SCIENCE CHINA
Information Sciences

April 2020, Vol. 63 140305:1–140305:12
https://doi.org/10.1007/s11432-019-2791-7

• RESEARCH PAPER •
Special Focus on Deep Learning in Remote Sensing Image Processing

Hybrid first and second order attention Unet for building segmentation in remote sensing images

Nanjun HE[1], Leyuan FANG[1*] & Antonio PLAZA[2]

[1] *College of Electrical and Information Engineering, Hunan University, Changsha* 410082*, China;*
[2] *Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politecnica, University of Extremadura, Extremadura* E-10003*, Spain*

Received 1 November 2019/Revised 31 December 2019/Accepted 11 February 2020/Published online 9 March 2020

Abstract Recently, building segmentation (BS) has drawn significant attention in remote sensing applications. Convolutional neural networks (CNNs) have become the mainstream analysis approach in this field owing to their powerful representative ability. However, owing to the variation in building appearance, designing an effective CNN architecture for BS still remains a challenging task. Most of CNN-based BS methods mainly focus on deep or wide network architectures, neglecting the correlation among intermediate features. To address this problem, in this paper we propose a hybrid first and second order attention network (HFSA) that explores both the global mean and the inner-product among different channels to adaptively rescale intermediate features. As a result, the HFSA can not only make full use of first order feature statistics, but also incorporate the second order feature statistics, which leads to more representative feature. We conduct a series of comprehensive experiments on three widely used aerial building segmentation data sets and one satellite building segmentation data set. The experimental results show that our newly developed model achieves better segmentation performance over state-of-the-art models in terms of both quantitative and qualitative results.

Keywords building segmentation (BS), convolutional neural networks (CNNs), remote sensing, high order pooling, attention

Citation He N J, Fang L Y, Plaza A, et al. Hybrid first and second order attention Unet for building segmentation in remote sensing images. Sci China Inf Sci, 2020, 63(4): 140305, https://doi.org/10.1007/s11432-019-2791-7

1 Introduction

With the development of imaging techniques, high resolution remote sensing images have become more accessible and affordable. Therefore, the automatic segmentation of buildings from high resolution images has gained considerable attention. A well established building segmentation (BS) map can not only be used for urban planning, disaster management, population estimation, etc., but also for many other geospatial applications such as socio-economics [1, 2]. However, the variation within the types of buildings is usually large (e.g., the shape and color of the different buildings are usually different) and the backgrounds within the images are extremely complex, which makes BS a challenging task.

During the past two decades, a considerable number of techniques have been developed for BS [2–9]. Generally, existing BS methods can be roughly divided into two categories, i.e., hand-crafted feature-based methods and deep neural network-based methods. The former utilize some priors of buildings, such as the shape and shadow, to obtain an initial segmentation result followed by some classical segmentation

[*] Corresponding author (email: Leyuan_fang@hnu.edu.cn)

methods, such as region growing. In [3], the authors utilized edge detection to locate shaded boundaries, and then exploited region growing to find boundaries without shadows in order to determine the location of the buildings. Moreover, by further considering the directional spatial relationship between buildings and their shadows, they developed a probabilistic landscape-based approach to post-process the segmentation results [4]. In [6], a novel recognition-driven variational framework was proposed to combine two kinds of competing shape priors, i.e., pose and affine-invariant for accurate BS. In general, the aforementioned (traditional) methods take some characteristics of the buildings into consideration. However, as mentioned above, the appearance of buildings can vary dramatically owing to illumination conditions and architectural style, especially in large areas. Under this context, image priors may not be general enough. Thus, the generalization and performance of traditional methods are often limited.

Over the past ten years, we have witnessed a tremendous growth of neural networks [10–17]. Because the Alexnet [10] was proposed in 2012 (with outstanding image classification performance on ImageNet), deep neural networks in general, and convolutional neural networks (CNNs) in particular, have dominated many image classification and segmentation tasks. Long et al. [12] proposed a fully convolutional neural (FCN) network for image segmentation, which is the first successful case of an application of a deep CNN model for image segmentation, achieving state-of-the-art performance. Inspired by the FCN, a series of deep neural networks have been developed for image segmentation. The authors in [14] proposed a Unet with an encode-decode architecture for biomedical image segmentation. A similar idea can be found in [16], called SegNet. The main difference between the Unet and SegNet is the upsampling strategy adopted during the decoding stage. Unet usually adopts interpellation (or transform convolutional operation) for upsampling, while SegNet introduces an up-pooling method for the same purpose [14, 16].

More recently, CNNs have also been extended for remote sensing image classification and segmentation [2, 9, 18–20]. Specifically, Yuan et al. [2] designed a CNN with a simple structure that integrates activation of multiple layers, and introduces a signed distance function for representing building edges. In [9], a classical semantic segmentation model called SegNet is applied for BS over the whole United States. Generally, existing CNN-based BS models can achieve promising segmentation performance. However, most of them mainly focus on deep or wider architecture design, neglecting to explore the correlations among different feature channels of intermediate features, which does not fully explore the representative ability of CNNs.

To address this problem, in this paper we develop a new model called hybrid first and second order attention network (HFSA) that makes full use of correlations among intermediate features to obtain more powerful and representative features. Specifically, the proposed HFSA consists of two basic components: first order channel attention (FOCA) and second order attention (SOCA). The FOCA module first utilizes global average pooling (GAP) to extract first-order feature statistics. Then, these statistics are fed into a tiny fully connected (FC) network and the output is used to rescale the input intermediate feature. On the other hand, bilinear pooling is used to exploit the second-order feature statistics among different channels of intermediate features in SOCA. The second-order feature statistics are then fed into another FC network. The output of the network is also used to rescale the input intermediate feature. The final output of the HFSA is the element-wise addition of the outputs of FOCA and SOCA. As a result, the HFSA can fully explore the correlation among intermediate features, including first- and second-order feature statistics. Here, we use a classical segmentation model (Unet) as the basic network and then insert the proposed HFSA into the Unet to construct our HFSA-Unet.

The reminder of this paper is organized as follows. In Section 2, we detail the proposed HFSA-Unet for BS. Section 3 provides experimental results using two aerial images and one satellite image, and presents a comprehensive comparison of our newly developed approach to other state-of-the art segmentation models. Section 4 summarizes the paper.

