Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images

Ruigang Niu\textsuperscript{\textcopyright}, Graduate Student Member, IEEE, Xian Sun\textsuperscript{\textcopyright}, Senior Member, IEEE, Yu Tian\textsuperscript{\textcopyright}, Wenhui Diao\textsuperscript{\textcopyright}, Member, IEEE, Kaiqiang Chen\textsuperscript{\textcopyright}, Member, IEEE, and Kun Fu\textsuperscript{\textcopyright}, Member, IEEE

Abstract—Semantic segmentation in very-high-resolution (VHR) aerial images is one of the most challenging tasks in remote sensing image understanding. Most of the current approaches are based on deep convolutional neural networks (DCNNs). However, standard convolution with local receptive fields fails in modeling global dependencies. Prior research works have indicated that attention-based methods can capture long-range dependencies and further reconstruct the feature maps for better representation. Nevertheless, limited by the mere perspective of spatial and channel attention and huge computation complexity of self-attention (SA) mechanism, it is unlikely to model the effective semantic interdependencies between each pixel pair of remote sensing data with complex spectra. In this work, we propose a novel attention-based framework named hybrid multiple attention network (HMANet) to adaptively capture global correlations from the perspective of space, channel, and category in a more effective and efficient manner. Concretely, a class augmented attention (CAA) module embedded with a class channel attention (CCA) module can be used to compute category-based correlation and recalibrate the class-level information. In addition, we introduce a simple yet effective region shuffle attention (RSA) module to reduce feature redundant and improve the efficiency of SA mechanism via regionwise representations. Extensive experimental results on the ISPRS Vaihingen, Potsdam benchmark, and iSAID data set demonstrate the effectiveness and efficiency of our HMANet over other state-of-the-art methods.

Index Terms—Aerial imagery, deep convolution neural networks (DCNNs), self-attention (SA) mechanism, semantic segmentation.

I. INTRODUCTION

S

I 

EMANTIC segmentation, also known as semantic labeling, is one of the fundamental and challenging tasks in remote sensing image understanding. The goal is to assign pixelwise semantic class labels for a given image. In particular, semantic segmentation in very-high-resolution (VHR) aerial images plays an increasingly significant role for its widespread applications, such as road extraction [1], urban planning [2], and land cover classification [3].

In recent years, deep convolutional neural networks (DCNNs) have demonstrated the powerful capacity of feature extraction and object representations compared with traditional methods in machine learning, such as random forests (RF) [4], support vector machine (SVM) [5], and conditional random fields (CRFs) [6]. In particular, state-of-the-art methods based on the fully convolutional network (FCN) [7] have made great progress. However, due to the fixed geometry structured, they are inherently limited by local receptive fields and short-range context information. This task is still very challenging.

To capture long-range dependencies, such as correlation coefficients between long-distance pixels, Chen et al. [8] proposed atrous spatial pyramid pooling (ASPP) with multiscale dilation rates to aggregate contextual information. Zhao et al. [9] further introduced the pyramid pooling module (PPM) to represent the feature map via multiple regions with different sizes. ScasNet [10] aggregates multiscale contexts in a self-cascade manner. Nevertheless, the context aggregation methods above are still unable to extract global contextual information, that is, it is unsatisfactory to cover global receptive fields by stacking and aggregating convolutional layers.

Furthermore, in order to generate dense and pixelwise contextual information, Nonlocal neural networks [11] utilize a self-attention (SA) mechanism, which enables a single feature from any position to perceive features from all other positions. It can be seen as a matter of feature reconstruction, that is, the feature representation of each position is a weighted sum of all other counterparts. DANet [12] introduces spatialwise and channelwise attention modules to enrich the feature representations. Besides, several works [13]–[15] improve the efficiency of the SA mechanism to some extent.

Semantic segmentation is essentially a pixel-by-pixel classification task, which requires the network to have a large field of view. Attention-based methods have been proved to be...
effective ways to obtain global fields of view and contexts in semantic segmentation. However, the standard SA mechanism has many limitations in modeling effective semantic dependencies between each pixel pair of remote sensing data of complex spectra. Hence, inspired by the success of attention-based methods above, and considering its limitations, we introduce multiple attention modules into a segmentation network to enrich the perspective of attention extraction and optimize the huge computational complexity of the SA mechanism.

Concretely, pixelwise attention approaches need to generate a dense attention map to measure the relationships between each pixel pair, which has a high computation complexity and occupies a huge number of GPU memory. Recent works [14], [16] have shown the fact that information redundancy is not conducive to feature representations. What is more, attention-based methods are restricted to the perspective of space and channel, ignoring category-based information, which is a key factor for semantic segmentation task. The category-based information is directly related to the last convolution of the network. In general, lack of category-based information and huge computation complexity of SA mechanism are two tough problems and will be elaborated in the following.

On the one hand, for remote sensing images with complex spectral signatures, the class-level information is usually directly reflected between different spectra, and the input data itself has sufficient class-level information. The category-based information is embedded in different spectra, namely different channels of the input feature. However, in previous works [7]–[9], [17], [18], the category-based information in the general segmentation network is only reflected in the last convolutional layer. Only the score map directly represents the probability that each pixel belongs to each category. In other words, through the complex feature extraction and representation of the middle stage of the network, the category-based information of the input data is already ambiguous or missing. Empirically, the lack of class-level information leads to poor object classification capabilities. Hence, different from other attention-based methods, we argue that retaining the category-based information in the middle stage of the network and extracting the corresponding attention representations is also crucial for semantic segmentation. We propose a so-called category-based correlation that explicitly models class-level representation of each pixel and further calculates the relationships between each category and each channel of the input feature. As shown in Fig. 1, category-based correlation mainly focuses on exploiting contextual information from a categorical perspective, which pays more attention to the pixels of the same category during the feature reconstruction. Besides, considering the limitations (such as the strictly fixed number of channels, see Section III-B for details) of nonlocal operation in channel dimension, we introduce a class channel attention (CCA) to further adaptively recalibrate the obtained category-based correlation.

On the other hand, a tricky problem in remote sensing images is that the feature representations of objects with the same category are quite different in complex scenes. Therefore, the dense pixelwise attention tends to extract the wrong similarity relationship between pixels, leading to serious classification errors. Besides, as shown in Fig. 2(b), it has a high computation complexity and occupies a huge number of GPU memory. Several works [14], [16] have proved that the invalid redundant information is not conducive to the feature representations. For example, as for a single feature belonging to “car” in Fig. 2(a), the pixelwise attention method usually extracts features of all other positions, among which we actually do not need to focus on the “building” and “impervious surface,” and it is more likely to extract the wrong similarity because of the complex scenes (such as in the shadow or overlapping). Aiming at the problems of feature redundancy and high computational complexity of SA, we employ a sparse regionwise attention mechanism to exploit a wider range of correlations. Experiments show that the
regionwise representation can capture long-range contextual information between pixels in a more efficient and robust manner.

