IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2020, accepted January 21, 2020, date of publication January 27, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969812

Vegetation Land Use/Land Cover Extraction From High-Resolution Satellite Images Based on Adaptive Context Inference

ZONGQIAN ZHAN, XIAOMENG ZHANG, YI LIU, XIAO SUN, CHAO PANG, AND CHENBO ZHAO
School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

Corresponding author: Xiaomeng Zhang (xiaomengzhang@whu.edu.cn)

This work was supported in part by the National Key Technologies Research and Development Program of China under Grant 2016YFB0501403, and in part by the Chinese National Natural Science Foundation under Grant 61871295.

ABSTRACT In this paper, automatic extraction of multi-context and multi-scale land use/land cover vegetation from high-resolution remote sensing images is tackled, aiming to solve typical challenges in classifying remote sensing images at a pixel level. To solve small inter-class differences and large intra-class differences between the vegetation and background, we introduce a vegetation-feature-sensitive focus perception (FP) module. Considering the intrinsic properties of vegetation objects, we established an adaptive context inference (ACI) model under a supervised setting that includes a context model to represent relationships between a center pixel and its neighbors under semantic constraints, as well as the spatial structures of vegetation features. Comparative experiments on the ZY-3 and Gaofen Image Dataset (GID) datasets demonstrate the effectiveness of our proposed automatic vegetation extraction model against the baseline Deeplab v3+ model. Taking precision, kappa coefficient, mean intersection over union (miou), precision rate, and F1-score as the evaluation indexes, the results showed an improvement in the precision by at least 1.44% and miou by 2.47%, over the baseline Deeplab v3+ model. In addition, the ACI module improved the precision and miou by 2% and 3.88%, and the FP module improved the precision and miou by 1.13% and 1.65%. These results and statistics of these comprehensive experiments illustrated that our adaptive and effective vegetation extraction model could satisfy different requirements of land use/land cover mapping applications.

INDEX TERMS Context inference, focus perception, high-resolution remote sensing images, land use, land cover, image segmentation, vegetation mapping.

I. INTRODUCTION

A. MOTIVATION AND OBJECTIVE

Currently, heterogeneous high-resolution remote sensing images (HRRSI) acquired from different geographical areas promote the development of large-coverage and multi-temporal land use/land cover mapping. Particularly, the automatic extraction of vegetation using the HRRSI dataset, can be used to overcome problems caused by deteriorating environmental quality [1], [2], the loss of prime agricultural lands [3]–[5], the destruction of important wetlands, and so on [6]–[9]. However, classification of HRRSI is exposed to new challenges and potentials in different applications, such as urban planning, precision agriculture, and resource management.

We investigated vegetation land use/land cover mapping problems as well as their related challenges. Land cover commonly represents the physical properties of a land's surface, while land use corresponds to the activities or functions for which humans utilize land, which are inherently related, but are nevertheless conceptually distinct [10]. Therefore, images are always ambiguously classified base on land use criterion rather than uniquely classified base on land cover criterion during mapping. However, to satisfy requirements to infer the land properties for land use or land cover applications, it's indispensable to combine heuristic, empirical, or physically-based models integrated with ground-knowledge or user interpretation [11]. Diverse imaging conditions such as

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

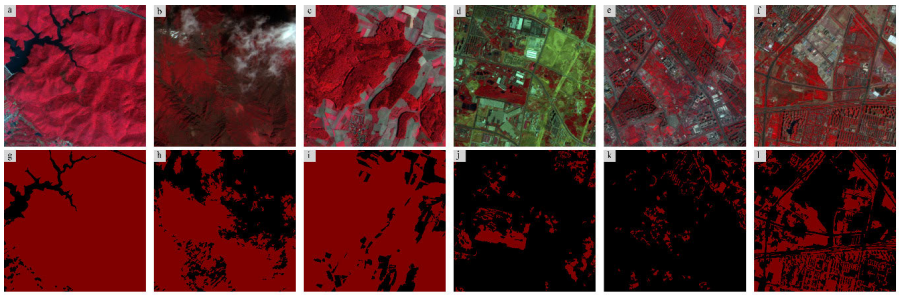


FIGURE 1. Illustration of the ZY-3 dataset. (a), (b), (c) and (d), (e), (f): vegetation examples in mountainous and urban areas, respectively; (g), (h), (i) and (j), (k), (l): binary labels in mountainous and urban areas, respectively.

photographic distortion, variations in scales, and changes of illuminations [9] affect HRRSI datasets. In the meantime, there exists diverse patterns of spectral reflection, temporal changes, and spatial distributions of vegetation objects represented by multiple spatial positions, sizes, shapes, and mutual relationships [8]. Thus, it is necessary to analyze the attributes of images before conducting vegetation extraction.

Based on shape, size, boundary, spatial context relationships, and seasonal changes in vegetation, we classify vegetation objects in images into mountainous areas, urban areas and plain areas and describe them according to their surroundings and location. Taking Figure 1 for example, forest and cultivated land with different levels of coverage in mountainous (Figure 1a, 1b, and 1c) and urban areas (Figure 1d, 1e and 1f), exhibits complex and confusing spectral, shape, and texture attributes. Forest and cultivated lands in mountainous areas have irregular geometric shapes with a large coverage, with different spectral reflectance in shady and sunny parts. Cultivated land in urban areas, however, is relatively fragmented with imprecise boundaries. Figure 2 shows cultivated land in plain areas containing regular geometric shapes with smooth and delicate textures, including long stripped fields, small roads, and canals without typical vegetation features. In Figure 2 vegetation (e.g. farmland, forest, meadows) and non-vegetation (e.g. eutrophic waters) objects possess similar spectral responses in the same image, which are hardly distinguishable. The spectral hue value of vegetation varies with soil, humidity, crop type and changes seasonally, although textures and shapes in the multi-temporal images almost never change. Therefore, intra-class differences sometimes may be large and inter-class differences may be small. This phenomenon together with intrinsically complicated spatial distribution patterns of vegetation objects make vegetation extraction challenging [12].

To solve the problems evident in the Figure 1 and Figure 2 examples, Dusseux *et al.* [13] used a time series of HRRSI to precisely identify land cover and land use classes at the field scale for inter-annual and intra-annual grassland monitoring in agricultural areas. This will also be useful in change detection applications [14]. Furthermore, data fusion based on temporal series from Synthetic aperture radar (SAR) and optical images are widely used [15]–[19]. Optical images

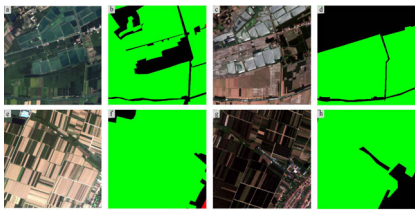


FIGURE 2. Illustration of the Gaofen Image Dataset (GID) [8]. (a), (e) and (c), (g): lake and farmland areas around Wuhan, Hubei Province, acquired on September 2, 2015 and June 26, 2016, respectively; (b), (f) and (d), (h): corresponding multi-class labels (The green is farmland, the red is building, and the black is unknown areas).

capture chemical, physical, and biological attributes of objects, while radar images reflect the shape, textural, structural, and dielectric properties. Consequently, methods using information from different sources will increase the discriminability of land cover/land use objects.

Besides, improvements in the spatial resolution simplify the mixed pixel problem inevitable in standard multi-spectral image processing. However, it also makes the problem of high intra-class and low inter-class variability more serious, which in turn involves a high level of classification errors. Therefore, a tradeoff between the spatial and spectral resolution should be considered [20], [21]. Methods integrating spectral and spatial information have already been verified as useful [22]–[24]. We inferred that integrating spectral and spatial context information might more accurately distinguish between various types of vegetation and other land use/land cover features.

Hence, HRRSI classification methods became a hot research direction. Ma *et al.* [25] developed a detailed review of supervised methods for land cover mapping approaches in land use and land cover classification. And deep learning methods are also used for classification of remote sensing images [26]–[28]. Zhu *et al.* [29] developed a detailed review of deep convolutional neural networks (CNNs). As we know, Traditional and deep learning-based methods have revealed their widespread usage in land use classification.

Therefore, we introduce several typical deep learning methods aimed at solving semantic segmentation task for public benchmark datasets. Up to now, several benchmarks

datasets for evaluating the land use/land cover mapping methods have been proposed [8], [30], [31], but these still cannot fully satisfy the requirements of practical land use/land cover applications. To better solve the HRRSI vegetation extraction application, we reviewed HRRSI classification methods based on deep learning tools below, which can provide some insights for constructing our vegetation extraction model.

To make better use of contextual relationship of HRRSI, there are many researches on public datasets. For ISPRS datasets [31], Marmanis *et al.* [32] set up two separate paths for intensity and range data with the same layer architecture, verified the effectiveness of integration of several networks; Volpi and Tuia [26] presented a CNN-based system relying on a downsample-then-upsample architecture; Chen *et al.* [33] proposed SNFCN and SDFCN frameworks with dense-shortcut connection structures; Zhang *et al.* [34] proposed dual dilated and non-dilated networks using multi-label manifold ranking (MR) method and embedded single stream optimization method, which illustrated the effectiveness of spatial context modeling and position information in HRRSI classification. For DeepGlobe datasets [30], Sun *et al.* [35] proposed a stacked U-Nets with multiple output and a hybrid loss function to address the problem of unbalanced classes of training data for road extraction from satellite imagery. Ghosh *et al.* [36] presented a dilated stacked U-Nets. However, improvement can be realized if more prior information such as orientation and texture might be considered.

Particularly in vegetation classification, since vegetation objects have relatively higher near infrared (NIR) reflectance and low visible reflectance, a vegetation extraction technique using conventional red and NIR bands to calculate the Normalized Difference Vegetation Index (NDVI) index has been well-established and applied for vegetation monitoring [37]. Furthermore, remote sensing of vegetation is realized using passive sensors to obtain electromagnetic wave reflectance information from canopies [38].