2 Network architecture

Figure 1 shows a block diagram of the HFSA-Unet. As it can be seen in the figure, the HFSA is equipped with skip connections to adaptively rescale intermediate features with the weight learned from the gating

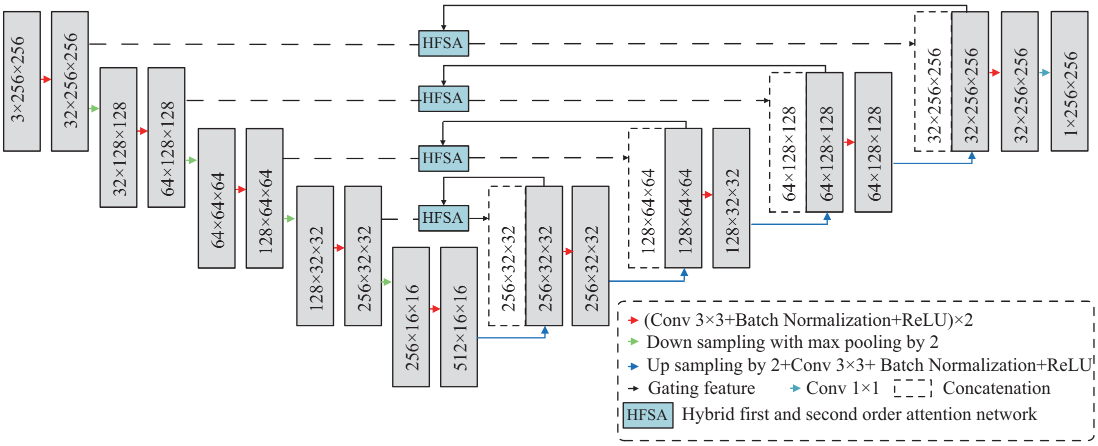

Figure 1 (Color online) Block diagram of the proposed HFSA-Unet for building segmentation. The HFSA is equipped with skip connections to adaptively rescale intermediate features in the encoding stage with weights learned from the correlation of intermediate features in the decoding stage (i.e., the gating feature).

feature. In addition, Figure 2 shows a graphical illustration of the HFSA network. Specifically, the HFSA consists of two basic modules: FOCA and SOCA. These two modules are described in detail below.

2.1 FOCA module

In order to utilize the first-order information to rescale the input features, we firstly use GAP to shrink the corresponding gating feature through its spatial dimensions. Assuming $\mathcal{X} = [X_1, \ldots, X_c, \ldots, X_C] \in \mathbb{R}^{H\times W\times C}$ is the input feature, $\mathcal{Y} = [Y_1, \ldots, Y_c, \ldots, Y_C] \in \mathbb{R}^{H\times W\times C}$ is the gating feature, the GAP is conducted as follows, where $y = [y_1, \ldots, y_c, \ldots, y_C]^{\mathrm{T}} \in \mathbb{R}^C$ is the output (pooled) feature. Note that, although the max pooling can also shrink the input through its spatial dimensions, it only extracts the maximum value of each channel, which may not fully utilize the global information. Therefore, the GAP is used here,

$$y_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Y_c(i,j). \tag{1}$$

Then, the pooled feature y is fed to a tiny network with two FC layers. The network can be designed with more layers, however, to make a better tradeoff between computation cost and performance, a tiny network with two FC layers is adopted here. Let $\theta_1\,(W_1, b_1)$ denote the FC$_1$, and $\theta_2\,(W_2, b_2)$ denote the FC$_2$. The output first order channel attention weights $f = [f_1, \ldots, f_c, \ldots, f_C] \in \mathbb{R}^C$ is derived as follows:

$$f = \sigma\left(W_2^{\mathrm{T}}\left(\delta\left(W_1^{\mathrm{T}} y + b_1\right)\right) + b_2\right), \tag{2}$$

where δ stands for the activation function of rectified linear unit (ReLU) and σ denotes the Sigmoid activation function. Using ReLU firstly is to enhance the nonlinearity of the tiny network, while the subsequent Sigmoid function is to scale the input into $(0, 1)$ to obtain the weights. With such design, the tiny network can make full use of the correlation among different channels to learn meaningful and useful weights for the rescaling of intermediate features. With the attention weights, we can obtain the final output of FOCA module (i.e., $\hat{\mathcal{X}}_f = [\hat{X}_1, \ldots, \hat{X}_c, \ldots, \hat{X}_C]$) by rescaling the input feature $\mathcal{X}$ as follows:

$$\hat{X}_c = f_c \cdot X_c. \tag{3}$$

2.2 SOCA module

We recall that FOCA can only make use of the mean (i.e., first-order statistics) of the the gating features to rescale the input feature, which does not incorporate the correlation among intermediate features. To this end, we introduce the SOCA module in this subsection. Given the same input feature $\mathcal{X}$ =

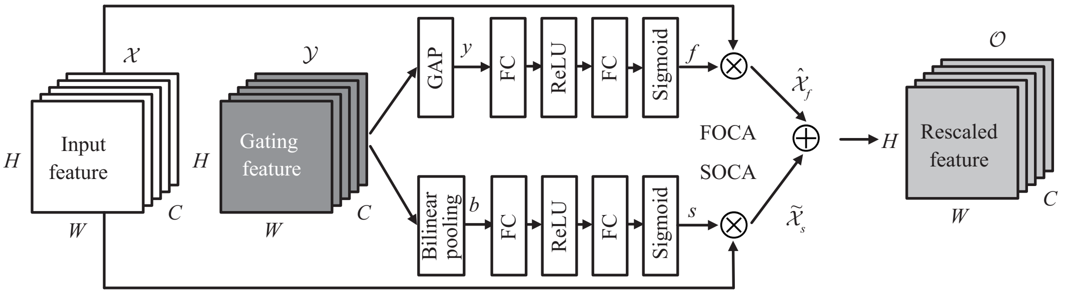

Figure 2 Block diagram of the proposed HFSA network. The HFSA consists of two basic modules: first order channel attention (FOCA) and second order channel attention (SOCA). The GAP denotes global average pooling. $\otimes$ denotes element-wise multiplication, while $\oplus$ denotes element-wise addition.