Toward the above two issues and our corresponding solutions, we propose a novel framework, named hybrid multiple attention network (HMANet). The HMANet mainly consists of two parallel branches, one of which is the class augmented attention (CAA) module embedded with the CCA module. Given the input feature, the CAA module first calculates category-based correlation and further generates the weighted class representation via a dense class affinity map, whereas the CCA module is added to adaptively recalibrate the class-level information through two linear scaling transformation functions, which efficiently helps to enhance the discriminative abilities for each class with a few parameters. The other branch of our network is the region shuffle attention (RSA) module, which aims to capture regionwise global information with a shuffling operator and obtain more robust correlation between objects. Notably, compared with pixelwise SA methods, the grouped regionwise representation requires 20 times less GPU memory usage and significantly reduces floating point operations (FLOPs) by about 77% with a few parameters. Finally, we concatenate the output features from each branch and the local representation and then feed them into a convolutional layer to further generate the fine segmentation map.

Our contributions can be summarized as follows.
1) We present a CAA module to exploit the category-based correlation between pixels and enhance the discriminant ability for each class. A CCA module is embedded to adaptively recalibrate the class-level information for better representations.
2) The RSA module is proposed to capture regionwise global information and obtain more robust relationships between objects in a more efficient and effective manner.
3) We propose a novel HMANet by taking advantage of the three attention modules above, which comprehensively captures feature correlations from the perspective of space, channel, and category.

The remainder of this article is arranged as follows. Related work is briefly introduced in Section II. Section III presents the details of our proposed method, including three attention modules. Experimental evaluations between our HMANet and the state-of-the-art methods, as well as ablation studies on the Vaihingen data set, are provided in Section IV. The conclusion is outlined in Section V.

II. RELATED WORKS

A full review is beyond the scope of this article. Here, we review some recent works on semantic segmentation of nature scenes and remote sensing images. Then, we turn to attention-based approach that is more relevant to our work.

A. Semantic Segmentation

It is one of the fundamental tasks of image understanding. FCN-based [7] methods have made great progress in semantic segmentation by leveraging the powerful representation abilities of classification networks [19], [20] pretrained on large-scale data [21]. Several model variants are proposed to aggregate multiscale contextual information that is vital for object perception. Concretely, DeepLabv2 [17] and DeepLab v3+ [26] employ ASPP to embed contextual representation, which consists of parallel convolutions with different dilated rates. PSPNet [9] proposes a PPM to extract the contextual information with different scales, each of which can be considered as the global representation. UNet [22], RefineNet [23], DFN [24], SegNet [25], DeepLab v3+ [26], and SPGNet [27] adopt an encoder–decoder structure to carefully recover the location information while retaining high-level semantic features. GCN [28] utilizes a global convolutional module and global pooling to harvest context information for global representations. BiSeNet [29] adopts efficient spatial and context path to achieve real-time semantic segmentation. In addition,
Yuan et al. [30] introduced a spatial structure preserving feature pyramid network to improve the spatial resolution of the category label map for semantic segmentation. Wang et al. [31] proposed a weakly supervised adversarial domain adaptation to enhance the segmentation performance from synthetic data to real scenes.

B. Semantic Segmentation of Aerial Imagery

Semantic segmentation in VHR aerial images benefits a lot from deep learning methods. For example, Mou et al. [32] proposed two network units, spatial relation module and channel relation module, to learn relationships between any two positions. TreeUNet [33] adopts a Tree-CNN block to transmit feature maps via concatenating connections and further fuse multiscale representations. ScasNet [10] proposes an end-to-end self-cascade network to improve the labeling coherence with sequential global-to-local contexts aggregation. SDNF [18] combines DCNNs and traditional decision forests algorithm in an end-to-end manner to achieve better classification accuracy. Marmanis et al. [34] focused on semantically edge detection to restore high-frequency details and further obtained fine object boundaries. DSMFNet [35] proposes a lightweight DSM fusion module to effectively aggregate depth information for semantic segmentation. LSLRR [36] improves the discriminative ability of low-rank representation in hyperspectral image (HSI) classification while considering the local geometric structure within data. Feng et al. [37] introduced a neighboring pixel affinity loss to improve the discriminability of semantic heterogeneous regions.

C. Attention-Based Methods

Attention is widely used for various tasks, such as machine translation [38], scene classification, and semantic segmentation. Squeeze-and-excitation networks [40] recalibrated the feature representations by modeling the dependencies between channels. Nonlocal [11] first adopts SA mechanism as a submodule for computer vision tasks, i.e., video classification, object detection, and instance segmentation. CCNet [13] harvests the contextual information of all the positions by stacking two serial criss-cross attention module. DANet [12] adopts a similar spatial and channel attention module to generate information from all pixels, which costs even more computation and GPU memory than the nonlocal operator [11], $A^2$-Nets [16] and expectation–maximization attention networks [14] sample sparse global descriptors to reconstruct the feature maps in an SA mechanism. ACFNet [41] proposes a coarse-to-fine segmentation network based on the attention class feature module, which can be embedded in any base network. Huang et al. [15], Yuan et al. [42], and Zhu et al. [43] further improved the efficiency of SA mechanism for semantic segmentation.

Motivated by the success of the attention-based methods above, and considering its limitations, we rethink the attention mechanism from the view of different perspectives and computation cost. Different from the previous works, we propose a hybrid multiple attention to capture global contexts from the perspective of space, channel, and category for better feature representations. Moreover, benefiting from the multiperspective attention mechanism and regionwise representations, HMANet is more efficient and effective than other attention-based methods. Comprehensive experimental results and empirical comparisons verify the superiority of our proposed method.

III. METHODOLOGY

A. Overview

As shown in Fig. 3, the network architecture mainly consists of three attention modules, CAA module, CCA module, and RSA module. The CAA module and the CCA module are embedded together as the upper branch of the network. The proposed CAA module aims to extract the class-level information, whereas the CCA module improves the process of
feature reconstruction via class channel weighting for better contextual representation. The lower branch of the network is the RSA module, accordingly, which greatly decreases the computational consumption and memory footprint in contrast to the original nonlocal block in computing long-range dependencies.

Concretely, given an input image, we first feed it into a convolutional neural network (CNN) to adaptively extract features for better representation, which is designed in a fully convolutional manner [7]. We take ResNet-101 pretrained on the ImageNet data set as our backbone. In particular, we remove the last two downsampling operations and use dilated convolutions in stage-3 and stage-4, which is also called a multigrid (MG) strategy for the latter, thereby retaining more spatial information and enlarging the output feature map X to 1/8 of the input image without adding extra parameters. Then, the features X from stage-4 of the backbone would be fed into two parallel attention branches.

