Lightweight Attention Network for Very High Resolution Image Semantic Segmentation

Renchu Guan, *Senior Member, IEEE*, Mingming Wang, Lorenzo Bruzzone, *Fellow, IEEE*, Haishi Zhao, *Member, IEEE* and Chen Yang, *Senior Member, IEEE*

***Abstract*—Semantic segmentation is one of the most challenging tasks for very high resolution (VHR) remote sensing applications. Deep convolutional neural networks (CNN) based on the attention mechanism have shown outstanding performance in VHR remote sensing images semantic segmentation. However, existing attention-guided methods require the estimation of a large number of parameters that are affected by the limited number of available labeled samples that results in underperforming segmentation results. In this paper, we propose a multi-scale feature fusion lightweight model (MSFFL) to greatly reduce the number of parameters and improve the accuracy of semantic segmentation. In this model, two parallel enhanced attention modules, i.e., the spatial attention module (SAM) and the channel attention module (CAM) are designed by introducing encoding position information. Then a covariance calculation strategy is adopted to recalibrate the generated attention maps. The integration of enhanced attention modules into the proposed lightweight module results in an efficient lightweight attention network (LiANet). The performance of the proposed LiANet is assessed on two benchmark datasets. Experimental results demonstrate that LiANet can achieve promising performance with a small number of parameters.**

***Index Terms*—Semantic segmentation, very high resolution (VHR) images, position information, covariance, lightweight, remote sensing.**

I. Introduction

VERY high resolution (VHR) remote sensing images with spatial resolution from meter to submeter have offered tremendous opportunities to distinguish and identify objects at a fine spatial scale. Semantic segmentation is an indispensable process for VHR remote sensing images in many applications including land use/cover mapping [1], urban planning [2], land/marine ecosystem processes [3] and environment monitoring [4]. The rich details and structural information of VHR remote sensing images lead to a dramatic increase in the spectral heterogeneity of the same geographical entity. In addition, the phenomenon of having different objects with similar spectral signatures caused by the few spectral information provided in VHR remote sensing images reduces the separability of different ground objects. Both these pose the challenge of fine-grained segmentation of VHR remote sensing imagery [5].

In recent years, deep convolutional neural networks (DCNNs), which have powerful feature extraction capabilities, have become one of the most popular methods for semantic segmentation [6]. The DCNN-based semantic segmentation models can be classified into different types. One type of network is the DeepLab series [7-10]. It can extract global feature information based on a large receptive field, also modeling multi-scale features efficiently [11]. The most recent DeepLab v3+ achieved excellent performance in various segmentation tasks by combining the astral spatial pyramid pooling (ASPP) and a powerful variant of Xception [10], [12]. Zhao et al. [13] proposed a pyramid pooling module that aggregates contextual information at different scales and improves the ability to obtain global semantic information. Another type of network is related to models with encoder-decoder structure. The most typical one is UNet, which can make full use of the context information of each stage by combining feature maps from the extraction path with the corresponding expanding path [14]. Given the simplicity of the encoder-decoder style, researchers proposed many improved UNet models [15-17] for semantic segmentation tasks.

However, DeepLab series models and architectures based on the encoder-decoder structure are not effective in taking into account the global semantic information at high spatial resolution, which is critical for segmentation performance. Recently, the attention mechanism gained huge success in machine translation tasks and has been introduced in VHR remote sensing image semantic segmentation. In this context, the spatial information related to the relative position of each pixel is particularly important, as it plays a guiding role in the attention strategy. The attention mechanism-based methods can efficiently obtain global dependencies while preserving the spatial information of features.

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0112501 and 2021ZD0112502, National Natural Science Foundation of China under Grant 42272340, 42241163, U22A2098, 61972174 and 62172187, the Science and Technology Planning Project of Guangdong Province under Grant 2020A0505100018, Guangdong Universities' Innovation Team Project under Grant 2021KCXTD015 and Guangdong Key Disciplines Project under Grant 2021ZDJS138, the Science-Technology Development Plan Project of Jilin Province under grants 20230101311JC. (*Corresponding authors: Haishi Zhao and Chen Yang*)

R. C. Guan M. M Wang and H. S. Zhao are with the Key Laboratory for Symbol Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, 130012, China (e-mail: guanrenchu@jlu.edu.cn, zhaohs18@ mails.jlu.edu.cn).

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

C. Yang is with the College of Earth Sciences, Jilin University, Changchun, 130061, China, and Key Laboratory of Lunar and Deep Space Exploration, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, China (e-mail: yangc616@jlu.edu.cn).

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

In the literature, there are many attention mechanism-based VHR semantic segmentation methods. Liu et al. [18] proposed an adaptive fusion network (AFNet), in which a scale-feature attention module (SFAM) and a scale-layer attention module (SLAM) are designed to fuse multiscale feature maps and segment the ground objects with high intra-class difference and various scales. Although AFNet achieved competitive performance on two International Society for Photogrammetry and Remote Sensing (ISPRS) datasets, it requires the estimation of a huge number of parameters. Peng et al. [19] used channel-wise attention to select informative feature maps and fused them with different level feature maps. Li et al. [20] used the convolutional block attention module (CBAM) [21] for ultra-high resolution remote sensing imagery semantic segmentation, which performed well on two ISPRS datasets. Zhao et al. [22] proposed a pyramid attention pooling module, in which attention mechanism was introduced to adaptively refine the features in the multiscale module. Besides, the self-attention, which can model the relationship among features at the global scale, has also been introduced for VHR semantic segmentation. Non-Local [23], the first self-attention method in computer vision, used a dot-product between query feature map and key feature map to gather information and achieved a great advantage in building channels and spatial attention. Fu et al. [24] proposed a parallel self-attention network to construct global context from spatial and channel domains. The combination of DCNN with the attention mechanism, especially self-attention, to generate global attention maps can effectively achieve fine-grained rich contextual information. This addresses the effects of large spectral variability within the same category and the tendency of homogeneity among different categories in VHR remote sensing image. However, this results in a sharp increase in the number of parameters of the model and in the computational time. In general, existing attention mechanism-based VHR semantic segmentation methods are limited by the large number of parameters and the high computational cost and cannot be applied to scenarios with limited resources. In addition, on-orbit real-time image segmentation is an urgent requirement for intelligent remote sensing systems, which can extract target information from images in real time to better serve tasks, such as pollution monitoring, forest fire early warning, and land cover type change monitoring [25]. Thus, it is necessary to design lightweight models to reduce the number of both parameters and calculations.

Existing lightweight semantic segmentation models [26-35] can be divided into two categories. One embeds lightweight convolutional neural networks directly as feature extractors in the previous models [36], [37]. The other one exploits the idea of designing lightweight modules for feature extraction to reduce the computational complexity [38] or redesigns the attention mechanism computation to reduce the quadratic complexity to linear complexity [26], [34], [35]. Although these methods can effectively reduce the number of parameters and computational complexity, the lightweight operation usually makes their feature representation capability reduced, which in turn results in a significant reduction in segmentation performance. Therefore, it is critical to design a lightweight attention-based VHR semantic segmentation model that can maintain competitive segmentation performance.

To address the above problems, in this paper, an efficient lightweight attention mechanism-based network (LiANet) is proposed for VHR remote sensing images semantic segmentation. In order to guarantee the segmentation performance and reduce the number of parameters, we propose to embed the spatial information, i.e., the position information, into the channel attention to simplify the conventional attention and adopt covariance matrix for modelling local and global dependency to strengthen the representation of attention maps. The experiments conducted on two ISPRS datasets verified the effectiveness of the proposed method. The main contributions of this work are as follows.

1) We design a multi-stage feature fusion lightweight (MSFFL) model, which can effectively fuse the high-level and low-level feature maps and greatly reduce the calculation parameters.
2) We present two parallel enhanced attention modules, i.e., a spatial attention module (SAM) and a channel attention module (CAM) by introducing the encoding position information to guarantee the performance of the proposed lightweight attention network.
3) We adopt covariance matrix to partially recalibrate the generated attention map and mitigate the degradation of segmentation performance in the lightweight networks. It allows the LiANet to achieve competitive performance with fewer parameters than the compared methods.

The remainder of this paper is organized as follows. Section II introduces the related work, including attention mechanism, position information, covariance and lightweight network with semantic segmentation. Section III presents the details of our proposed model. First, this section illustrates the multi-stage features fusion lightweight module. Second, it presents the spatial attention module and the channel attention module. Then, the proposed lightweight network embedded with the two attention modules is introduced. Experiments of results on two VHR data sets and discussion are provided in Section IV. Finally, Section V draws the conclusion of the paper.