$[X_1, \ldots, X_c, \ldots, X_C] \in \mathbb{R}^{H \times W \times C}$, and the same gating feature $\mathcal{Y} = [Y_1, \ldots, Y_c, \ldots, Y_C] \in \mathbb{R}^{H \times W \times C}$ as in Subsection 2.1, the SOCA is implemented as follows.

Firstly, bilinear pooling is conducted on the gating feature as shown below. B denotes the obtained pooling matrix, and vec denotes a vectorization operation. As shown in (4), the matrix B is semi-symmetric positive definite (SPD) and each item in B stands for the inner-product of two different feature channels, by means of which high-order information among gating features can be fully incorporated. There is a alternative model called covariance pooling [20] which can also explore the high order information. However, when the spatial dimension of input is large, the forward propagation of covariance pooling needs to calculate and storage a huge affine matrix. For example, with the input size 256×256, the affine matrix will have more than four billions items. Thus, the covariance pooling is not suitable here.

$$B = \begin{bmatrix} \mathrm{vec}(Y_1) \\ \cdots \\ \mathrm{vec}(Y_c) \\ \cdots \\ \mathrm{vec}(Y_C) \end{bmatrix} \cdot \begin{bmatrix} \mathrm{vec}(Y_1) \\ \cdots \\ \mathrm{vec}(Y_c) \\ \cdots \\ \mathrm{vec}(Y_C) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{C \times C}. \tag{4}$$

As shown in [21], matrix normalization plays a crucial role for discriminative feature learning to a semi-SPD matrix. For this reason, here we also employ matrix normalization on B. Generally, there are two common matrix normalization strategies, i.e., matrix logarithm and matrix power [21]. Base on the results in [21], the matrix power has better performance in items of classification task, and therefore the matrix power normalization is adopted here. Because B is a semi-SPD matrix, it can be eigen-decomposed as follows:

$$B = U \Sigma U^{\mathrm{T}}, \tag{5}$$

where U denotes the matrix of eigen-vectors, while $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_c, \ldots, \sigma_C)$ is a diagonal matrix with eigenvalues (in non-increasing order). Then matrix normalization can be converted to the power of eigenvalues:

$$\hat{B} = B^{\alpha} = U \Sigma^{\alpha} U^{\mathrm{T}} \in \mathbb{R}^{C \times C}, \tag{6}$$

where α is a positive real number, and $\Sigma^{\alpha} = \mathrm{diag}\,(\sigma_1^{\alpha}, \ldots, \sigma_c^{\alpha}, \ldots, \sigma_C^{\alpha})$. As reported in [21], $\alpha = 1/2$ works well for discriminative representations. Thus, we set $\alpha = 1/2$. Note that, because Eq. (6) heavily depends on the eigen-decomposition (which is not well-supported on graphics processing units), we use the Newton-Schulz iteration to speed up the computation of matrix normalization, i.e., calculating the square root of matrix B. Given $B_0 = B$ and $Z_0 = I$, and assuming that $I \in \mathbb{R}^{C \times C}$ is an identity matrix, the Newton-Schulz iteration can be conducted as follows:

$$\begin{aligned} B_n &= \frac{1}{2} B_{n-1} \left(3I - Z_{n-1} B_{n-1}\right), \\ Z_n &= \frac{1}{2} \left(3I - Z_{n-1} B_{n-1}\right) Z_{n-1}. \end{aligned} \tag{7}$$

After several iterations of (7), we can calculate the square root of matrix B and $\hat{B} = B^{\frac{1}{2}} = B_n$. In practice, we set n to be 5. In [21, 22], the L_2 normalization is followed above step to ensure the stability of neural network. However, we do not observe unstable situation during the training stage without L_2 normalization for our task. For the sake of simplicity, we do not use L_2 normalization here. After obtaining the normalized matrix $\hat{B} = [b_1, \ldots, b_c, \ldots, b_C]$, we shrink $\hat{B}$ to extract the channel-wise statistics $d \in \mathbb{R}^C$ as follows:

$$d = \frac{1}{C} \sum_{i}^{C} b_c \in \mathbb{R}^C. \tag{8}$$

Finally, similar to FOCA, the d is fed into a tiny network with two FCs, i.e., $\theta_3\left(W_3, b_3\right)$ and $\theta_4\left(W_4, b_4\right)$, to make full use of the second-order statistics. The output attention weight $s \in \mathbb{R}^C$ is learned as follows:

$$s = \theta_4\left(\theta_3\left(d\right)\right) = \sigma\left(W_4^{\mathrm{T}}\left(\delta\left(W_3^{\mathrm{T}} d + b_3\right)\right) + b_4\right), \tag{9}$$

where δ stands for the activation function of ReLU and σ denotes the Sigmoid activation function. The output of SOCA $\tilde{\mathcal{X}}_s = [\tilde{X}_1, \ldots, \tilde{X}_c, \ldots, \tilde{X}_C]$ is then obtained by rescaling $\mathcal{X}$. The final output of HFSA $\mathcal{O}$ is the element-wise addition of $\hat{\mathcal{X}}_f$ and $\tilde{\mathcal{X}}_s$, i.e., $\mathcal{O} = \hat{\mathcal{X}}_f \bigoplus \tilde{\mathcal{X}}_s$.

$$\tilde{X}_c = s_c \cdot X_c. \tag{10}$$

2.3 Back propagation of FC layers in HFSA

Let l be the loss of whole network (l is a scalar), $\frac{\partial l}{\partial \mathcal{O}}$ be the derivative of the l respect to the $\mathcal{O}$, and $\mathcal{X}$ be the input intermediate features. Because $\mathcal{O} = \hat{\mathcal{X}}_f \bigoplus \tilde{\mathcal{X}}_s$, we can obtain the derivative of the l respect to the $\hat{\mathcal{X}}_f$ as follows:

$$\frac{\partial l}{\partial \hat{\mathcal{X}}_{ijk}^f} = \frac{\partial l}{\partial \mathcal{O}_{ijk}} \frac{\partial \mathcal{O}_{ijk}}{\partial \hat{\mathcal{X}}_{ijk}^f} = \frac{\partial l}{\partial \mathcal{O}_{ijk}}. \tag{11}$$