Apart from NDVI index, Li *et al.* [39] proposed a temporal-attention CNN-GRU approach to differentiate subtle phenological differences between crops. Zhong *et al.* [40] developed DNNs to classify summer crops using EVI time series images. Sidike *et al.* [41] proposed a deep progressively expanding neural network for mapping different types of vegetation objects including various crops, weeds, and crop residues. Farooq *et al.* [42] used a Convolutional Neural Network (CNN) to learn middle and high level spatial features for weed classification. Zhang and Verma [43] presented an adaptive texton clustering model and ANN classifiers for segmenting vegetation from real-world roadside image scenes. Chen *et al.* [44] proposed an improved CNN to extract fine spatial distribution information. There are few vegetation context-reasoning models for fully automatic vegetation extraction pipelines; particularly cases in which only several bands of information are available. In addition, integrating attention mechanisms with high-level and low-level semantic information from features maps for vegetation extraction remains to be studied further.

In our research, we integrated adaptive context inference (ACI) and focus perception (FP) modules into a semantic segmentation framework to automatically and adaptively extract vegetation in HRRSI, based on the various requirements of land use/land cover vegetation mapping applications.

B. CONTRIBUTIONS

To make full use of limited information to solve the vegetation land use/land cover problem, we extracted the agricultural forestry land and urban green space from HRRSI, using our proposed vegetation land use/land cover extraction method. We also conducted comprehensive experiments on ZY-3 and GF-2 HRRSI datasets to validate the effectiveness of our proposed approach. In summary, the main contributions of this paper are as follows:

(1) We introduced a FP module to extract sensitive features from different types of vegetation and an integrated attention mechanism containing high-level and low-level semantic information to solve the problem of small inter-class and large intra-class differences;
(2) We established an ACI model under a supervised setting to satisfy the inference of spatial structure relations, based on the data-driven pattern recognition methodology;
(3) We conducted comparative experiments to analyze the influences of different modules on vegetation extraction results, as well as an ablation study and parameter sensitivity analysis on ACI module, which finally gave some conclusions about different modules' influences.

The remainder of the paper is organized as follows. The existing work is described in Section 2, including the development of semantic segmentation, structural reasoning, and the attention mechanism. The baseline model Deeplab v3+ [45], the ACI module, and the FP module of our proposed method are described in Section 3. The experimental setup, dataset description, and our evaluation metrics are presented in Section 4. The vegetation extraction results and experimental performance metrics for full-image and difficult local areas on ZY-3 and GF-2 HRRSI are illustrated in Section 5 and Section 6. Section 7 concludes the study.

II. RELATED WORK

A. DEVELOPMENT OF SEMANTIC SEGMENTATION

Fully convolutional neural network-based approaches have made remarkable progress in semantic segmentation tasks. Long *et al.* [46] conducted a deconvolution operation and pre-defined parameters to fuse feature maps from different intermediate layers for final prediction refinement, but this could not simultaneously aggregate local and contextual information in convolutional feature maps. Subsequently, Noh *et al.* [47] learned the parameters of deconvolution layers

from training data in an end-to-end system. Wang *et al.* [48] proposed dense up-sampling convolution (DUC) to directly predict the full-resolution probability map by convolution.

To alleviate the low resolution of feature maps, Yu and Koltun [49] conducted dilated convolution to increase the resolution of CNN feature maps and improve semantic segmentation accuracy. Yu *et al.* [50] identified the gridding problems of dilated convolution and proposed hybrid dilated convolution (HDC) to remove abnormal artifacts. PSPNet [51] and DeepLab system [52], [53] perform spatial pyramid pooling at different gird scales or dilation rates (called Atrous Spatial Pyramid Pooling, or ASPP). However, Chen *et al.* [54] pointed out that the pyramid pooling module in PSPNet might lose pixel-level localization information. Moreover, dilated convolution calculated sparsely in an ASPP module might cause grid artifacts. The existence of objects at multiple scales and the resolution reduction of a feature map remain to be further investigated.

B. STRUCTURAL REASONING

To incorporate structural reasoning into semantic segmentation, conditional random field (CRF) methods have been proposed that consider image segmentation predictions among the highly correlated pixels [54]–[57]. To introduce prior information into semantic segmentation, Chen *et al.* [54] used a CRF to model the compatibility between the predicted labels. Liu *et al.* [58] pointed out that a simple global average pooling operation significantly improves the accuracy. Pinheiro *et al.* [59] proposed DeepMask to utilize global information through a fully connected layer. Zhao *et al.* [51] combined separable convolution operation using a pyramid pooling module in PSPNet to approximate large convolution kernels with an enlarged receptive field. Liu *et al.* [60] incorporated high-order relations and a mixture of label contexts into a Markov Random Field, this proposed ParseNet delivers accurate performance results, which can represent various types of pair-wise functions. Zheng *et al.* [57] integrated the desirable properties of both CNNs and CRFs, and proposed CRF-RNN to express a CRF as a recurrent neural network (RNN) and plug it into a deep convolutional neural network (CNN) as an end-to-end system. Marvin and Cipolla [61] proposed reformulating the inference model in terms of convolutions, which solved the slow training and inference speeds of CRFs, as well as the difficulty of learning the internal CRF parameters. In this way, all the parameters of a convolutional CRF could be optimized through back propagation, although a more sophisticated CRF architecture might capture additional global context information. Deeplab v3 + encodes multi-scale contextual information at multiple rates and effective field-of-views in the spatial pyramid pooling module. and also captures sharper object boundaries by gradually recovering spatial information through the encoder-decoder structure [45]. Effective spatial context inference models remain to be studied further.

C. ATTENTION MECHANISM

An attention mechanism captures visual feature dependencies in the spatial and channel dimensions. Specifically, approaches adopting an attention mechanism usually consist of two stages: descriptor extraction and feature aggregation. For a given size of image, the descriptor extraction stage extracts convolution feature maps expressed as C-dimension and HW-size descriptors, each of which captures local details, but lacks a global view. The feature aggregation stage often uses different pooling strategies to aggregate local and contextual information to generate another set of descriptors. Finally, each descriptor contains global contextual information, as well as local details.

Recently, many attention models have been proposed for various tasks. Squeeze-and-excitation networks model channel-wise relationships to enhance the representational power of a neural network [62]. Non-local neural networks have been employed to calculate a correlation matrix between each spatial point-pair in a feature map to guide dense contextual information aggregation [63]. A self-attention mechanism was first introduced in DANet as a position and channel attention module to model spatial and channel interdependencies and the rich contextual dependencies among local features, which significantly improved segmentation results [64]. Therefore, to tackle the small inter-class differences and large intra-class differences caused by the deformation of object features and interference with noise, we can strengthen the distinguishability of object and background features, as well as the similarity within the same category.

III. METHODOLOGY

As shown in Figure 3, the proposed approach consists of two novel components, a FP module to extract sensitive features from different types of vegetation and an ACI module to refine the boundary location of extracted segments in the framework of a fully convolutional neural network. The upper part of our framework shows the integration of the Deeplab v3+ model and FP module. The bottom half of the figure illustrates how the ACI module is embedded into the semantic segmentation framework. Furthermore, the original baseline Deeplab v3+ model is introduced in Section 3, part A, details of the FP module and ACI module are illustrated in Section 3, part B and Section 3, part C.

A. DEEPLAB v3+ MODEL

Deeplab v3+ [45] is one of the recent state-of-the-art approaches based on the feature net (e.g., ResNet101 [65]). We refer to its architecture and set the output resolution of the feature net ''Res4'' in Figure 3 as one in sixteen of the original image size. Since each descriptor in ''Res4'' lacks contextual information, it applies the ASPP module on ''Res4'' and gets new feature maps f_{ASPP}. To incorporate global information, it also applies global average pooling, 1×1 convolution, and bilinearly up-sampling on ''Res4'' in sequence to get an image-level feature.

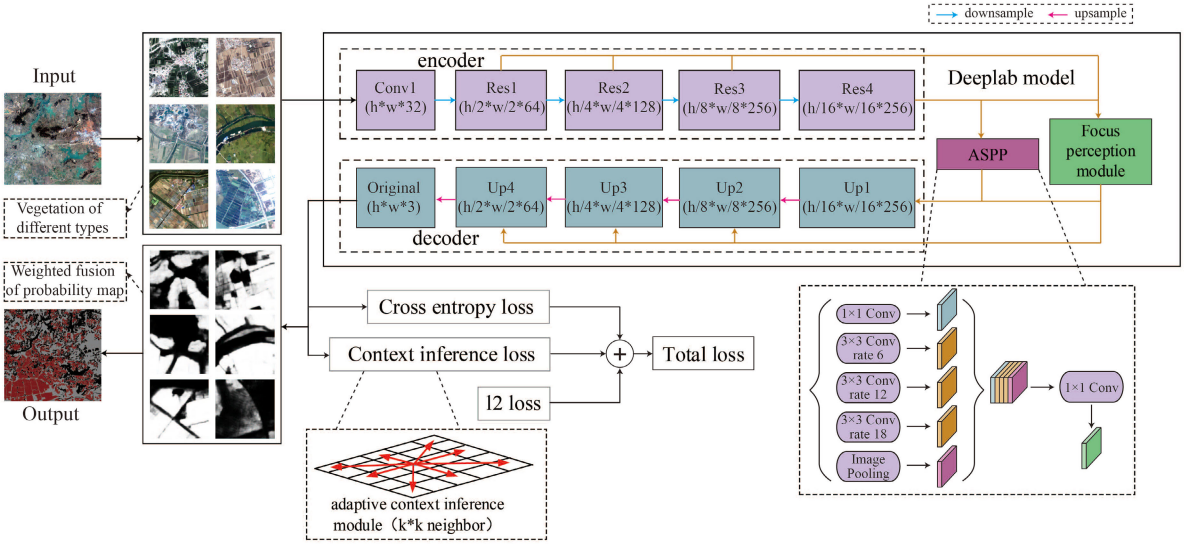


FIGURE 3. Flowchart of the ACI and FP vegetation land use/ land cover extraction method.