The upper branch is the CAA module with the embedded CCA module. The CAA module is designed to model the dependencies between specific categories and the corresponding features after the dimension reduction. It extracts the similarity relationships between each category and each channel of the input feature through matrix operations. It helps to obtain a fine-grained feature representation that is more sensitive to object category information and enhance the discriminative ability of the network. The CCA module can be defined as the adaptive feature reconstruction [see (4)] of class channel information, which can effectively improve the feature representation of category information. It is worth mentioning that the CCA module takes the class affinity matrix [see (1)] and class attention map as the input features, both of which are generated by the CAA module. Then, it obtains the adaptive weighted class affinity matrix. Ideally, given the input feature map $X \in \mathbb{R}^{C \times H \times W}$, in which $C$, $H$, and $W$ denote the number of channels, height, and width of feature map, respectively, and the CAA embedded with CCA module can effectively extract the class channel correlation and adaptively aggregate long-range contextual information from a category view. Eventually, it outputs the same size feature map $Y \in \mathbb{R}^{C \times H \times W}$ following the SA scheme [11]. In the experiments, we crop the image into 512 $\times$ 512 slices. Accordingly, $H$ and $W$ of the input feature are both 64, and we set the input channel $C = 512$ to retain enough input information.

The lower branch of the network, RSA module, is proposed with the intuition of decomposing the dense pointwise affinity matrix into two sparse region-based counterparts, either of which could efficiently capture the global context in a sparser way via adaptive average pooling method. With the combination of the two affinity matrices, the RSA module could capture abundant spatial contextual information of the local input feature $X$ and then output feature $Z \in \mathbb{R}^{C \times H \times W}$. Finally, we concatenate the output features of the two branches $\{Y, Z\}$ and the local feature representation $X$ to obtain better feature representations, and then, the fused features are fed into a convolutional layer to generate the fine segmentation map.

B. Class Augmented Attention

The SA mechanism is essentially a kind of matrix multiplication operation in mathematics, in which the two dimensions are the number of channels $C$ and the product of height and width $[H \times W]$ of the input feature map. The standard channel affinity matrix of size $C \times C$ can be obtained by the matrix multiplication of two inputs with dimension $C \times H \times W$ and $H \times W \times C$, such as channel attention module in DANet [12]. Intuitively, the value of the channel dimension is strictly limited in such kind of channel attention module. The query, key, and value functions are defined as $1 \times 1$ convolution to compress the channel dimension and enhance the feature representation in nonlocal operator [11]. However, it is totally eliminated in the module. Nevertheless, it leads into category information when one of the channels $C$ is replaced by the number of the categories $N$. Accordingly, the obtained class attention map of size $N \times H \times W$ can be regarded as a coarse segmentation map. We utilize the ground truth as the supervision, so that the class attention map has the similar distribution with the final segmentation map. Note that the operations retain the query, key, and value transformation functions in the meantime.

The intuition of the proposed CAA is to capture long-range contextual information from the perspective of category information. We explicitly model the relationships between each category in the data set and each channel of the input feature cube. Next, we will elaborate the process to capture class-level contextual information.

As shown in Fig. 4, the local feature $X \in \mathbb{R}^{C \times H \times W}$ is the output from the 3 $\times$ 3 conv after stage-4 of ResNet in our implementation. The CAA module first applies two convolutional layers to generate two feature maps $X' \in \mathbb{R}^{C' \times H \times W}$ and $P' \in \mathbb{R}^{N \times H \times W}$, $C'$ is the reduced channel number of the local feature for less computational cost and $P'$ is the class attention map from the ground-truth supervised segmentation. For each channel $k$ in $P' \in \mathbb{R}^{N \times H \times W}$, $P_k \in \mathbb{R}^{H \times W}$ represents the confidence of pixels of all position $i$ belongs to class $k$, where $N$ is the number of categories. $X'_i \in \mathbb{R}^{H \times W}$ represents the $u$th element of $X'$ long channel dimension. Then, we can further generate the class affinity map $A \in \mathbb{R}^{C' \times N}$ via aggregating all the position $i$ in spatial dimension of $X'$ and $P'$. The class affinity operation is defined as follows:

$$a_{u,k} = \sum_{i=1}^{HW} (x'_{u,i} \cdot p'_{k,i})$$

(1)

where $a_{u,k} \in A$ denotes the explicit class correlation between feature $X'_u$ and $P_k$, $u = [1, 2, \ldots, C']$, $k = [1, 2, \ldots, N]$, and $A \in \mathbb{R}^{C' \times N}$.

The final class augmented object representation can be formulated as follows:

$$Y = \rho(\delta(A \cdot P') \parallel X)$$

(2)

in which $\parallel$ denotes the concatenation operation and $Y \in \mathbb{R}^{C \times H \times W}$ is the output feature map. Here, $\delta(\cdot)$ and $\rho(\cdot)$ are both transformation functions implemented by $1 \times 1$ conv $\rightarrow$ BN $\rightarrow$ ReLU. The original local feature $X$ is concatenated to enhance the feature representation. Equation (2) indicates that
the final representation of each channel is a class weighted sum of all channels in class attention map, which models the category-based semantic dependencies between feature maps. Intuitively, the proposed CAA module improves the perception and discriminability of the model for class-level information in a straightforward manner.

C. Class Channel Attention

The high-level semantics of CNNs are empirically considered to be embedded in the channel dimension. Each channel map of deep features can be regarded as a class-related response. In addition, recent works [40], [44] have demonstrated the effectiveness of modeling channel correlation in classification and segmentation tasks. Therefore, we propose a CCA module to exploit class channel dependencies and generate a new class affinity map with rich and adaptive contextual information. It is effectively embedded in the CAA module with a few parameters.

The main structure of CCA module is shown in Fig. 5. Given the class attention map $P \in \mathbb{R}^{N \times H \times W}$ and class affinity map $A \in \mathbb{R}^{C \times N}$ output from the CAA module above, the $k$th adaptive class channel statistical representations can be formulated as follows:

$$W_k = \sigma(f_{W_1, W_2}(\text{GAP}(P_k)))$$

where $\text{GAP}(P_k) = (1/HW) \sum_{i=1}^{H} \sum_{j=1}^{W} P_k(i, j)$ is the channelwise global average pooling (GAP) to generate class-related statistics, and $\sigma$ is the sigmoid activation. If $x = \text{GAP}(P) \in \mathbb{R}^N$, then the key adaptive feature recalibration function is defined as follows:

$$f_{W_1, W_2}(x) = W_2 \eta(W_1 x)$$

in which $W_2 \in \mathbb{R}^{N \times \alpha N}$ and $W_1 \in \mathbb{R}^{\alpha N \times N}$ and $\eta$ denotes the ReLU function. Concretely, $W_1(\cdot)$ and $W_2(\cdot)$ are two linear

fully connected transformations, i.e., dimensionality adjustment layers with ratio $\alpha$ (this parameter value will be discussed in Section IV-D3) to augment and squeeze the representations of category information in the channel dimension. Note that we opt to employ a simple ReLU function to ensure the nonlinearity of the model and limit the complexity following the squeeze-and-excitation networks [40].

The final output of the CCA module is obtained by recalibrating $A$ with the weighted factor $W$ and the original class affinity map

$$A' = \text{softmax}(\gamma W \cdot A + A)$$

where $\gamma$ is a learnable parameter initialized to 0. The residual connection $A$ is added to retain the original representation. It can be integrated into the standard CAA module above.
D. Region Shuffle Attention

Attention-based neural networks, especially for spatial pointwise attention methods, mainly aim to capture long-range contextual dependencies through a SA mechanism or its variants. Eventually, they generate a dense affinity matrix. Even for smaller feature maps, the obtained pointwise affinity matrix will take up a lot of (GPU) memory. Hence, the key point of the proposed RSA is to harvest the regionwise dependencies as well as its counterparts after recombination in a sparse and efficient manner. We illustrate our approach via a simple schematic in Fig. 6.