Then, because $\hat{X}_c = f_c \cdot X_c$ and $\hat{\mathcal{X}}_f = [\hat{X}_1, \ldots, \hat{X}_c, \ldots, \hat{X}_C]$, the derivative of the l respect to the $f = [f_1, \ldots, f_c, \ldots, f_C] \in \mathbb{R}^C$ is derived based on the chain rule as follows:

$$\frac{\partial l}{\partial f_c} = \sum_{i=1}^{H} \sum_{j=1}^{W} \left[\frac{\partial l}{\partial \hat{\mathcal{X}}_f} \bigotimes \mathcal{X}\right]_c. \tag{12}$$

Then, we can obtain the derivative of the l respect to input of Sigmoid function σ_{in} as follows:

$$\frac{\partial l}{\partial \sigma_{\text{in}}} = \sigma_{\text{in}}(1 - \sigma_{\text{in}}) \bigotimes \frac{\partial l}{\partial f}. \tag{13}$$

Subsequently, we can obtain the derivative respect to W_2^{T} and b_2 as follows:

$$\frac{\partial l}{\partial W_2^{\mathrm{T}}} = \frac{\partial l}{\partial \sigma_{\text{in}}} \frac{\partial \sigma_{\text{in}}}{\partial W_2^{\mathrm{T}}} = \frac{\partial l}{\partial \sigma_{\text{in}}} \delta_{\text{out}}^{\mathrm{T}}, \tag{14}$$

$$\frac{\partial l}{\partial b_2} = \frac{\partial l}{\partial \sigma_{\text{in}}} \frac{\partial \sigma_{\text{in}}}{\partial b_2} = \frac{\partial l}{\partial \sigma_{\text{in}}}, \tag{15}$$

where $\delta_{\text{out}}^{\mathrm{T}}$ stands for the output of ReLU. Because $\sigma_{\text{in}} = W_2^{\mathrm{T}} \delta_{\text{out}} + b_2$, we can obtain the derivative respect to the δ_{out} as follows:

$$\frac{\partial l}{\partial \delta_{\text{out}}} = \frac{\partial l}{\partial \sigma_{\text{in}}} \frac{\partial \sigma_{\text{in}}}{\partial \delta_{\text{out}}} = W_2 \frac{\partial l}{\partial \sigma_{\text{in}}}. \tag{16}$$

Because $\delta_{\text{out}} = \delta\left(\delta_{\text{in}}\right) = \delta\left(W_1^{\mathrm{T}} y + b_1\right)$, we can deduce the derivative respect to the first FC layer $\theta_1\left(W_1, b_1\right)$ as follows:

$$\frac{\partial l}{\partial W_1^{\mathrm{T}}} = \frac{\partial l}{\partial \delta_{\text{out}}} \frac{\partial \delta_{\text{out}}}{\partial \delta_{\text{in}}} \frac{\partial \delta_{\text{in}}}{\partial W_1^{\mathrm{T}}} = \left(W_2 \frac{\partial l}{\partial \sigma_{\text{in}}} \bigotimes \left(\delta_{\text{in}} > 0\right)\right) y^{\mathrm{T}}, \tag{17}$$

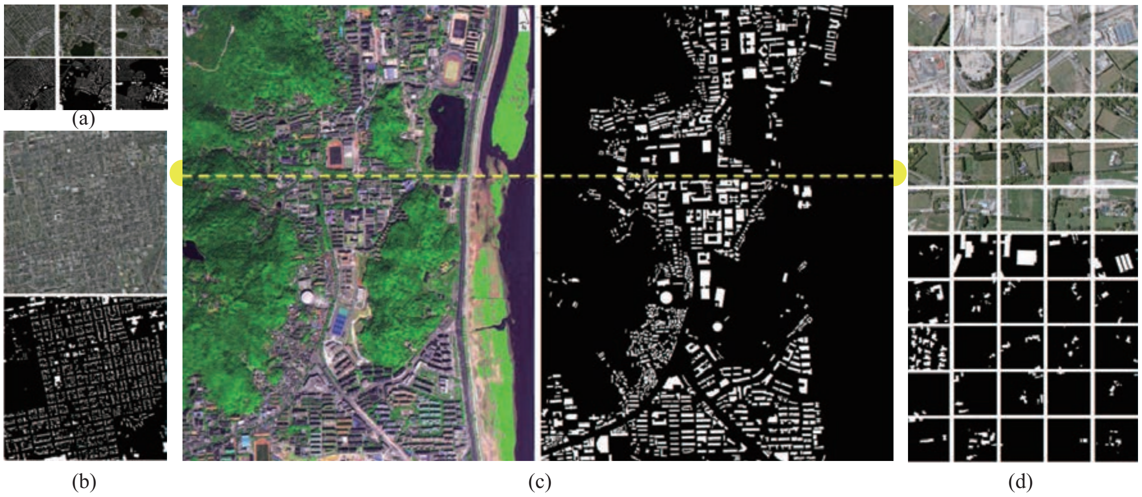

Figure 3 (Color online) Some samples in three considered test image data sets. (a) Massachusetts buildings data set. (b) Inria building data set. (c) Hunan university building data set. The part above the yellow dotted line (i.e., Pailou road) is the test set, while the part below is the training set. (d) Wuhan University building data set.

$$\frac{\partial l}{\partial b_1} = \frac{\partial l}{\partial \delta_{\text{out}}}\frac{\partial \delta_{\text{out}}}{\partial \delta_{\text{in}}}\frac{\partial \delta_{\text{in}}}{\partial b_1} = \left(W_2\frac{\partial l}{\partial \sigma_{\text{in}}} \bigotimes (\delta_{\text{in}} > 0)\right), \tag{18}$$

where δ_{in} denotes the input of the ReLU and $\delta_{\text{in}} > 0$ denotes that setting all the items in δ_{in} greater than 0 to be 1, while the rest of them to be 0.

Eqs. (11)–(18) show the back propagation of FCs in the FOCA module. Similarly, we can also obtain the back propagation of FCs in the SOCA module. For the sake of simplicity, the equations are omitted. So far, the backpropagation calculation of FC layers in proposed HFSA is illustrated.

3 Experimental results

3.1 Data sets

(1) Massachusetts buildings data set (MBD) [23]. The MBD consists of 151 aerial images collected over the Boston area, with each of the images being 1500 × 1500 pixels and 1-meter spatial resolution. In general, the building types in the MBD are individual houses and garages. The images are in color with three RGB bands. The ground-truth is obtained by rasterizing building footprints obtained from the OpenStreetMap project. Following the experiment setup in [24], we randomly split the MBD dataset into a training set of 111 images, and a test set of 40 images. Some samples in the MBD are shown in Figure 3(a).