II. Related Work

A. Attention Mechanism

The attention mechanism has been proven powerful in the field of image semantic segmentation. Generally, two attention structures, i.e., spatial and channel attention modules can be constructed in a CNN network. The spatial attention module generates a spatial attention map by utilizing the inter-spatial relationship of features. The channel attention module produces a channel attention map by using the inter-channels relationship of features and assigning different weight coefficients to each channel to measure its importance.

According to the calculation method of attention map, attention mechanism can be divided into soft attention and self-attention. A typical soft-attention approach is the Squeeze-and-Excitation (SE) Network [39], which assigns

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

different weight coefficients to each channel. CBAM [21] is another commonly used soft-attention method that extends SE by adding the attention to the spatial dimensions. Moreover, Zhang et al. [40] proposed a channel-wise attention, which considers contextual semantics and category information to improve the accuracy of segmentation.

However, soft attention cannot express the interdependence among channels or spatial features. Spatial relevance is very important for semantic segmentation tasks. Therefore, the self-attention mechanism that can model the relationship across features has received a large attention from researchers. Moreover, various Non-Local based models have been proposed to improve the performance of the semantic segmentation. An example is the point-wise spatial attention network (PSANet) [41], in which the prediction of one position can be aided by information from other positions. Yuan et al. [42] proposed a pixelated object context to approximate the object by learning the pixel-wise similarity map. Li et al. [43] introduced self-attention into UNet [14] and achieved good performance in fine-resolution remote sensing image semantic segmentation tasks.

Although Non-Local (i.e., self-attention) based models are widely used in semantic segmentation tasks and have achieved good results, the above methods have high computational complexity and usually introduce a large number of parameters.

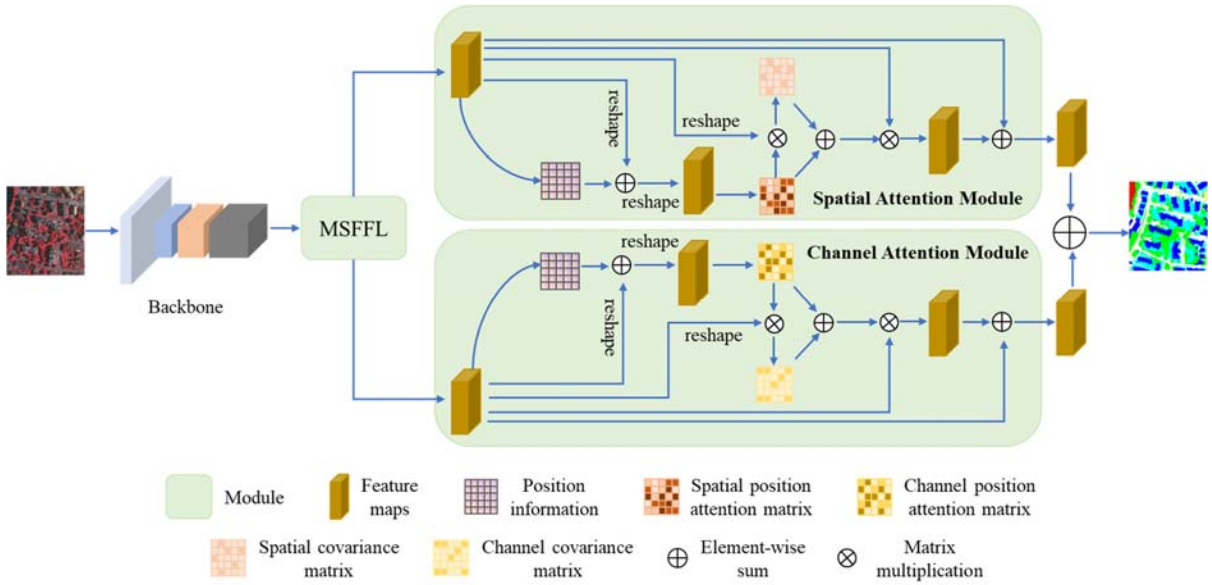

Fig. 1. Overview of the proposed network. The MSFFL module is the multi-stage feature fusion lightweight module.

B. *Position Information Used in Semantic Segmentation*

Location as position information is vital for VHR semantic segmentation. The conventional convolution operations can obtain local position information. However, a limited receptive field makes CNN fails to model long-range dependencies and makes it difficult to obtain the complete location of the global image. In [44], Hou et al. encoded the channel attention along the two dimensions of the space to obtain accurate position information when capturing the long-range dependencies. Wu et al. [45] combined position embedding information with a residual network and a bidirectional long short-term memory network to achieve unconstrained off-line handwritten word recognition.

For VHR remote sensing images, the coordinate of each pixel records its spatial position in the scene. However, when a CNN gets deeper, the semantic information of features extracted by it becomes stronger, while the position information results weaker. To address such a limitation, in this paper we design a powerful spatial and channel attention module for mitigating the loss of position information.

C. *Covariance Combined with Semantic Segmentation*

The covariance matrix reflects the correlations between the multi-dimensional data. It is an important tool in pattern recognition, computer vision, and signal processing [46]. It has been used in image semantic segmentation to construct feature context that dependencies with effects similar to the attention mechanism models. Therefore, it can be used as a way to enhance the expression ability of the attention mechanism and to strengthen correlation, thus weaken irrelevance and correct the expression of relevance to some extent in order to obtain richer semantic information.

In earlier research, Yang et al. [47] proposed the two-dimensional principal component analysis (2DPCA) for image representation, which preserves the spatial information of the image when calculating the covariance matrix. Recently, Liu et al. [48] introduced the covariance matrix into the attention mechanism to strengthen the details in the image, and significantly improved the performance of semantic segmentation.

However, for VHR remote sensing datasets, a troublesome problem is that the feature representations of objects with the same category are quite different in complex scenes. This tends

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

to extract the wrong similarity relationship between pixels for the pixel-wise attention. To solve this problem, we introduce the covariance matrix to enhance and correct the expression of the attention map.

D. Lightweight Network with Semantic Segmentation

Lightweight networks are particularly important for model deployment. As mentioned in the introduction, there are mainly two types lightweight methods. The first type adopts existing embedded mobile lightweight models as feature extraction network, such as MobileNet [36] and ShuffleNet [37]. The other type is based on a self-designed network as the backbone to capture low-level and high-level feature maps. At present, the application of the self-designed lightweight networks is still at the stage of rapid development. Cai et al. [29] adopted the depth-wise separable convolution to design the attention enhancement module and reduce the number of parameters. In [31], the encoder-decoder style convolutional neural network and the asymmetric depth-wise separable convolution units are designed to fully extract different level feature maps and reduce the number of parameters. Differently from [29], [31], the dilated convolution strategy is also adopted in autonomously designed network to further achieve lightweight models [28], [30]. In addition, in [38] the criss-cross attention captures the contextual information of each pixel on its criss-cross path to achieve high computational efficiency and less GPU memory requirements.

III. Proposed Approach

In this section, we present the proposed LiANet framework for VHR scene semantic segmentation. The overall framework of the proposed LiANet is shown in Fig. 1. Our proposed approach consists of three modules aimed at: 1) fusing multi-stage features with fewer parameters by MSFFL; 2) capturing global context information from the perspective of SAM and CAM; 3) using covariance matrix to further improve segmentation performance in the lightweight networks.

A. Multi-stage Features Fusion Lightweight Module

In classical convolution operation, only local features can be extracted for the limited receptive field size. To obtain global context, a common strategy is to stack convolutional layers continuously to extract global high-level semantic features. This is widely used in pre-training backbone and the related output is exploited as the input to downstream tasks to improve the performance of the model. Most of existing methods simply fuse and concatenate different feature maps, which leads to a large increase in the number of parameters model. In general, the low-level feature map obtained by the shallow layer contains more texture and structure information, while the deep layer extracts a high-level feature map with rich global and abstract semantic information. Both texture and deep semantic information are important for accurate segmentation of objects. Accordingly, we combine the high-level feature map with the low-level feature map to get features with better discriminative ability. More specifically, the high-level feature maps are up-sampled to the same size of the low-level maps. This allows

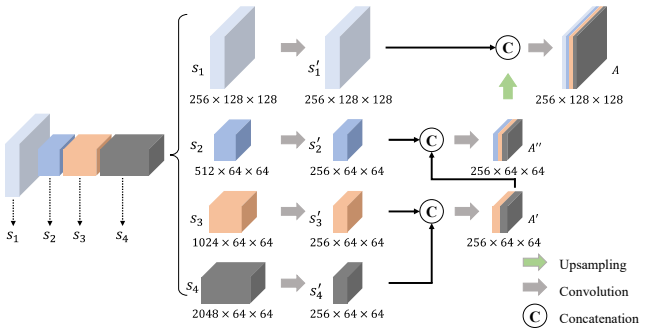

Fig. 2. Architecture of the MSFFL module.

low-level feature maps to indirectly obtain global context from high-level feature maps, by reducing significantly the number of parameters. Thus, a multi-stage feature fusion lightweight (MSFFL) module is designed to maintain segmentation performance and compress the number of model parameters. The architecture of MSFFL module is shown in Fig. 2.

