HMANet: Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images

Ruigang Niu[a,b,c,1], Xian Sun[a,b,c,*], Wenhui Diao[a,c], Kaiqiang Chen[a,c] and Kun Fu[a,b,c]

[a]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.
[b]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China.
[c]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.

ARTICLE INFO

Keywords:
Semantic segmentation
Aerial imagery
Deep convolution neural networks
Self-attention mechanism

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 researches 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 spacial and channel attention and huge computation complexity of self-attention mechanism, it is unlikely to model the effective semantic interdependencies between each pixel-pair of remote sensing data of 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. Additionally, we introduce a simple yet effective region shuffle attention (RSA) module to reduce feature redundant and improve the efficiency of self-attention mechanism via region-wise representations. Extensive experimental results on the ISPRS Vaihingen and Potsdam benchmark demonstrate the effectiveness and efficiency of our HMANet over other state-of-the-art methods.

1. Introduction

Semantic segmentation, also known as semantic labeling, is one of the fundamental and challenging tasks in remote sensing image understanding, whose goal is to assign pixel-wise semantic class labels for a given image. In particular, semantic segmentation in very high resolution (VHR) aerial images plays an increasingly significance for its widespread applications, such as road extraction [21], urban planning [45] and land cover classification [23].

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) [27], Support Vector Machine (SVM) [11] and Conditional Random Fields (CRFs) [47]. Particularly, state-of-the-art methods based on the Fully Convolutional Network (FCN) [20] 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.* [6] proposed atrous spatial pyramid pooling (ASPP) with multi-scale dilation rates to aggregate contextual information. Zhao *et al.* [46] further introduced the pyramid pooling module (PPM) to represent the feature map via multiple regions with different sizes. ScasNet [19] aggregates multi-scale 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 pixel-wise contextual information, Non-local Neural Networks [38] utilizes a self-attention 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 [10] introduces spatial-wise and channel-wise attention modules to enrich the feature representations. Besides, several works [16, 17, 15] improve the efficiency of the self-attention mechanism to some extent. However, pixel-wise attention approaches still 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 [8, 17] have shown the fact that information redundancy is not conducive to feature representations. What's 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 classifier of the network. In general, lack of category-based information and huge computation complexity of self-attention mechanism are two tough problems and will be elaborated below.

*Corresponding author.
ORCID(s): 0000-0003-3161-7313 (R. Niu)

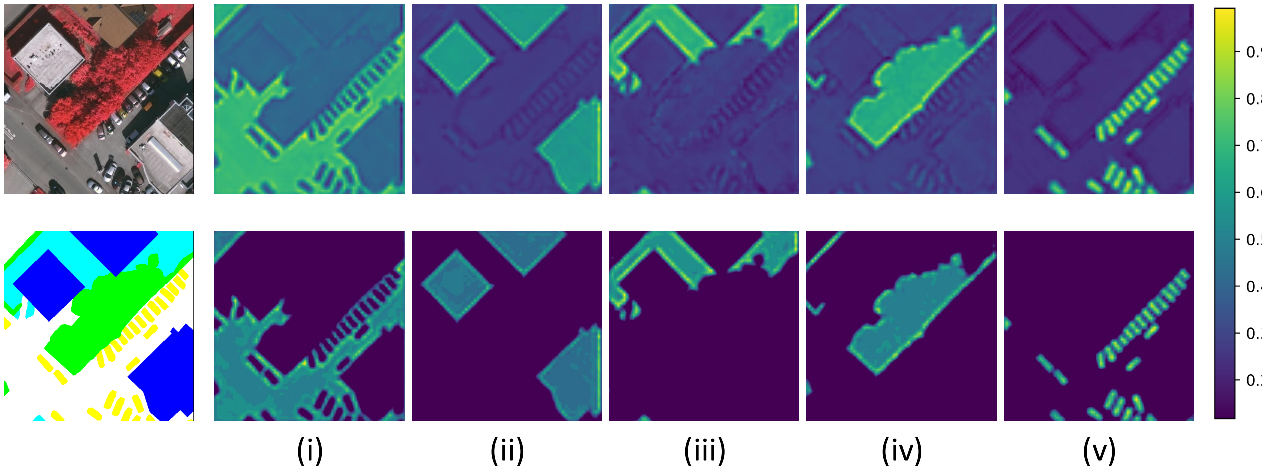


Figure 1: Visualization of Class Attention Maps corresponding to each class: (i) Imprevious surfaces. (ii) Building. (iii) Low vegetation. (iv) Tree. (v) Car. The upper row tends to represent the class attention response while the lower counterpart represents the non-negative response after non-linear activation. (Best viewed in color).

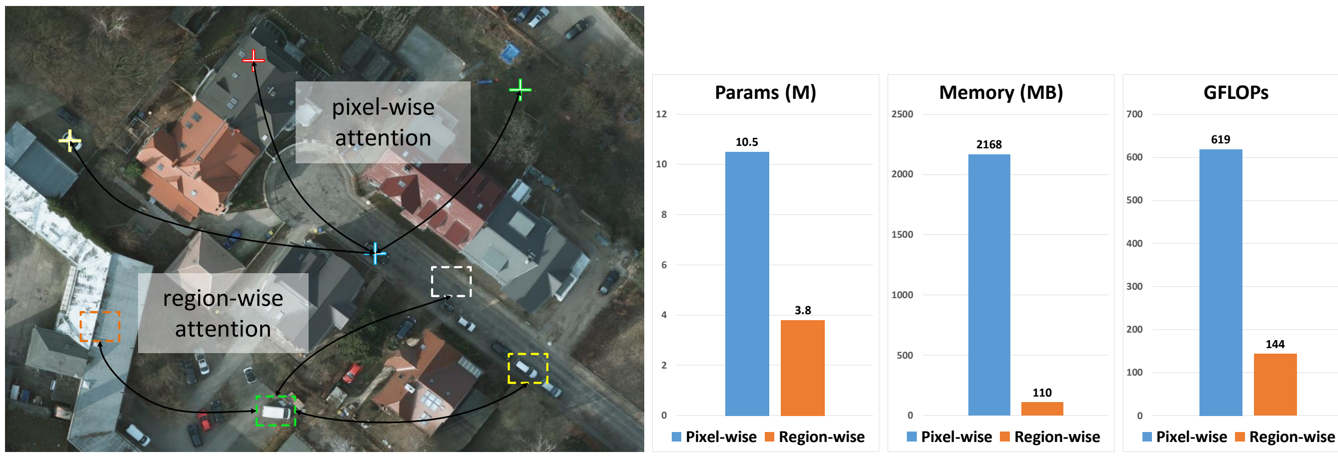
On one hand, in previous works [20, 5, 6, 46, 25], category-based information is only reflected in the last convolutional layer, that is, the score map representing the probability that each pixel belong to each category. Empirically, the lack of class-level information in the middle stage of the network leads to poor object classification capabilities. Different from the methods above, we argue that exploiting class-level information is also vital for semantic segmentation task. The class-level response is closely relevant to the output of the classifier. Thus, we propose a so-called category-based correlation that models class-level representation of each pixel and further calculates the relationships between categories and corresponding channels of the feature cube. 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.