1) Region Representations: We partition the input feature maps into regions via a permutation operation, and each one is fed into an adaptive GAP layer to obtain the region representations afterward. Then, we merge the pointwise representations of the regions to generate a sparse representation of the whole input feature. Therefore, the SA on the original input features can be effectively replaced with the same attention toward the merged sparse representation for convenience.

2) Shuffle Attention Representations: Despite the SA on the merged feature that can empirically capture long-range contextual information from all positions, the pixel-to-pixel connections are still ambiguous. In order to exploit more explicit contextual dependencies from a regional perspective, we apply shuffle attention to alternately pool the corresponding subregions and compute its SA representations, respectively. In theory, it achieves a complementary representation of spatial information. Further experiments show that the cascade of attention-weighted representations of the two subregions can effectively enhance the contextual dependencies. The performance is superior to the pixelwise nonlocal operator.

As shown in Fig. 6, we first divide the input feature X into G partitions and each partition contains P positions, where each \( X_{G,p} \in \mathbb{R}^{C \times P} \) is a subset of \( X_G \). Then, we merge the point statistics after GAP to obtain the sparse representation \( X_m \in \mathbb{R}^{C \times G} \). We apply SA on \( X_m \) following the nonlocal operation [11] as given next:

\[
A_m = \text{softmax} \left( \frac{\theta(X_m)^T \phi(X_m)}{\sqrt{d}} \right)
\]

\[
Z_m = w A_m g(X_m) + X_m
\]

where \( A_m \in \mathbb{R}^{G \times G} \) is a sparse affinity matrix based on global information and \( Z_m \in \mathbb{R}^{C \times G} \) is the weighted output features. Here, \( \theta(\cdot) \) and \( \phi(\cdot) \) are both transformation functions implemented by \( 1 \times 1 \) conv \( \rightarrow \) BN \( \rightarrow \) ReLU, whereas \( g(\cdot) \) is a learnable parameter initialized to 0.

The regional weighted representation \( X'_G \) can be obtained by regionwise multiplication of \( Z_m \) and \( X_G \) (best viewed in color). We apply another permutation to regroup the representations, and then, the feature \( X'_G \) would be fed into the same regionwise attention block to generate the final representations \( Y \).

Compared with standard SA mechanism, our approach greatly reduces the complexity in time and space from \( O((H \times W)^2C) \) to \( O(2((1/(G_h^2G_w^2)) + (1/(P_h^2P_w^2))) (H \times W)^2C) \), where \( G_h \) and \( G_w \) are the number of partitions along height and width dimensions, while each partition contains \( P_h \) and \( P_w \) pixels, respectively.

In general, the proposed RSA module makes up for the deficiency of nonlocal block that it is a huge consumption of memory footprint. In addition, it can be plugged into any existing architectures at any stage without breaking its initial performance and optimized in an end-to-end manner.

E. Hybrid Multiple Attention Network

1) Integration of Attention Module: In order to take full advantage of these proposed attention modules, we further aggregate the CAA module embedded with the CCA module (the upper branch shown in Fig. 3) and the RSA module (the lower branch) in cascading and parallel manner, and both are concatenated with the local feature. Eventually, the feature after concatenation would be fed into the last convolution to generate the final segmentation map.

2) Algorithm Description: To better understand the workflow of our proposed HMANet, we further put an algorithm describing the method with detailed parameters in Algorithm 1.

3) Loss Function: Besides the conventional multiclass cross-entropy loss \( \mathcal{L}_{ce} \), we use the auxiliary supervision \( \mathcal{L}_{aux} \) after stage-3 to improve the performance and make it easier to optimize following PSPNet [9]. The auxiliary loss can be formulated as

\[
\mathcal{L}_{aux} = -\frac{1}{BN} \sum_{i=1}^{B} \sum_{j=1}^{N} \sum_{k=1}^{K} \mathbb{1}(g_i^j = k) \times \log \left( \frac{\exp(p_{i,j,k})}{\sum_{m=1}^{K} \exp(p_{i,j,m})} \right)
\]
Algorithm 1 Workflow of the Proposed HMANet

**Input:** The input image $x$ and label data $y$, the network weight parameters $w$, number of iterations $n$, and learning rate $l$.

**Output:** The learned weight parameters $w$ of the proposed HMANet.

1. Initialize $w$ and learning rate $l$.
2. for $i = 1$ to $n$ do
3. Obtain the basic feature map $X \in \mathbb{R}^{C \times H \times W}$ from the backbone.
4. **Branch 1:**
5. Calculate the class affinity map $A$ by Eq. 1.
6. Adaptively recalibrate the class-level information in channel dimension ($A \leftarrow A'$) via Eq. 4.
7. Extract the class augmented attention feature representation $Y \in \mathbb{R}^{C \times H \times W}$ by Eq. 2 ($A \leftarrow A'$).
8. **Branch 2:**
9. Obtain the shuffle attention feature representation $Z \in \mathbb{R}^{C \times H \times W}$ between regions (Each region is calculated by Eq. 6 and Eq. 7).
10. Then:
11. Aggregate the two features above and calculate the final score map $f(x) = Cls(Aggregate(Y, Z))$.
12. Calculate $Loss(w) = L(y, f(x))$ by Eq. 11.
13. Calculate the back propagation gradient $\frac{\partial Loss(w)}{\partial w}$.
14. Update $w \leftarrow w - l \frac{\partial Loss(w)}{\partial w}$.
15. end for

$\mathbb{I}(g_j^i = k) = \begin{cases} 1, & g_j^i = k \\ 0, & \text{otherwise} \end{cases}$ (9)

where $B$ is the mini batch size, $N$ is the number of pixels in each batch, $K$ is the number of categories, and $p_j^k$ is the prediction after ResNet-stage-3 of the $j$th pixel in the $i$th patch for the $k$th class. $\mathbb{I}(g_j^i = k)$ is an indicator function as illustrated in (9), and it takes 1 when the ground truth of the $j$th position in the $i$th patch (i.e., $g_j^i$) belongs to the $k$th class and 0 in other cases.

The class attention loss $L_{cls}$ from CAA module is also employed as an extra auxiliary supervision. Likewise, the class attention loss can be formulated as

$L_{cls} = -\frac{1}{BN} \sum_{i=1}^{B} \sum_{j=1}^{N} \sum_{k=1}^{K} \mathbb{I}(g_j^i = k) \log \left( \frac{\exp(a_j^i,k)}{\sum_{m=1}^{K} \exp(a_j^i,m)} \right)$ (10)

where $a_j^i,k$ is the response value of the class attention map of the $j$th pixel in the $i$th patch for the $k$th class; other definitions are the same as above.