Let us assume that the different colored rectangles (i.e., s_1, s_2, s_3 and s_4) in Fig. 2 represent multi-stage feature maps from the ResNet-50 [49] backbone. Note that we eliminate the downsampling operation in the last two stages of ResNet and adopt the atrous convolutions, resulting in enlarging the size of the final feature map to 1/8 of the input image, and this allows the feature map to retain more details. Inspired by UNet [14], it consists of a channel reduction path, which includes a 1×1 convolution to alleviate the number of channels, and a feature fusion path which is used to combine the different level feature maps. For the MSFFL module, we first apply the 1×1 convolution to reduce the number of channels of the two high-level feature maps (i.e., s_3 and s_4) to obtain s_3' and s_4', and then they are concatenated together, followed by a 1×1 convolution operation to fuse the concatenated feature maps. Next, the feature map A' obtained in the previous step is concatenated with the feature map s_2' after reducing the number of channels, and the feature map A'' is subsequently derived by fusion with a convolution operation. Finally, A'' is upsampled and concatenated with the feature map s_1', and the final feature map A with reduced number of channels and rich semantic information is obtained after a 1×1 convolution. The proposed MSFFL makes full use of the guiding effects of high-level feature maps and enhances the diversity of features. In this module, the number of parameters can be reduced effectively without affecting the VHR segmentation accuracy.

B. Spatial Attention Module

In image semantic segmentation, the acquired long-range semantic information becomes more abundant with the increase in the number of convolutional layers, but the spatial information between pixels becomes blurred. Spatial information is vital in semantic segmentation as it represents the relative position between pixels. Therefore, we design an attention mechanism module that models spatial information. It incorporates a position index into feature maps. In addition, we use the covariance matrix to correct the attention for making the generated attention maps more accurate in reflecting the correlation between pixels.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

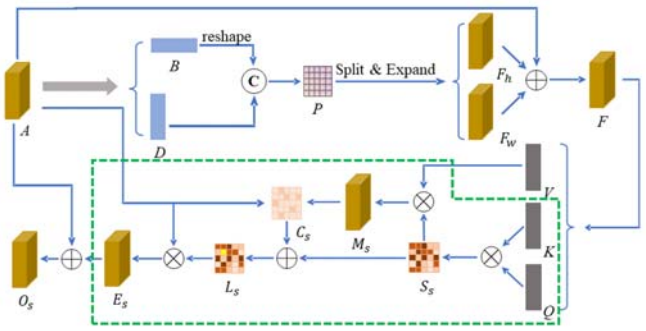

Fig. 3. Structure of the spatial attention module (SAM).

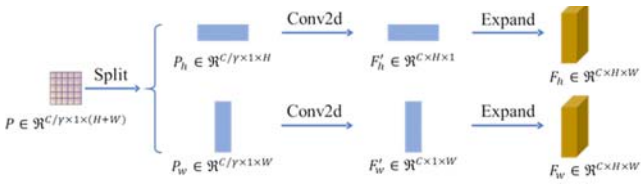

Fig. 4. Details of the Split & Expand operator.

Inspired by [44], we add the position information of pixels to the conventional attention to establish a rich context relationship model in feature maps, which enhances the ability of attention representation. Meantime, the spatial attention and the channel attention are independent each other, which keeps the spatial dimension in the spatial attention module unchanged while compresses the number of channels. Therefore, the number of parameters of the model can be further reduced.

Fig. 3 illustrates the spatial attention module. Given the output of the multi-stage feature fusion lightweight module $A \in \Re^{C \times H \times W}$, we first use the pooling operation to obtain the position information B and D from the H and W dimensions, where $B \in \Re^{C \times H \times 1}$ and $D \in \Re^{C \times 1 \times W}$. Next we reshape B in $\Re^{C \times 1 \times H}$ and concatenate it with D, and use a convolutional layer for the fusion. We also exploit a ReLU function and batch normalization to obtain the feature map $P \in \Re^{C/\gamma \times 1 \times (H+W)}$ with the position information on both H and W dimensions, where γ is the compression ratio for reducing the module size. Then the Split & Expand operator (see in Fig.4) is adopted to obtain the feature maps F_h and F_w in the horizontal and vertical directions, respectively. From the Fig.4 one can see that $P \in \Re^{C/\gamma \times 1 \times (H+W)}$ is first split into two parts, i.e., $P_h \in \Re^{C/\gamma \times 1 \times H}$ and $P_w \in \Re^{C/\gamma \times 1 \times W}$. After that, we use the 1×1 convolution operation to convert the number of channels of P_h and P_w to be the same as that of A. Then, the outputs P'_h and P'_w from the last step are extended to the size of A according the broadcast mechanism. Thus, we can obtain the comprehensive feature map $F \in \Re^{C \times H \times W}$ by the element-wise add operation among F_h, F_w and A (see in Fig.3). In this way, the position information of each feature can be recorded. Then we feed F into the 1×1 convolution layer to generate three new feature maps Q, K and V with reduced number of channels, where $\{Q, K, V\} \in \Re^{C/\gamma \times H \times W}$. Q and K are used to calculate the spatial attention map S^s. Specifically, Q and K are first reshaped to $\Re^{C/\gamma \times N}$, where $N = H \times W$ is the number of pixels in the feature map, and then the spatial attention map $S^s \in \Re^{N \times N}$ can be obtained via a matrix multiplication between Q and the transpose of K followed by a softmax layer:

$$s_{ij}^s = \frac{\exp(Q_i \cdot K_j^T)}{\sum_{j=1}^{N} \exp(Q_i \cdot K_j)} \tag{1}$$

where s_{ij}^s measures the correlation between i and j.

Meanwhile, we perform covariance operation to effectively correct the expression ability of spatial attention map S^s. To be specific, we perform a matrix multiplication between V (which is first reshaped to $\Re^{C/\gamma \times N}$) and the transpose of S^s to get $M_s \in \Re^{C/\gamma \times N}$. Then we calculate the covariance $C^s \in \Re^{N \times N}$ between transpose of M_s and A as follows:

$$C^s = cov(A, M_s) = \frac{1}{N}\left[(A - \mu_A) \cdot \left(M_s - \mu_{M_s}\right)\right] \tag{2}$$

where μ_A and μ_{M_s} represent the mean value vectors at dimension $N = H \times W$.

Let $L^s \in \Re^{N \times N}$ be the modified attention map obtained by performing element-wise sum operation between S^s and the covariance C^s. We reshape A to $\Re^{C \times N}$ and apply a matrix multiplication between A and L^s and reshape the result to $E^s \in \Re^{C \times H \times W}$. Finally, to prevent the vanishing of the gradient and make better use of the original input A, we multiply E^s by a scale parameter α and perform an element-wise adding operation with A to obtain the final output $O^s \in \Re^{C \times H \times W}$ as follows:

$$O^s = \alpha \cdot E^s + A \tag{3}$$

where α is initialized to 0 and gradually learns to assign weight.

Through above enhancement, the final feature map O^s will contains the position information of the global context based on the attention map recalibrated by the covariance result.

C. Channel Attention Module

Each channel map can be regarded as a different response to the input. Modelling the interdependency between channels can help to emphasize the more important semantic features. Accordingly, we design an enhanced channel attention module to explicitly model dependencies between channels.

The SENet [37] proposed in the literature firstly compresses the feature map of each channel to a value, then calculates the weight vector of all channels, and finally applies the weight value to the feature map in each channel. In this way, the feature map is completely compressed to the size of 1×1. Thus, the spatial size of the feature map in each channel and the influence of the position information on the channel weight are not considered. Moreover, it complete use of the spatial size of the entire feature map to model the channel context information leads to a huge number of parameters.

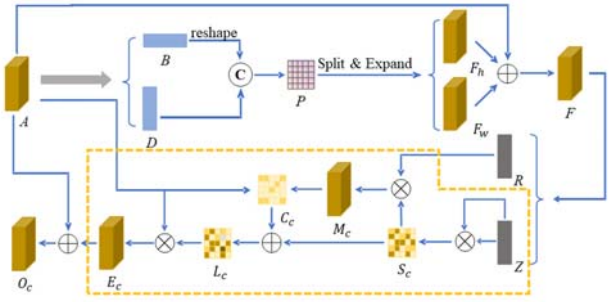

Fig. 5. Structure of the channel position attention module (CAM).