(2) Inria building data set (IBD) [25]. The IBD data set includes 180 aerial images with public segmentation labels over five different areas, i.e., Austin in TX, Chicago in IL, Kitsap County in WA, Vienna in Austria, and West Tyrol also in Austria. Each area has 36 images. Each image has 5000×5000 pixels with 0.3-meter spatial resolution. All the images are RGB with values from 0 to 255. Following the experimental setup in [24], 27 images in Austin area are used for training, while the rest for test. Some samples in the IBD are shown in Figure 3(b).

(3) Hunan University building data set (HNUBD). The dataset is collected from the GaoFen-2 satellite over Hunan University, Changsha, China. It contains one panchromatic image with 1-meter spatial resolution and one multispectral image (i.e., RGB and Near infrared) with 4-meter spatial resolution. We used the RGB bands of the pansharpened image with values ranging from 0 to 255. HNUBD contains one image with 12536 × 9898 pixels. We split the image using a road called Pailou as a reference. The part above the Pailou road comprises about 4600 × 9898 pixels, while the part below the Pailou road comprises 7867×9898 pixels. The latter image is used as the training set. The HNUBD dataset is shown in Figure 3(c).

(4) Wuhan University building data set (WHUBD) [24]. The original aerial data comes from the New Zealand land information services website. The authors manually edited Christchurch's building vector data, with about 22000 independent buildings. The original ground resolution of the images is 0.075 m. WHUBD consists of 5772 images with size 512 × 512. Following the experimental setup in [24], 4736 images are used for training, while the rest for test. Some samples in the WHUBD are shown in Figure 3(d).

3.2 Implementation details

The samples in MBD, IBD and HNUBD are cropped into image patches with size of 256 × 256 pixels and stride of 128, respectively, before feeding the network. The learning rate is set to 10E−3 with an Adam optimizer. No weight decay is used. The batch size is set to 8. The iteration times for MBD, IBD and HNUBD are 50 epochs, respectively. The network is initialized from a normal distribution $\mathcal{N}(0, 0.001)$. All experiments are conducted with Pytorch on a PC with one GTX-1080 Ti graphics processing unit. Binary cross-entropy loss and dice loss are added for optimization purposes. Let $\hat{p} \in \mathbb{R}^{H\times W}$ be the probability prediction map of the network, and $p \in \mathbb{R}^{H\times W}$ be the ground truth, the loss function is shown as follows:

$$l = -\sum (p_{ij}\log(\hat{p}_{ij}) + (1-p_{ij})\log(1-\hat{p}_{ij})) + \left(1 - \frac{2\sum p_{ij}\hat{p}_{ij}}{\sum p_{ij} + \sum \hat{p}_{\,ij}}\right). \tag{19}$$

3.3 Comparison with state-of-the-art methods

In this subsection, we compare our proposed model[1)] with five the-state-of-art segmentation models, i.e., multi-layer perception (MLP) [2], FCN [12], Unet [14], SegNet [16], and SiU-net [24]. Specifically, the FCN is one of the first deep models that are used for image segmentation purposes. Following the experiment setup in [12], pretrained-VGG16 is used as the backbone. The Unet has been selected as a basic model in which our model is inspired. The MLP is a classical deep model that is used in [2] for BS, achieving very good performance. SegNet is a classical semantic segmentation model in computer vision field. SiU-net uses Siamese Unet to explore mutiscale information of images to enhance segmentation performance. MLP, FCN, Unet, SegNet and SiU-net are conducted on the same PC with implementation details strictly following the proposed model, while the results of SiU-net are directly duplicated from [24]. In addition, in order to evaluate the performance of the four considered models, the following metrics have been adopted:

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \text{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{20}$$

$$F_1 = \frac{2\times \text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}, \quad \mathrm{IoU} = \frac{\text{prediction} \cap \text{groundtruth}}{\text{prediction} \cup \text{groundtruth}}, \tag{21}$$

where TP stands for true positive rate and FN stands for false negative rate.

The segmentation performance comparison is reported in Table 1. As can be seen, the proposed HFSA-Unet outperforms FCN, MLP, SegNet and Unet on the four considered data sets in terms of F1-score and IoU, which are two important metrics for evaluation of the segmentation models with large margins. Moreover, the proposed model is also obviously superior to the latest model, i.e., Siamese Unet (SiU-net) on WHUBD. For example, on the WHUBD, the IoUs obtained by the FCN, MLP, Unet, SegNet and SiU-net are 85%, 83.56%, 87.66%, 85.10%, and 88.40%, respectively, while the IoU obtained by the proposed model is 90.72%. The improvement in terms of IoU over the other models is 5.72%, 7.16%, 3.06%, 5.62%, and 2.32%, respectively, which demonstrates the effectiveness of the proposed model. In addition, Figures 4–7 show a visual comparison between the baseline model and the proposed HFSA on the four test data sets. From Figures 4–7, we can observe that the proposed model can achieve better segmentation performance than the baseline model Unet. For example, from the first column of Figure 7, there are many bright non-buildings areas that are very similar to buildings in the images. Unet

[1)] The code of proposed model will be released on our homepage: http://www.escience.cn/people/henanjun/index.html.

Table 1 Comparison of segmentation results on four different data sets[a)]

Data	Method	Precision	Recall	F1-score	IoU
MBD	FCN [12]	77.22	71.19	73.96	58.75
	MLP [2]	75.26	76.69	75.87	61.20
	Unet [14]	81.76	77.45	79.36	65.95
	Unet[b)] [24]	68.10	74.60	–	55.20
	SegNet [16]	69.82	75.21	72.08	56.57
	HFSA-Unet	**84.75**	**79.08**	**81.75**	**69.23**
IBD	FCN [12]	88.15	88.47	88.29	79.07
	MLP [2]	85.46	87.88	86.62	76.43
	Unet [14]	91.57	86.08	88.68	79.72
	Unet[b)] [24]	84.60	82.10	–	71.40
	SegNet [16]	88.97	89.30	89.12	80.41
	HFSA-Unet	**92.30**	**89.89**	**91.07**	**83.63**
HNUBD	FCN [12]	72.94	71.14	72.03	56.29
	MLP [2]	68.50	67.74	68.12	51.66
	Unet [14]	76.01	66.83	71.13	55.19
	SegNet [16]	69.40	68.51	68.95	52.61
	HFSA-Unet	**76.31**	**71.65**	**73.90**	**58.61**
WHUBD	FCN [12]	91.25	92.56	91.89	85.00
	MLP [2]	90.84	91.25	91.04	83.56
	Unet [14]	94.73	92.15	93.42	87.66
	Unet[b)] [24]	90.03	94.50	–	86.80
	SegNet [16]	91.93	91.97	91.95	85.10
	SiU-net[b)] [24]	93.80	93.90	–	88.40
	HFSA-Unet	**95.09**	**95.18**	**95.13**	**90.72**

a) The best value is highlighted in bold. The average values over the whole data set are reported.

b) The results are directly duplicated from that paper, while others are implemented by ourselves.