In this paper, to model the relationships between the center pixel and its neighbors under semantic constraints as well as different spatial structures of vegetation features, we considered the intrinsic property of vegetation objects and established an ACI model, integrating with the ASPP module and the global average pooling into an encoder-decoder framework, shown in Figure 3.

B. FOCUS PERCEPTION

Taking the Resnet [65] as the backbone, our vegetation extraction model is designed according to the configuration in Deeplab v3+ [45], which consists of an encoder part, decoder part, ASPP part, and global image-level feature extraction. The ASPP module used in the Deeplab v3+ [45] extracts different scales of feature information to incorporate neighbor scales of context features, but cannot select features channel-wise, as in SENet [62] and EncNet [66]. The original pixel-based scene context used to encode high-dimensional representations may suffer from spatial resolution loss. Therefore, we adopted the attention mechanism and spatial pyramid to extract precise dense features for pixel labeling, and the whole flowchart of our FP module is illustrated in Figure 4.

Since target scales of extracted vegetation are diverse and fuzzy, we can excavate vegetation-sensitive attributes of different layers of neural networks. Through the feature pyramid attention (FPA) module on top of the high-level output of fully convolutional neural network architecture, we embedded different levels of context information, and combined the global attention up-sampling module as the decoder module of the segmentation model.

In detail, we applied channel reduction (CR) on low-level image features (green solid lines) and attention-weighted multiplication followed by global average pooling (GA) on

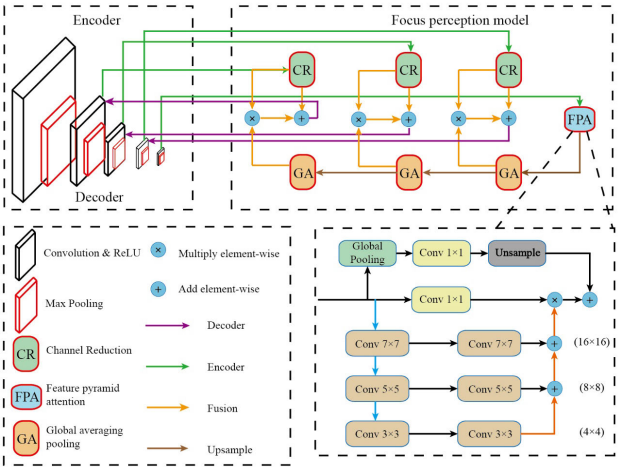


FIGURE 4. Illustration of the FP module.

the high image feature (brown solid lines), and then concatenated these global contexts and local contexts (range solid lines) to learn the vegetation-sensitivity features (purple solid line). We performed 3*3 convolution on the low-level features to reduce the channels of feature maps from CNN (CR) and performed 1*1 convolution with batch normalization and ReLU non-linearity on the high-level features (GA), and then multiplied the global context with the low-level features to produce the weighted low-level features, which were added with a gradually up-sampled global context. To avoid too high computational burdens, we conducted a global perception deconvolution operation on each decoder layer to pay attention to vegetation-sensitivity features at a pixel level after extracting high-level features from CNNs. Then, the extracted global context of high-level features can serve as guidance for low-level features to weight and select category localization details.

C. CONTEXT ADAPTIVE INFERENCE

When modeling the contextual information between the actual land cover types, boundary confusion exists between the vegetation feature and other features, so it is hard to obtain accurate boundary location information. In the meanwhile, the spatial distribution of vegetation is variable in different images, as shown in the Figures 1 and 2. To solve this problem, we model relationships between neighboring pixels in the label space, and operate context inference during the training stage. An ACI loss function is introduced to automatically obtain segments respectful of spatial structures and small details, inspired by the adaptive affinity field (AAF) loss function [67]. The total loss function consists of the unary cross-entropy loss (introduced by Deeplab v3+ model) and the ACI loss, which are denoted by L_{unary} and L_{ACI}, respectively [67].

$$S^* = \arg\min_S \max_w L_{unary} + L_{ACI} \tag{1.1}$$

Firstly, to model the spatial structures with diversity in size, shape, and context, we constructed a context model to represent relationships between the center pixel and its neighbors. Consequently, the ACI loss depend on the weighted summation of the KL (Kullback-Leibler) divergence of both the intra-class and inter-class part [67]:

$$\begin{aligned} L_{AAF} &= \sum_c \sum_k (w_{\overline{bck}} L^{\overline{bck}}_{affinity} + w_{bck} L^{bck}_{affinity}) \\ s.t. \sum_k w_{\overline{bck}} &= \sum_k w_{bck} = 1 \; and \; w_{\overline{bck}}, w_{bck} \geq 0 \end{aligned} \tag{1.2}$$

$$L^{ic}_{affinity} = \left\{ \begin{array}{c} L^{i\overline{bc}}_{affinity} = D_{KL}(\widehat{y}_j(c) || \widehat{y}_i(c)) \\ if \; y_i(c) = y_j(c) \\ L^{ibc}_{affinity} = \max\{0, m - D_{KL}(\widehat{y}_j(c) || \widehat{y}_i(c))\} \\ otherwise \end{array} \right\} \tag{1.3}$$

However, for the intra-class and inter-class part, we adopted inverse strategies. We expected that the intra-class part would be as small as possible, while the inter-class part was as large as possible. $L^{\overline{bck}}_{affinity}$ and $L^{bck}_{affinity}$ have been defined by [67], where $w_{\overline{bck}}$ and w_{bck} denote weight coefficient for the non-boundary and boundary part, and m in (1.3) is the threshold of KL divergence.

Secondly, to solve the intrinsic property of vegetation objects whose spatial structure cannot be exhaustively expressed in fixed patterns, we established an ACI model under a supervised setting to make the segmentation network adapt to the various sizes of different objects and satisfy the inference of spatial structure relations based on a data-driven pattern recognition methodology. In Figure 5, squares filled with white or yellow stars, round rectangles, or squares are neighboring pixels of the yellow or white hexagon center pixel, with the size of 1, 2, and 3, respectively, each of which contains eight corners. Since there are infinite relations between the center pixel and its neighbors, we cannot list all the patterns to provide a fixed representation, so it is indispensable for our context model to adaptively implement structure reasoning during the training stage.

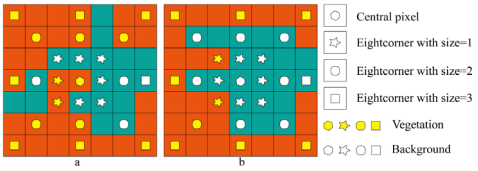


FIGURE 5. Illustration of reasoning process of ACI module with different values of neighborhood size.

In Figure 5, we illustrate two cases of ACI for the vegetation and background:

(1) The center pixel of the left one is yellow, and the majority of neighborhood pixels of the first, second, and third circle are the background (6 vs. 2), vegetation (5 vs. 3), and vegetation (7 vs.1), so a vegetation object will be misclassified as background if we only use an affinity field of size 1. On the contrary, we can strengthen the integrity of the vegetation object if we use an affinity field of size 1, 2, and 3 and adaptively implement context reasoning;

(2) As for the right one, the center background pixel is denoted as a white hexagon and its majority neighborhood pixels of the first, second, and third circle are background (6 vs. 2), background (7 vs. 1), and vegetation (7 vs.1), so if we only choose size 3, we may misclassify this object as vegetation.

Since updating the parameters of the network depends on the back propagation of gradients, which can be calculated by derivatives, we conducted a detailed analysis of the KL divergence associated with the loss function and derivation of differential formulas of the ACI loss function in terms of variables p and q, which stand for the prediction and the ground truth label, respectively.

We derived the first derivative of KL divergence as in Equation (1.4), and obtained Equation (1.5) and (1.6), in which p and q stand for the distribution of two independent variables.

$$\begin{aligned} D_{KL}(P||Q) &= p \log \frac{p}{q} + \overline{p} \log \frac{\overline{p}}{\overline{q}} \\ &= p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)} p, \quad q \in (0,1) \end{aligned} \tag{1.4}$$

$$\frac{\partial f(p,q)}{\partial p} = \log(\frac{p}{q} \cdot \frac{(1-q)}{(1-p)}) \tag{1.5}$$

$$\frac{\partial f(p,q)}{\partial q} = -\frac{p}{q} + \frac{1-p}{1-q} \tag{1.6}$$

Through analyzing Equation (1.5) and (1.6), when p and q are dissimilar, we can draw the conclusions presented below:

(1) If $p > q$, then $\frac{\partial f(p,q)}{\partial p} > 0$, $\frac{\partial f(p,q)}{\partial q} < 0$, and we can increase the p and decrease the q to make the difference between p and q more obvious, leading to the KL divergence being larger;

(2) If $p < q$, then $\frac{\partial f(p,q)}{\partial p} < 0$, $\frac{\partial f(p,q)}{\partial q} > 0$, and vice versa.

Based on these proofs, we can verify the effectiveness of KL divergence's influence on the distinguishing of the two distributions.

Furthermore, we can use the mathematical principle to verify the influence of the ACI loss function [67] on the boundary location accuracy's improvement. We conducted comprehensive experiments using different hyper parameters of the loss function, to verify and analyze the parameters' effects on the vegetation extraction performance.

IV. EXPERIMENTAL SETUP AND DATASET DESCRIPTION

A. IMPLEMENTATION DETAILS

Due to the limited GPU memory and lack of finely-annotated labels, we decided to trade-off the receptive field size in favor of a larger batch size. We used the poly learning rate policy and set the base learning rate as 0.001. We used the Nadam Optimizer (Adam with Nesterov momentum) and trained the network for 30,000 iterations, with a learning rate of 2.5×10^{-4}. To speed up the experiments for the validation of ZY-3 datasets and GID datasets, we downsized the crop-size to 256*256 (512*512) and batch-size to 8, so that a single NVIDIA Quadro P5000 GPU and 16s Inter(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHZ were sufficient for training. The training iterations for all experiments can be further improved by increasing the iteration number. Momentum and weight decay were set to 0.9 and 0.0005, respectively.