On the other hand, aiming at the feature redundancy and heavy computation complexities of the pixel-wise attention mechanism, we propose a simple yet efficient scheme to address the issue. For example, as for a single feature belonging to 'car' in Fig. 2 (a), the pixel-wise attention method usually extracts features of all other positions, among which we actually do not need to focus on the 'building' and 'impervious surface'. But it is hard for the network to learn such precise features due to the similarity of different categories under some circumstance (such as in the shadow or overlapping). Besides, several works [17, 8] have proved that the invalid redundant information is not conducive to the feature representations, and it is a waste of GPU memory in the meantime. It is illustrated in Fig. 2 (b). Therefore, we employ a more robust region-wise attention mechanism to exploit a wider range of correlations. Empirically, region-wise representation can extract long-range contextual information between pixels in a more efficient manner.

Towards 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 Class Channel Attention (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. While 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 region-wise global information with a shuffling operator and obtain more robust correlation between objects. Besides, compared with pixel-wise self-attention methods, the grouped region-wise representation requires 20× less GPU memory usage and significantly reduces 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 classifier to further generate the fine segmentation map.

Our contributions can be summarized as follows:

- We present a Class Augmented Attention (CAA) module to exploit the category-based correlation between pixels and enhance the discriminant ability for each class, within which a Class Channel Attention (CCA) module is embedded to recalibrate the class-level information for better representations adaptively.
- The Region Shuffle Attention (RSA) module is proposed to capture region-wise global information and obtain more robust relationships between objects in a more efficient and effective manner.
- We propose a novel Hybrid Multiple Attention Network (HMANet) by taking advantage of the three at-



(a) (b)

Figure 2: Intuitive understanding of pixel-wise attention and region-wise attention. The right side shows the the advantage of region-wise attention over the standard pixel-wise attention in terms of the Parameters (measured by M), GPU memory cost (measured by MB) and computation cost (measured by GFLOPs). It can be seen that the region-wise attention requires 20× less GPU memory usage and reduces FLOPs by about 77%.

tention modules above, which comprehensively captures feature correlations from the perspective of space, channel and category.

The reminder of this paper is arranged as follows. Related work is briefly introduced in Section. 2. Section. 3 presents the details of our proposed method, including three attention modules, respectively. Experimental evaluations between our HMANet and the state-of-the-art methods, as well as ablation studies on Vaihingen dataset are provided in Section. 4. Finally, the conclusion is outlined in Section. 5

2. Related Work

A full review is beyond the scope of this paper. 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.

Semantic Segmentation. Semantic segmentation is one of the fundamental tasks of image understanding. Fully Convolutional Networks (FCNs) [20] based methods have made great progress in semantic segmentation by leveraging the powerful representation abilities of classification networks [12, 14] pretrained on large-scale data [31]. Several model variants are proposed to aggregate multi-scale contextual information that is vital for object perception. Concretely, DeepLabv2 [5] and DeepLabv3 [6] employ atrous spatial pyramid pooling (ASPP) to embed contextual representation, which consists of parallel convolutions with different dilated rates. PSPNet [46] proposes a pyramid pooling module (PPM) to extract the contextual information with different scales, each of which can be considered the global representation. UNet [29], RefineNet [18], DFN [40], SegNet [2], DeepLabv3+ [7] and SPGNet [9] adopt encoder-decoder structure to carefully recover the location information while retaining high-level semantic features. GCN [28] utilizes global convolutional module and global pooling to harvest context information for global representations. In addition, BiSeNet [39] adopts efficient spatial and context path to achieve real-time semantic segmentation.

Semantic Segmentation of Aerial Imagery. Semantic segmentation in VHR aerial images benefits a lot from deep learning methods. For example, Mou *et al.* [26] propose two network units, spatial relation module and channel relation module, to learn relationships between any two positions. TreeUNet [43] adopts a Tree-CNN block to transmit feature maps via concatenating connections and further fuse multi-scale representations. ScasNet [19] proposes an end-to-end self-cascade network to improve the labeling coherence with sequential global-to-local contexts aggregation. SDNF [25] combines DCNNs and traditional decision forests algorithm in an end-to-end manner to achieve better classification accuracy. Marmanis *et al.* [24] focus on semantically edge detection to restore high-frequency details and further obtain fine object boundaries. DSMFNet [4] proposes a lightweight DSM fusion module to effectively aggregate depth information, within which Cao *et al.* [4] investigate four fusion strategies corresponding to different scenarios.

Attention-based Methods. Attention is widely used for various tasks, such as machine translation [3, 35], scene classification and semantic segmentation. Squeeze-and-Excitation Networks recalibrated the feature representations by modeling the dependencies between channels. Non-local [38] first adopts self-attention mechanism as a submodule for computer vision tasks, *i.e.*, video classification, object detection and instance segmentation. CCNet [16] harvests the contextual information of all the positions by stacking two serial criss-cross attention module. DANet [10] adopts similar spatial and channel attention module to generate information from all pixels, which costs even more computation and GPU memory than the Non-local operator [38]. A^2-

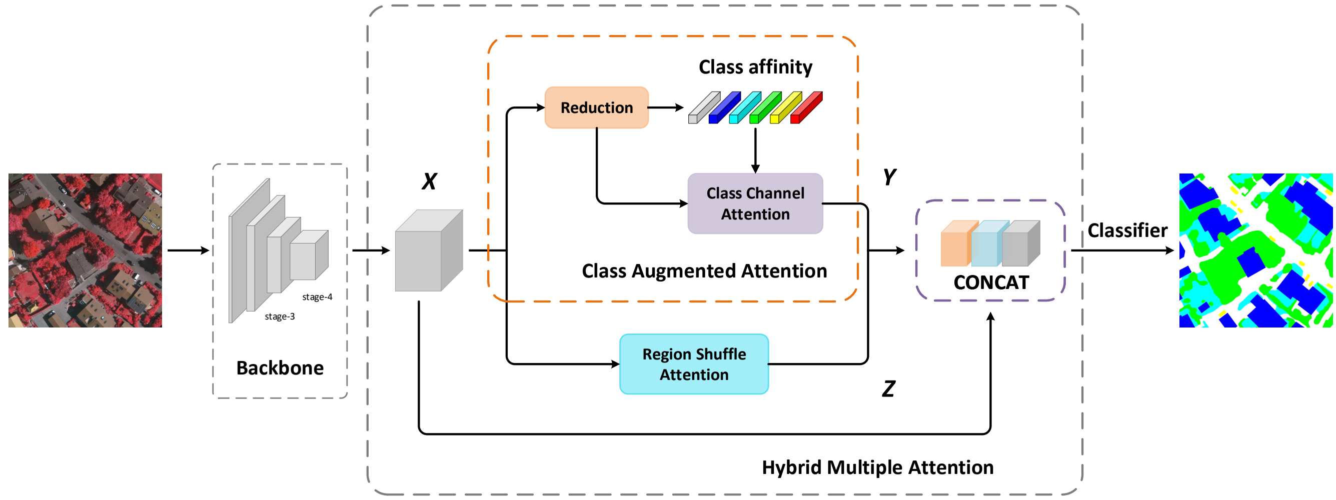


Figure 3: The pipeline of the proposed Hybrid Multiple Attention Network (HMANet). The key components are the two parallel branches, Class Augmented Attention (CAA) module embeded with Class Channel Attention (CCA) module and Region Shuffle Attention (RSA) module, which obtain the category-based correlation and region-wise contextual dependencies, respectively. Empirically, we concatenate the two output feature maps $\{Y, Z\}$ and the local representation X to further generate the final segmentation map (Best viewed in color).