In this work, we design an enhanced channel attention module (CAM) (see Fig.5) for considering the influence of spatial size and position information on channel attention without using a large amount of calculation and parameters. We use the same method of SAM to generate the position information of the *H* and *W* dimensions, and superimpose them with the input $A \in \Re^{C\times H\times W}$ to get $F \in \Re^{C\times H\times W}$. With this operation, we focus on the construction of attention between channels, and further lightweight model in *H* and *W* dimensions. Then, we scale the size in dimensions of *H* and *W* to half of the original with the pooling method resulting in $F' \in \Re^{C\times H/2\times W/2}$. Next we reshape F' to generate $R \in \Re^{C\times T}$ and $Z \in \Re^{C\times T}$, where $T = H/2 \times W/2$. After that, we generate the channel attention map $S^c \in \Re^{C\times C}$ by matrix multiplication between Z and the transpose of Z using the follow formula:

$$s_{ij}^{c} = \frac{exp\left(Z_i \cdot Z_j^T\right)}{\sum_{j=1}^{C} exp\left(Z_i \cdot Z_j^T\right)} \tag{4}$$

where the s_{ij}^{c} measures the impact of channel i on channel j. In addition, we perform a matrix multiplication between S^c and *R*, and reshape the result to $M_c \in \Re^{C\times H/2\times W/2}$. Next, we use the bilinear interpolation to recover the size of M_c to $\Re^{C\times H\times W}$. To correct the result of the attention map S^c, we first reshape both M_c and A to $\Re^{C\times N}$, where $N = H \times W$. Then we calculate the covariance matrix $C^c \in \Re^{N\times N}$ using (2) between M_c and transpose of *A*. After that, an element-wise adding operation is applied to $C^c$and S^c to generate the final channel attention map $L^c \in \Re^{C\times C}$. Finally, we perform matrix multiplication between L^c and A to get the feature map $E^c \in \Re^{C\times H\times W}$. At last, a residual style design is used between E^c and A to obtain the final output $O^c \in \Re^{C\times H\times W}$. In this process, β is also used to measure the importance of the attention map, according the following equation:

$$O^c = \beta \cdot E^c + A \tag{5}$$

where β is initialized to 0 in this module.

CAM models the long-range semantic dependencies between channels and boosts the feature expression in channel dimension.

D. Lightweight Attention Mechanism-based Network (LiANet)

Combining the designed multi-stage feature fusion lightweight (MSFFL) module, the spatial attention module (SAM) and the channel attention module (CAM) with position information and covariance recalibration, we obtain the lightweight attention mechanism-based network (LiANet). First, the MSFFL module is designed to compress the model and effectively reduce the parameters without reducing the performance of the model. Then, in order to fully model the long-range dependence information, we use the attention mechanism in the spatial and channel directions to construct the context relationship. The position information is introduced in our modules to improve the generality of attention. At the same time, in order to partially modify the effect of the attention map, a covariance matrix is produced to enrich the content of attention. Finally, we perform an element-wise sum operation between spatial attention and channel attention to generate the final prediction map.

IV. Experimental Results

To assess the effectiveness of the proposed model, extensive experiments have been conducted on two different public data sets, i.e., the Vaihingen and Potsdam that are 2D semantic labeling challenging benchmarks provided by ISPRS. First, the description of the two data sets, the implementation details and the evaluation metrics are provided. Then the results for each data sets are reported and analyzed.

A. Description of Data Sets

1) Vaihingen Data Set: This data set is a subset of the data used for the test of digital aerial cameras carried out by the German Association of Photogrammetry and Remote Sensing (DGPF). It was captured over Vaihingen in Germany. The spatial resolution of the images is varied from 8 m to about half a meter. 33 VHR patches are considered, 16 images are used for training, and the others for testing. Each image contains red, green and near-infrared spectral bands, as well as a digital surface model (DSM). In our experiment, DSM information was not used. Six land-cover categories are included in this dataset, i.e., impervious surfaces, building, low vegetation, tree, car and clutter/background.

2) Potsdam Data Set: The second data set chosen for evaluation is the public Potsdam data set. This dataset consists of 38 patches. All patches contain red, green, blue and near-infrared spectral bands and a DSM channel. In this benchmark, 24 images are used for training and the others for testing. This dataset contains six land-cover classes, i.e., impervious surfaces, building, low vegetation, tree, car, and clutter/background.

B. Experimental Setup

1) Compared methods and implementation details

To assess the effectiveness of the proposed LiANet, two widely used methods (i.e., DeepLabV3+ [10] and PSPNet [13]) and three approaches integrating attention mechanism (i.e., DANet [24], PSANet [41], and MAResUNet [43]) are chosen as benchmarks for comparisons. Meanwhile, four lightweight models based on the attention mechanism (i.e., CoordAttention [44] ABCNet [26], MANet [34], A2FPN [35]) are also adopted. The comparative algorithms are briefly described below.

1) DeepLabV3+ [10]: This is a typical semantic segmentation method based on ASPP enlarging receptive field size.
2) PSPNet [13]: This is a classical approach with pyramid pooling module that merges multi-scale contextual information used for semantic segmentation.
3) DANet [24]: This is an attention-based method with parallel self-attention module to capture global context information.
4) PSANet [41]: Point-wise spatial attention module collects information on other positions for accurate semantic segmentation.
5) MAResUNet [43]: ResNet and multistage attention are incorporated into the UNet for VHR remote sensing image segmentation.
6) CoordAttention [44]: Coordinate attention, embedding

positional information into channel attention, is designed to enhance the representation of target objects. It should be noted that CoordAttention is a lightweight semantic segmentation method.

7) ABCNet [26]: This is a lightweight semantic segmentation model with a spatial path to retain the abundant spatial detail and a contextual path to capture the global contextual information.
8) MANet [34]: Kernel attention with linear complexity is adopted to alleviate the need for a large amount of computation in attention.
9) A2FPN [35]: This conbines Feature Pyramid Network (FPN) and Attention Aggregation Module (AAM) enhanced multiscale feature learning through attention-guided feature aggregation.

For all the experiments, the ResNet-50 [49] pretrained on the ImageNet [50] is used as the backbone, a poly learning rate policy is adopted, in which the initial learning rate is set to 0.004, and the power is set to 0.9. The stochastic gradient descent (SGD) is used for training the model, with momentum and weight decay set to 0.9 and 0.0001, respectively. For data augmentation, the original large images are random cropped to 512×512 to create the training sets; then rotation is adopted to increase the diversity of training data. In the experiment, 4× NVIDIA Tesla V100 GPUs are used to perform all the experiments. The batch size is set to 8.

2) Evaluation metrics

Following the evaluation method provided by the data publisher [51], two metrics are used to validate the performance of proposed model, i.e. the overall accuracy (OA) and the F1 score. According to the ISPRS evaluation rules, the clutter/background category is not involved in the model performance comparison. In addition, we also provide the per-class F1 score. The OA can be defined as follows:

$$OA = \frac{TP}{TP + FP + FN + TN} \tag{6}$$

where TP, FP, FN and TN are the representations of true positives, false positives, false negatives, and true negatives.

The F1 score is defined as:

$$F1 = 2\frac{P \cdot R}{P + R} \tag{7}$$

where the precision (P) and the recall (R) can be calculated as

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{8}$$

To further evaluate the performance of the proposed model, we also compare the number of parameters, computation complexity, and the computational time taken by each method on the test set.

C. Results on the Vaihingen Data Set

The first set of experiments is conducted on the Vaihingen data set. Table I reports the quantitative results in term of OA, F1-Score, number of parameters (unit is M = 10^6) and test speed (time taken to process an image) of all models. One can see that the number of parameters used in the proposed LiANet (24.59M) is only slightly hihger than that of A2FPN (24.20M) and smaller than other comparison methods. LiANet achieves segmentation performance similar to DeepLabV3+, but using only half of the parameters. Compared with PSPNet, LiANet achieves a superior performance, with a 2.64% and 4.15% improvement in OA and F1 respectively, and with slightly more than half the number of parameters. With respect to the attention-based methods, LiANet achieves better performance than the DANet and PSANet with about half of the number of parameters. MAResUNet obtains better accuracy but with a number of parameters that is more than four times higher than that of the proposed LiANet. For the lightweight comparison models, the CoordAttention, ABCNet and MANet achieve satisfactory segmentation accuracy with 26.93M, 30.80M and 35.90M parameters, respectively. However, LiANet achieves better segmentation performance with only 24.59M parameters. For this dataset, the A2FPN combining FPN and AAM achieves slightly better performance than LiANet with 24.2M parameters, with 0.42% and 0.51% improvement in OA and F1, respectively. The results reported in Table I demonstrate that the proposed LiANet can achieve a good segmentation performance with a reduced number of parameters.

Fig. 6 shows the semantic segmentation results of different methods in two different scenarios (left: a large building with large internal variations; right: small cars difficult to be segmented) in the test set of Vaihingen dataset. As one can see, for the contextual information aggregation-based methods the PSPNet and DeepLabV3+ can neither segment large building objects well nor classify small car objects correctly.

