boldsymbol{V}$, resulting in $O(D_k N^2)$ time complexity and $O(N^2)$ space complexity. Thus, to compute the similarity between each pair of positions, the dot-attention would occupy at least $O(N^2)$ memory and require $O(D_k N^2)$ computation.

For kernel attention, as the softmax function is substituted for kernel smoothers, we can alter the order of the commutative operation and avoid multiplication between the reshaped key matrix $\boldsymbol{K}$ and query matrix $\boldsymbol{Q}$. Therefore, we can calculate the product between $\text{softplus}(\boldsymbol{K})^T$ and $\boldsymbol{V}$ first and then multiply the result and $\boldsymbol{Q}$ with only $O(dN)$ time complexity and $O(dN)$ space complexity.

Dot-attention and kernel attention are compared in terms of resource consumption in Fig. 5. For a 64 × 64 × 64 input, the kernel attention yields a 21-fold saving of memory (69MB to 3MB) and an 89-fold saving of computation (3 GMMACC to 37 MMACC). With increasing input size, the gap widens. For a 64 × 256 × 256 input, the dot-attention requires unreasonable memory (17 GB) and computation (829 GMACC), while the kernel attention utilizes merely 1/340 memory (51MB) and 1/1417 computation (585 MMACC).

B. Ablation Study

In the proposed MANet, attention blocks are used to exploit global contextual representations and enhance the capability for feature extraction. To evaluate the performance of each attention block, we conduct ablation experiments using different settings listed in Table I and Table II.

Table I shows the comparison of the ablation study on the Vaihingen dataset which demonstrates that the utilization of attention blocks increases the accuracy significantly compared with the baseline FCN with DeBlocks (ResNet-50), particularly for small objects, i.e., the Car. Even using a single attention block to enhance the context information could gain at least 1.44% improvement in OA, 2.42% in mean F1-score, and 3.46% in mIoU. Moreover, low-level attention blocks contribute more than those in high-levels as the former contains rich context information. When all attention blocks are attached, the remarkable 6.18%increase in OA, 5.63% in mean F1-score, and 7.53% in mIoU are achieved. These results demonstrate that our attention block brings significant breakthough to semantic segmentation by exploiting global context information from different perspectives.

The ablation study results of the Potsdam dataset are reported in Table II. The utilization of a single attention block increases > 2.50% in mean F1-score, 0.66% in OA, and 4.72% in mIoU, while the accuracy increase brought in by all attention blocks

TABLE I
QUANTITATIVE COMPARISON RESULTS ON THE VAIHINGEN TEST SET.

Method	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
FCN	89.731	93.169	80.569	88.890	71.552	84.782	87.987	75.872
FCN+Attention1	91.379	94.271	82.757	89.337	78.267	87.202	89.424	79.330
FCN+Attention2	91.831	94.612	82.791	89.671	83.543	88.490	89.703	80.714
FCN+Attention3	91.898	94.801	83.692	89.268	83.019	88.536	89.895	81.072
FCN+Attention4	91.854	94.787	83.867	89.855	86.045	89.282	90.202	81.729
Proposed MANet	93.024	95.471	84.637	89.978	88.945	90.411	90.963	83.397

TABLE II
QUANTITATIVE COMPARISON RESULTS ON THE POTSDAM TEST SET.

Method	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
FCN	90.839	95.591	84.097	84.75	84.952	88.046	88.022	81.419
FCN+Attention1	90.880	95.267	85.845	87.113	93.682	90.557	88.682	86.140
FCN+Attention2	91.471	94.855	85.719	88.153	96.013	91.242	89.134	86.681
FCN+Attention3	92.036	95.207	86.820	87.446	95.155	91.333	89.558	87.107
FCN+Attention4	92.949	96.749	87.115	87.701	95.785	92.060	90.493	87.940
Proposed MANet	93.254	96.632	87.991	88.948	96.387	92.642	91.054	89.012