Nets [8] and Expectation-Maximization Attention Networks [17] sample sparse global descriptors to reconstruct the feature maps in an self-attention mechanism. ACFNet [44] proposes a coarse-to-fine segmentation network based on attention class feature module, which can be embedded in any base network. Huang *et al.* [15], Yuan *et al.* [41] and Zhu *et al.* [48] further improve the efficiency of self-attention mechanism for semantic segmentation.

Motivated by the success of the attention-based methods above, we rethink the attention mechanism from the view of different perspectives and computation cost. Different from previous works, we propose a Hybrid Multiple Attention to capture global correlations from the perspective of space, channel and category respectively for better feature representations. Moreover, benefiting from the multi-perspective attention mechanism and region-wise representations, HMANet is more efficient and effective than other attention-based methods. Comprehensive empirical results verify the superiority of our proposed method.

3. Approach

3.1. Overview

As shown in Fig. 3, the network architecture mainly consists of three attention modules, Class Augmented Attention (CAA) module, Class Channel Attention (CCA) module and Region Shuffle Attention (RSA) module, among which CAA module and CCA module are embedded together as the upper branch of the network. The proposed CAA module aims to extract the class-based correlation of the feature map while 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 no-local 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 [20]. We take ResNet-101 pre-trained on ImageNet dataset as our backbone. In particular, we remove the last two down-sampling operations and use dilated convolutions in stage-3 and stage-4, which is also called a multi-grid 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 the stage-4 of the backbone would be fed into two parallel attention branches.

The upper branch is the CAA module embedded with the CCA module, while the CAA module is designed to model the dependencies between specific categories and the corresponding features after the dimension reduction and the CCA module can be defined as the adaptive feature reconstruction of class channel information. It is worth mentioning that the CCA module takes the class affinity matrix and class attention map as the input features, both of which are generated by the CAA module, then, 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, 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, outputs the same size feature map $Y \in \mathbb{R}^{C \times H \times W}$ following the self-attention scheme [38].

The lower branch of the network, RSA module, is proposed with the intuition of decomposing the dense point-

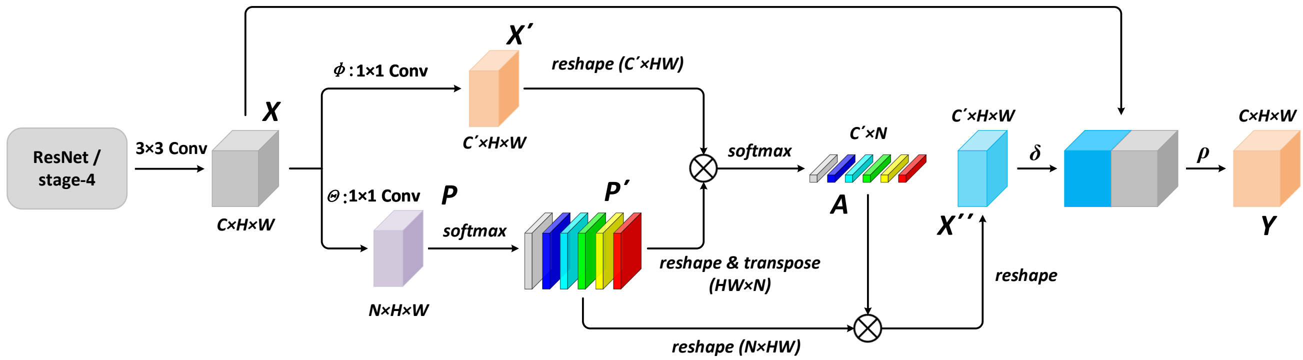


Figure 4: The details of class augmented attention module (Best viewed in color).

wise 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 then output feature $\mathrm{Z} \in \mathbb{R}^{C\times H\times W}$. Finally, we concatenate the output features of the two branches $\{\mathrm{Y}, \mathrm{Z}\}$ and the local feature representation X to obtain better feature representations, then, the fused features are fed into a classifier to generate the fine segmentation map.

3.2. Class Augmented Attention

The self-attention 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 respectively. The standard channel affinity matrix of size $C \times C$ can be obtained by merging the HW dimension, thereby generating channel attention map, such as channel attention module in DANet [10]. Intuitively, the definition of non-local operation constrains the scaling of the channel in such kind of channel attention module, that is, the query, key and value operation. Nevertheless, it leads into category information when one of the channels C is replaced by the channel corresponding to the segmentation map supervised by the ground truth, retaining the query, key and value transformation functions in the meantime.

The intuition of the proposed class augmented attention is to capture long-range contextual information from the perspective of category information, that is, to explicitly model the relationships between each category in the dataset 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, given a local feature $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$, output from the 3×3 conv after stage-4 of ResNet in our implementation, the class augmented attention module first applies two convolutional layers to generate two feature maps $\mathrm{X}' \in \mathbb{R}^{C'\times H\times W}$, and $\mathrm{P} \in \mathbb{R}^{N\times H\times W}$, respectively, where C' is the reduced channel number of the local feature for less computational cost and P is the class attention map supervised by the ground-truth segmentation. For each channel k in $\mathrm{P} \in \mathbb{R}^{N\times H\times W}$, $\mathrm{P_k} \in \mathbb{R}^{H\times W}$ is available to represent the confidence that pixels of all position i belongs to class k, where N is the number of categories. $\mathrm{X'_u}$ represents the uth element of X' along channel dimension. Then, we can further generate the class affinity map $\mathrm{A} \in \mathbb{R}^{C'\times N}$ via aggregating all the position i in spatial dimension of X' and P after a softmax layer. The class affinity operation is defined as follows:

$$s_{u,k} = \sum_i x'_{u,i} \frac{e^{p_{k,i}}}{\sum_{j=1}^{N} e^{p_{j,i}}} \tag{1}$$

where $s_{u,k} \in \mathrm{S}$ denotes the explicit class correlation between feature $\mathrm{X'_u}$ and $\mathrm{P_k}$, $u = [1, 2, ..., C']$, $k = [1, 2, ..., N]$, $\mathrm{S} \in \mathbb{R}^{C'\times N}$. Then, we apply a softmax operation along the class dimension to generate the class affinity map A.