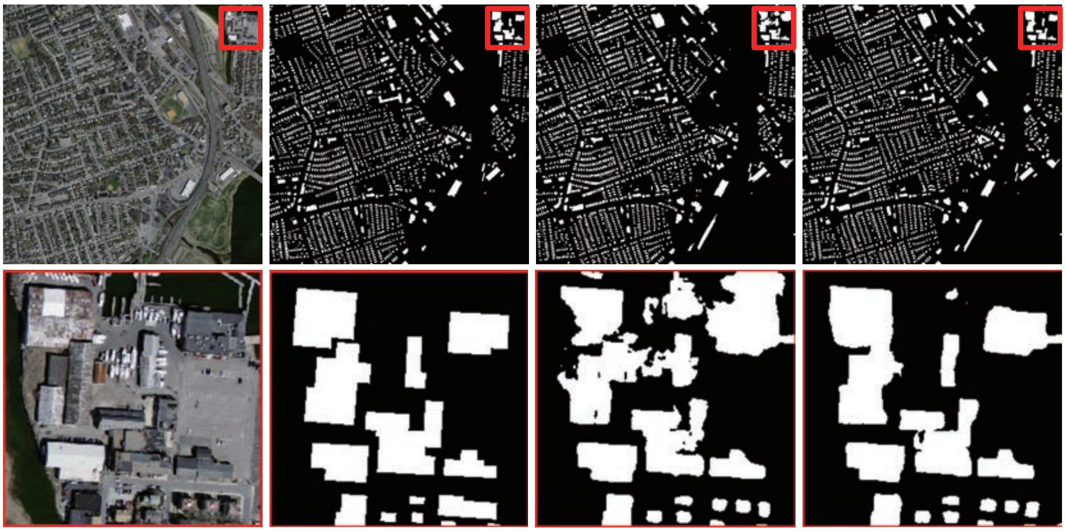

Figure 4 (Color online) Segmentation maps obtained by Unet and the proposed model of one representative image on the MBD dataset. From the first column to the last column, we display the input image, the ground-truth, the segmentation result obtained by Unet, and the segmentation result obtained by the proposed HFSA.

misidentifies them into building, while our proposed HFSA-Unet can still distinguish those areas. The main reason is that, the proposed method can effectively explore the correlations among intermediate layers, which enables it to learn more discriminative feature for building segmentation.

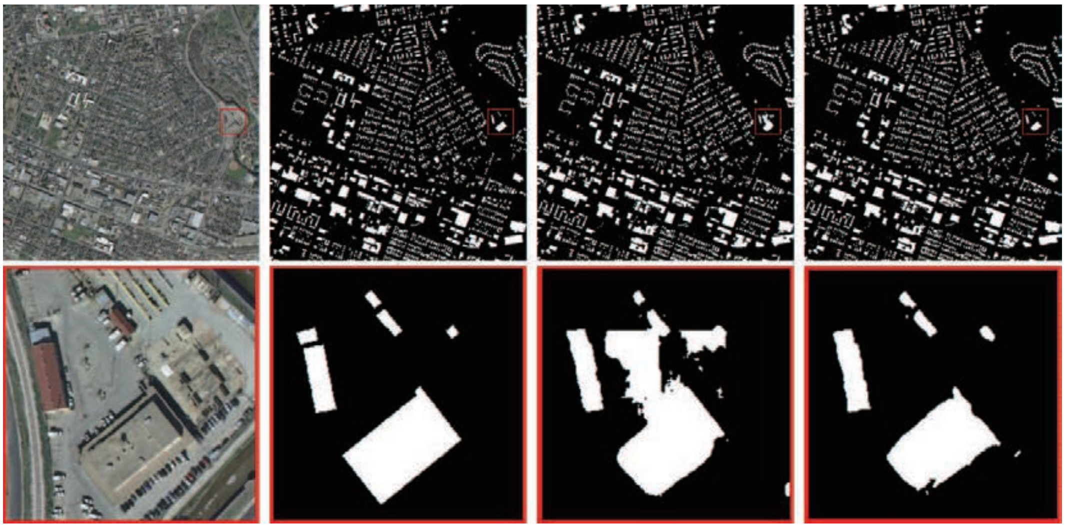

Figure 5 (Color online) Segmentation maps obtained by Unet and the proposed model of one representative image on the IBD dataset. From the first column to the last column, we display the input image, the ground-truth, the segmentation result obtained by Unet, and the segmentation result obtained by the proposed HFSA.

3.4 Ablation study

In order to further verify the effectiveness of proposed HFSA-Unet, we conduct a comprehensive alation study here. Specifically, we compare the proposed HFSA-Unet with the baseline as well as its two variants. Specifically, the baseline is original Unet. The first variant is the Unet sole with FOCA module, while the second variant is the Unet sole with SOCA module. The corresponding results are illustrated in Table 2, where '✓' means that module is adopted, while '–' means that module is discarded. From Table 2, the following two observations can be summarized. Firstly, introducing FOCA or SOCA into Unet can improve the segmentation performance. For example, on WHUBD, with FOCA, the IoU is improved from 87.66% to 89.12%, and with SOCA, the IoU is improved from 87.66% to 89.36%. Secondly, by using the FOSA and SOCA simultaneously, the IoU can be further improved from 89.12% and 89.36% to 90.72%, which verifies that the combination of FOSA and SOCA is necessary.