For data augmentation, we adopted random mirroring and random resizing between 0.5 and 2 for all datasets. We did not upscale the logits (prediction maps) back to the input image resolution; instead, we followed [45]'s approach by down-sampling the ground-truth labels for training (in this code, output_stride = 8). Since we aimed to obtain the different modules' impact on vegetation extraction, for inference, we did not average the scores from left-right flipped and multi-scale inputs (scales = 0.5, 0.75, 1, 1.25, 1.5, 1.75), but set the down-sample ratio before feeding it into the ACI module as 1/8, as [45] did.

B. DATASET DESCRIPTION

By comparing the features with different attributes of vegetation species in the images (ZY-3 and GF-2), we found that the second-level vegetation classes were impossible to distinguish accurately and vegetation was only classified into agricultural and forestry land and urban green space. Agriculture and forestry land include cultivated land, garden land, woodland, and grassland, whereas urban green space includes artificial green space, nurseries, flower gardens, and ribbon green trees, without further division.

1) ZY-3 DATASET

Experiment 1 used the ZY-3 multi-spectral images from abroad and areas, including green, red, and near-infrared bands, with a ground resolution of 5.8 meters. In detail, the overseas images cover the northern part of Laos, the Fengsali Province, and the German-Heidelberg area, and contain various representative land use categories of vegetation, buildings, roads, bare soil, etc. The images are evenly distributed, and the complexity of land cover types is lower than that of the territory images. Images cover the northern and southern areas of Qingdao, Changsha, Bohai, Harbin, Hefei, etc., and contain land use categories of vegetation, buildings, roads, and water bodies. Additionally, the images have abundant spectral information of different land cover types and obvious spectral information on vegetation.

We automatically annotated several ZY-3 remote sensing images using different colors (red for vegetation and black for background). We created vegetation datasets in which the vegetation labels covered farmland, forest, and meadows, as well as other sub-classes of vegetation defined in the Geographical National Survey Standard [12]. The background type covered build-up, waters, roads, and other land cover types without obvious vegetation features in images. Finally, we made binary datasets as follows: 27108 image patches of the size 256*256 were employed for training and 6778 for validation. An example of the dataset is shown in Figure 1.

2) GID DATASET

In this paper, we use partial GF-2 satellite images from the released GID dataset [8] to train an automatic vegetation extraction CNN model, only including red, blue and green bands (We did not use the infrared band as ZY-3 did due to the lack of available band data). In the GID dataset, five representative land use categories are annotated: built-up, farmland, forest, meadow, and waters. These land use categories are labeled with five different colors: red, green, blue, cyan, yellow, and blue, respectively. Areas that do not belong to the above five categories or cannot be artificially recognized are labeled as unknown, and are represented using a black color. We used the multispectral images following the same strategy as [8], ignoring the unknown area and only computing the accuracy of the deterministic area.

We created a vegetation binary label dataset for 129 GF-2 images, in which vegetation types covered farmland, forest, and meadow and background types covered built-up and waters. Considering the difference in the spatial resolution between ZY-3 (5.8m) and GF-2 (1m), we adopted larger image patch for GF-2 and smaller image patch for ZY-3, to make the real land-cover area of each image patch consistent. In total, we chopped 21,672 image patches of the size 512*512 without overlapping areas for training and 5418 for validation. An example of the dataset is shown in Figure 6.

Considering the inconsistence in bands used to train the vegetation extraction model, the model trained for ZY-3 is not available on the GID dataset, and vice versa.

Because the image size of HRRSI is usually 3–4 times the size of indoor/outdoor images, the HRRSI inevitably have to be cropped into little patches before feeding them into the network. Besides, we adopted an efficient inference pipeline to extract vegetation and significantly relieve the boundary effect caused by the cropping and sticking process. Here,

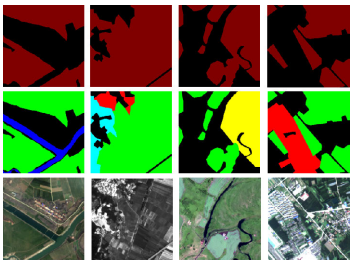


FIGURE 6. Illustration of GID training and validation dataset. The first, second, and third row illustrates the binary label, multi-class label, and original images, respectively.

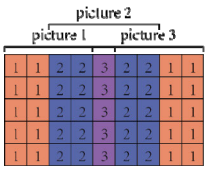


FIGURE 7. Illustration of the voting strategy used for sticking little patches into a full image.

we present an example to illustrate the weighted average strategy's efficiency in eliminating the stitching seam caused by stitching images. In Figure 7, we illustrate three adjacent probability maps' weighted average sticking process, where the numbers in squares represent the statistical overlapped times. Instead of the sticking strategy where the back predicted label covers the previous predicted label, we obtained the predicted probability map of each little image patch, and conducted a voting strategy which used the weighted average probability result as the final prediction. The weight matrix was composed of the statistical times standing for overlapped levels in order to eliminate the boundary effects caused by the disagreement of adjacent predictions.

C. EVALUATION METRICS

We assessed the experimental results with the Kappa coefficient (Kappa), overall accuracy (OA), and class-specific accuracy. The accuracy rate indicates the number of true positive samples in the samples that are predicted to be positive, which is also known as the precision rate.

$$precision\ rate = {}^{TP}/_{(TP+FP)} \tag{1.7}$$

The recall rate indicates how many positive examples in the samples are predicted correctly.

$$recall\ rate = {}^{TP}/_{(TP+FN)} \tag{1.8}$$

F-Measure is the weighted harmonic average of the precision and recall rate. In our research we use F1-score, which is derived from F-measure.

$$F1 = F_{\alpha^2=1} = \frac{(\alpha^2+1)P*R}{\alpha^2(P+R)}\ {}_{\alpha^2=1} = \frac{2PR}{(P+R)} \tag{1.9}$$

V. EXPERIMENTAL ANALYSIS OF THE ZY-3 DATASET AND ABLATION STUDY

In this part all experiments are based on original ZY-3 test images. To verify the effectiveness of methods integrated with the ACI module under different parameter configurations and the FP module, we showed the results of baseline and different methods in Figure 8. As for the ACI module, we set comparative experimental groups to investigate the influences of the ACI module on vegetation extraction results. Configurations of these groups are shown in Table 1. In this table, "Margin value" means the margin used to calculate the boundary loss value between pixels with different categories, denoted as m (in equation 1.3) and set as 3, 2, 1, or 0.5; "learning rate ratio" means the ratio between parameters' learning rate from traditional network and the ACI module, set as 1:1 or 1:0.1; "kernel size" means the size of the ACI module, set as "s3" and "s357". "s3" means radius distance's values of 1 (equal to kernel size 3) and "s357" means we averaged ACI losses of radius distance values of 1, 2, and 3 (equal to kernel size 3, 5 and 7).

TABLE 1. Definition of methods with or without an ACI module and FP module (in accordance with Figure 8, 9 and 10).

method	Adding ACI module			Adding FP module
	Margin value	kernel size	learning rate ratio	
a	2.0	3,5,7	1:1	no
b	2.0	3	1:1	no
c	2.0	3,5,7	1:0.1	no
d	0.5	3	1:1	no
e	1.0	3	1:1	no
f	3.0	3,5,7	1:1	no
g	no	no	no	no
h	no	no	no	yes
i	2.0	3,5,7	1:1	yes

To further analyze the detailed qualitative vegetation extraction of two challenging image patches extracted from the full test image, we illustrate their results in Figures 9 and 10, and reported statistical evaluation indexes of methods in Table 3 and 4, whose arranged sequence is also explained in Table 1.

For land use and land cover classification, land use areas are always more complete and based on a global perspective, while land cover areas are often incomplete and based on local details. Because land use is mainly based on human's utilization, and land cover corresponds to physical properties, which has been mentioned in Section 1.

In this section, we split the detailed analysis into three parts. We analyzed the effectiveness of the ACI module under different parameter configurations in part A and FP module in part B. In part C, we analyzed the interaction between these two modules.