In contrast, for the attention mechanism-based models (i.e., DANet, PSANet and MAResUNet), the MAResUNet provides good segmentation result on large building targets, but it shows under segmentation in small targets. On both classes the DANet achieves the best segmentation results after the MAResUNet. In comparison, the PSANet achieves poor results in the segmentation of large and small objects. The lightweight CoordAttention method also fails to accurately segment large and small targets at the same time. One can see that it mis-segmented too much small car targets. The ABCNet, MANet and A2FPN lightweight attention-based models also failed to accurately segment large building objects. Finally, ABCNet has the worst segmentation accuracy for Car among all comparison methods.

The proposed LiANet performs slightly worse than the MAResUNet and the DANet in the segmention of both large building with large internal variations and small cars. Nonetheless, the most segmentation results are better than those of the contextual information aggregation-based methods. This is due to the rich contextual semantic information that can be captured by the LiANet model, which integrates the MSFFL module and attention modules combined with the spatial location and the covariance.

For the number of parameters in the models, Table I reports the computational complexity (FLOPs, i.e., the number of floating-point operations of the models, unit is G = 10^9) and the speed on the test set for each method. Taking into account three metrics, i.e., the number of parameters, the computational complexity and the test speed, one can see from Table I that the typical non-lightweight models have a greater number of

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

TABLE I
QUANTITATIVE RESULTS IN TERMS OF PER-CLASS PIXEL ACCURACY, OA, MEAN F1 SCORE, NUMBER OF PARAMETERS, COMPUTATION COMPLEXITY AND TEST SPEED (VAIHINGEN DATA SET)

Method	Per-class F1 Score (%)					OA (%)	Average F1 Score (%)	Parameters (M)	FLOPs (G)	Test Speed (ms/image)
	Im. Surf.	Build.	L. Veg.	Tree	Car					
DeepLabV3+ [10]	94.63	95.71	85.65	90.29	80.70	91.93	89.39	53.35	248.12	93.17
PSPNet [13]	**98.21**	91.74	79.27	86.47	65.67	89.48	84.27	44.43	184.52	**74.42**
DANet [24]	95.67	96.07	85.21	89.03	77.73	91.89	88.74	45.36	205.18	83.14
PSANet [41]	95.38	95.30	85.39	88.82	78.49	91.60	88.68	53.04	238.20	88.22
MAResUNet [43]	95.62	**96.46**	**87.59**	**91.59**	**86.07**	**93.15**	**91.47**	97.96	85.22	82.29
CoordAttention [44]	94.71	95.07	85.81	90.32	70.56	91.70	87.29	26.93	124.42	77.46
ABCNet [26]	93.04	93.81	84.90	89.84	47.96	90.38	81.91	30.80	28.98	301.56
MANet [34]	94.12	94.16	86.74	91.34	78.03	91.80	88.88	35.90	33.63	569.35
A2FPN [35]	95.44	95.80	86.96	91.02	75.45	92.54	88.93	**24.20**	**25.63**	307.44
LiANet (Ours)	95.63	96.00	85.67	89.79	75.02	92.12	88.42	24.59	130.35	102.69

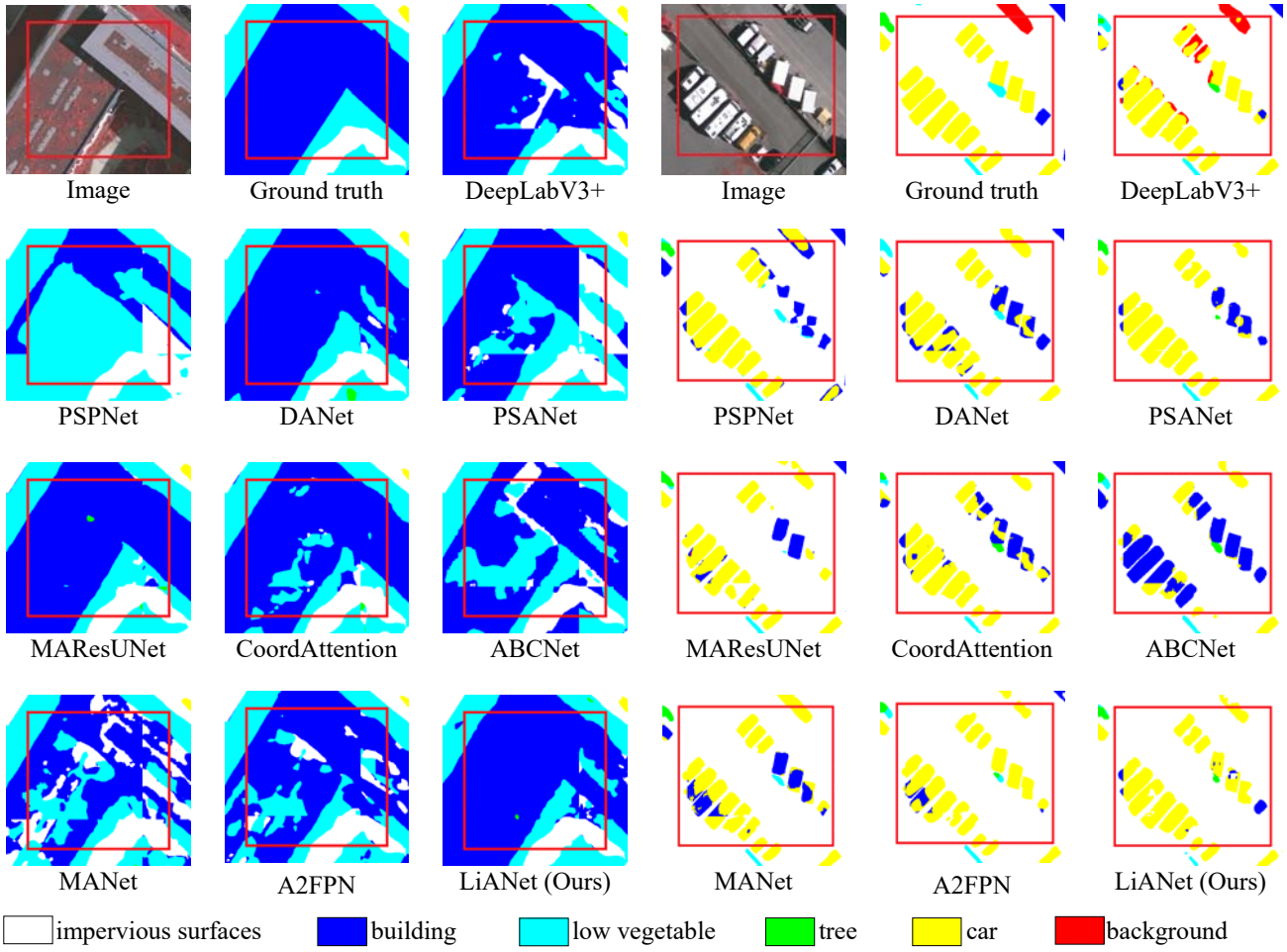

Fig. 6. Examples of qualitative semantic segmentation results on Vaihingen data set.

parameters and a higher computational complexity. For the compared lightweight models, ABCNet, MANet and A2FPN have a fewer number of parameters and least computational complexity. In contrast, the proposed model achieves a competitive computational complexity and test speed with a small number of parameters.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

TABLE II
QUANTITATIVE RESULTS IN TERMS OF PER-CLASS PIXEL ACCURACY, OA, MEAN F1 SCORE, NUMBER OF PARAMETERS, COMPUTATION COMPLEXITY AND TEST SPEED (POTSDAM DATA SET)

Method	Per-class F1 Score (%)					OA (%)	Average F1 Score (%)	Parameters (M)	FLOPs (G)	Test Speed (ms/image)
	Im. Surf.	Build.	L. Veg.	Tree	Car					
DeepLabV3+ [10]	94.53	94.02	81.16	75.38	**92.64**	91.09	87.55	53.35	248.12	88.14
PSPNet [13]	93.13	94.95	64.63	**78.48**	84.62	87.62	83.22	44.43	184.52	68.57
DANet [24]	95.50	95.57	**81.78**	77.74	91.62	92.23	88.44	45.36	205.18	74.66
PSANet [41]	94.94	95.27	79.84	75.97	91.86	91.42	87.58	53.04	238.20	75.27
MAResUNet [43]	**95.97**	**96.66**	81.42	77.59	92.58	**92.66**	**88.84**	97.96	85.22	70.68
CoordAttention [44]	95.32	95.44	81.06	77.18	90.23	91.91	87.85	26.93	124.42	**67.37**
ABCNet [26]	94.56	93.35	80.27	76.16	87.57	90.84	86.38	30.80	28.98	134.31
MANet [34]	94.52	92.81	80.60	77.85	90.96	90.87	87.35	35.90	33.63	264.74
A2FPN [35]	94.90	92.88	80.76	77.40	90.04	91.08	87.20	**24.20**	**25.63**	135.90
LiANet (Ours)	95.33	96.10	79.79	77.14	91.37	91.99	87.95	24.59	130.35	89.18