TABLE III
QUANTITATIVE COMPARISON RESULTS ON THE VAIHINGEN TEST SET

Method`	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
FCN [1]	89.731	93.169	80.569	88.890	71.552	84.782	87.987	75.872
FCN+SE [2]	91.886	94.604	83.185	89.379	77.084	87.228	89.711	80.560
FCN+CBAM [4]	91.592	94.766	84.195	89.494	80.877	88.185	89.956	80.577
FCN+EAM [3]	92.450	95.075	83.743	89.479	86.231	89.396	90.324	81.993
FCN+FAM [11]	92.605	94.214	84.154	**90.138**	84.897	89.202	90.304	81.401
FCN+LAM [15]	92.075	94.820	83.420	89.730	83.626	88.734	90.047	81.021
DANet [5]	91.384	94.100	83.086	89.015	76.794	86.876	89.473	78.050
PSPNet [12]	91.383	94.196	83.050	88.713	75.021	86.473	89.358	77.486
DeepLabV3+ [14]	91.630	94.086	82.505	87.991	77.656	86.774	89.124	78.722

TABLE IV
THE ABLATION STUDY ABOUT BACKBONES ON THE POTSDAM DATASET.

Method`	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
FCN [1]	90.839	95.591	84.097	84.750	84.952	88.046	88.022	81.419
FCN+SE [2]	91.647	96.118	86.078	88.009	95.077	91.386	89.598	87.140
FCN+CBAM [4]	92.719	96.127	85.773	88.217	95.827	91.733	89.898	87.722
FCN+EAM [3]	92.748	96.041	86.480	88.407	96.023	91.940	90.241	87.861
FCN+FAM [11]	92.580	96.127	86.787	88.165	95.792	91.890	90.179	87.875
FCN+LAM [15]	92.771	96.406	86.476	87.277	96.090	91.804	90.119	87.542
DANet [5]	91.944	96.348	86.003	87.673	86.010	89.596	89.728	83.710
PSPNet [12]	92.199	96.107	86.940	88.339	86.302	89.977	90.143	84.056
DeepLabV3+ [14]	92.093	95.282	85.549	86.537	94.813	90.855	89.176	86.331
EaNet [17]	**92.872**	96.302	86.163	87.991	95.303	91.726	90.154	87.594
Proposed MANet	93.254	**96.632**	**87.991**	**88.948**	**96.387**	**92.642**	**91.054**	**89.012**