Finally, we use three parameters to balance these losses as follows:

$L = \lambda_1 L_{ce} + \lambda_2 L_{cls} + \lambda_3 L_{aux}$ (11)

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are set as 1, 0.5, and 0.4 to balance the loss, respectively. Note that the ablation studies for the three loss functions and the sensitivity of the model to the choice of the weight values will be elaborated in Section IV-D1.

IV. EXPERIMENTS

To validate the effectiveness of our proposed method, we first conduct extensive experiments on two state-of-the-art aerial image semantic segmentation benchmarks, i.e., ISPRS 2-D Semantic Labeling Challenging for Vaihingen [45] and Potsdam [46], consisting of VHR true orthophotograph (TOP) tiles and corresponding digital surface models (DSMs) derived from dense image matching techniques. We further carry out experiments on the more challenging iSAID [47] data set to justify the validity of our model. In this section, we first introduce the data sets and implementation details, and then, we perform extensive ablation experiments on the ISPRS Vaihingen data set. Finally, we report our results on the three data sets.

A. Data Sets

1) Vaihingen: The Vaihingen data set contains 33 orthorectified image tiles (TOP) mosaic with three spectral bands (red, green, and near-infrared), plus a normalized DSM of the same resolution. The data set has a ground sampling distance (GSD) of 9 cm, with an average size of 2494 × 2064 pixels, which involves five foreground object classes and one background class. We use the benchmark organizer defined 16 images for training and 17 to test our model following the previous works [3], [32], [48]–[50]. Note that we do not use DSM in our experiments.

2) Potsdam: The Potsdam 2-D semantic labeling data set is composed of 38 high-resolution images of size 6000 × 6000 pixels, with a GSD of 5 cm. The data set offers NIR-R-G-B channels together with DSM and normalized DSM. There are 24 images in the training set and 14 images in the test set, which have the same six categories corresponding to the Vaihingen data sets.

3) iSAID: iSAID data set is based on the DOTA [51] data set for object detection, and it provides dense instance and semantic annotations for objects. The iSAID data set consists of 2806 high-resolution remote sensing images. The image sizes range from about 800 × 800 pixels to about 4000 × 13000 pixels. It contains 655451 instance annotations over 15 foreground categories and one background category of the objects. In this work, we only use semantic mask for object segmentation. Also, we use the predefined 1411 images for training and 458 images to validate our model. Note that the test set has 937 images, but they are unavailable.

B. Evaluation Metrics

To evaluate the performance of the proposed network, we calculate the $F_1$ score for the foreground object classes with the following formula:

$F_1 = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$ (12)

where $\beta$ is the equivalent factor between precision and recall and is set as 1. Intersection over union (IoU) and overall accuracy (OA) are defined as

$IoU = \frac{TP}{TP + FP + FN}$ (13)

where $TP$, $FP$, and $FN$ are the true positive, false positive, and false negative, respectively.
\[ \text{OA} = \frac{\text{TP} + \text{TN}}{N} \] (14)
in which TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively, and \( N \) is the total number of pixels.

Notably, OA is computed for all categories, including background for a comprehensive comparison with different models. In addition, the evaluation is carried out using ground truth with eroded boundaries provided in the data sets following previous studies.

C. Implementation Details

We use ResNet-101 [19] pretrained on ImageNet [21] as our backbone and employ a poly learning rate policy where the initial learning rate is multiplied by \( 1 - (\text{iter}/\text{max_iter})^{\text{power}} \) with power = 0.9 after each iteration following the prior works [13], [14], [17]. We utilize stochastic gradient descent (SGD) optimizer with the initial learning rate of 0.01 for training. The momentum and weight decay coefficients are set to 0.9 and 0.0005, respectively. We replace the standard BatchNorm with InPlace-ABNSync [52] to synchronize the mean and standard deviation of BatchNorm across multiple GPUs. For the data augmentation (DA), we apply random horizontal flipping, random scaling (from 0.5 to 2.0), and random crop over all the training images. The input size for all data sets is set to 512 \( \times \) 512. We employ 4 \times NVIDIA Tesla K80 GPU for 80k iterations and the batch size is 4. For semantic segmentation, we choose FCN (VGG-16 pretrained on ImageNet) [7] as our baseline, and we also utilize the ResNet-101 [19] baseline for further comparison experiments.

D. Experiments on Vaihingen Data Set

1) Ablation Study for Weight Parameters and Multiple Loss Functions: The proposed model utilizes multiple loss functions to optimize the learning process. We first conduct experiments to analyze the sensitivity of the model to the choice of the weight parameters \( \lambda_2 \) and \( \lambda_3 \). Concretely, we set the weight parameter of the main cross-entropy loss (i.e., \( \lambda_1 \)) as 1 and only preserve one of the auxiliary loss functions to further study the optimal value of \( \lambda_2 \) and \( \lambda_3 \). The experimental results for \( \lambda_2 \) and \( \lambda_3 \) are presented in Tables I and II. It can be seen that the choice of \( \lambda_2 = 0.5 \) and \( \lambda_3 = 0.4 \) yields the best result. Besides, it is worth mentioning that the model is not particularly sensitive to parameter selection. Thus, in order to avoid the influence of the training error of each experiment, we conduct five experiments for each value of the parameters and take the average value as the final result.

We further investigate the performance of the three loss functions following the optimal settings in Tables I and II. As shown in Table III, both auxiliary loss functions have certain improvement effects on model training optimization. It yields a result of 90.98\% in OA and 82.87\% in mean IoU when we utilize all the loss functions.

2) Ablation Study for Attention Modules: In the proposed HMANet, three attention modules are employed on the top of the dilation network to exploit global contextual representations from the perspective of space, channel, and category. To further verify the performance of attention modules, we conduct extensive experiments with different settings in Table IV. Note that for a fair comparison with the baseline model FCN (VGG-16), we also use VGG-16 as the backbone on HMANet in this experiment. Besides, we further investigate two integration patterns, that is, the parallel and cascading fashion, to adaptively accomplish information propagation.

As shown in Table IV, the proposed attention modules bring remarkable improvement compared with the baseline FCN (VGG-16). We can observe that the use of only CAA module yields a result of 89.15\% in OA and 79.56\% in mean I...
TABLE V

Comparison between different integration patterns. Cascade-C-R indicates that CAA embedded with CCA module is followed by the RSA module, and vice versa. Parallel-C-R represents CAA embedded with CCA and RSA that are appended on the top of the ResNet-101 in parallel.

| Method                  | OA(%) | mIoU(%) |
|-------------------------|-------|---------|
| ResNet-101 Baseline     | 90.12 | 80.81   |
| ResNet-101 + Cascade-C-R| 90.88 | 82.62   |
| ResNet-101 + Cascade-R-C| 90.76 | 82.45   |
| ResNet-101 + Parallel-C-R| 90.98 | 82.87   |

TABLE VI

Performance on the Vaihingen test set for different ascending ratios $\alpha$ in the CCA module.

| Ratio $\alpha$ | OA(%) | mIoU(%) |
|----------------|-------|---------|
| 50             | 90.78 | 82.48   |
| 75             | 90.80 | 82.49   |
| 100            | 90.82 | 82.52   |
| 125            | 90.84 | 82.53   |
| 150            | 90.85 | 82.54   |
| 175            | 90.83 | 82.52   |
| 200            | 90.81 | 82.50   |

IoU, which brings 2.64% and 6.87% improvement in OA and mIoU, respectively. Meanwhile, employing RSA individually outperforms the baseline by 2.72% in OA and 6.96% in mIoU. Furthermore, when we employ the integration of two corresponding attention modules together, the performance of our network is further boosted up. Finally, it behaves superiorly compared to other methods when we integrate the three attention modules, which improves the segmentation performance over baseline by 3.44% in OA and 7.99% in mIoU. In summary, it can be seen that our approach brings great benefit to object segmentation via exploiting global context from different perspectives.