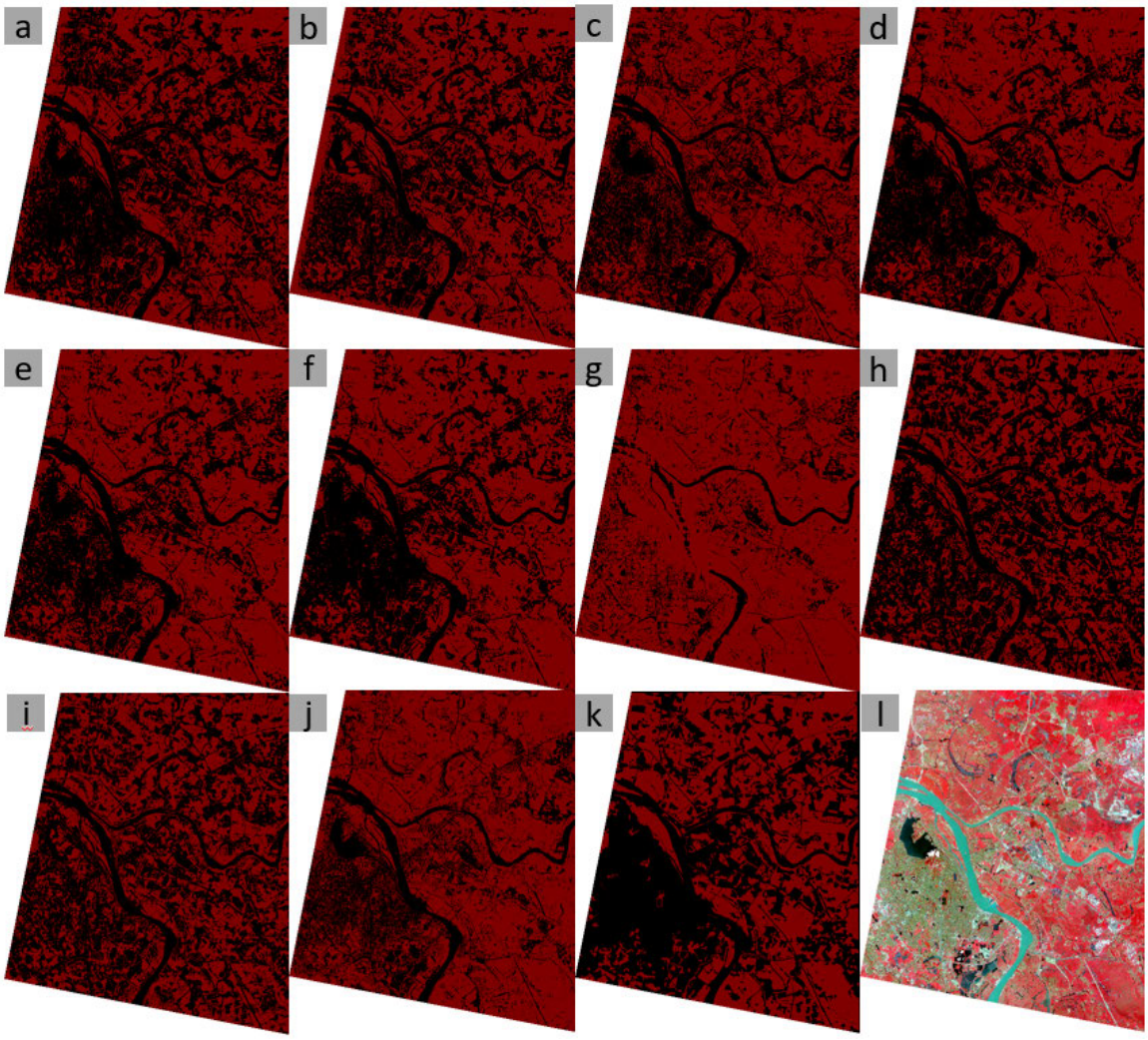


FIGURE 8. Complete vegetation extraction results of different methods based on the ZY-3 dataset (Yue Nan): (a) - (i) are results from methods described in Table 1; (j) parallelepiped classification method [68]; (k) ground truth label; (l) original image.

A. PARAMETER ANALYSIS OF THE ADAPTIVE CONTEXT INFERENCE MODULE

From a global perspective, though the baseline's overall performance in Figure 8(g) is relatively acceptable, in many cases, it may misclassify buildings and roads in urban areas into vegetation, as well as increasing intra-class difference and decreasing inter-class differences. Comparing Figures 8(f) and 8(i) with Figure 8(g), the ACI module can improved the baseline's performance in such situation. Furthermore, we compare the traditional parallelepiped classification method proposed by Q. Liu and M. Tang [72] in Figure 8(j) with our vegetation extraction result in Figure 8(f), which the former cannot avoid this phenomenon either described above. In Figures 8(a)–8(f), the results of methods with ACI module under different parameter configurations exhibit refinement of the segments' boundary, such that different requirements of land use/land cover applications can be satisfied.

1) RATIO BETWEEN PARAMETERS' LEARNING RATE

In Table 2-4, comparing the statistics of ''a'' and ''c'' configurations for the full-images in Figures 8(a) and 8(c), little patch 1 in Figures 9(a) and 9(c), and little patch 2 in Figures 10(a) and 10(c), the precision, kappa coefficient, miou, precision rate, and F1-score of method under ''a'' configuration are relatively higher than those of method under ''c'' configuration.

This indicates that a higher ratio (1:1) between the parameters' learning rate from traditional network and the ACI module can obtain a more complete boundary compared with the lower one (1:0.1). This infers that a higher ratio leads to a greater emphasis on the learning of the ACI module and directs the vegetation model to predict more complete segments under semantic constraints in the label space, making it more suitable for the land use principle, while the smaller ratio is more suitable for the land cover principle.

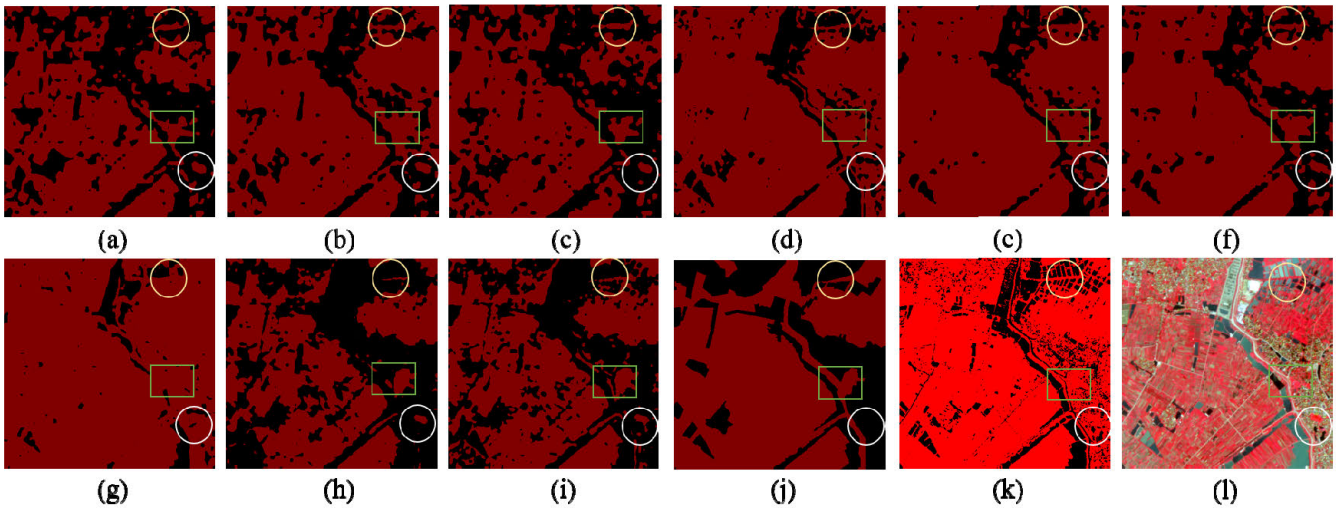


FIGURE 9. Vegetation extraction results of difficult image patch 1 from ZY-3 dataset (Yue Nan). (a) - (i) are results from methods described in Table 1; (j) ground truth label; (k) parallelepiped classification method [68]; (l) difficult area 1 cropped form original test image.

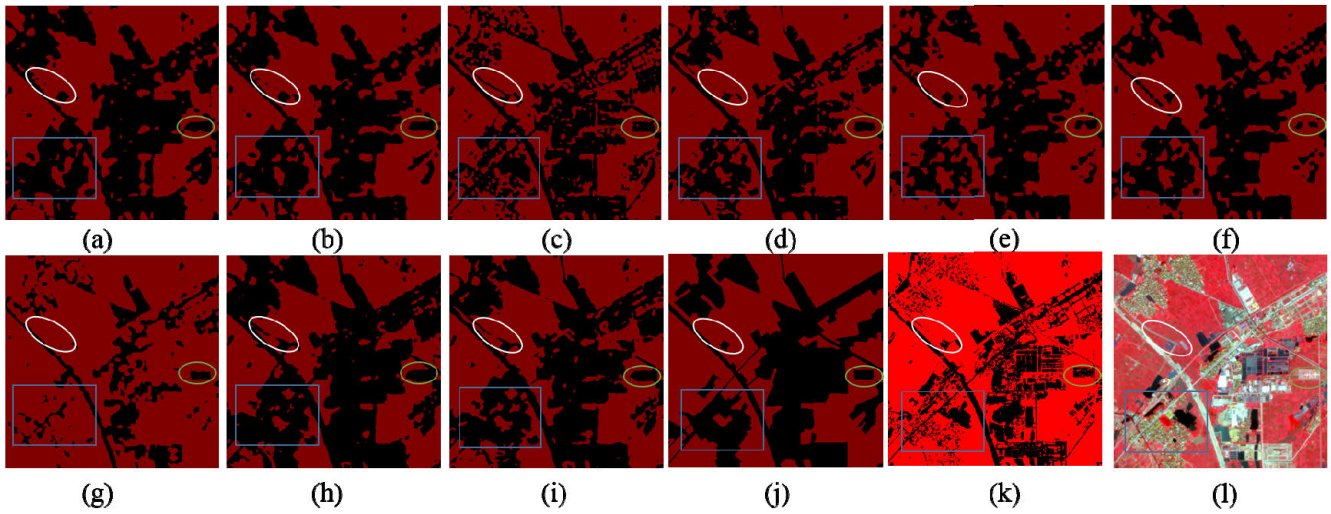


FIGURE 10. Vegetation extraction results of difficult image patch 2 from ZY-3 dataset (Yue Nan). (a) - (i) are results from methods described in Table 1; (j) ground truth label; (k) parallelepiped classification method [68]; (l) difficult area 2 cropped form original test image.

2) MARGIN TO CALCULATE THE BOUNDARY LOSS VALUE

For areas highlighted by a yellow circle, the distribution of vegetation shows a linear shape along with the road. In terms of the abundancy of details and completeness of segments, Figure 9(d) is better than Figure 9(e) and worse than Figure 9(b). Since margin values for Figures 9(b), 9(d), and 9(e) are 2, 0.5, and 1, respectively, a smaller margin value may cause the vegetation extraction results to be more scattered and less complete, leading to a reduction in the statistics of the ACI module in terms of the land use principle. As for Table 2-6, when comparing the statistics of "b", "d", and "e" configurations for the images in Figures 8-10(b), 8-10 (d) and 8-10 (e), configuration "d" ($m = 0.5$) and "e" ($m = 1$) exhibit a relatively worse performance than "b" ($m = 2$).

Here we can speculate this phenomenon based on the derivation formulas in Section 3.4, the ACI constraint calculated between pixels of different label categories is designed to distinguish the pixel label categories on both sides of the foreground and background targets. Given smaller value of parameter m in Equation (1.3), the value in equation (1.3) that used to impose boundary punishment of the inconsistency of the center pixel and the neighbor corner pixel in the label space will be clipped to zero. These phenomena will make the learning about these two kinds of parameters invalid, hence, the network will not excessively rely on the effects of the ACI module. Therefore, when we consider the land use principle, a relatively high value of margin m is more suitable. On the contrary, a relatively low value of margin m may lead to the vegetation predictions more scattered on the land cover principle.