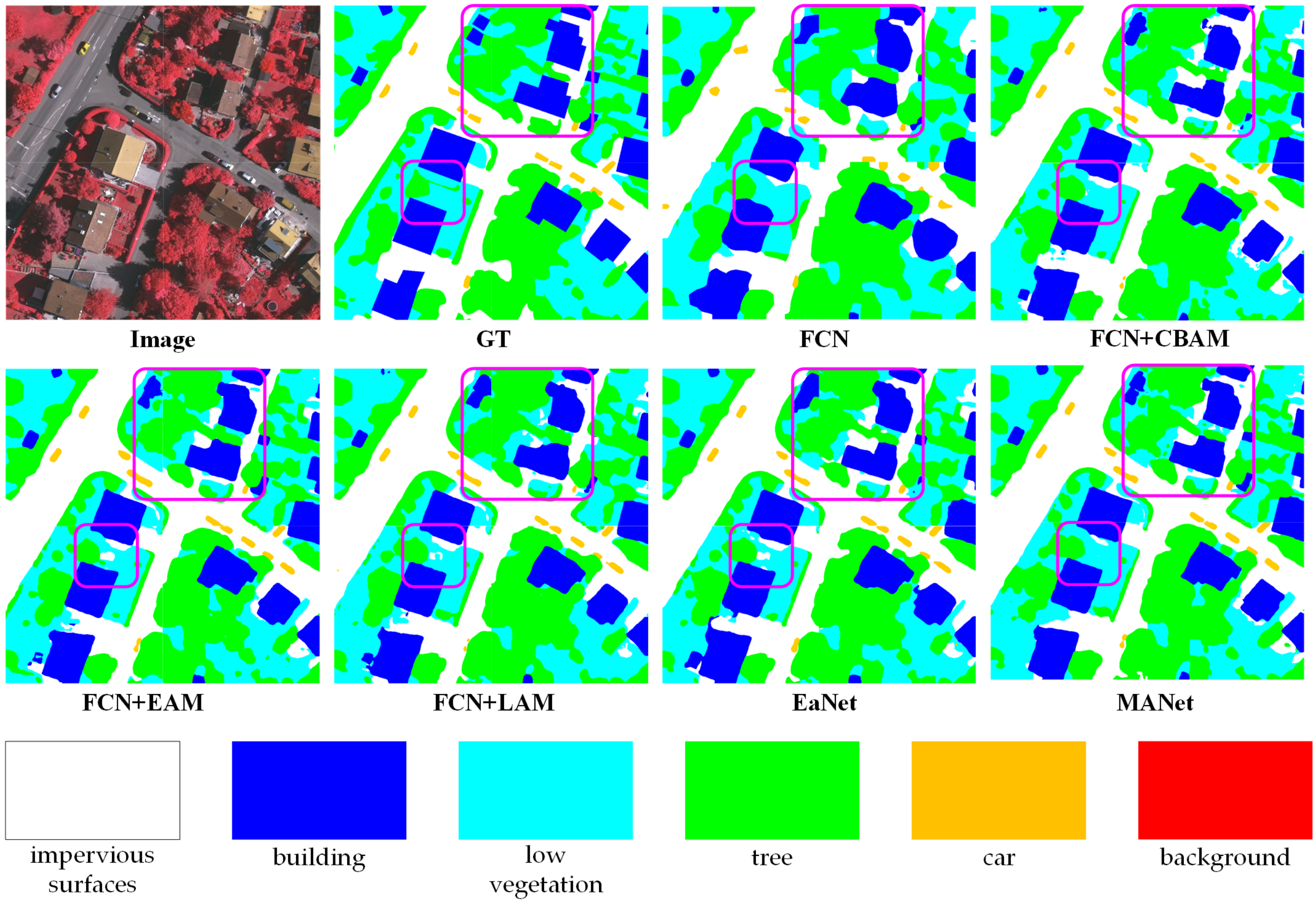


Fig. 7. Qualitative comparisons (1024 × 1024 patches) between our method and baseline on Vaihingen test set.

are 4.60% in mean F1-score, 3.03% in OA, and 7.6% in mIoU, respectively.

To validate the effectiveness visually and qualitatively, we present comparison of the segmented features generated by FCN and our MANet in Fig. 6. Due to the limited receptive field, the FCN generates the category of a specific pixel in consideration of its a few neighborhood only, leading to visually fragmented maps and confusion of objects. By contrast, the proposed attention block can model the global dependency of all pixels in the input features, and capture the global context information with enhanced segmentation accuracy. Particularly, the complex contour of the Low vegetation is preserved completely by our MANet (Fig. 6 (d)). Meanwhile, the category of Car generated by the proposed MANet is classified effectively and superior to the FCN as shown in Fig. 6 (b) and Fig. 6 (e).

C. Quantitative Comparison Diverse Methods

To further confirm the effectiveness of the proposed MANet, we compare our method with state-of-the-art approaches presented in the literature. Specifically, the comparative methods not only include the scaling attention mechanism i.e., SE module [2] and CBAM [4] but also consider the simplified dot-product attention mechanism i.e., EAM [3], FAM [11], and LAM [15]. Besides, several comparative networks are also taken into comparison, including the DANet [5] which utilizes the conventional dot-product attention mechanism and other receptive-field-enlarging, i.e., PSPNet [12], DeepLabV3+ [14], as well as EaNet [17]. For fair comparison, all experiments are conducted under the same setting for training and testing. All methods are implemented based on the same ResNet-50 backbone while the FCN-based methods are equipped with DeBlocks. The detailed segmentation accuracy on the Vaihingen dataset and Potsdam dataset of each network is listed in TABLE III and TABLE IV, respectively.

TABLE V
COMPARISON WITH SCALING ATTENTION.