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

$$\mathrm{Y_u} = \rho(\delta \sum_{k=1}^{N} (a_{u,k} \cdot \frac{e^{P_k}}{\sum_{j=1}^{N} e^{P_j}}) + \mathrm{X_u}) \tag{2}$$

in which $\mathrm{Y_u}$ denotes the uth feature plane of the output feature map $\mathrm{Y} \in \mathbb{R}^{C\times H\times W}$. $a_{u,k}$ is a scalar value of $s_{u,k}$ after softmax layer. Here, $\delta(\cdot)$ and $\rho(\cdot)$ are both transformation functions implemented by 1×1 conv $\rightarrow$ BN $\rightarrow$ ReLU. The original local feature X is added to enhance the adaptive feature representation. The Eq. (2) indicates that the final representation of each channel is a category-based weighted sum of all channels in class attention map, which models the category-based semantic dependencies between feature maps. That is to say, the proposed CAA module improves the perception and discriminability of class-level information in a straightforward manner.

3.3. Class Channel Attention

The high-level semantics of CNNs are empirically considered to be embedded in the channel dimension, among

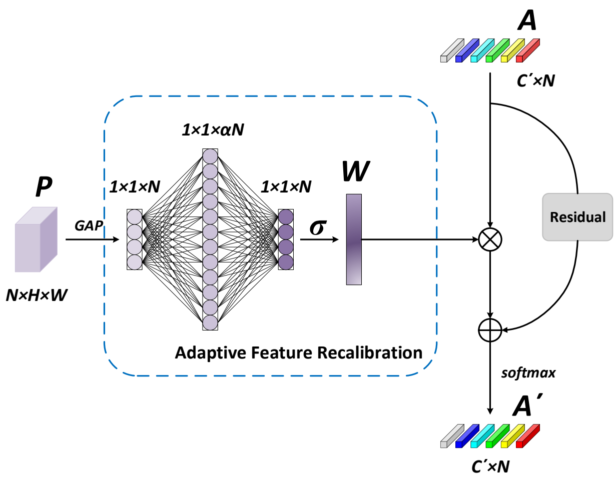


Figure 5: Diagram of class channel attention module.

which each channel map of deep features can be regarded as a class-related response. Additionally, recent works [13, 37] have demonstrated the effectiveness of modeling channel correlation in classification and segmentation tasks. Therefore, we propose a class channel attention (CCA) module to exploit class channel dependencies and generate a new class affinity map with rich and adaptive contextual information, which is effectively embedded in the CAA module with a few parameters.

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

$$\mathrm{W_k} = \sigma(f_{\{\mathrm{W_1,W_2}\}} \cdot (GAP(\frac{e^{P_k}}{\sum_{j=1}^{N} e^{P_j}}))) \tag{3}$$

where $GAP(P_k) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} P_k(i,j)$ is the channel-wise global average pooling to generate class-related statistics and σ is the Sigmoid activation. Let $x = GAP(P_k)$, the key adaptive feature recalibration function is defined as:

$$f_{\{\mathrm{W_1,W_2}\}}(x) = \mathrm{W_2}\theta(\mathrm{W_1}x) \tag{4}$$

in which $\mathrm{W_2} \in \mathbb{R}^{N\times\alpha N}$ and $\mathrm{W_1} \in \mathbb{R}^{\alpha N\times N}$ are two linear transformations, *i.e.*, a dimensionality ascending layer with ratio α (this parameter value will be discussed in Section 4.4.2) to augment and squeeze the feature representations, respectively, and θ denotes the ReLU function.

The final output of the CCA module is obtained by recalibrating A with the weighted factor $\mathrm{W_k}$ and the original class affinity map:

$$\mathrm{A}' = softmax(\sum_{i=1}^{N}(\gamma \mathrm{W_i} + 1)\mathrm{A_i}) \tag{5}$$

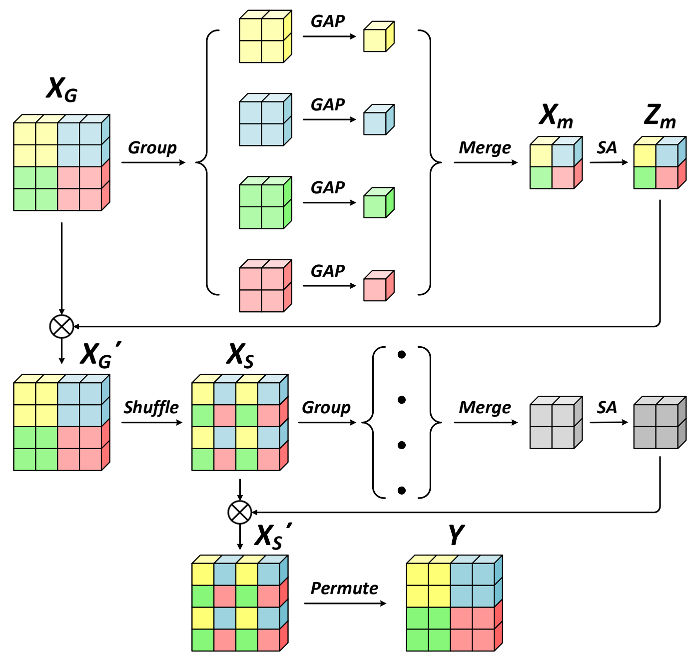


Figure 6: An example of region shuffle attention when the numbers of partitions G and positions in each patition P are both 2.

where γ is a learnable parameter initialized to 0. The residual connection is added to retain the original representation, thus, it can be integrated into the standard CAA module above without breaking its initial behavior, which efficiently helps to enhance the feature adaptive recalibration of class information.

3.4. Region Shuffle Attention

Attention-based neural networks, in terms of spatial point-wise correlation representations, mainly aim to capture long-range contextual dependencies through a self-attention mechanism or its variants, eventually, generating a dense affinity matrix. Even for smaller feature maps, the point-wise affinity matrix obtained will take up a lot of (GPU) memory. Hence, the key point of the proposed region shuffle attention is to harvest the region-wise 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.

Region Representations. We partition the input feature maps into regions via a permutation operation, each of which is fed into an adaptive global average pooling layer to obtain the region representations afterwards. Then, we merge the point-wise representations of the regions to generate a sparse representation of the whole input feature. Therefore, the self-attention on the original input features can be effectively replaced with the same attention towards the merged counterparts for convenience.

Shuffle Attention Representations. Despite the self-attention 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 a shuffle attention to alternately pool the corresponding sub-regions and compute its self-attention representations, respectively, achieving a complementary representation of spatial information. Further experiments show that the cascade of attention weighted representations of the two sub-regions can effectively enhance the contextual dependencies, superior to the pixel-wise non-local operator.

As illustrated in Fig. 6, we first divide the input feature X into G partitions and each partition contains P positions, where each $\mathrm{X_{G,p}} \in \mathbb{R}^{C\times P}$ is a subset of $\mathrm{X_G}$. Then, we merge the point statistics after global average pooling to obtain the sparse representation $\mathrm{X_m} \in \mathbb{R}^{C\times G}$. We apply self-attention on $\mathrm{X_m}$ following the non-local operation [38] as below:

$$\mathrm{A_m} = softmax(\frac{\theta(\mathrm{X_m})^{\mathrm{T}}\phi(\mathrm{X_m})}{\sqrt{d}}) \tag{6}$$

$$\mathrm{Z_m} = w\mathrm{A_m}g(\mathrm{X_m}) + \mathrm{X_m} \tag{7}$$

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

The regional weighted representation $\mathrm{X'_G}$ can be obtained by region-wise multiplication of $\mathrm{Z_m}$ and $\mathrm{X_G}$. We apply another permutation to regroup the representations, then, the feature $\mathrm{X_S}$ would be fed into the same region-wise attention block to generate the final representations Y.

