

Improved object-based convolutional neural network (IOCNN) to classify very high-resolution remote sensing images

Xianwei Lv, Zhenfeng Shao, Dongping Ming, Chunyuan Diao, Keqi Zhou & Chengzhuo Tong

To cite this article: Xianwei Lv, Zhenfeng Shao, Dongping Ming, Chunyuan Diao, Keqi Zhou & Chengzhuo Tong (2021) Improved object-based convolutional neural network (IOCNN) to classify very high-resolution remote sensing images, *International Journal of Remote Sensing*, 42:21, 8318-8344, DOI: [10.1080/01431161.2021.1951879](https://doi.org/10.1080/01431161.2021.1951879)

To link to this article: <https://doi.org/10.1080/01431161.2021.1951879>



Published online: 02 Oct 2021.



[Submit your article to this journal](#)





[View related articles](#)



[View Crossmark data](#)



Improved object-based convolutional neural network (IOCNN) to classify very high-resolution remote sensing images

Xianwei Lv¹, Zhenfeng Shao¹, Dongping Ming², Chunyuan Diao³, Keqi Zhou² and Chengzhuo Tong⁴

¹State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China; ²School of Information Engineering, China University of Geosciences (Beijing), Beijing, China; ³Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA; ⁴The Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

The land cover classification of very high-resolution (VHR) remote sensing images is a challenging task. VHR images depict many complex objects with various shapes in complicated contexts. The deep learning-based method is a solution for such difficult task and feature extraction. Nevertheless, this method cannot efficiently deal with images with complex scene structures. An improved object-based convolutional neural network (IOCNN) is designed to classify VHR images with zone division and convolutional position sampling techniques in this study. The method can achieve the best performance of each zone at its own optimized scales. Based on multi-scale convolutional deep features extracted from VHR images, the objects with irregular shapes can be classified using the approach. In this study, the zone-level scale adaption and multi-scale recognition of complex objects are achieved. The performance of IOCNN is compared with the state-of-the-art methods for feature extraction, including five object-based CNN approaches and two fully convolutional networks (FCNs). The results show that the classification performance of IOCNN is considerably stronger than that of state-of-the-art methods. The overall accuracies of the land cover classification in IOCNN are 91.65% and 93.49% on two tested images. The results demonstrate the practicability of IOCNN.

ARTICLE HISTORY

Compiled 20 June 2021

1. Introduction

Land cover classification is an essential but arduous task in remote sensing applications, especially for very high-resolution (VHR) images. Conventional land cover classification approaches using object-based image analysis (OBIA) require image segmentation, feature selection, classifier training, and image classification, in which objects are meaningful segmented units in images (Blaschke (2010); Blaschke et al. (2014); Chen et al. (2018); Ma, Tengyu, and Manchun (2018)). Images can be classified by pre-trained classifiers (Bayesian

CONTACT Zhenfeng Shao  shaozhenfeng@whu.edu.cn  State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China

classifiers, random forests, and support vector machines) based on selected features, such as object textures, contexts, spectral information, shapes, and other morphological features (Ma et al. (2017); Mui, Yuhong, and Weng (2015); Sun et al. (2019)). To date, various methods have been developed and applied in diverse study areas, including urban areas (Fu et al. (2019); Shao et al. (2019a); Zhang et al. (2018a)) and natural areas (Chen et al. (2017); He et al. (2019a), (2019b))). However, classifying remote sensing images with complex objects and land surfaces becomes more challenging, as the image resolution becomes finer, especially for aerial and satellite images (Zhao and Du (2016)).