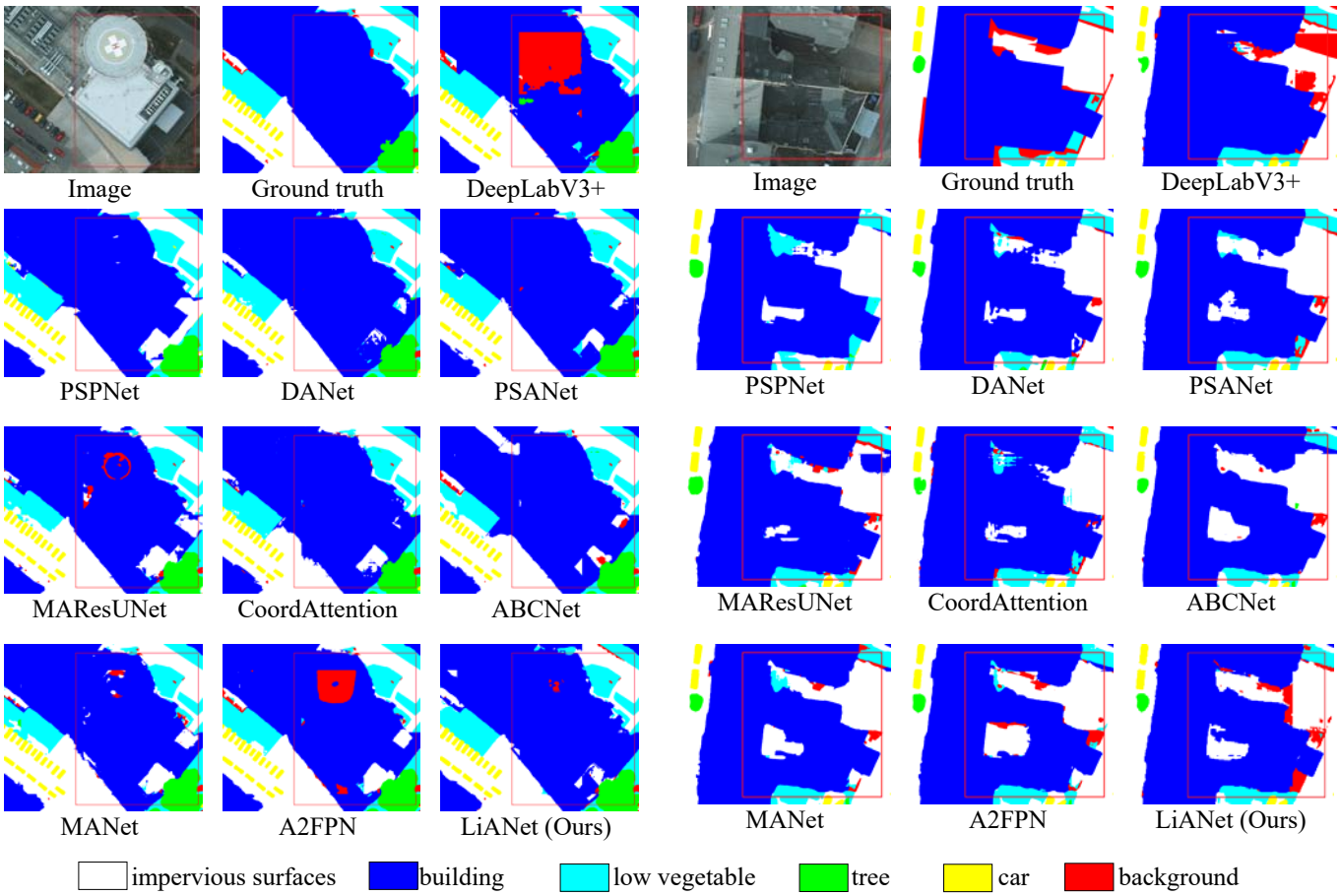

Fig. 7. Examples of qualitative semantic segmentation results on Potsdam data set.

D. *Results on the Potsdam Data Set*

Table II reports the accuracy metrics and the number of parameters used in different methods.

From the table, on can see that the proposed LiANet obtains 0.9% and 4.37% improvement of OA with respect to DeepLabV3+ and PSPNet respectively. Moreover, the number of parameters of LiANet is only the 46.09% and 55.35% of the number of parameters of DeepLabV3+ and PSPNet respectively. For the attention-based methods, the MAResUNet and the DANet both achieve good segmentation performance, which are slightly higher than that of the proposed LiANet in term of OA (0.67% and 0.24%) and F1 (0.89% and 0.49%).

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

However, the number of parameters of MAResUNet (97.96M) and DANet (45.36M) are 3.98 and 1.84 times higher than that of the LiANet (24.59M). The PSANet obtains the lowest accuracy with more parameters. Comparing the lightweight models, CoordAttention, ABCNet and MANet all achieve worse segmentation performance than the proposed LiANet, and their numbers of parameters (2.34M, 6.21M and 11.31M, respectively) are larger than that of LiANet. Despite the fact that the number of parameters of A2FPN is slightly less than that of LiANet, its OA and F1 are 0.91% and 0.75% lower than those of LiANet, respectively. From the results reported in Table II, the proposed LiANet achieves competitive performance with the smallest number of parameters. This is consistent with the original motivation of reducing the number of parameters in the model as much as possible while maintaining high performance.

The computation complexity of each model and the time required to process an image are reported in Table II. It should be noted that since both datasets take image patch with size of 512×512 as input, the computation complexity of each model is the same on both datasets. From Table II, one can draw a conclusion similar to that of the Vaihingen dataset. Although the computational complexity of the proposed model is higher than those of other lightweight models (such as ABCNet, MANet, and A2FPN), the test speed of the proposed model is fastest. Experiments on both datasets indicate that our proposed model is able to achieve a better trade-off in terms of the number of parameters, computational complexity and test speed.

Fig. 7 shows some examples of image from the test set of the Potsdam dataset. The qualitative analysis points out that the LiANet is only inferior to MAResUNet and better to other comparison methods. In particular, it can reduce the misclassification of land-cover types, such as building, compared with other methods.

E. Ablation Study

In this part, we conducted comprehensive experiments to verify the effect of the MSFFL module, the position information and the covariance added in attention. The experiments have been conducted on the two considered data sets. Considering that our proposed LiANet model is inspired by DANet, the DANet is selected as the baseline network for comparison. OA and F1 score are used to evaluate the performance of different modules, also with the number of parameters to demonstrate the degree of lightweight. The experimental results on the Vaihingen and the Potsdam data sets are shown in Table III and Table IV respectively, in which M represents the MSFFL module, P means adding position information to the attention module, and C is the covariance.

For the Vaihingen data set, we can see that the baseline model (i.e., DANet) achieves 91.89% and 88.74% in OA and F1 score respectively, which are competitive compared with others from Table III. However, the model with the lightweight module (i.e., MSFFL) significantly reduces the number of parameters of the baseline model from 45.36M to 24.48M. In addition, the OA increases by 0.05% and the F1 score reduces by 0.05%. In terms of position information, the incorporation with MSFFL in the baseline increases the number of parameters of 0.04M, but achieves better OA and F1 score. This confirms the importance of integration of the position information. Finally, when we add the covariance to adjust the attention map, the OA is increased of 0.14% with only an increase of 0.07M parameters and decrease of F1 score of 0.33%. Therefore, in general, our improvements to the model are valid and can reduce the number of parameters while guaranteeing no degradation in model performance.

For the Potsdam data set (Table IV), the baseline DANet achieves the best performance in terms of OA and F1 score. We point out that adding the MSFFL module into the baseline did not increase the OA and F1 score, but the number of parameters of the model was significantly reduced. This is mainly due to the feature representation capability reduced in lightweight models when using a small number of parameters, which results in degraded segmentation performance on the suburban scene datasets with complex ground objects. When the position information is introduced, the OA slightly increases whereas the F1 score decreases. The use of covariance enhances both the OA and F1 of the model to a level almost comparable to the performance of the baseline. However, this is achieved with a sharply smaller number of parameters that largely compensates the slight decline in model performance.

The proposed model greatly reduces the number of parameters by significantly reducing the number of channels during feature fusion, as well as by compressing the spatial size by half when calculating the channel attention that captures the contextual relationships among channels. Then, the position information and covariance matrix operations with only a small number of parameters are introduced to improve the model performance. The results in Table III and Table IV show that the proposed model is able to significantly reduce the number of parameters of the model while taking into account the segmentation performance of the model, which confirms that effectiveness of proposed modules.