Compared with standard self-attention mechanism, our approach greatly reduces the complexity in time and space from $\mathcal{O}((H\times W)^2C)$ to $\mathcal{O}(2(\frac{1}{G_h^2G_w^2}+\frac{1}{P_h^2P_w^2})(H\times W)^2C)$, 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 region shuffle attention module makes up for the deficiency of non-local block that it is a huge consumption of memory footprint. Additionally, it can be plugged into any existing architectures at any stage without breaking its initial performance, and optimized in an end-to-end manner.

3.5. Hybrid Multiple Attention Network

Integration of Attention Module. In order to take full advantage of three proposed attention modules, we further aggregate the CAA module embedded with the CCA module (the upper branch illustrated in Section. 3.1) and the RSA module (the lower branch) in an cascading and parallel manner, both of which is concatenated with the local feature. Eventually, the feature after concatenation would be fed into a classifier to generate the final segmentation map.

Loss Function. Besides the conventional multi-class 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 [46]. The class attention loss $\mathcal{L}_{cls}$ from CAA module is also employed as an extra auxiliary supervision. Finally, we use three parameters to balance these loss as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{ce} + \lambda_2\mathcal{L}_{cls} + \lambda_3\mathcal{L}_{aux} \tag{8}$$

where λ_1, λ_2 and λ_3 are set as 1, 0.5 and 0.4 to make these loss value ranges comparable.

4. Experiments

To validate the effectiveness of our proposed method, we conduct extensive experiments on two state-of-the-art aerial image semantic segmentation benchmarks, *i.e.*, ISPRS 2D Semantic Labeling Challenging for Vaihingen and Potsdam, consisting of very high resolution true ortho photo (TOP) tiles and corresponding digital surface models (DSMs) derived from dense image matching techniques. In this section, we first introduce the datasets and implementation details, then we perform extensive ablation experiments on the ISPRS Vaihingen dataset. Finally, we report our results on the two datasets.

4.1. Datatsets

Vaihingen. The Vaihingen dataset contains 33 orthorectified image tiles (TOP) mosaic with three spectral bands (red, green, near-infrared), plus a normalized digital surface model (DSM) of the same resolution. The dataset has a spatial resolution of 9 cm, with an average size of 2494×2064 pixels, which involves five foreground object classes and one background class. We select 16 images for training and 17 to test our model following the previous works [26, 22, 36, 32, 23]. Noted that we do not use DSM in our experiments.

Potsdam. The Potsdam 2D semantic labeling dataset is composed of 38 high resolution images of size 6000×6000 pixels, with a spatial resolution of 5 cm. The dataset offers NIR-R-G-B channels together with DSM and normalized DSM. There are 24 images in training set and 16 images in test set, which have 6 foreground classes corresponding to the Vaihingen benchmark.

4.2. 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:

$$F_1 = (1\times\beta^2)\cdot\frac{precision\cdot recall}{\beta^2\cdot precision + recall} \tag{9}$$

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

$$\mathrm{IoU} = \frac{TP}{TP+FP+FN} \tag{10}$$

$$\text{OA} = \frac{TP + FN}{N} \tag{11}$$

in which TP, FP and FN are the number of true positives, false positives and false negatives, respectively. N is the total number of pixels.

Notably, overall accuracy is computed for all categories including background for a comprehensive comparison with different models. Additionally, the evaluation is carried out using ground truth with eroded boundaries provided in the datasets following previous studies.

4.3. Implementation Details

We use ResNet-101 [12] pretrained on ImageNet as our backbone and employ a poly learning rate policy where the initial learning rate is multiplied by $1-(\frac{iter}{max\_iter})^{power}$ with $power = 0.9$ after each iteration following the prior works [5, 17, 16]. The initial learning rate is set to be 0.01 for all datasets. Momentum and weight decay coefficients are set to 0.9 and 0.0005 respectively. We replace the standard BatchNorm with InPlace-ABNSync [30] to synchronize the mean and standard-deviation of BatchNorm across multiple GPUs. For the data augmentation, 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 datasets is set to 512×512. We employ 4× NVIDIA Tesla K80 GPU for $80k$ iterations and batch size is 4. For semantic segmentation, we choose FCN [20] as our baseline.

4.4. Experiments on Vaihingen Dataset

4.4.1. 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 Tab. 1. Besides, we further investigate two integration patterns, that is, the parallel and cascading fashion, to adaptively accomplish information propagation.

As illustrated in Tab. 1, the proposed attention modules bring remarkable improvement compared with the baseline FCN. We can observe that the use of only class augmented attention module yields a result of 90.77% in overall accuracy and 82.45% in mean IoU, which brings 4.26% and 9.76% improvement in OA and mIoU, respectively. Meanwhile, employing region shuffle attention individually outperforms the baseline by 4.28% in OA and 9.8% 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 4.47% in OA and 10.18% in mIoU. In summary, it can be seen that our approach brings great benefit to object segmentation via exploiting global context from different perspectives.

Table 1
Ablation study for attention modules on Vaihingen test set. CAA represents class augmented attention module, CCA represents channel class attention module, RSA represents region shuffle attention module.

Method	CAA	CCA	RSA	OA(%)	mIoU(%)
Baseline [20]				86.51	72.69
HMANet	✓			90.77	82.45
HMANet			✓	90.79	82.49
HMANet	✓	✓		90.85	82.54
HMANet	✓		✓	90.87	82.61
HMANet	✓	✓	✓	**90.98**	**82.87**

Table 2
Comparison between different integration patterns. Cascade-C-R indicates that CAA embedded with CCA module is followed by RSA module, and vice versa. Parallel-C-R represents CAA embedded with CCA and RSA are appended on the top of the ResNet-101 in parallel.

Method	OA(%)	mIoU(%)
ResNet-101 Baseline	86.51	72.69
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**

We further investigate the effect of different aggregation methods of the three attention modules. As shown in Tab. 2, the ResNet101 +Parallel-C-R, corresponding to the schematic diagram in Fig. 3, achieve the best performance, *i.e.*, 90.98% in overall accuracy, as well as 82.27% in mean IoU. While the two cascading integration patterns, "+Cascade-C-R" and "+Cascade-R-C" achieve 90.88% and 90.76% in overall accuracy, respectively. It shows that the cascading integration patterns lead to a decline in experimental results. The reason may be that the region-wise attention representation is not conducive to the extraction of category information only in the case of direct serial connection.

4.4.2. Ablation Study for Sub-parameters

Ascending ratio. The ascending ratio α introduced in Eq. (4) is a hyper-parameter 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 α values. As shown in Tab. (3), we can conclude that our approach consistently outperforms the baseline under different choices of hyper-parameters, among which the choice ratio $\alpha = 150$ achieves slightly better results than others. Qualitatively, the ratio α is the scaling factor of category information, which can takes a moderate value while controlling the computational cost.