Deep learning has produced a new paradigm in the field of remote sensing (Ma et al. (2019); Heydari and Mountrakis (2019)). Fully convolutional networks (Ronneberger, Fischer, and Brox (2015); Long, Shelhamer, and Darrell (2015); Badrinarayanan, Kendall, and Cipolla (2017)) and convolutional neural networks (LeCun, Bengio, and Hinton (2015)) with multiple hidden layers have led to dramatic progress in remote sensing image applications, such as image fusion (Shao and Cai (2018); Song et al. (2018)), imagery registration (Hughes et al. (2018); Merkle et al. (2017); Shao et al. (2020b)), scene classification (Zou et al. (2015)), object detection (Zhong, Han, and Zhang (2018); Shao et al. (2019b), (2019c), 2020a)), land cover and land use classification (Liu, Zhang, and Eom (2016); Pan, Shi, and Xia (2018); Zhang et al. (2019); Zheng et al. (2020)), semantic segmentation (Maggiori et al. (2016); Zhang et al. (2017)), and OBIA (Liu and Abd-Elrahman. (2018); Zhao, Du, and Emery (2017)). Deep learning techniques have been particularly effective in land cover classification and OBIA, in which abstract deep features hidden in images can be extracted and fused. Deep features extracted by deep learning are more powerful than artificially designed features used in traditional classification approaches (Zhao and Du (2016); Zhang et al. (2018b)). FCNs with end-to-end structures can precisely predict labels of every pixel in images and have achieved progress in semantic segmentation. In the field of remote sensing, FCNs are often used to detect objects from remote sensing imageries (Wurm et al. (2019)). Researchers have attempted to use FCNs in hyperspectral image classification (Zheng et al. (2020)). In the field of CNNs, some previous studies have generated multi-scale features by combining single-scale features extracted by CNNs and subsequently demonstrated high classification accuracies and robustness (Zhao and Du (2016); Zhao et al. (2015)). Various CNN-based applications of remote sensing have attained progress by optimizing network structures (Chen, Zhao, and Jia (2015); Marcos et al. (2018); Zhu et al. (2018)), introducing multi-source data into VHR image processing (Pan, Shi, and Xia (2018); Chen, Huang, and Bing (2017); Zhang et al. (2017)), dynamic monitoring of time series (Cai et al. (2018)), and reducing the number of training samples (Pan, Shi, and Xia (2018)).

Although object-based CNNs (OCNNs) has greatly improved the overall accuracy of the land cover classification of VHR images (Mui, Yuhong, and Weng (2015); Fu et al. (2018); Huang, Zhao, and Song (2018); Lv et al. (2019), Lv et al. (2018)); Tong et al. (2020); Zhou et al. (2020)), OCNNs still suffer from performance issues at fine scales. Specifically, the classification accuracy at the object and zone levels can vary greatly.

At the object level, Most OCNN methods fail to consider the influence of the convolutional positions (centre points of convolutional windows or respective fields) on classification performance. Moreover, the research focusing on the generation strategies of convolutional positions is rare. Figure 1 illustrates how CNNs use convolutional positions to extract sub-image blocks as the basic units.

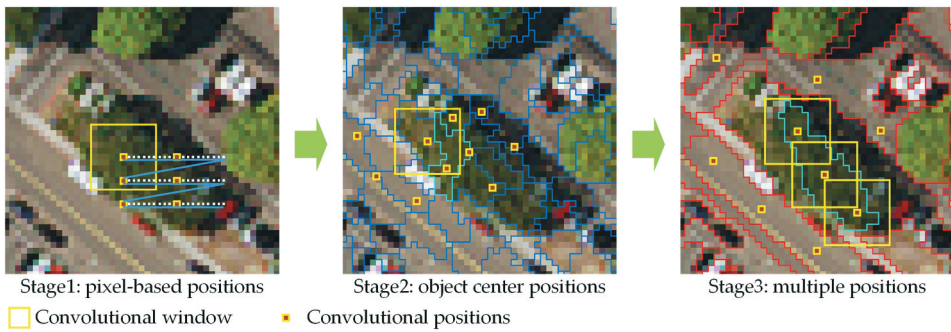


Figure 1. The convolutional positions in different CNN methods for remote sensing image classification.