Dataset	Method	Mean F1	OA	mIoU
Vaihingen	FCN [1]	84.782	87.987	75.872
	+ SE [2]	87.228	89.711	80.560
	+ CBAM [4]	88.185	89.956	80.577
	+ Ours	**90.411**	**90.963**	**83.397**
Potsdam	FCN [1]	88.046	88.022	81.419
	+ SE [2]	91.386	89.598	87.140
	+ CBAM [4]	91.733	89.898	87.722
	+ Ours	**92.642**	**91.054**	**89.012**

1) Comparison with Scaling Attention

The scaling attention mechanisms are designed to reinforce informative features and reduce information-lacking features, instead of capturing global context information such as dot-product attention mechanism. Hence, the scale attention and dot-product attention are not identical. In our experiments, we compare the performance of our method with two well-verified scaling attention mechanisms, i.e., SE module [2] and CBAM

[4], andthe results are shown in Table V. As the CBAM [4] introduces the extra channel scaling attention block compared with the SE module [2], "+CBAM" achieves higher accuracies compared with "+SE". In contrast, our MANet extracts global context correlation from the feature maps. Experimental results demonstrate the superiority of our method compared with scaling attention mechanism.

TABLE VI
COMPARISON WITH SIMPLIFIED DOT-PRODUCT ATTENTION.

Dataset	Method	Mean F1	OA	mIoU
Vaihingen	FCN [1]	84.782	87.987	75.872
	+ EAM [3]	89.396	90.324	81.993
	+ FAM [11]	89.202	90.304	81.401
	+ LAM [15]	88.734	90.047	81.021
	+ Ours	**90.411**	**90.963**	**83.397**
Potsdam	FCN [1]	88.046	88.022	81.419
	+ EAM [3]	91.940	90.241	87.861
	+ FAM [11]	91.890	90.179	87.875
	+ LAM [15]	91.804	90.119	87.542
	+ Ours	**92.642**	**91.054**	**89.012**

2) Comparison with Simplified Dot-product Attention

As both space and time consumption of the standard dot-product attention mechanism increase quadratically with the input size, several research has devoted to simplify the complexity of the attention mechanism, including the efficient attention mechanism (EAM) [3], the fast attention mechanism (FAM) [11], and the linear attention mechanism (LAM) [15]. As shown in Table VI, the proposed KAM achieves the best accuracy compared with other simplified dot-product attention mechanism, due to the appropriate simplified scheme adopted.

TABLE VII
COMPARISON WITH OTHER COMPARATIVE NETWORKS.

Dataset	Method	Mean F1	OA	mIoU
Vaihingen	FCN [1]	84.782	87.987	75.872
	+ DAB [5]	86.876	89.473	78.050
	+ PPM [12]	86.473	89.358	77.486
	+ ASPP [14]	86.774	89.124	78.722
	+ LKPP [17]	88.578	90.252	80.580
	+ Ours	**90.411**	**90.963**	**83.397**
Potsdam	FCN [1]	88.046	88.022	81.419
	+ DAB [5]	89.596	89.728	83.710
	+ PPM [12]	89.977	90.143	84.056
	+ ASPP [14]	90.855	89.176	86.331
	+ LKPP [17]	91.726	90.154	87.594
	+ Ours	**92.642**	**91.054**	**89.012**

3) Comparison with other Comparative Networks

The conventional dot-product attention mechanism is introduced in DANet [5] to capture feature dependencies both in spatial and channel dimensions, while PSPNet [12], DeepLabV3+ [14], and EaNet [17] employ variants of spatial pyramid pooling (SPP) to enlarge the receptive field. The proposed MANet models the global context information in the input features instead of expanding finite receptive fields by convolution layers with different kernel sizes (e.g. SPP). Besides, we capture the context information in multi-layers rather than in the lowest layer only (e.g. DANet). Hence, the performance of our MANet exceeds these comparative networks with a large margin.

D. Evaluation in Efficiency

We evaluate our kernel attention mechanism not only with the standard dot-product attention mechanism but also the scaling attention mechanism and the receptive-field-enlarging modules in terms of the computation complexity measured with GFLOPs (G), the number of parameters measured with Millions (M), as well as the memory consumption measured with Megabytes (MB). Note, we evaluate the consumption of the modules with the cost of 3×3 convolution for dimension reduction, and we do not consider the cost of backbone to ensure the fairness of the comparison. As illustrated in Table VIII, for input in the size of 2048 × 128 × 128, our KAM requires 10× less GPU memory usage and significantly reduces about 78% parameters and computation complexity when compared with the DAB [5] based on the dot-product attention mechanism. Besides, it can be seen that our KAM is more efficient than other specially-designed modules when processing fine-resolution feature maps.