Effect of the Partition numbers. We further investigate the effect of different partition numbers of the proposed region shuffle attention module, *i.e.*, G and P. We conduct extensive experiments with various choices of G and P, and

Table 3
Performance on Vaihingen test set for different ascending ratio α in CCA module.

Ratio α	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

Table 4
Effect of partition numbers G_h and G_w within region shuffle attention module.

Method	G_h	G_w	OA(%)	mIoU(%)
Baseline [20]	-	-	86.51	72.69
RSA	16	16	90.70	82.35
	16	8	90.75	82.44
	8	16	90.77	82.47
	8	**8**	**90.79**	**82.49**
	8	4	90.78	82.47
	4	8	86.76	82.46
	4	4	90.75	82.44

present the corresponding results in Tab. (4). Noted 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 $G_h = G_w = 8$ achieve the best 90.79% in overall accuracy 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 is 64 pixels in our experiments, thus eclectic choice of grouping is more conducive to self-attention 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 FCN 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.

4.4.3. *Comparison with Context Aggregation Approaches*

,

We compare the performance of several well verified context aggregation approaches, *i.e.*, Atrous Spatial Pyramid Pooling (ASPP) in DeepLabv3 [6], Pyramid Pooling Module (PPM) in PSPNet [46], RCCA in CCNet [16] and Self-Attention in non-local networks [38]. All the experiments above are conducted under the same training/testing settings for fairness. We report the related results in Tab. (5). Concretely, "+PPM" achieves better performance compared with "+ASPP" in terms of expanding local receptive fields. Both "+Self-Attention" and "+RCCA" generate contextual information

Table 5
Comparison with context aggregation approaches.

Method	OA(%)	mIoU(%)
Baseline [20]	86.51	72.69
+ ASPP (Our impl.) [6]	90.51	81.39
+ PPM (Our impl.) [46]	90.82	82.52
+ Self-Attention (Our impl.) [38]	90.62	82.17
+ RCCA (Our impl.) [16]	90.76	82.45
+ **Ours**	**90.98**	**82.87**

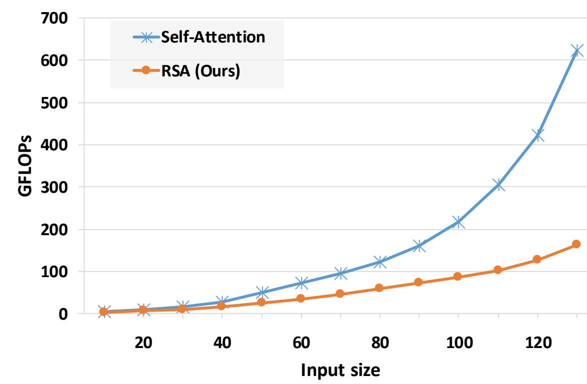


Figure 7: Comparison of numerical complexity. The x-axis represents the height and width of the input feature map and the y-axis represents the computation cost measured with GFLOPs.

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.

4.4.4. *Efficiency Comparison*

Comparison with Self-attention. As illustrated in Fig. 7. We first compare our RSA module with the standard self-attention mechanism in terms of the computation cost measured with GFLOPs. As the size of input feature map increases, the GFLOPs of self-attention mechanism gradually increases exponentially while the counterparts of our RSA module is almost linearly increasing. It can be seen that the RSA module is much more efficient than the self-attention mechanism when processing high-resolution feature maps.
Comparison with Context Aggregation modules and Attention modules. We further compare our proposed class augmented attention module and region shuffle attention module with ASPP [5, 6], PPM [46], SA [38], RCCA [16], OCR [41] and ISA [15] in terms of efficiency, including parameters, GPU memory and computation cost (GFLOPs). We report the results in Tab. (6). Notably, we evaluate the cost of all above methods without considering the cost of backbone

Table 6
Efficiency comparison with ASPP, PPM, Self-Attention, RCCA, OCR and ISA when processing input feature map of size $[1 \times 2048 \times 128 \times 128]$ during inference stage.

Method	Params(M▲)	Memory(MB▲)	GFLOPs(▲)
ASPP [6]	15.1	284	503
PPM [46]	22.0	792	619
SA [38]	10.5	2168	619
RCCA [16]	10.6	427	804
OCR [41]	10.5	202	354
ISA [15]	11.8	252	386
CAA(Ours)	9.3	283	148
RSA(Ours)	**3.8**	**110**	**144**

and include the cost of 3×3 convolution for dimension reduction to ensure the fairness of the comparison. As shown in Tab. (6), compared with standard Self-Attention (SA) mechanism, our RSA module requires 20× less GPU memory usage and significantly reduce FLOPs by about 77% with a few parameters, which proves the efficiency of region-wise representations in capturing long-range contextual information.

4.4.5. Comparison with State-of-the-art

We first adopt some common strategies to improve performance following [10, 42, 17]. (1) DA: Data augmentation with random scaling (from 0.5 to 2.0) and random left-right flipping. (2) Multi-Grid: We employ hierarchical grids of different sizes (1,2,4) within stage-4 of ResNet-101. (3) MS + Flip: We average the segmentation score maps from 5 image scales $\{0.5, 0.75, 1.0, 1.25, 1.5\}$ and left-right flipping counterparts during inference.

Experimental results are shown in Tab. (7). We successively adopt the above strategies to obtain better object representations, which achieves 0.19% ,0.11% and 0.16% improvements respectively in overall accuracy.

We further compare our method with existing methods on Vaihingen test set. Notably, most of the methods adopt the same backbone (ResNet-101) as ours. Results are shown in Tab. (8). It can be seen that our HMANet 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, it improves the second best CCNet by 0.93%, which demonstrates the effectiveness of capturing category-based information and global region-wise correlation.

4.4.6. Visualization Results

We provide qualitative comparisons between our HMANet and baseline network in Fig. 8, including 512 × 512 and 1024×1024 patches. In particular, we leverage the red dashed box to mark those challenging regions that are easily 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 region-wise representations.

Table 7
Performace comparison between data augmentation (DA), multi-grid (MG) and multi-scale with horizontal flipping (MS + Flip). We report the results on the test set of Vaihingen.

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**

4.5. Experiments on Potsdam Dataset

We carry out experiments on ISPRS Potsdam benchmark to further evaluate the effectiveness of HMANet. Empirically, we adopt the same training and testing settings on Potsdam dataset. Numerical comparisons with state-of-the-art methods are shown in Tab. (9). Remarkably, HMANet achieve 92.21% in overall accuracy 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. 9. 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).

5. Conclusion

In this paper, we propose a novel attention-based framework for dense prediction tasks in the field of remote sensing, namely Hybrid Multiple Attention Network (HMANet), which adaptively captures global contextual information from the perspective of space, channel and category. In particular, we introduce a class augmented attention module embedded with a class channel attention module to compute category-based correlation and further adaptively recalibrate the class-level information. Additionally, to address the feature redundancy and improve the efficiency of self-attention mechanism, a region shuffle attention module is presented to obtain robust region-wise representations. Extensive experiments on ISPRS Vaihingen and Potsdam benchmark demonstrate the effectiveness and efficiency of the proposed HMANet.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 61725105 and 41701508.

References

[1] Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140, 20–32.
[2] Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation.