3.5 Generalization ability comparison

In practice, the generalization ability is curial to a deep model for automatical building segmentation. In this subsection, we evaluate the generalization ability of all the test models, including proposed HFSA-Unet via the transfer learning from aerial images to satellite images. Specifically, we train the model on the two representative aerial image data sets (i.e., MBD and WHUBD), respectively and then test the segmentation performance of the model on a satellite image data set (i.e., HNUBD). The quantitative segmentation results are shown in Table 3. As can be observed, the proposed HFSA-Unet shows much better performance than its counterpart in terms of F1-score and IoU. For example, when we train a model on the MBD, and then test the model on HNUBD, the F1-scores achieved by FCN, MLP, Unet and SegNet are 20.48%, 30.69%, 20.45%, and 22.44%, respectively, while the F1-score achieved by proposed HFSA-Unet is 35.88%. The average improvement is over 12%, which indicates that the proposed HFSA-Unet has more powerful generalization ability. This is mainly owing to the reason that, the proposed HFSA can make fully use of the correlation among the different channels of intermediate feature and thus enhance its generalization ability. However, we also would like to emphasize that the above segmentation performance is still much lower than the situation that the training set and test are from the same data set. In other word, cross-dataset or cross-domain building segmentation still remains a challenging problem and it would be a very interesting work in the future.

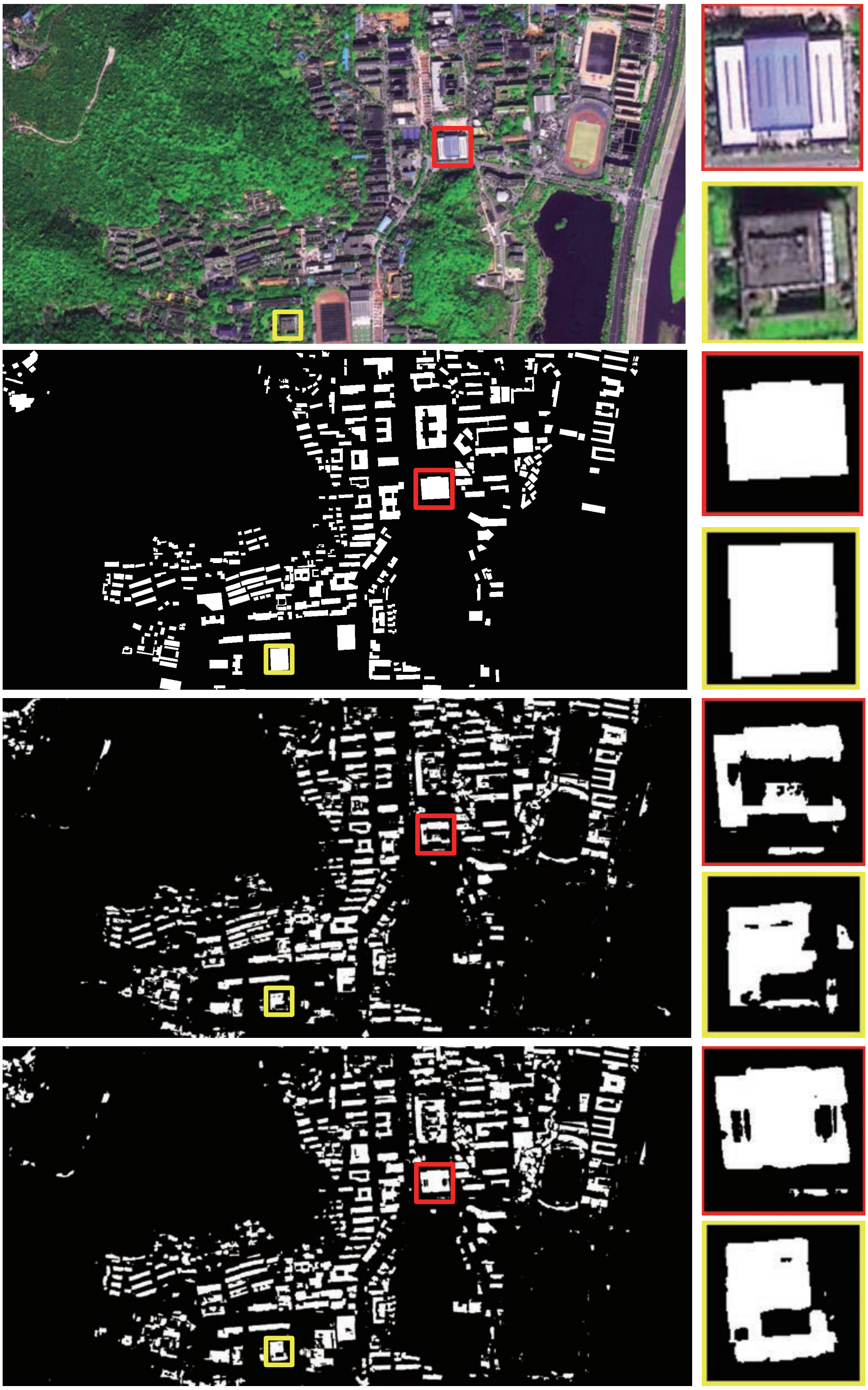

Figure 6 (Color online) Segmentation maps obtained by Unet and the proposed model on the HNUBD dataset. From the first line to the last line we display the input image, the ground-truth, the segmentation result obtained by Unet, and the segmentation result obtained by the proposed HFSA.

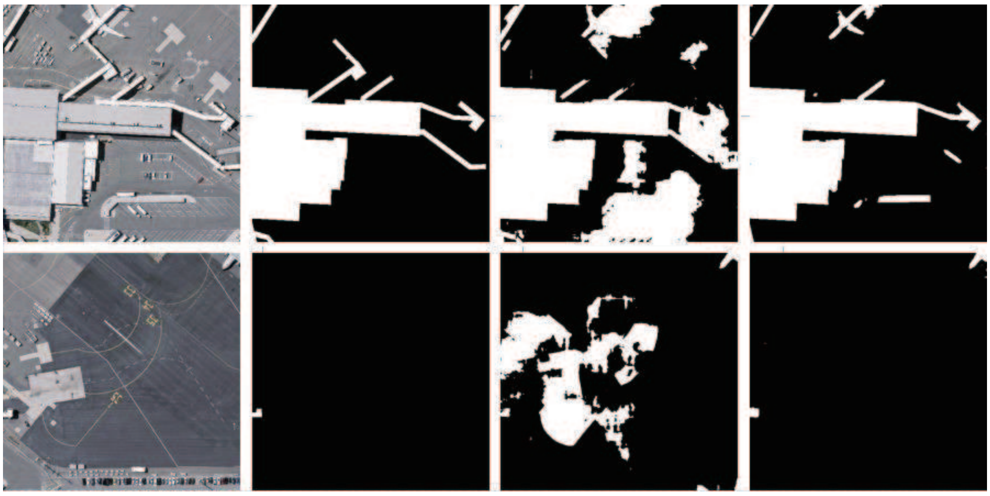

Figure 7 Segmentation maps obtained by Unet and the proposed model of two representative images on the WHUBD dataset. From the first column to the last column we display the input image, the ground-truth, the segmentation result obtained by Unet, and the segmentation result obtained by the proposed HFSA-Unet.