Selection strategies for convolutional positions have changed over time along with the advancements in CNN-based methods. The pixel-based CNN considers pixels as convolutional positions and requires massive runtime and a huge amount of storage space (Zhao and Du (2016)), while the superpixel-based CNN utilizes the centres of super-pixels (a segmentation result with familiar object shapes) as the convolutional positions (Lv et al. (2018)). Object-based CNN can further reduce the number of convolutional positions by utilizing multiple convolutional positions within objects (Zhang et al. (2018b); Lv et al. (2018); Zhou et al. (2020)).

However, existing strategies for selecting convolutional positions are mostly unreliable because of the uncertainty of sample locations. Some research directly takes centroids of objects as convolutional positions, regardless of the shape of the objects (Fu et al. (2018); Lv et al. (2019); Zhang et al. (2020); Chen, Ming, and Xianwei (2019)). In other research, randomly generated positions in objects are used as the convolutional positions (Lv et al. (2018)). To address this issue, some object-based morphological methods are proposed to generate multiple convolutional positions in objects to address this issue (Zhang et al. (2018b)). Nevertheless, these existed morphological methods still cannot represent the whole objects. The generation methods cannot select reasonable and suitable convolutional positions for OCNNs. An ideal algorithm for convolutional position sampling needs to satisfy the following conditions: the number of positions within an object should be precisely controlled by users; the positions cannot be located on the boundary of the object; and the positions are expected to be evenly distributed within the object. Thus, a binary tree sampling method (the process of sampling is the process of selecting convolutional positions) based on object morphological attributes is used to solve the problem in this study.

The essence of the zone-level issue is the problem of scale effects. The zone-level issue is that CNN-based methods typically study remote sensing images at a single scale (objects in industrial, agriculture, and residential scenes are often studied at the same scale in a remote sensing image), ignoring the hierarchical and stratified natures of geographic phenomena in VHR images. Besides, the scale effect remains a challenge in OCNNs. The scale effect in remote sensing image classification implies that the classification results tend to vary considerably at different scales (window sizes of the study). Objects vary widely in terms of size and have different optimized scales. Scale effect at

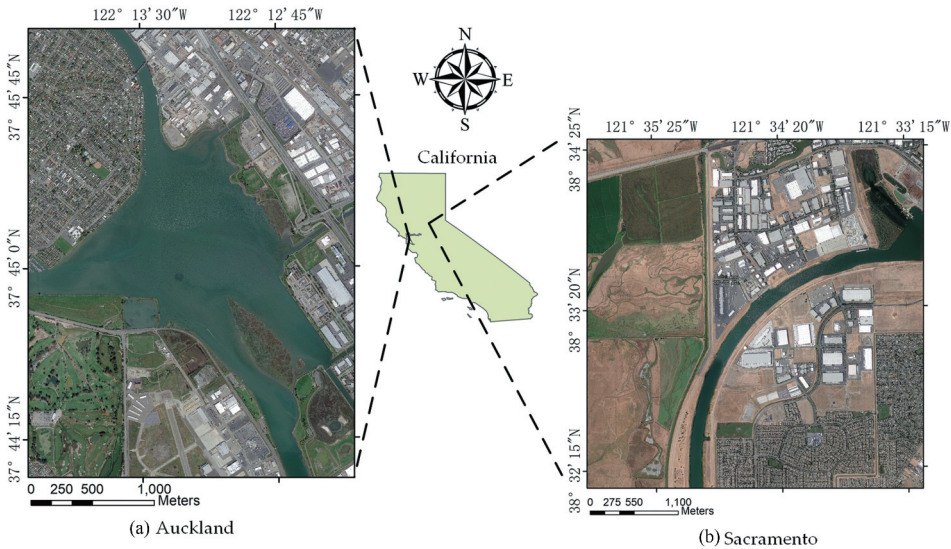


Figure 2. Studied VHR images.

zone-level is a result of the interaction of the object-level scale effect and the spatial distribution of the objects in the image. Given this condition, OCNNs cannot achieve the desired classification performance of remote sensing images by using single scales. In contrast, multi-scale methods can precisely identify objects. Nevertheless, these methods lead to a compromise between different optimized scales (Lv et al. (2018)). According to Tobler's First Law of Geography (everything is related to everything else, but near things are more related to each other), geographical objects or attributes are related to one another. In areas where the spatial distribution of geographical objects is heterogeneous, the scale effects at different zones can be diverse and the optimized scales for each zone may vary. Therefore, zone division is necessary and can be used in VHR images to achieve zone-level scale-adaptive classifications.