TABLE VIII
EFFICIENCY COMPARISON WITH DIFFERENT MODULES WHEN PROCESSING INPUT FEATURE MAP OF SIZE [1 × 2048 × 128 × 128] DURING INFERENCE STAGE.

Method	Complexity (G)	Parameters (M)	Memory (MB)
SE [2]	618.6	38.3	256
CBAM [4]	618.6	38.3	256
EAM [3]	154.7	9.4	288
FAM [11]	**85.9**	**5.3**	**160**
LAM [15]	**85.9**	**5.3**	**160**
DAB [5]	392.2	23.9	1546
PPM [12]	309.5	23.1	257
ASPP [14]	503.0	15.1	284
LKPP [17]	884.2	54.5	818
Ours	**85.9**	**5.3**	**160**

E. Qualitative Analysis of the Segmentation Results

Examples of the predicted patches in the size of 1024 × 1024 are provided in Fig. 7 and Fig. 8, where regions with obvious improvement are highlighted by red boxes. Due to the loss of spatial information, the segmentation maps generated by FCN are ambiguous, particularly at the contour of objects. The utilization of scaling attention mechanisms, i.e., SE [2] and CBAM [4] brings limited accuracy increase. Although receptive-field-enlarging networks like PSPNet [12] and DeepLabV3+ [14] demonstrate enhanced segementation in confusing areas, the complex contour of the low vegetation is not generated completely shown in Fig. 8. With attention blocks extracting global context information in multi-layers, the proposed MANet not only reduces the incomplete and irregular semantic objects, but also better preserves the geometric details and complex contours. Specifically, the geometry of buildings in Fig. 7 as well as the edges of the low vegetation in Fig. 8 are preserved. Besides, there are significant improvement in preserving the boundaries and reducing fragmented segments.

F. Discussion on the Attention Mechanism

Selective visual attention endows humans with the ability to orientate towards conspicuous objects over the visual scene in a computationally efficient manner. Thus, the attention mechanism, inspired by the biological mechanism, is intended as a computationally efficient structure with configurable flexibility. By representing the concept of attention via the lens of the kernel [55], we design a kernel attention module with $O(N)$ complexity. The effectiveness and efficiency of the proposed kernel attention is demonstrated consistently across a