Table 2 Comparison of segmentation performance comparison between proposed model and its variants

Data	FOCA	SOCA	Precision	Recall	F1-score	IoU
MBD	–	–	81.76	77.45	79.36	65.95
	✓	–	82.80	77.92	80.19	67.02
	–	✓	80.84	80.92	80.69	67.79
	✓	✓	**84.75**	**79.08**	**81.75**	**69.23**
IBD	–	–	91.57	86.08	88.68	79.72
	✓	–	90.74	90.25	90.24	82.64
	–	✓	91.13	88.72	89.89	81.66
	✓	✓	**92.30**	**89.89**	**91.07**	**83.63**
HNUBD	–	–	76.01	66.83	71.13	55.19
	✓	–	76.21	69.21	72.54	56.91
	–	✓	75.52	69.32	72.28	56.60
	✓	✓	**76.31**	**71.65**	**73.90**	**58.61**
WHUBD	–	–	94.73	92.15	93.42	87.66
	✓	–	93.17	95.34	94.25	89.12
	–	✓	94.42	94.34	94.38	89.36
	✓	✓	**95.09**	**95.18**	**95.13**	**90.72**

Table 3 Generalization ability comparison via transfer learning from source dataset to (→) target dataset[a)]

Method	MBD→ HNUBD		WHUBD→ HNUBD	
	F1-score	IoU	F1-score	IoU
FCN [12]	20.48	11.41	20.91	11.59
MLP [2]	30.69	18.13	34.21	20.64
Unet [14]	20.45	11.38	39.26	24.42
SegNet [16]	22.44	12.64	37.37	22.97
HFSA-Unet	**35.88**	**21.86**	**42.04**	**26.62**

a)The metrics of F1-score & IoU are reported for evaluation.

4 Conclusion

In this paper, a new HFSA-Unet is proposed to explore the correlation among immediate layers for automatic building segmentation in remote sensing images. Specifically, the HFSA model explores both first- and second-order statistics to adaptively rescale the features, thus obtaining more representative and discriminative features. Our experimental results demonstrate that the newly developed HFSA network can not only achieve better segmentation performance than the baseline model, but also outperform several state-of-the-art approaches in the task of BS.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61922029, 61771192), National Natural Science Foundation of China for International Cooperation and Exchanges (Grant No. 61520106001), and Huxiang Young Talents Plan Project of Hunan Province (Grant No. 2019RS2016).

References

1 Jensen J R, Cowen D C. Remote sensing of urban suburban infrastructure and socio-economic attributes. Photogramm Eng Remote Sens, 1999, 65: 611–622

2 Yuan J. Learning building extraction in aerial scenes with convolutional networks. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 2793–2798

3 Liow Y T, Pavlidis T. Use of shadows for extracting buildings in aerial images. Comput Vision Graph Image Process, 1990, 49: 242–277

4 Ok A O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS J Photogrammetry Remote Sens, 2013, 86: 21–40

5 Inglada J. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS J Photogrammetry Remote Sens, 2007, 62: 236–248

6 Karantzalos K, Paragios N. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. IEEE Trans Geosci Remote Sens, 2009, 47: 133–144

7 Kim T, Muller J. Development of a graph-based approach for building detection. Image Vision Comput, 1999, 17: 3–14

8 Li E, Femiani J, Xu S, et al. Robust rooftop extraction from visible band images using higher order CRF. IEEE Trans Geosci Remote Sens, 2015, 53: 4483–4495

9 Yang H L, Yuan J, Lunga D, et al. Building extraction at scale using convolutional neural network: mapping of the united states. IEEE J Sel Top Appl Earth Observ Remote Sens, 2018, 11: 2600–2614

10 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105

11 Zhou Q, Wang Y, Liu J, et al. An open-source project for real-time image semantic segmentation. Sci China Inf Sci, 2019, 62: 227101

12 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 3431–3440

13 Wang W, Gao W, Hu Z Y. Effectively modeling piecewise planar urban scenes based on structure priors and CNN. Sci China Inf Sci, 2019, 62: 029102

14 Ronneberger O, Fischer P, Brox T. Unet: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Berlin: Springer, 2015. 234–241

15 Lu Y H, Zhen M M, Fang T. Multi-view based neural network for semantic segmentation on 3D scenes. Sci China Inf Sci, 2019, 62: 229101

16 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495

17 Geng Q C, Zhou Z, Cao X C. Survey of recent progress in semantic image segmentation with CNNs. Sci China Inf Sci, 2018, 61: 051101

18 Haut J M, Paoletti M E, Plaza J, et al. Visual attention-driven hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2019, 57: 8065–8080

19 He N, Fang L, Li S, et al. Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans Geosci Remote Sens, 2018, 56: 6899–6910

20 He N, Fang L, Li S, et al. Skip-connected covariance network for remote sensing scene classification. IEEE Trans Neural Netw Learn Syst, 2019. doi: 10.1109/TNNLS.2019.2920374

21 Lin T Y, Maji S. Improved bilinear pooling with CNNs. In: Proceedings of British Machine Vision Conference (BMVC), 2017

22 Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of Internation Conference of Computer Vision (ICCV), 2015. 1449–1457

23 Mnih V. Machine learning for aerial image labeling. Dissertation for Ph.D. Degree. Toronto: University of Toronto, 2013

24 Ji S, Wei S, Lu M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans Geosci Remote Sens, 2019, 57: 574–586

25 Maggiori E, Tarabalka Y, Charpiat G, et al. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, 2017. 3226–3229