TABLE III
QUANTITATIVE COMPARISONS AMONG ABLATION STUDIES ON THE VAIHINGEN DATA SET. M IS MSFFL, P IS POSITION INFORMATION AND C IS COVARIANCE

Method	OA (%)	F1 (%)	Parameters (M)
DANet (baseline)	91.89	88.74	45.36
baseline + M	91.94	88.69	24.48
baseline + M + P	91.98	88.75	24.52
baseline + M + P + C (LiANet)	92.12	88.42	24.59

TABLE IV
QUANTITATIVE COMPARISONS AMONG ABLATION STUDIES ON THE POTSDAM DATA SET. M IS MSFFL, P IS POSITION INFORMATION AND C IS COVARIANCE

Method	OA (%)	F1 (%)	Parameters (M)
DANet (baseline)	92.23	88.44	45.36
baseline + M	88.60	84.63	24.48
baseline + M + P	88.65	83.64	24.52
baseline + M + P + C (LiANet)	91.99	87.95	24.59

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3272614

F. Analysis and Discussion

In this section, the effects of the two parameters, i.e., the size of the input image and the number of output channels, on the model performance are analyzed in detail. In the following experiments, to minimize the impact of input image size on model performance, three different crop sizes (i.e., 128×128, 256×256 and 512×512), have been considered. In order to better balance the relationship between the model accuracy and the amount of model parameters, the number of output channels of the MSFFL is selected from {256, 512, 1024, 2048} to analyze the proposed model.

1) Size of input image

In the experiments, we crop the original large image to image patches given as input to the model for adapting to the GPU memory. However, different image patch sizes may impact the performance of the model. This is because image contains objects with different scales, and differences in crop size can reduce the variety of objects in each image. The OA versus the different crop sizes and the number of channels of MSFFL on the two considered data sets are shown in Fig. 8.

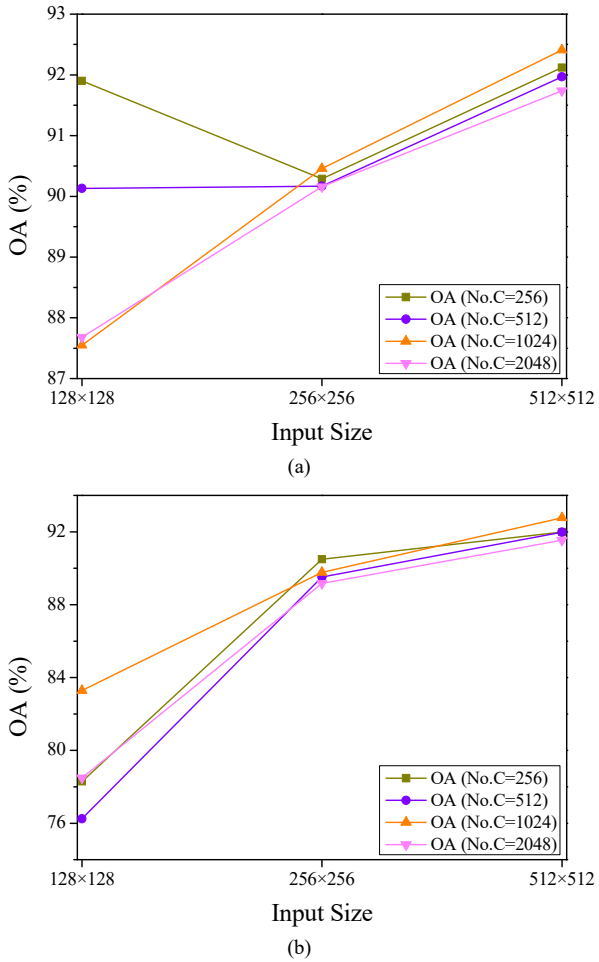

Fig. 8. Performance of the proposed LiANet versus different crop sizes: (a) Vaihingen data set, (b) Potsdam data set.

Fig. 8 (a) shows the OA versus different cropping sizes of on the Vaihingen data set. One can see that OA increases by increasing the input image size when the number of output channels is 512, 1024 and 2048. Moreover, although the model achieves good performance with the input image size of 128×128 for 256 channels, the best performance is still achieved with the input image size of 512×512. Therefore, the input size of 512×512 was chosen for the Vaihingen dataset

As shown in Fig. 8 (b), on the Potsdam dataset the OA increases by increasing the input image size for any number of output channels. In the range between 128 and 512, the model is very sensitive to the input size. When the input size is set to 128×128, the lack of context information affects the information available in the model training. When the input size is 512×512, the accuracy of the model reaches the maximum value. Thus, we used image patches of size of 512×512 for training and testing on the Potsdam dataset.

We can also observe from Fig. 8 that the performance of the proposed model shows a generally consistent trend with the input image size for different output channels on both datasets, indicating that the choice of the input image size is not affected by the number of output channels of MSFFL.

2) Number of output channels

In the multi-scale feature fusion model, the number of output channels of the feature map can be changed by fusing with multi-stage information. A different number of channels in the feature maps has an impact on the performance of the model. Meanwhile, it is worth noting that the number of parameters of the model increases exponentially by increasing the number of output channels. In order to better balance the relationship between the model accuracy and the amount of model parameters, we performed experiments by training and testing the model with different degrees of reduction on the output channels of the MSFFL. Fig. 9 shows the relationship between OA and the number of output channels on the Vaihingen and Potsdam data sets, respectively, under different input sizes.

From Fig. 9 (a), one can observe that the segmentation performance of the model slightly fluctuates by reducing the number of MSFFL output channels when the input sizes are 256×256 and 512×512. When the input size is equal to 128×128, the OA shows a tendency to increase as the number of output channels decreases. For the cases in which the input size is 256×256 and 512×512, although the model accuracy is optimal when the number of output channels is 1024, the accuracy is only slightly below the optimal accuracy when the number of output channels is 256, and the number of parameters is reduced by almost half. In addition, when the number of channels continue to increase over 1024, the OA and F1 Score do not improve or even decline. This is because the excessive number of channels causes the model to learn redundant features from the data and to loose generalization capabilities. To find a trade-off between the segmentation performance and the amount of model parameters, we selected 256 channels as the final channel size of the fused feature map in the LiANet on the Vaihingen dataset. Under this combination, our model not only has a significant improvement of the segmentation performance, but also achieves the goal of the lightweight model.

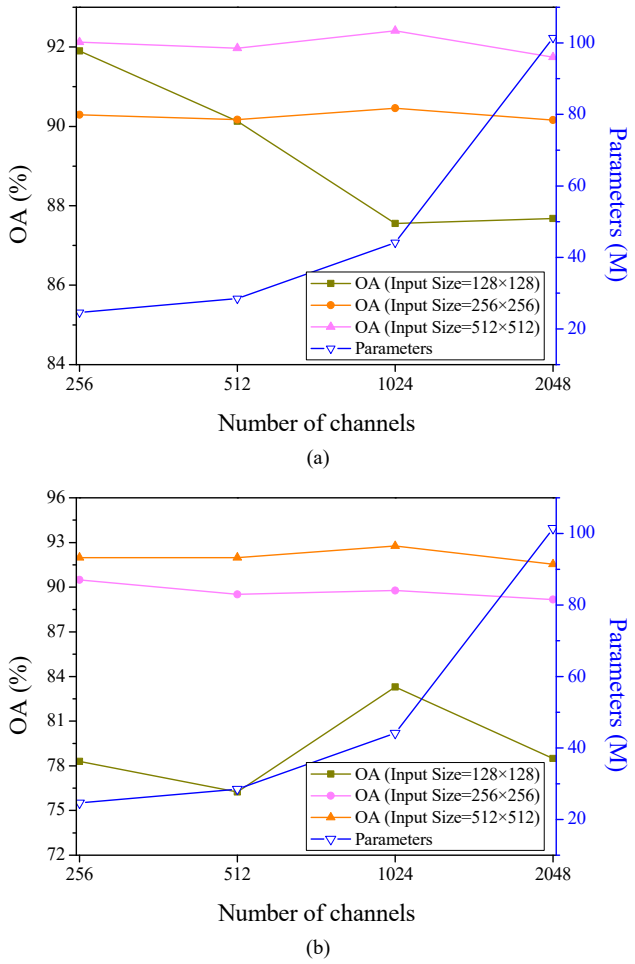

Fig. 9. Performance and number of parameters versus the level of model reduction (different output channels of MSFFL module): (a) Vaihingen data set, (b) Potsdam data set.