We further investigate the effect of different aggregation methods of the three attention modules. As shown in Table V, the ResNet101 + Parallel-C-R, corresponding to the schematic in Fig. 3, achieves the best performance, i.e., 90.98% in OA as well as 82.27% in mean IoU, whereas the two cascading integration patterns, “+Cascade-C-R” and “+Cascade-R-C,” achieve 90.88% and 90.76% in OA, respectively. It shows that the cascading integration patterns lead to a decline in experimental results. The reason may be that the regionwise attention representation is not conducive to the extraction of category information only in the case of direct serial connection.

3) Ablation Study for Subparameters:

a) Ascending ratio: The ascending ratio $\alpha$ introduced in (4) is a hyperparameter, which allows us to control the scale of feature transformations. As the choice of ascending ratio does not have much effect on the computational cost, we only investigate the performance between a range of different $\alpha$ values. As shown in Table VI, we can conclude that our approach consistently outperforms the baseline under different choices of hyperparameters, among which the choice ratio $\alpha = 150$ achieves slightly better results than others. Qualitatively, the ratio $\alpha$ is the scaling factor of category information, which can take a moderate value while controlling the computational cost.

b) Effect of the partition numbers: We further investigate the effect of different partition numbers of the proposed RSA module, i.e., $G_h$ and $G_w$. We conduct extensive experiments with various choices of $G$ and $P$ and present the corresponding results in Table VII. Note that $G$ and $P$ are mutually constrained, namely, we just need to determine the values of $G_h$ and $G_w$. We can see that the performance is robust for a range of partition numbers, among which the choice $G_h = G_w = 8$ achieves the best 90.79% in OA and 82.49% in mean IoU. Empirically, the output stride of the backbone is set to 8, that is, the height and width of the input feature are 64 pixels in our experiments, and thus, eclectic choice of grouping is more conducive to SA weighted representations of each region. In practice, using an identical partition number may not be optimal (due to the distinct roles performed by different base network and different training settings, e.g., output stride and input size), so further improvements may be achievable by tuning the partition numbers to meet the needs of the given base architecture.

4) Comparison With Context Aggregation Approaches:

We compare the performance of several well-verified context aggregation approaches, i.e., ASPP in DeepLab v3 [8], PPM
in PSPNet [9], RCCA in CCNet [13], and SA in nonlocal networks [11]. All the experiments above are conducted under the same training/testing settings for fairness. We report the related results in Table VIII. Concretely, “+PPM” achieves better performance compared with “+ASPP” in terms of expanding local receptive fields. Both “+Self-Attention” and “+RCCA” generate contextual information from all spatial positions in the feature maps, leading to limited object contexts. In contrast, our HMANet calculates global correlations from the perspective of space, channel, and category. Results show that HMANet outperforms other context aggregation approaches, which demonstrates the effectiveness of capturing global contextual information from different perspectives.

5) Efficiency Comparison:

a) Comparison with SA: As shown in Fig. 7. We first compare our RSA module with the standard SA mechanism in terms of the computation cost measured with GFLOPs. As the size of input feature map increases, the GFLOPs of SA mechanism gradually increase exponentially, whereas the counterpart of our RSA module is almost linearly increasing. It can be seen that the RSA module is much more efficient than the SA mechanism when processing high-resolution feature maps.

b) Comparison with context aggregation modules and attention modules: We further compare our proposed CAA module and RSA module with ASPP [8], [17], PPM [9], SA [11], RCCA [13], OCR [42], and ISA [15] in terms of efficiency, including parameters, GPU memory, and computation cost (GFLOPs). We report the results in Table IX. Notably, we evaluate the cost of all the above methods without considering the cost of backbone and include the cost of 3 × 3 convolution for dimension reduction to ensure the fairness of the comparison. As shown in Table IX, compared with standard SA mechanism, our RSA module requires 20 times less GPU memory usage and significantly reduces FLOPs by about 77% with a few parameters, which proves the efficiency of regionwise representations in capturing long-range contextual information.

6) Comparison With State of the Art: We first adopt some common strategies to improve the performance following [12], [14], and [53].

![Fig. 7. Comparison of numerical complexity. The x-axis represents the height and width of the input feature map and the y-axis represents the computation cost measured with GFLOPs.](image)

### Table IX

| Method | Params(M\^2) | Memory(MB\^2) | GFLOPs(\^2) |
|--------|--------------|---------------|--------------|
| ASPP [8] | 15.1 | 284 | 503 |
| PPM [9] | 22.0 | 792 | 619 |
| SA [11] | 10.5 | 2168 | 619 |
| RCCA [13] | 10.6 | 427 | 804 |
| OCR [42] | 10.5 | 202 | 354 |
| ISA [15] | 11.8 | 252 | 386 |
| CAA(Ours) | 9.3 | 283 | 148 |
| RSA(Ours) | 3.8 | 110 | 144 |

### Table X

| Method | DA | MG | MS + Flip | OA(%) | mIoU(%) |
|--------|----|----|-----------|-------|--------|
| HMANet | ✓  |    |           | 90.98 | 82.87  |
| HMANet | ✓  | ✓  |           | 91.17 | 83.11  |
| HMANet | ✓  | ✓  | ✓         | 91.28 | 83.27  |
| HMANet | ✓  | ✓  | ✓         | 91.44 | 83.49  |

1) DA: DA with random scaling (from 0.5 to 2.0) and random left-right flipping.
2) MG: We employ hierarchical grids of different sizes (1, 2, and 4) within stage-4 of ResNet-101.
3) MS + Flip: We average the segmentation score maps from five image scales \{0.5, 0.75, 1.0, 1.25, 1.5\} and left-right flipping counterparts during inference.

Experimental results are shown in Table X. We successively adopt the above strategies to obtain better object representations, which achieves 0.19%, 0.11%, and 0.16% improvements in OA.

We further compare our method with existing methods on the Vaihingen test set. Notably, most of the methods adopt ResNet-101 as their backbone. The results are shown in Table XI. It can be seen that our HMANet (ResNet-101) outperforms other context aggregation methods and attention-based methods by a large margin. Moreover, our HMANet is much more efficient in parameters, memory, and GFLOPs. Especially, our $F_1$ score of Car is much higher than other approaches, and it improves the second best CCNet by 0.93%, which demonstrates the effectiveness of capturing category-based information and global regionwise correlation.