TABLE 2. Overall statistics of predictions of different methods (in accordance with Figure 8).

	precision (%)	kappa	miou (%)	precision rate (%)	recall rate(%)	F1
8a	95.83	0.9137	90.29	**95.84**	93.97	0.9490
8b	95.54	0.9087	89.96	92.61	96.92	0.9472
8c	94.80	0.8935	88.38	91.83	95.93	0.9383
8d	95.48	0.9071	89.71	93.67	95.50	0.9457
8e	95.41	0.9058	89.67	92.46	96.75	0.9455
8f	**96.06**	**0.9191**	**91.03**	93.64	97.02	**0.9530**
8g	94.06	0.8791	87.15	88.92	**97.76**	0.9313
8h	95.19	0.9003	88.80	95.75	92.44	0.9407
8i	95.50	0.9070	89.62	94.73	94.33	0.9453

TABLE 3. Overall statistics of predictions of different methods (in accordance with Figure 9).

(in accordance with Figure 9).

	precision (%)	kappa	miou (%)	precision rate (%)	recall rate(%)	F1
9a	93.45	0.8032	73.07	91.00	78.76	0.8444
9b	93.04	0.8059	74.11	82.16	88.33	0.8513
9c	92.66	0.7790	70.24	91.12	76.80	0.8252
9d	92.02	0.7872	72.37	76.80	92.63	0.8397
9e	92.01	0.7885	72.57	76.34	93.63	0.8411
9f	**93.61**	**0.8234**	**76.23**	82.58	90.84	**0.8651**
9g	90.57	0.7609	69.94	71.37	**97.22**	0.8231
9h	92.20	0.7538	66.79	**94.51**	69.48	0.8009
9i	93.08	0.7897	71.36	91.59	76.37	0.8329

3) SIZE OF THE ACI MODULE

For areas consisting vegetation and buildings (white circle and green rectangular areas), the baseline in Figure 9(g) and parallelepiped classification method in Figure 9(j) misclassify most of the background as vegetation. The integrity of the vegetation boundary can be arranged from low to high as Figures 9(f), 9(a), 9(d), 9(b), 9(e), 9(c), and 9(g), so improvement of the baseline's performance can be verified effectively. For areas highlighted by yellow and white circles as well as green rectangular areas, when the boundary between the foreground and background objects is ambiguous, a wide transition bond exists in these two objects with different categories.

For areas consisting of vegetation and background of roads, water, and buildings (blue rectangular areas), Figure 10(a)'s result is more similar to the land use principle and the boundaries are more complete, while Figure 10(c) is more compatible in terms of the land cover principle. For background areas consisting of linear roads highlighted by white and green ovals, Figures 10(c), 10(f)'s results are more similar to the original image in terms of land cover and others are more consistent with the land use principle.

TABLE 4. Overall statistics of predictions of different methods (in accordance with Figure 10).

	precision (%)	kappa	miou (%)	precision rate (%)	recall rate (%)	F1
10a	95.29	0.8413	76.98	91.64	82.79	0.8699
10b	95.39	0.8488	78.12	88.98	86.49	0.8771
10c	93.94	0.8079	73.24	82.02	87.24	0.8455
10d	95.37	**0.8536**	**78.96**	85.23	91.48	**0.8824**
10e	94.84	0.8368	76.82	83.98	90.01	0.8689
10f	95.19	0.8478	78.21	84.84	90.92	0.8777
10g	90.66	0.7405	66.50	67.62	**97.58**	0.7988
10h	95.14	0.8301	75.30	**95.66**	77.97	0.8591
10i	**95.62**	0.8534	78.60	91.59	84.71	0.8802

We can conclude that the optimal size of the ACI varies with the size of the target to be extracted. By learning the ACI constraint, we can guarantee label consistency between the center pixel and the distant corner pixel. The collection of pairwise bonds inside a segment ensures that all the pixels are predicted for the same category and pushes network predictions on two pixels of different ground-truth labels apart to obtain clear segmentation boundaries.

As for Table 2-4, when comparing the statistics of configuration "a" (s = 3, 5, 7) and "b" (s = 3), "a" achieves a relatively better performance than "b". In other words, a larger radius distance will give rise to enhancement of the pixel consistency in the label space. In this situation, the ACI model identifies obvious boundaries, while ambiguous boundaries or those with less details may be ignored. Therefore, the radius distance should be adjusted according to the considered object, and empirically, the core size that can obtain the best performance is "s357", means we need to averaged ACI losses of radius distance values of 1, 2, and 3.

4) CONCLUSIONS ABOUT THE PARAMETER CONFIGURATIONS

Because vegetation extraction results can be evaluated on the basis of land cover, or land use standard, our adaptive and effective vegetation extraction model can be applicable to satisfy various requirements of real applications through controlling the parameter configurations in Table 5.

B. EFFECTIVENESS OF THE FOCUS PERCEPTION MODULE

In Table 1-4, we can verify the effectiveness of the FP module, comparing the evaluation indexes for Figures 8(g) and 8(h), Figures 9(g) and 9(h), and Figures 10(g) and 10(h).

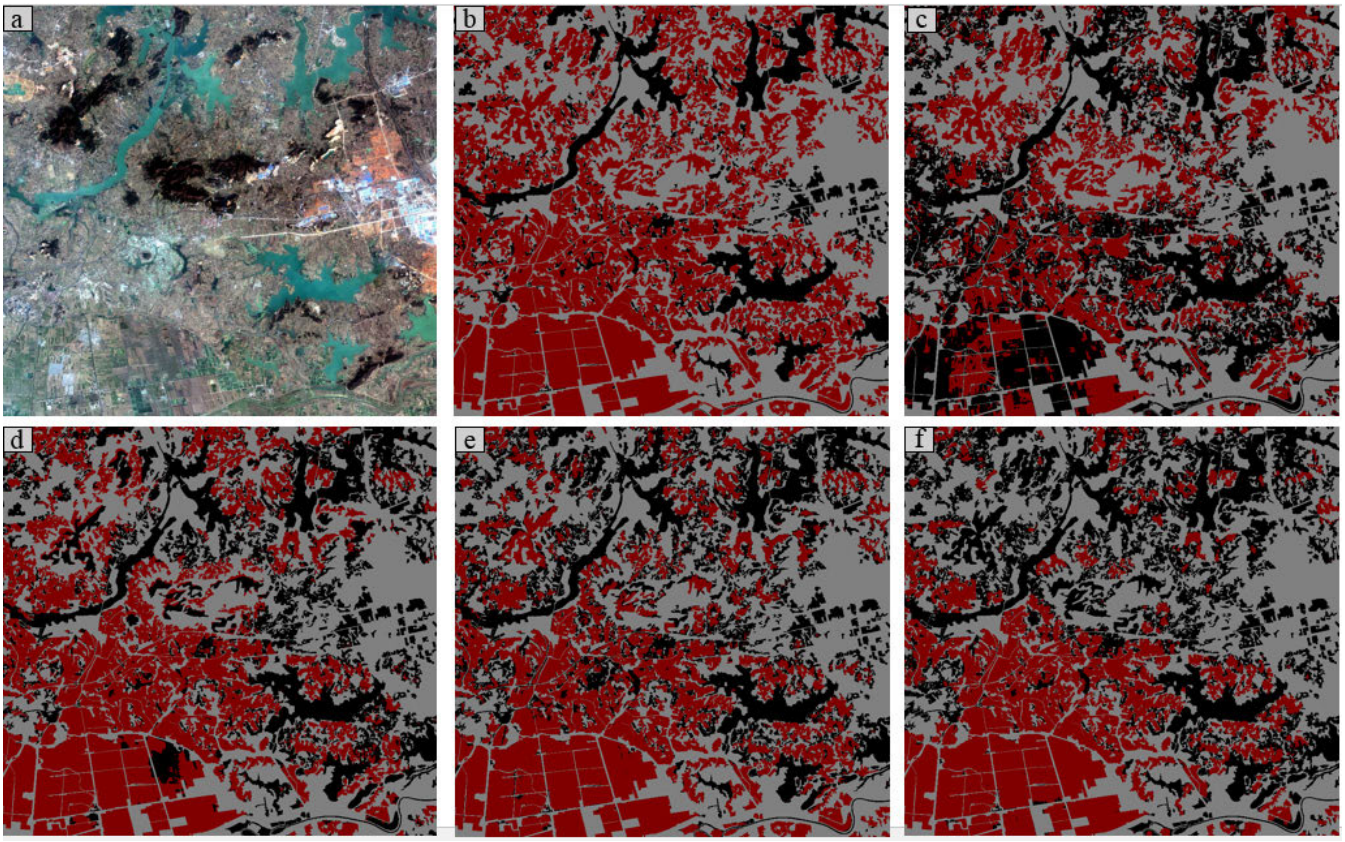


FIGURE 11. Full-image vegetation extraction results of different methods based on the GID dataset (E114.0, N30.5 in Wuhan, Hubei Province, China). (a) Original image; (b) ground truth label; (c) - (f) are results from methods described in Table 6.

TABLE 5. Analysis of parameter configurations.

parameter	Margin	Neighborhood size	Learning rate ratio	Add or do not add
Land use	Higher (2–3)	Larger (s357)	Larger (1:1)	Do not add
Land cover	Lower (0.5–2)	Smaller (s3)	Smaller (1:0.1)	Add

The configurations with ''g'' are better than the configurations with ''f''. Taking Figures 8(g) and 8(h) as example, the method with the FP module improve the baseline's evaluation indexes by 1.13%, 2.12%, 1.65%, 6.83%, and 0.94%, respectively, due to its ability to place an emphasis on the evident spatial position and channel and make full use of effective information, without causing too high computational burdens. It can consequently improve the vegetation extraction results of images with in-evident vegetation features.