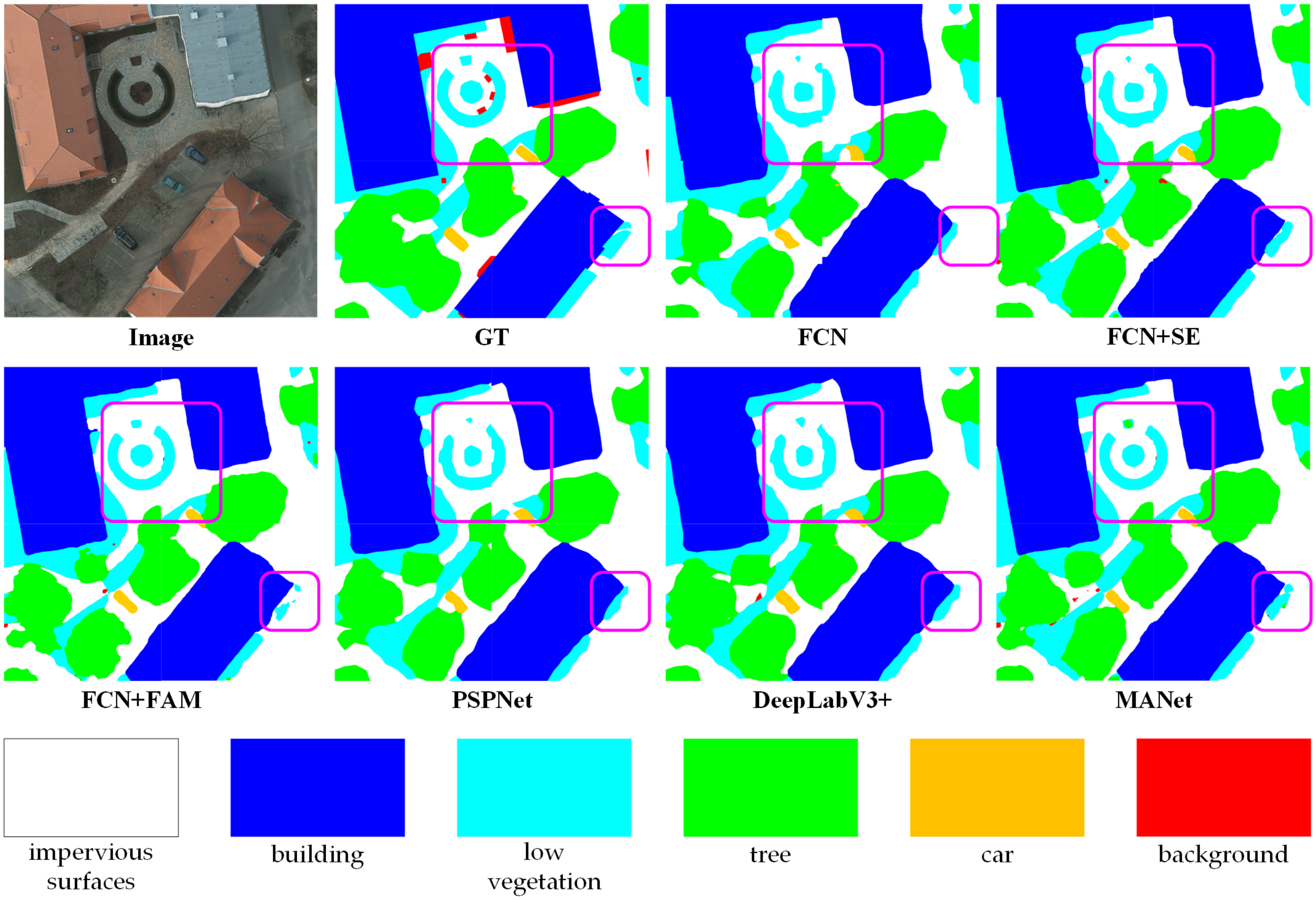


Fig. 8. Qualitative comparisons (1024 × 1024 patches) between our method and baseline on Potsdam test set.

wide range of quantitative experiments. We envisage the demonstrated resource efficiency will encourage more pervasive and flexible combinations between attention mechanisms and networks.

VI. Conclusion

This paper proposes kernel attention to reduce the complexity of the dot-product attention mechanism into $O(N)$. By integrating kernel attention and ResNet-50, we design a Multi-Attention-Network (MANet) comprised of a multi-scale strategy to incorporate semantic information at different levels, together with self-attention modules to aggregate relevant contextual features hierarchically. MANet exploits contextual dependencies over local features producing increased accuracy and computational efficiency. We implement a series of experiments involving the complex task of semantic segmentation of fine-resolution remote sensing images. MANet produces consistently the best classification performance with the highest accuracy. An extensive ablation study is conducted to evaluate the impact of the individual components of the proposed framework. Experimental results on ISPRS Vaihingen and Potsdam datasets demonstrate that the performance of the proposed framework greatly exceeds comparative benchmark methods.

References

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

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.

[3] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient Attention: Attention with Linear Complexities," *arXiv preprint arXiv:1812.01243,* 2018.

[4] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3-19.

[5] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146-3154.

[6] J. Zhang, L. Feng, and F. Yao, "Improved maize cultivated area estimation over a large scale combining MODIS–EVI time series data and crop phenological information," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 94, pp. 102-113, 2014.

[7] C. Zhang *et al.*, "Joint Deep Learning for land cover and land use classification," *Remote sensing of environment,* vol. 221, pp. 173-187, 2019.

[8] D. Sulla-Menashe, J. M. Gray, S. P. Abercrombie, and M. A. Friedl, "Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product," *Remote Sensing of Environment,* vol. 222, pp. 183-194, 2019.

[9] C. Zhang, P. A. Harrison, X. Pan, H. Li, I. Sargent, and P. M. Atkinson, "Scale Sequence Joint Deep Learning (SS-JDL) for land

use and land cover classification," *Remote Sensing of Environment,* vol. 237, p. 111593, 2020.