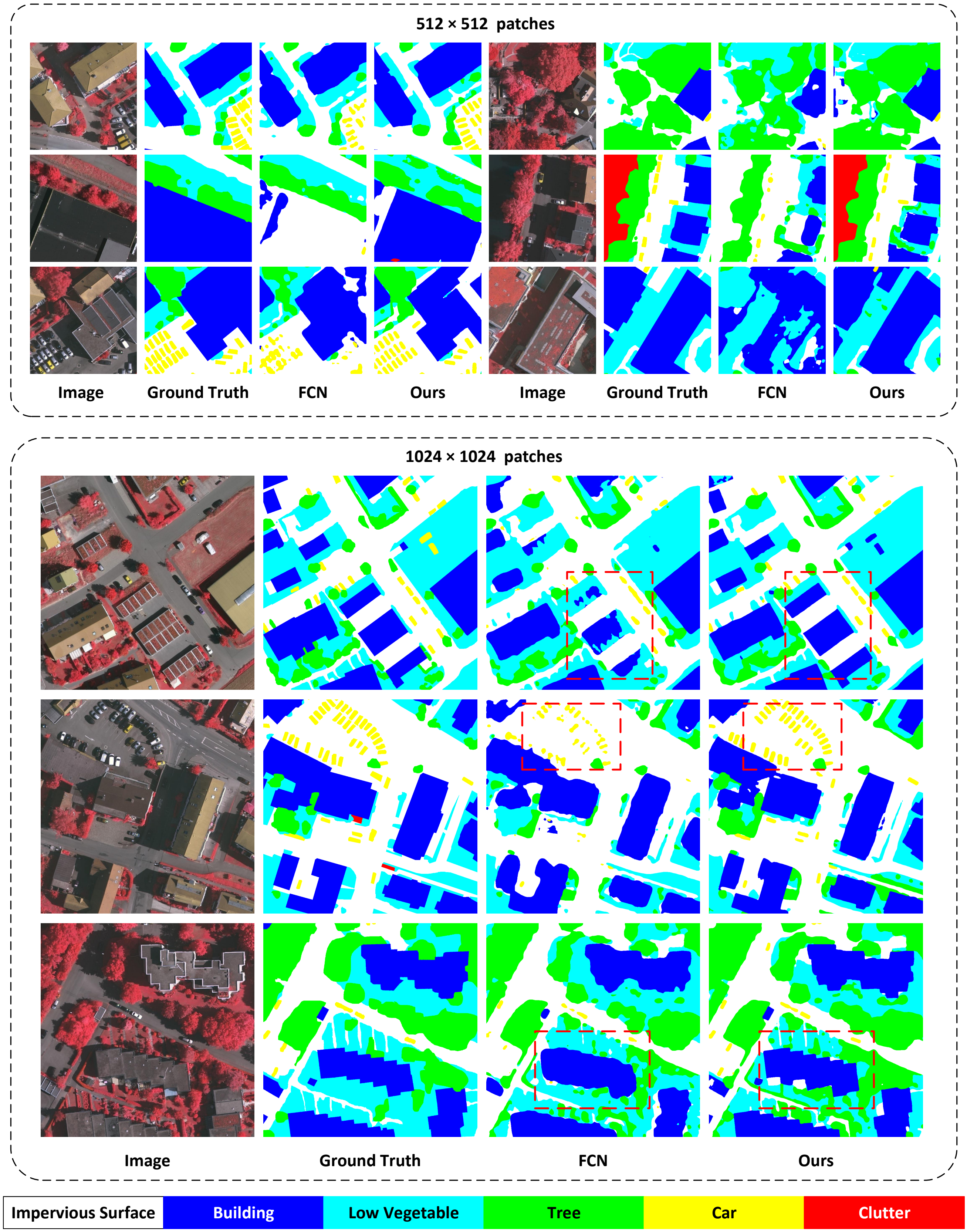


Figure 8: Qualitative comparisons between our method and baseline on Vaihingen test set.

Table 8
Comparisons with state-of-the-arts on Vaihingen test set.

Method	Imp. surf.	Building	Low veg.	Tree	Car	mean F_1	OA(%)	mIoU(%)
FCN [20]	88.67	92.83	76.32	86.67	74.21	83.74	86.51	72.69
UZ_1 [36]	89.20	92.50	81.60	86.90	57.30	81.50	87.30	-
RoteEqNet [23]	89.50	94.80	77.50	86.50	72.60	84.18	87.50	-
S-RA-FCN [26]	91.47	94.97	80.63	88.57	87.05	88.54	89.23	79.76
DANet [10]	91.63	95.02	83.25	88.87	87.16	89.19	89.85	80.53
V-FuseNet [1]	92.00	94.40	84.50	89.90	86.30	89.42	90.00	-
DLR_9 [24]	92.40	95.20	83.90	89.90	81.20	88.52	90.30	-
TreeUNet [43]	92.50	94.90	83.60	89.60	85.90	89.30	90.40	-
DeepLabV3+ [6]	92.38	95.17	84.29	89.52	86.47	89.57	90.56	81.47
PSPNet [46]	92.79	95.46	84.51	89.94	88.61	90.26	90.85	82.58
ACFNet [44]	92.93	95.27	84.46	90.05	88.64	90.27	90.90	82.68
BKHN11	92.90	**96.00**	84.60	89.90	88.60	90.40	91.00	-
CASIA2 [19]	93.20	**96.00**	84.70	89.90	86.70	90.10	91.10	-
CCNet [16]	93.29	95.53	85.06	90.34	88.70	90.58	91.11	82.76
HMANet (Ours)	**93.50**	95.86	**85.41**	**90.40**	**89.63**	**90.96**	**91.44**	**83.49**

Table 9
Numerical comparisons with state-of-the-arts on Potsdam test set.

Method	Imp. surf.	Building	Low veg.	Tree	Car	mean F_1	OA(%)	mIoU(%)
FCN [20]	88.61	93.29	83.29	79.83	93.02	87.61	85.59	78.34
UZ_1 [36]	89.30	95.40	81.80	80.50	86.50	86.70	85.80	-
S-RA-FCN [26]	91.33	94.70	86.81	83.47	94.52	90.17	88.59	82.38
DANet [10]	91.50	95.83	87.21	88.79	95.16	91.70	90.56	83.77
V-FuseNet [1]	92.70	96.30	87.30	88.50	95.40	92.04	90.60	-
Multi-filter CNN [34]	90.94	96.98	76.32	73.37	88.55	85.23	90.65	-
TreeUNet [43]	93.10	97.30	86.60	87.10	95.80	91.98	90.70	-
DeepLabV3+ [6]	92.95	95.88	87.62	88.15	96.02	92.12	90.88	84.32
CASIA3 [19]	93.40	96.80	87.60	88.30	96.10	92.44	91.00	-
PSPNet [46]	93.36	96.97	87.75	88.50	95.42	92.40	91.08	84.88
BKHN3	93.30	97.20	88.00	88.50	96.00	92.60	91.10	-
AMA_1	93.40	96.80	87.70	88.80	96.00	92.54	91.20	-
CCNet [16]	93.58	96.77	86.87	88.59	96.24	92.41	91.47	85.65
HUSTW4 [33]	93.60	**97.60**	88.50	88.80	94.60	92.62	91.60	-
SWJ_2	**94.40**	97.40	87.80	87.60	94.70	92.38	91.70	-
HMANet (Ours)	93.85	97.56	**88.65**	**89.12**	**96.84**	**93.20**	**92.21**	**87.28**

IEEE transactions on pattern analysis and machine intelligence 39, 2481–2495.