C. ANALYSIS OF THE INTERACTION BETWEEN TWO MODULES

From statistical viewpoint, Table 2-4 illustrated overall statistics of methods with or without an ACI module and FP module, and the ground-truth labels are provided under the land cover principle. Taking the precision, kappa coefficient, miou, precision rate, and F1-score as the evaluation indexes, method with the ACI module and the FP module in Figure 8(i) can improve the baseline in Figure 8(g) by 1.44%, 2.79%, 2.47%, 5.81%, and 1.4%, respectively. For the ACI module under the ''f'' configuration in Figure 8(f), it can improve the baseline by 2%, 4%, 3.88%, 4.72%, and 2.17%, respectively, even though the recall rate is relatively decreased. For the FP module alone in Figure 8(h), it can improve the baseline's precision, kappa coefficient, miou, precision rate, and F1-score by 1.13%, 2.12%, 1.65%, 6.83%, and 0.94%, respectively, while decrease the recall rate by 5.32%. In summary, except for the recall rate, all of the evaluation indexes are improved when using methods with ACI module or FP module.

Although the effectiveness of the ACI module and the FP module, the performance is not likely to increase when integrating these two modules in terms of all measures, as we can find the evaluation indexes for Figures 8(f), 9(f), and 10(f) is higher than those for Figure Figures 8(i), 9(i), and 10(i) which is explained as follows. And the same phenomenon can be found for Figure 9(f) and 9(i) while the evaluation indexes for Figure 10(f) are lower than those for Figure 10(i). Thus, we can conclude that the integration between the ACI module and the FP module are not always positive, which we try to analyze as follows.

The FP module focuses on vegetation features while the ACI model focuses on the integrity and regularity of the boundary. These two modules can dependently improve the

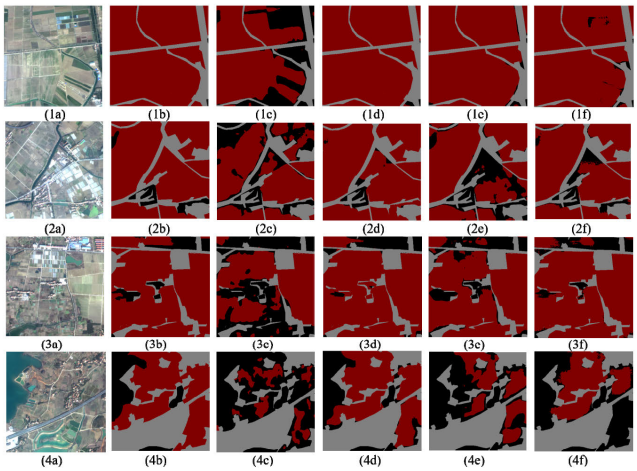


FIGURE 12. To illustrate the effectiveness of the ACI module. (a) original image of GID dataset (E114.0, N30.5 in Wuhan, Hubei Province, China); (b) ground truth label; (c) - (f) are results from methods described in Table 6.

TABLE 6. Definition of methods with or without an ACI module and FP module.

method	Adding ACI module			Adding FP module
	Margin value	kernel size	learning rate ratio	
11c	no	no	no	no
11d	3.0	3,5,7	1:0.1	no
11e	3.0	3,5,7	1:1	no
11f	no	no	no	yes

overall accuracy in particular aspects, but may conflict with each other when used together. However, based on the statistics of the results compared with those of the baseline, the data are still valid and consistent with previous conversations on qualitative detail analysis.

VI. EXPERIMENTAL ANALYSIS OF THE GID DATASET

In this section, we aim to verify the effectiveness of our methods on GID dataset. This section is split into three parts, global and local vegetation results are shown in Figure 11 (Part A) and Figure 12 (Part B), respectively, and we illustrate the global statistics of methods in Table 7 (Part C). Totally, the arranged sequence of Figures 11 and 12 as well as Table 7 is explained in Table 6.

A. FULL-IMAGE VEGETATION ANALYSIS OF THE GID DATASET

For the GID dataset, we took experimental analysis of the ZY-3 dataset as a guideline, in order to choose appropriate parameter configurations of the ACI module and evaluate the vegetation extraction results. Based on different comparative experimental groups of the ZY-3 dataset, we finally chose the learning rate ratio value of 1:0.1 and 1:1, the margin value of 3 and the size value of 3, 5 and 7.

In Figure 11(c), the vegetation extraction results of the baseline in some complicated urban areas are not completely correct compared with the land use ground truth label, and some omission areas (especially the left bottom part) exist, which can be improved to some degree through adding the FP module in Figure 11(f) and ACI module in Figures 11(d) and 11(e).

B. DIFFICULT AREA'S VEGETATION EXTRACTION OF THE GID DATASET

In Figure 12, further detailed qualitative vegetation extraction analysis of four image patches are shown. The sequence of these little images corresponds to Table 6. Through interpreting the vegetation extraction results of the difficult local image patches extracted from the original test GF-2 HRRSI, we can verify our proposed approach's ability to improve the performance of vegetation land use mapping, compared with the baseline method.

From the local perspective, our vegetation results of these difficult and challenging areas can better satisfy the requirements of land use or land cover applications. Taking the baseline's results in the third column for example, Figure 1(c) illustrates scattered segments, the top part of Figure 2(c) is omitted, and Figures 3(c) and 4(c) misclassify some building and road objects into vegetation.

Besides, it is difficult for the baseline method to accurately extract urban green space and urban grassland in Figure 12 (2a) and (3a), which are intermingled with roads and residential buildings in terms of their regular shapes and textural features. Thus, it is necessary to model different spatial structures of vegetation features, such as the size, shape, and context.

After introducing a FP module, we extracted sensitive features of different types of vegetation and decoupled the evident features from the external features of vegetation, and the problem of small inter-class differences and large intra-class differences between vegetation and background could be eliminated.

After constructing an ACI model, the integrity of pixels with a consistent label category and the distinguishability of pixels with different label categories improved, which can be verified through comparing the results of the last two columns with the third column in Figure 12. Taking vegetation binary classification for example, we can make full use of the textural and shape information of non-vegetation samples to improve the boundary location accuracy and then indirectly improve the accuracy of the urban green space and urban grassland extraction results.

C. STATISTICS OF THE GID DATASET

We have not reported the results of the method integrating the ACI module and FP module. In Table 7, the overall accuracy of the GF-2, the precision, kappa coefficient, miou, precision rate, recall rate, and F1-score are improved by 3.16%, 17.99%, 21.19%, 3.31%, 20.39%, and 16.63%, respectively, under the "f" parameter configuration. As for the FP module, it can improve the baseline's precision, kappa coefficient,

TABLE 7. Statistics of predictions of methods with or without an ACI module (in accordance with Figure 11).

	precision (%)	kappa	miou (%)	precision rate (%)	recall rate(%)	F1
11c	92.63	0.6246	49.41	96.49	50.31	0.6614
11d	95.70	0.8036	70.59	97.06	72.13	0.8276
11e	**95.79**	**0.8045**	**70.60**	**99.80**	**70.70**	**0.8277**
11f	93.38	0.6667	53.88	99.51	54.10	0.7003

miou, precision rate, recall rate, and F1-score by 0.75%, 4.21%, 4.47%, 3.02%, 3.79%, and 3.89%, respectively.

As demonstrated by the comparison above, our adaptive inference module is applicable to effectively refining the boundary location of vegetation.

VII. CONCLUSION

In this paper, we present a novel method for the automatic vegetation extraction of HRRSI through an adapted state-of-the-art fully convolution neural network integrating an ACI module and a FP module. Different from the traditional methods using the NDVI index to extract the vegetation, our methods can be employed effectively, even without sufficient spectral information.

The proposed vegetation land use/land cover extraction from HRRSI based on the ACI method has been evaluated with two typical remote sensing datasets to prove its meanings and effectiveness in automatic vegetation land use/land cover extraction from HRRSI. Several conclusions can be drawn from the experiments.

Compared with the baseline Deeplab v3+ [45] for ZY-3 and GID, the precision, kappa coefficient, miou, precision rate, and F1-score of the method integrating the ACI module are improved, at most, by 2%, 4%, 3.88%, 6.92%, and 2.17%, and at least, by 2.14%, 9.17%, 13.52%, 0.61%, and 7.97%, respectively, even though the recall rate is relatively decreased for ZY-3, while increased by at least 13.15% for GID, due to our context model's ability to represent relationships between the center pixel and its neighbors under semantic constraints. The FP module has been verified to improve the precision, kappa coefficient, miou, precision rate, and F1-score by at most 1.13%, 2.12%, 1.65%, 6.83%, and 0.94% and by at least 0.75%, 4.21%, 4.47%, 0.20%, 3.79%, and 3.89%, respectively. This phenomenon can be attributed to the FP module's capability to eliminate the problem of small inter-class differences and large intra-class differences.

When we merged the ACI and FP module into the baseline Deeplab v3+ [45], we could increase the precision, kappa coefficient, miou, precision rate, and F1-score by 1.44%, 2.79%, 2.47%, 5.81%, and 1.4% for ZY-3 dataset, respectively. Even though we cannot always guarantee a positive interaction between these two modules due to their intrinsic differences in improving the vegetation extraction performance, we can still verify our method's effectiveness for improving the baseline Deeplab v3+ [45] through analyzing the details of segments extracted in Figures 8–12.

Nevertheless, although the proposed method can perform well for the vegetation extraction of HRRSI in an automatic pipeline, it still has some unavoidable limitations. When applied to new HRRSI whose distribution of the feature space is significantly different with the training data, it may not guarantee an obvious improvement of the performance. This will be explored in our future research by incorporating the scene constraint or relation-augmented information [69]–[71] during the training stage or reducing the distribution differences between the target domain and the source domain with transfer learning methods [72].

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers for their comments and suggestions. We will also thank Ke, Tsung-Wei and Hwang, Jyh-Jing from the International Computer Science Institute in UC Berkeley University and Rishizek for providing the source code and implementation details.

REFERENCES

[1] J. P. Ardila, V. A. Tolpekin, W. Bijker, and A. Stein, "Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images," *ISPRS J. Photogram. Remote Sens.*, vol. 66, no. 6, pp. 762–775, Nov. 2011.