7) Visualization Results: We provide qualitative comparisons between our HMANet and baseline network in Fig. 8, including $512 \times 512$ and $1024 \times 1024$ patches. In particular, we leverage the red dashed box to mark those challenging regions that are easy to be misclassified. It can be seen that our method outperforms the baseline by a large margin. HMANet predicts more accurate segmentation maps, that is, it can obtain finer boundary information and maintain the object...
coherence, which demonstrates the effectiveness of modeling category-based correlation and regionwise representations.

8) Visualization of Attention Module: To get a deeper understanding of our proposed attention modules, we visualize the intermediate outputs of the important stage of the network. Note that we take the average value in the channel dimension for visualization. In Fig. 9, we visualize the learned class attention map in the CAA module. Fig. 9(b)–(d) shows the response value of the class attention map to the “Impervious Surface,” “Building,” and “Car,” respectively. We further show the direct feature output before and after the dual branches of the network. As shown in Fig. 10(b), the feature response extracted by the backbone is not clear. It is not focused on the objects, and the internal features are fragmented. Fig. 10(c) shows the output feature through the upper branch (CAA embedded with CCA module) mentioned above. It can be observed that objects of the same category have a similar response and maintain the coherence of the internal features. Fig. 10(d) shows the output of the lower branch (RSA module), and note that the numbers of the partitions \( G \) and positions in each partition \( P \) are both 8.
Fig. 8. Qualitative comparisons between our method and baseline on the Vaihingen test set.
We can see that the heat maps are regionalized. In large-scale scenes, the regionwise attention mechanism is more robust and efficient.

E. Experiments on Potsdam Data Set

We carry out experiments on the ISPRS Potsdam benchmark to further evaluate the effectiveness of HMANet. Empirically,
TABLE XIII
Numerical Comparisons With State of the Arts on the iSAID Validation Set

| Method          | Backbone          | IoU per category (%) | mIoU (%) |
|-----------------|-------------------|-----------------------|----------|
|                 |                   | PL       | BD     | BR     | GTF    | SV     | LV     | SH     | TC     | BC     | ST     | SBF    | RA     | HA     | SP     | IHC    |        |
| UNet [22]       | -                 | 74.74   | 6.51   | 7.48   | 5.52   | 35.62  | 49.89  | 49.0   | 78.60  | 22.89  | 0      | 9.67   | 46.49  | 45.64  | 38.03  |        | 37.39  |
| FCN-8s [7]      | VGG-16            | 62.90   | 26.44  | 8.17   | 27.65  | 37.05  | 49.35  | 51.74  | 74.81  | 30.24  | 22.91  | 52.07  | 51.91  | 42.02  | 30.74  | 0      | 41.66  |
| DenseASPP [59]  | DenseNet-121      | 78.12   | 67.54  | 29.61  | 52.28  | 38.44  | 57.10  | 61.15  | 86.09  | 56.56  | 50.05  | 74.10  | 64.8   | 51.09  | 43.26  | 0      | 56.81  |
| CCNet [13]      | ResNet-50         | 78.82   | 72.37  | 30.50  | 52.20  | 44.40  | 53.09  | 56.25  | 85.26  | 49.02  | 63.01  | 68.68  | 51.78  | 48.66  | 43.59  | 24.26  | 56.86  |
| Non-local [11]  | ResNet-50         | 80.96   | 71.85  | 29.07  | 49.75  | 47.80  | 59.09  | 62.36  | 84.86  | 48.39  | 69.21  | 41.58  | 53.52  | 50.84  | 45.12  | 29.89  | 57.27  |
| DANet [12]      | ResNet-50         | 80.93   | 71.41  | 28.55  | 54.54  | 42.12  | 57.50  | 60.24  | 84.73  | 52.92  | 63.01  | 63.25  | 40.61  | 48.82  | 46.05  | 32.30  | 57.49  |
| Semantic FPN [60]| ResNet-50         | 80.83   | 71.75  | 33.99  | 51.64  | 45.14  | 59.15  | 63.68  | 86.61  | 57.78  | 59.49  | 73.58  | 58.71  | 51.27  | 46.42  | 0      | 59.31  |
| DeepLab v3+ [8] | ResNet-50         | 75.70   | 75.94  | 32.11  | 59.24  | 33.79  | 54.54  | 59.02  | 84.18  | 58.52  | 55.15  | 73.78  | 67.51  | 45.76  | 44.24  | 31.14  | 59.33  |
| PSPNet [9]      | ResNet-50         | 79.50   | 75.70  | 32.46  | 60.15  | 42.96  | 58.03  | 65.20  | 85.57  | 61.12  | 52.10  | 71.90  | 68.60  | 54.26  | 46.78  | 10.89  | 60.25  |

| HMANet (Ours)   | ResNet-50         | **83.79** | 74.71  | 28.98  | 54.57  | **50.28** | **59.74** | **65.38** | **88.69** | 60.51 | **70.92** | 70.20 | 62.88 | 51.91 | **51.41** | **32.58** | **62.64** |
| HMANet (Ours)   | ResNet-101        | **84.48** | 76.93  | 37.13  | 57.31  | **49.11** | 58.78  | 65.60  | 88.34  | 63.82  | 71.44  | 70.98  | 62.71  | 51.01  | 50.70  | 33.36  | **63.85** |

* iSAID includes 15 foreground categories, i.e., plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), baseball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP) and helicopter (IHC).

Fig. 11. Visualization results of HMANet on the Potsdam test set.
we adopt the same training and testing settings on the Potsdam data set. Numerical comparisons with state-of-the-art methods are shown in Table XII. Remarkably, HMANet (ResNet-101) achieves 92.21% in OA and 87.28% in mean IoU. Notably, we compare the two types of available input images, i.e., RGB and IRRG color modes. Results show that the former can obtain better segmentation maps.

In addition, qualitative results are presented in Fig. 11. It can be seen that HMANet produces better segmentation maps than baseline. We mark the improved regions with red dashed boxes (best viewed in color).

F. Experiments on iSAID Data Set

To verify the validity of the proposed model, we further conduct experiments on the more challenging iSAID data set. The biggest challenge of the data set is the extreme foreground–background imbalance problem, and the background is much more complex.

For a fair comparison, we also choose ResNet-50 as the backbone in the experiment. The numerical results are shown in Table XIII. It can be seen that our HMANet achieves an mIoU of 62.64%, which outperforms other well-verified
methods (such as DeepLab v3+ [8] and PSPNet [9]). It is worth mentioning that ordinary pixelwise attention methods (nonlocal [11], CCNet [13], and DANet [12]) do not work well in such extreme scenarios. Too much redundant and invalid information in the background area is not conducive to the extraction of attention value between pixels. In contrast, the regionwise attention mechanism is much more effective and robust due to the region representations. The performance is further boosted up to 63.85% with backbone ResNet-101.

Some visualized results of HMANet (ResNet-50) on the iSAID validation set are shown in Fig. 12. The last column visualizes the segmentation results of the foreground objects.