[10] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016,* vol. 3, pp. 473-480, 2016.

[11] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention," *arXiv preprint arXiv:2006.16236,* 2020.

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

[13] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," 1978.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587,* 2017.

[15] R. Li, S. Zheng, C. Duan, J. Su, and Z. Ce, "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters,* vol. 10.1109/LGRS.2021.3063381, 2021.

[16] A. Huete, "Huete, AR A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment," *Remote sensing of environment,* vol. 25, pp. 295-309, 1988.

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

[18] H. Zhao and X. Chen, "Use of normalized difference bareness index in quickly mapping bare areas from TM/ETM+," in *International geoscience and remote sensing symposium*, 2005, vol. 3, p. 1666.

[19] H. Li *et al.*, "Mapping urban bare land automatically from Landsat imagery with a simple index," *Remote Sensing,* vol. 9, no. 3, p. 249, 2017.

[20] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment,* vol. 58, no. 3, pp. 257-266, 1996.

[21] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International journal of remote sensing,* vol. 27, no. 14, pp. 3025-3033, 2006.

[22] G. Rutherford, A. Guisan, and N. Zimmermann, "Evaluating sampling strategies and logistic regression methods for modelling complex land cover changes," *Journal of Applied Ecology,* vol. 44, no. 2, pp. 414-424, 2007.

[23] Q. Du and C.-I. Chang, "A linear constrained distance-based discriminant analysis for hyperspectral image classification," *Pattern Recognition,* vol. 34, no. 2, pp. 361-373, 2001.

[24] U. Maulik and I. Saha, "Automatic fuzzy clustering using modified differential evolution for image classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 48, no. 9, pp. 3503-3510, 2010.

[25] Y. Fu *et al.*, "An improved combination of spectral and spatial features for vegetation classification in hyperspectral images," *Remote Sensing,* vol. 9, no. 3, p. 261, 2017.

[26] K. Tatsumi, Y. Yamashiki, M. A. C. Torres, and C. L. R. Taipe, "Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data," *Computers and Electronics in Agriculture,* vol. 115, pp. 171-179, 2015.

[27] C. Adede, R. Oboko, P. W. Wagacha, and C. Atzberger, "A Mixed Model Approach to Vegetation Condition Prediction Using Artificial Neural Networks (ANN): Case of Kenya's Operational Drought Monitoring," *Remote Sensing,* vol. 11, no. 9, p. 1099, 2019.

[28] C. Zhang *et al.*, "A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 140, pp. 133-144, 2018.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730-3738.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[31] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779,* 2019.

[32] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of Hyperspectral Image Based on Double-Branch Dual-Attention Mechanism Network," *Remote Sensing,* vol. 12, no. 3, p. 582, 2020.

[33] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, no. 12, pp. 2481-2495, 2017.

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

[35] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925-1934.

[36] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics,* 2020.

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

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122,* 2015.

[39] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151-7160.

[40] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv preprint arXiv:1903.11816,* 2019.

[41] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579,* 2015.

[42] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," ed, 2018.

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog,* vol. 1, no. 8, p. 9, 2019.

[44] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794-7803.

[46] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354-7363.

[47] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3286-3295.

[48] Y. Li *et al.*, "Attention-guided unified network for panoptic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7026-7035.

[49] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509,* 2019.

[50] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451,* 2020.

[51] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience,* vol. 2, no. 3, pp. 194-203, 2001.

[52] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, no. 6, pp. 989-1005, 2009.

[53] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, no. 5, pp. 802-817, 2006.

[54] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing,* vol. 27, no. 5, pp. 2368-2378, 2017.

[55] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel," *arXiv preprint arXiv:1908.11775,* 2019.

[56] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[57] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[58] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

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

[60] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534-11542.

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

[62] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: Springer, 2018, pp. 3-11.

[63] H. Huang *et al.*, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 1055-1059.

[64] R. Li, C. Duan, and S. Zheng, "MACU-Net Semantic Segmentation from High-Resolution Remote Sensing Images," *arXiv preprint arXiv:2007.13083,* 2020.

[65] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognition,* vol. 106, p. 107404, 2020.