[2] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[3] A. Karnieli, N. Agam, R. T. Pinker, M. Anderson, M. L. Imhoff, G. G. Gutman, N. Panov, and A. Goldberg, "Use of NDVI and land surface temperature for drought assessment: Merits and limitations," *J. Climate*, vol. 23, no. 3, pp. 618–633, Feb. 2010.

[4] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems Eng.*, vol. 114, no. 4, pp. 358–371, Apr. 2013.

[5] C. Zhang and J. M. Kovacs, "The application of small unmanned aerial systems for precision agriculture: A review," *Precis. Agricult.*, vol. 13, no. 6, pp. 693–712, Dec. 2012.

[6] A. Kumar, A. C. Pandey, and A. Jeyaseelan, "Built-up and vegetation extraction and density mapping using worldview-II," *Geocarto Int.*, vol. 27, no. 7, pp. 557–568, Nov. 2012.

[7] Q. J. Liu, T. Takamura, N. Takeuchi, and G. Shao, "Mapping of boreal vegetation of a temperate mountain in China by multitemporal Landsat TM imagery," *Int. J. Remote Sens.*, vol. 23, no. 17, pp. 3385–3405, Jan. 2002.

[8] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," Jul. 2018, *arXiv:1807.05713*. [Online]. Available: https://arxiv.org/abs/1807.05713

[9] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[10] P. H. Verburg, K. Kok, R. G. Pontius, and A. Veldkamp, *Modeling Land-Use and Land-Cover Change, in Land-Use and Land-Cover Change*. Berlin, Germany: Springer, 2006, pp. 117–135.

[11] C. P. Giri, *Remote Sensing Of Land Use And Land Cover: Principles And Applications*. Boca Raton, FL, USA: CRC Press, 2016.

[12] *The National Survey of Geographical Conditions Leading Group Office*, State Council, P.R.C. Gen. Situation Index Geographical Conditions (Chin. Manual, GDPJ 01-2013), Beijing, China, 2014.

[13] P. Dusseux, L. Hubert-Moy, R. Lecerf, X. Gong, and T. Corpetti, "Identification of grazed and mown grasslands using a time series of high-spatial-resolution remote sensing images," in *Proc. 6th Int. Workshop Anal. Multi-Temporal Remote Sens. Images*, Jul. 2011, pp. 145–148.

[14] Z. Y. Lv, T. F. Liu, P. Zhang, J. A. Benediktsson, T. Lei, and X. Zhang, "Novel adaptive histogram trend similarity approach for land cover change detection by using bitemporal very-high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9554–9574, Dec. 2019.

[15] Y. Ban, H. Hu, and I. M. Rangel, "Fusion of quickbird MS and RADARSAT SAR data for urban land-cover mapping: Object-based and knowledge-based approach," *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1391–1410, Mar. 2010.

[16] P. Dusseux, T. Corpetti, L. Hubert-Moy, and S. Corgne, "Combined use of multi-temporal optical and radar satellite images for grassland monitoring," *Remote Sens.*, vol. 6, no. 7, pp. 6163–6182, Jun. 2014.

[17] L. D. O. Pereira, C. da Costa Freitas, S. J. S. S. Anna, D. Lu, and E. F. Moran, "Optical and radar data integration for land use and land cover mapping in the Brazilian Amazon," *GISci. Remote Sens.*, vol. 50, no. 3, pp. 301–321, Jun. 2013.

[18] W. S. Walker, J. M. Kellndorfer, E. Lapoint, M. Hoppus, and J. Westfall, "An empirical in SAR-optical fusion approach to mapping vegetation canopy height," *Remote Sens. Environ.*, vol. 109, no. 4, pp. 482–499, Aug. 2007.

[19] B. Waske and S. Van Der Linden, "Classifying multilevel imagery from SAR and optical sensors by decision fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1457–1466, May 2008.

[20] E. Binaghi, I. Gallo, and M. Pepe, "A neural adaptive model for feature extraction and recognition in high resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 24, no. 20, pp. 3947–3959, Jan. 2003.

[21] A. Carleer, O. Debeir, and E. Wolff, "Assessment of very high spatial resolution satellite image segmentations," *Photogramm Eng. Remote Sens.*, vol. 71, no. 11, pp. 1285–1294, Nov. 2005.

[22] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[23] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, Jun. 2009.

[24] L. Zhang, X. Huang, B. Huang, and P. Li, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.

[25] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogram. Remote Sens.*, vol. 130, pp. 277–293, Aug. 2017.

[26] 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.

[27] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016.

[28] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3368–3379, Jul. 2015.

[29] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[30] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[31] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vols. I–3, pp. 293–298, Oct. 2012.

[32] D. Marmanis, "Semantic segmentation of aerial images with an ensemble of CNNs. ISPRS Annals of the Photogrammetry," *Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Mar. 2016

[33] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018.

[34] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 500, May 2017.

[35] T. Sun, Z. Chen, W. Yang, and Y. Wang, "Stacked U-nets with multi-output for road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 202–206.

[36] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Chellappa, "Stacked U-nets for ground material segmentation in remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 257–261.

[37] T. M. Lillesand and R. W. Kiefer, *Remote Sensing and Image Interpretation*. Hoboken, NJ, USA: Wiley, 2000.

[38] E. Honkavaara, H. Saari, J. Kaivosoja, I. Pölönen, T. Hakala, P. Litkey, J. Mäkynen, and L. Pesonen, "Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture," *Remote Sens.*, vol. 5, no. 10, pp. 5006–5039, Oct. 2013.

[39] Z. Li, G. Chen, and T. Zhang, "Temporal attention networks for multitemporal multisensor crop classification," *IEEE Access*, vol. 7, pp. 134677–134690, 2019.

[40] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, Feb. 2019.

[41] P. Sidike, V. Sagan, M. Maimaitijiang, M. Maimaitiyiming, N. Shakoor, J. Burken, T. Mockler, and F. B. Fritschi, "DPEN: Deep progressively expanded network for mapping heterogeneous agricultural landscape using Worldview-3 satellite imagery," *Remote Sens. Environ.*, vol. 221, pp. 756–772, Feb. 2019.

[42] A. Farooq, X. Jia, J. Hu, and J. Zhou, "Multi-resolution weed classification via convolutional neural network and superpixel based local binary pattern using remote sensing images," *Remote Sens.*, vol. 11, no. 14, p. 1692, Jul. 2019.

[43] L. Zhang and B. Verma, "Roadside vegetation segmentation with adaptive texton clustering model," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 159–176, Jan. 2019.

[44] Y. Chen, C. Zhang, S. Wang, J. Li, F. Li, X. Yang, Y. Wang, and L. Yin, "Extracting crop spatial distribution from Gaofen 2 imagery using a convolutional neural network," *Appl. Sci.*, vol. 9, no. 14, p. 2917, Jul. 2019.

[45] L.-C. Chen, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. 2018, pp. 801–818.

[46] 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.

[47] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[48] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," Nov. 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[50] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[51] H. Zhao, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[52] 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.

[53] L.-C. Chen, "Rethinking atrous convolution for semantic image segmentation," Jun. 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," Dec. 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[55] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 582–589.

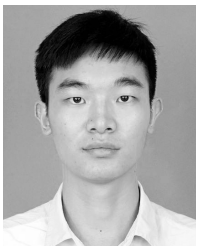
[56] F. Shen, R. Gan, S. Yan, and G. Zeng, "Semantic segmentation via structured patch prediction, context CRF and guidance CRF," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1953–1961.

[57] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.
[58] W. Liu and A. A. C. Rabinovich Berg, "Parsenet: Looking wider to see better," Jun. 2015, *arXiv:1506.04579*. [Online]. Available: https://arxiv.org/abs/1506.04579
[59] P. O. Pinheiro and R. P. C. Dollár, "Learning to segment object candidates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1990–1998.
[60] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1337–1385.
[61] M. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," May 2018, *arXiv:1805.04777*. [Online]. Available: https://arxiv.org/abs/1805.04777
[62] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[63] X. Wang, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[64] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[65] 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.
[66] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.
[67] T.-W. Ke, J.-J. Hwang, Z. Liu, and S. X. Yu, "Adaptive affinity fields for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV*, 2018, pp. 605–621.
[68] Q. Lü and M. Tang, "Detection of hidden bruise on kiwi fruit using hyperspectral imaging and parallelepiped classification," *Procedia Environ. Sci.*, vol. 12, pp. 1172–1179, 2012.
[69] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.
[70] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.
[71] L. Shi-Hua, "A review of methods for classification of remote sensing images," *Remote Sens. Land Resour.*, vol. 17, no. 2, pp. 1–6, 2009.
[72] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1134–1148, Jun. 2014.

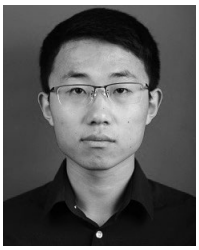
ZONGQIAN ZHAN received the M.A. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2003 and 2007, respectively. He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include camera calibration, close-range photogrammetry, and unmanned aerial vehicle photogrammetry.

XIAOMENG ZHANG received the bachelor's degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2017, where she is currently pursuing the master's degree with the School of Geodesy and Geomatics. Her research interests include semantic segmentation and object detection based on deep learning, photogrammetry, and remote sensing.

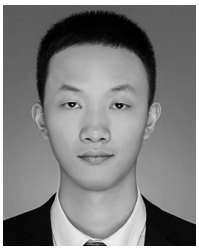
YI LIU received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009. She is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. Her research interests include remote sensing image interpretation and deep learning.



XIAO SUN received the bachelor's degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2017, where he is currently pursuing the master's degree. His current research mainly focuses on relativistic geodesy based on microwave links between satellites and ground station.



CHAO PANG received the bachelor's degree in cartography and geographic information system from the China University of Mining and Technology, Beijing, China, in 2017. He is currently pursuing the master's degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China. His research interests include semantic segmentation and change detection based on deep learning and remote sensing.



CHENBO ZHAO received the bachelor's degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2017, where he is currently pursuing the master's degree with the School of Geodesy and Geomatics. His research interests include semantic segmentation based on deep learning.