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

[4] Cao, Z., Fu, K., Lu, X., Diao, W., Sun, H., Yan, M., Yu, H., Sun, X., 2019. End-to-end dsm fusion networks for semantic segmentation in high-resolution aerial images. IEEE Geoscience and Remote Sensing Letters .

[5] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

[6] Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

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

[8] Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018b. Aˆ 2-nets: Double attention networks, in: Advances in Neural Information Processing Systems, pp. 352–361.

[9] Cheng, B., Chen, L.C., Wei, Y., Zhu, Y., Huang, Z., Xiong, J., Huang, T.S., Hwu, W.M., Shi, H., 2019. Spgnet: Semantic prediction guidance for scene parsing, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5218–5228.

[10] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.

[11] Gualtieri, J.A., Cromp, R.F., 1999. Support vector machines for hyperspectral remote sensing classification, in: 27th AIPR Workshop: Advances in Computer-Assisted Recognition, International Society for Optics and Photonics. pp. 221–232.

[12] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

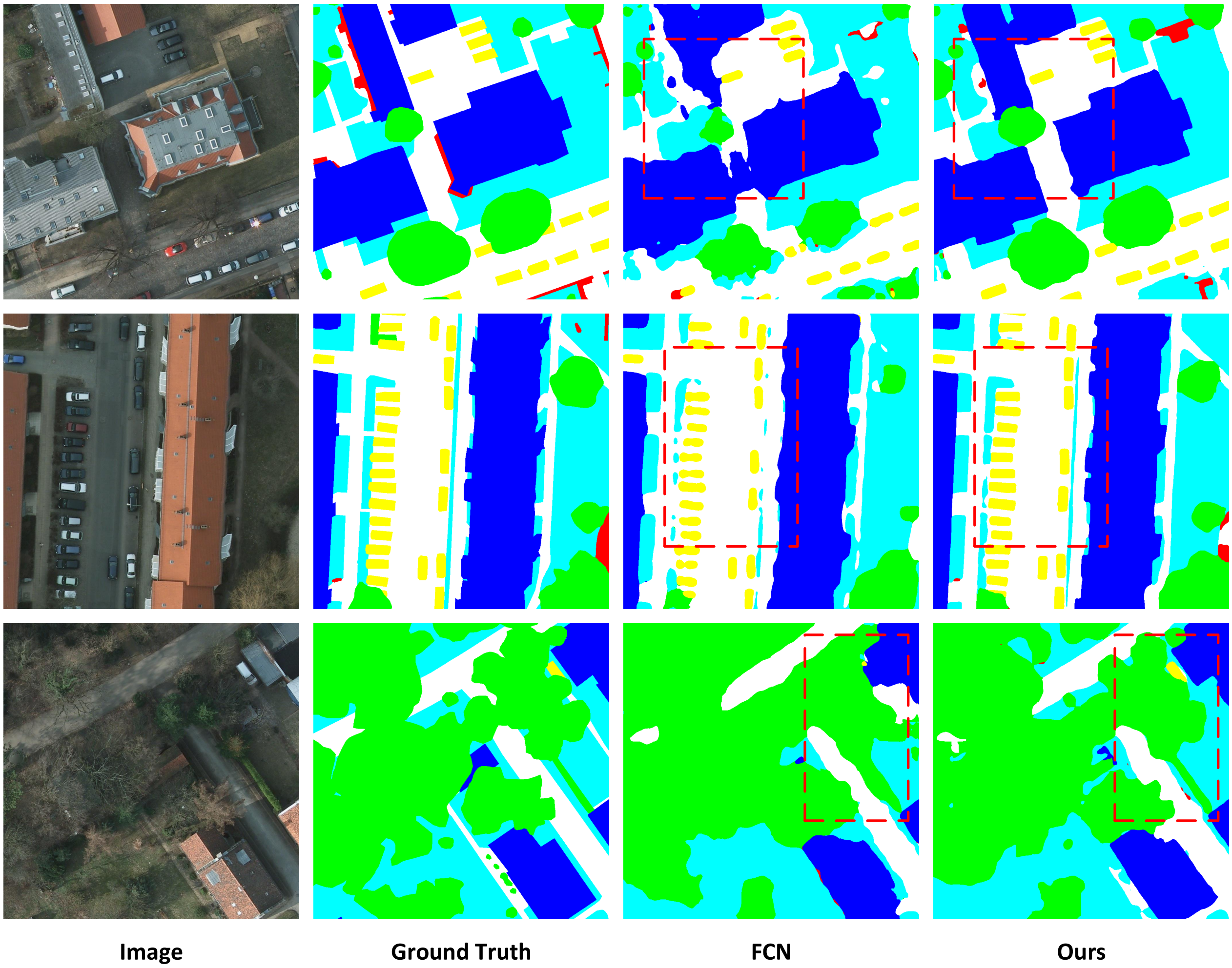


Figure 9: Visualization results of HMANet on Potsdam test set.

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

[14] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

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

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

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

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

[19] Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C., 2018. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS Journal of Photogrammetry and Remote Sensing 145, 78–95.

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

[21] Maboudi, M., Amini, J., Malihi, S., Hahn, M., 2018. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. ISPRS Journal of Photogrammetry and Remote Sensing 138, 151–163.

[22] Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. High-resolution aerial image labeling with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing 55, 7092–7103.

[23] Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. ISPRS journal of photogrammetry and remote sensing 145, 96–107.

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

[25] Mi, L., Chen, Z., 2020. Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 159, 140–152.

[26] Mou, L., Hua, Y., Zhu, X.X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12416–12425.

[27] Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26, 217–222.

[28] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4353–4361.

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

[30] Rota Bulò, S., Porzi, L., Kontschieder, P., 2018. In-place activated batchnorm for memory-optimized training of dnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5639–5647.

[31] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252.

[32] Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv:1606.02585 .

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

[34] Sun, Y., Zhang, X., Xin, Q., Huang, J., 2018. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and lidar data. ISPRS journal of photogrammetry and remote sensing 143, 3–14.

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

[36] Volpi, M., Tuia, D., 2016. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing 55, 881–893.

[37] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2019. Eca-net: Efficient channel attention for deep convolutional neural networks. arXiv preprint arXiv:1910.03151 .

[38] Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.

[39] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018a. Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341.

[40] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018b. Learning a discriminative feature network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1857–1866.

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

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

[43] Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., Li, W., 2019. Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 156, 1–13.

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

[45] Zhang, Q., Seto, K.C., 2011. Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime light data. Remote Sensing of Environment 115, 2320–2329.

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

[47] Zhong, P., Wang, R., 2007. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Transactions on Geoscience and Remote Sensing 45, 3978–3988.

[48] Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X., 2019. Asymmetric non-local neural networks for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 593–602.