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

[67] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks,* vol. 121, pp. 74-87, 2020.

[68] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Transactions on Image Processing,* vol. 28, no. 5, pp. 2545-2557, 2018.

[69] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065,* 2019.

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

[71] H. Zhao *et al.*, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267-283.

[72] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916,* 2018.

[73] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 548-557.

[74] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603-612.

[75] F. Zhang *et al.*, "Acfnet: Attentional class feature network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6798-6807.

[76] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," *arXiv preprint arXiv:1907.12273,* 2019.

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

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

[79] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 433-442.

[80] S. Zhang, X. He, and S. Yan, "Latentgnn: Learning efficient non-local relations for visual recognition," in *International Conference on Machine Learning*, 2019, pp. 7374-7383.

[81] B. Wu *et al.*, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," *arXiv preprint arXiv:2006.03677,* 2020.

[82] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 1853-1863.

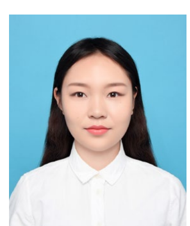
[83] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 9225-9235.

Rui Li received a bachelor's degree from the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China in 2019. He is currently pursuing a master's degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include semantic segmentation, hyperspectral image classification, and deep learning.

Shunyi *Z*heng received the Post-Doctorate from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2002. He is currently a Professor at the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. His research interests include remote sensing data processing, digital photogrammetry, and three-dimensional reconstruction. Prof. Zheng received the First Prize for Scientific and Technological Progress in Surveying and Mapping, China, in 2012 and 2019.

Ce Zhang received Ph.D. Degree in Geography from Lancaster Environment Centre, Lancaster University, U.K. in 2018. He was the recipient of a prestigious European Union (EU) Erasmus Mundus Scholarship for a European Joint MSc programme between the University of Twente (The Netherlands) and the University of Southampton (U.K.). Dr. Zhang is currently a Lecturer in Geospatial Data Science at the Centre of Excellence in Environmental Data Science (CEEDS), jointly venture between Lancaster University and UK Centre for Ecology & Hydrology (UKCEH). His major research interests include geospatial artificial intelligence, machine learning, deep learning, and remotely sensed image analysis.



Chenxi Duan received a bachelor's degree from the College of Geology Engineering and Geomatics, Chang'an University, Xi'an, China in 2019. She is currently pursuing a master's degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include cloud removal, numerical optimization, and machine learning.

Jianlin Su received the master's degree from the School of Mathematics, Sun Yat-sen University, Guangzhou, China. He is currently the senior researcher in the Shenzhen Zhuiyi Technology Co., Ltd

His research interests are focused on the generative model, including the language model and Seq2Seq in natural language processing, and the GAN, VAE, and flow in the computer vision. Besides, he is also interested in the basic theory of machine learning. His homepage is http://jianlin.su.

Peter M. Atkinson received the MBA degree from the University of Southampton, Southampton, U.K. in 2012, and the Ph.D. degree from the University of Sheffield, Sheffield, U.K. (NERC CASE Award with Rothamsted Experimental Station), in 1990.

He is currently the Dean of the Faculty of Science and Technology with Lancaster University, U.K. He was previously a Professor of geography with the University Southampton, where he is currently a Visiting Professor. He is also a Visiting Professor with Queen's University Belfast, U.K., and with the Chinese Academy of Sciences, Beijing, China. He has authored more than 270 peer-reviewed articles in international scientific journals and around 50 refereed book chapters. He has also edited nine journal special issues and eight books. His research interests include remote sensing, geographical information science, and spatial (and space-time) statistics applied to a range of environmental science and socio-economic problems.

Prof. Atkinson is an Editor-in-Chief of Science of Remote Sensing, a sister journal of Remote Sensing of Environment. He is also an Associate Editor for the Computers and Geosciences and sits on the editorial boards of several further journals including Geographical Analysis, Spatial Statistics, the International Journal of Applied Earth Observation and Geoinformation, and Environmental Informatics. He sits on various international scientific committees. He previously held the Belle van Zuylen Chair with Utrecht University, The Netherlands and is recipient of the Peter Burrough Award of the International Spatial Accuracy Research Association.