Similar results are obtained on the Potsdam data set [Fig. 9 (b)] when the input size is 256×256 and 512×512. The model accuracy fluctuates with the number of output channels when the input size is 128×128. The best OA is achieved with 1024 output channels. This may be due to the small input size of 128 ×128, which makes the ground objects information contained in the different cropped image patches more variable and finally causes the model performance to be unstable. As a result, considering all the cases together and to balance between the segmentation performance and the number of model parameters, we used 256 channels as the final channel size of the output feature map of MSFFL module.

As one can observed in Fig. 9, the trend of the performance of the proposed model with the number of output channels remains basically the same for different input sizes on both datasets. This again indicates that there is not mutual influence between the input image size and the number of output channels.

V. Conclusion

In this paper, a new lightweight attention mechanism-based network (LiANet) for VHR images semantic segmentation has been presented. In the proposed network, an effective MSFFL module is designed to fuse the high-level and low-level feature maps and sharply reduce the calculation parameters. Meanwhile, two parallel enhanced attention modules, i.e., a spatial attention module (SAM) and channel attention module (CAM), are designed by introducing position information to enhance the ability to express attention. To further improve the segmentation performance, covariance calculation is used to correct the expression of the attention mechanism. Experimental results on two benchmark VHR data sets show that the proposed LiANet can greatly reduce the number of model parameters while maintaining the stable performance of the model with respect to other literature methods.

In the future developments of this work, we will focus on exploring lightweight convolution operations to further reduce the number of parameters and computational complexity of the model while maintaining high segmentation performance. This can be important for extreme scenarios in terms of few available resources like on-board of satellites. Moreover, we will embed proposed lightweight approach in the segmentation models designed for large-size remote sensing images (such as MFVNet [52]) to achieve good accuracy with lower number of parameters.

Acknowledgment

The authors would like to thank the International Society for Photogrammetry and Remote Sensing (ISPRS) for providing the Vaihingen and Potsdam datasets, and the editors and reviewers for their constructive comments, which greatly improved the quality of this paper.

References

[1] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, “Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models,” *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018.

[2] Q. Zhang and K. C. Seto, “Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime light data,” *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011.

[3] A. Collin, B. Long, P. Archambault. “Merging land-marine realms: Spatial patterns of seamless coastal habitats using a multispectral LiDAR,” *Remote Sens. Environ.*, vol. 123, pp. 390-399, Aug. 2012.

[4] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, “Dynamic multicontext segmentation of remote sensing images based on convolutional networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

[5] G. Deng, Z. Wu, C. Wang, et al. “CCANet: Class-Constraint Coarse-to-Fine Attentional Deep Network for Subdecimeter Aerial Image Semantic Segmentation,” *IEEE Trans. Geosci. Remote Sens.*, 2021.

[6] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *arXiv:2001*, Sep. 2020.

[7] 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.

[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *In IEEE Trans. Pattern Anal. Mach. Intell.(TPAMI)*, vol. 40, no. 4, pp. 834-848, Apr. 2018.

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

[10] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image

segmentation," *In Proc. Euro. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801-818.

[11] B. Quan, B. Liu, D. Fu, H. Chen and X. Liu, "Improved Deeplabv3 For Better Road Segmentation In Remote Sensing Images," *Int. Conf. Comput. Engineering Artif. Intell. (ICCEAI)*, 2021, pp. 331-334, doi: 10.1109/ICCEAI52939.2021.00066.

[12] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251-1258.

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

[14] O. Ronneberger, P. Fischer, and T. Brox. "U-net: convolutional networks for biomedical image segmentation," *In Med. Image Comput. Comput-Ass. Inte. (MICCA)*, Nov. 2015, pp. 234–241.

[15] H. H. Ding, X. D. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multiscale aggregation for scene segmentation," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2393–2402.

[16] Badrinarayanan V, Kendall A, and Cipolla R, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *In IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 39, no. 12, pp. 2481-2495, Dec. 2017.

[17] L. Jiang, H. S. Zhao, S. Liu, X. Y. Shen, C. W. Fu, and J. Y. Jia, "Hierarchical point-edge interaction network for point cloud semantic segmentation," *In Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10433-10441.

[18] R. Liu, L. Mi, Z. Chen. "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, Sep. 2021.

[19] C. Peng, K. Zhang, Y. Ma, et al. "Cross Fusion Net: A Fast Semantic Segmentation Network for Small-Scale Semantic Information Capturing in Aerial Scenes," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[20] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong and C. Tao. "SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905-909, May 2021.

[21] S. Woo, J. Park, J. Y. Lee, et al. "Cbam: Convolutional block attention module," *In Proc. Euro. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3-19.

[22] Q. Zhao, J. Liu, Y. Li and H. Zhang, "Semantic Segmentation With Attention Mechanism for Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[23] X. L. Wang, R. Girshick, A. Gupta, K. He, "Non-local neural networks," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794-7803.

[24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation." *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146-3154.

[25] B. Zhang et al., "Progress and Challenges in Intelligent Remote Sensing Satellite Systems," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 1814-1822, 2022

[26] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remote Sensing Images." *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84-98, 2021.

[27] Y. Zhang, Y. Chen, Q. Ma, C. He and J. Cheng, "Dual Lightweight Network with Attention and Feature Fusion for Semantic Segmentation of High-Resolution Remote Sensing Images," *IEEE Int. Geosci. Remote Sens. Symposium IGARSS*, 2021, pp. 2755-2758.

[28] R. Gomes, P. Rozario and N. Adhikari, "Deep Learning optimization in remote sensing image segmentation using dilated convolutions and ShuffleNet," *IEEE Int. Conf. Electro Inf. Technol. (EIT)*, 2021, pp. 244-249.

[29] J. Cai *et al*., "Real-Time Semantic Segmentation of Remote Sensing Images Based on Bilateral Attention Refined Network," *IEEE Access*, vol. 9, pp. 28349-28360, 2021.

[30] G. Zhang, T. Lei, Y. Cui, et al. "A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation", *ISPRS Int. J. Geoinf*, vol. 8, no. 12, 2019, pp.582.

[31] L. Lv, Y. Guo, T. Bao, et al. "MFALNet: A Multiscale Feature Aggregation Lightweight Network for Semantic Segmentation of High-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2172-2176, Dec. 2021.

[32] X. Hu and L. Jing, "LDPNet: A Lightweight Densely Connected Pyramid Network for Real-Time Semantic Segmentation," *IEEE Access*, vol. 8, 2020, pp. 212647-212658.

[33] S. Liu, B. Li, J. Xiong, F. Wang, Z. Zhuo and X. Ren, "A Lightweight and Efficient Network for Logistics Truck Scene Semantic Segmentation," *IEEE Int. Conf. Comput. Comm. (ICCC)*, 2020, pp. 2156-2160.

[34] R. Li, S. Zheng, C. Zhang, et al. "Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[35] R. Li, L. Wang, C. Zhang, C. Duan and S. Zheng. "A2-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images." *Int. J. Remote Sens.*, vol. 43, no.3, pp. 1131-1155, 2022.

[36] M. Sandler, A. Howard, M. Zhu, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2018, pp. 4510-4520.

[37] X. Zhang, X. Zhou, M. Lin, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2018, pp. 6848-6856.

[38] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," *In Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 603-612, doi: 10.1109/ICCV.2019.00069.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[40] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, "Context encoding for semantic segmentation," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7151–7160.

[41] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," *In Proc. Euro. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 267-283.

[42] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: object context network for scene parsing," *arXiv:1809.00916*, 2018.

[43] R. Li, S. Zheng, C. Duan, J. Su and C. Zhang, "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.,* doi: 10.1109/LGRS.2021.3063381.

[44] Q. Hou, D. Zhou and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13708-13717, doi: 10.1109/CVPR46437.2021.01350.

[45] X. Wu, Q. Chen, J. You, and Y. Xiao, "Unconstrained offline handwritten word recognition by position embedding integrated ResNets model," *In IEEE Signal Process. Lett., vol.* 26, no. 4, pp. 597-601, Apr. 2019.

[46] A. K. Jain, R. P. W. Duin and Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 22, no. 1, pp. 4-37, Jan. 2000, doi: 10.1109/34.824819.

[47] Jian Yang, D. Zhang, A. F. Frangi and Jing-yu Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 26, no. 1, pp. 131-137, Jan. 2004, doi: 10.1109/TPAMI.2004.1261097.

[48] Y. Liu, Y. Chen, P. Lasang and Q. Sun, "Covariance Attention for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* doi: 10.1109/TPAMI.2020.3026069.

[49] *K. He, X. Zhang, S. Ren, J. Sun*, "Deep Residual Learning for Image Recognition," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770-778.

[50] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[51] ISPRS. *International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Challenge.* [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html.

[52] Y. Li, W. Chen, X. Huang, et al. "MFVNet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation." *Sci. China Inf. Sci.*, vol. 66, no. 4, pp. 140305-, 2023.