V. Conclusion

In this article, we propose a novel attention-based framework for dense prediction tasks in the field of remote sensing, namely HMANet, which adaptively captures global contextual information from the perspective of space, channel, and category. In particular, we introduce a CAA module embedded with a CCA module to compute category-based correlation and further adaptively recalibrate the class-level information. In addition, to address the feature redundancy and improve the efficiency of SA mechanism, an RSA module is presented to obtain robust regionwise representations. Extensive experiments on the ISPRS Vaihingen, Potsdam benchmark, and iSAID data sets demonstrate the effectiveness and efficiency of the proposed HMANet. In the future, we would like to explore the complementary of the proposed method with other approaches focusing on graph convolutional networks. Global attention representation of the graph structure data is also another direction that is worthy of exploitation.

References

[1] M. Maboudi, J. Amini, S. Malihi, and M. Hahn, “Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images,” ISPRS J. Photogramm. Remote Sens., vol. 138, pp. 151–163, Apr. 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] 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.

[4] M. Pal, “Random forest classifier for remote sensing classification,” Int. J. Remote Sens., vol. 26, no. 1, pp. 217–222, Jan. 2005.

[5] J. A. Gualtieri and R. F. Crump, “Support vector machines for hyperspectral remote sensing classification,” Proc. SPIE, vol. 3584, pp. 221–232, Jan. 1999.

[6] P. Zhong and R. Wang, “A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 12, pp. 3978–3988, Dec. 2007.

[7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587. [Online]. Available: http://arxiv.org/abs/1706.05587

[9] H. Zhao, J. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2881–2890.

[10] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, “Semantic labeling in very high resolution images via a self-cascaded convolutional neural network,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 78–95, 2018.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[12] J. Fu et al., “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3146–3154.

[13] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 603–612.

[14] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9167–9176.

[15] L. Huang, Y. Yuan, C. Guo, C. Zhang, X. Chen, and J. Wang, “Interlaced sparse self-attention for semantic segmentation,” 2019, arXiv:1907.12273. [Online]. Available: http://arxiv.org/abs/1907.12273

[16] Y. Chen, Y. Kalamidis, J. Li, S. Yan, and J. Feng, “A*-nets: Double attention networks,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 352–361.

[17] L.-C. 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,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[18] L. Mi and Z. Chen, “Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation,” ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 140–152, Jan. 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[21] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[22] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Berlin, Germany: Springer, 2015, pp. 234–241.

[23] G. Lin, A. Milan, C. Shen, and I. Reid, “ReRefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1925–1934.

[24] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Song, “Learning a discriminative feature network for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1857–1866.

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

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 801–818.

[27] B. Cheng et al., “SPGNet: Semantic prediction guidance for scene parsing,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5218–5228.

[28] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters—Improve semantic segmentation by global convolutional network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4353–4361.

[29] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Song, “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 325–341.

[30] Y. Yuan, J. Fang, X. Lu, and Y. Feng, “Spatial structure preserving feature pyramid network for semantic image segmentation,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 15, no. 3, pp. 1–19, Sep. 2019.

[31] Q. Wang, J. Gao, and X. Li, “Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4370–4386, Sep. 2019.

[32] L. Mou, Y. Hua, and X. X. Zhu, “Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[33] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, “TreeUNet: Adapting a tree convolutional neural networks for subdecimeter aerial image segmentation,” ISPRS J. Photogramm. Remote Sens., vol. 156, pp. 1–13, Oct. 2019.
[34] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, “Classification with an edge: Improving semantic image segmentation with boundary detection,” ISPRS J. Photogramm. Remote Sens., vol. 135, pp. 158–172, Jan. 2018.

[35] Z. Cao et al., “End-to-End DSM fusion networks for semantic segmentation in high-resolution aerial images,” IEEE GeoSci. Remote Sens. Lett., vol. 16, no. 11, pp. 1766–1770, Nov. 2019.

[36] J. Wang, X. He, and X. Li, “Locality and structure regularized low rank representation for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 2, pp. 911–923, Feb. 2019.

[37] Y. Feng et al., “Nploss: Neighboring pixel affinity loss for semantic segmentation in high-resolution aerial imagery,” ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. 20, pp. 475–482, Aug. 2020.

[38] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473

[39] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[40] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[41] F. Zhang et al., “ACFNet: Attentional class feature network for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6798–6807.

[42] Y. Yuan, X. Chen, and J. Wang, “Object-contextual representations for semantic segmentation,” 2019, arXiv:1909.11065. [Online]. Available: http://arxiv.org/abs/1909.11065

[43] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, “Asymmetric non-local neural networks for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 593–602.

[44] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-net: Efficient channel attention for deep convolutional neural networks,” 2019, arXiv:1910.03151. [Online]. Available: http://arxiv.org/abs/1910.03151

[45] ISPRS 2D Semantic Labeling Contest-Vaihingen. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

[46] ISPRS 2D Semantic Labeling Contest-Potsdam. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/

[47] S. Waqas Zamil et al., “ISAID: A large-scale dataset for instance segmentation in aerial images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Dec. 2019, pp. 28–37.

[48] E. Magni, Y. Tarabalka, G. Charpiat, and P. Alliez, “High-resolution aerial image labeling with convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[49] M. Volpi and D. Tuia, “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 2, pp. 881–893, Feb. 2017.

[50] J. Sherrah, “Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,” 2016, arXiv:1606.02585. [Online]. Available: http://arxiv.org/abs/1606.02585

[51] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3974–3983.

[52] S. Rota Bulò, L. Porzi, and P. Kontschieder, “In-place activated batchnorm for memory-optimized training of dnns,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5639–5647.

[53] Y. Yuan and J. Wang, “OCNet: Object context network for scene parsing,” 2018, arXiv:1809.00916. [Online]. Available: http://arxiv.org/abs/1809.00916

[54] K. Nogueira, M. Dalla 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, 2019.

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

[56] L. Ding, J. Zhang, and L. Bruzzone, “Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[57] Y. Sun, X. Zhang, Q. Xin, and J. Huang, “Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LIDAR data,” ISPRS J. Photogramm. Remote Sens., vol. 143, pp. 3–14, Sep. 2018.

[58] Y. Sun, Y. Tian, and Y. Xu, “Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning,” Neurocomputing, vol. 330, pp. 297–304, Feb. 2019.

[59] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3684–3692.

[60] H. Liu et al., “An end-to-end network for panoptic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6172–6181.

Ruigang Niu (Graduate Student Member, IEEE) received the B.Sc. degree from Northwester Polytechnical University, Xi’an, China, in 2018. He is pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include computer vision, pattern recognition, and remote sensing image processing, especially on semantic segmentation.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009. He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Yu Tian received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 2018. He is pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include computer vision, pattern recognition, and remote sensing image processing, especially on instance segmentation.

Wenhui Diao (Member, IEEE) received the B.Sc. degree from Xidian University, Xi’an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016. He is an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Kaiqi Jiang (Member, IEEE) received the Ph.D. degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2019. He was a Visiting Scholar with the Karlsruhe Institute of Technology, Karlsruhe, Germany, from November 2017 to October 2018. He is an Assistant researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Kun Fu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively. He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.