This study aims to design a land cover classification method to address the issues existing in OCNNs. In particular, the method aims to achieve the following goals: 1) zone division is proposed for the pre- and post-processing of VHR image land cover classifications; 2) a binary tree sampling (BTS) method is applied to select the appropriate convolutional positions; 3) OCNN are combined with the zone division and the BTS to devise a land cover classification method for VHR images.

2. Study area

Two study areas in California, namely, San Leandro Bay suburbs on the west coast of Auckland (Figure 2a) and Washington Lake located on the southwest of Sacramento (Figure 2b), were selected for this study.

The two VHR images were captured on 17 August 2018 (Auckland) and 3 April 2018 (Sacramento). Both images were downloaded from Google Earth with a spatial resolution of 0.6 metres and three bands. The Auckland image contains $8,185 \times 7,850$ pixels, while the Sacramento image contains $4,960 \times 6,971$ pixels. There is obvious spatial aggregation

Table 1. The dominant categories of geographical objects in each zone.

Zone	Auckland	Sacramento
Residence	Asphalt roads, cement roads, Residential buildings, shadow, vegetation	Asphalt roads, bare soil, cement roads, Residential buildings, shadow,vegetation, water bodies
Industry	Asphalt road, parking space, Residential buildings, factory buildings, shadow, vegetation	Asphalt roads, bare soil, cement roads, factory buildings, shadow, trucks, vegetation, water bodies
Nature	Asphalt roads, vegetation, water bodies, wetlands	Asphalt roads, bare soil, vegetation, water bodies, wetlands

Table 2. The training and validation data for two studied images.

Category	Auckland		Sacramento	
	Training	Validation	Training	Validation
Asphalt roads	1,018	582	1,254	746
Bare soil	–	–	996	604
Cement roads	746	454	1,002	598
Factory buildings	985	615	978	620
Parking space	903	597	–	–
Residential buildings	1,011	589	973	627
Shadow	993	607	1,128	672
Trucks	–	–	929	543
Vegetation	895	505	1,145	655
Water bodies	1,010	590	1,001	599
Wetlands	314	186	445	247

phenomenon of geographical objects in the two images. Each image shows three types of zones including residential, natural, and industrial zones. There are clear boundaries among these zones. In addition, the categories of objects in zones vary greatly. The dominant categories of different zones in the study areas are shown in [Table 1](#).

The samples were selected from manually labelled data by using a stratified random scheme. As recommended by previous researchers (Zhang et al. (2018b); Chen et al. (2016)), the samples should be divided into two sets, namely, a training set containing 62.5% of all samples and a validation set containing the remaining 37.5% samples. Approximately 1,600 samples were detected per category for training and validation, which is sufficient to train the CNN models. The sample size of each category is shown in [Table 2](#).

A total of 12,600 samples (7,875 training samples and 4,725 validation samples) in the Auckland image and 15,762 samples (9,851 training samples and 5,911 validation samples) in the Sacramento image were obtained. The sizes of each category are listed in [Table 2](#). Besides these samples for training and validation, 3,884 and 4,813 samples were also collected independently for the two Auckland and Sacramento to supplement the accuracy assessment. The examples of the categories in the two images are shown in [Figure 3](#).

3. Methods

The workflow of IOCNN for the VHR image land cover classification is illustrated in [Figure 4](#). IOCNN is a supervised classification method. The three key components of IOCNN are pre-processing (zone division, training sample selection, and testing data



Figure 3. Examples of land cover categories in each image.

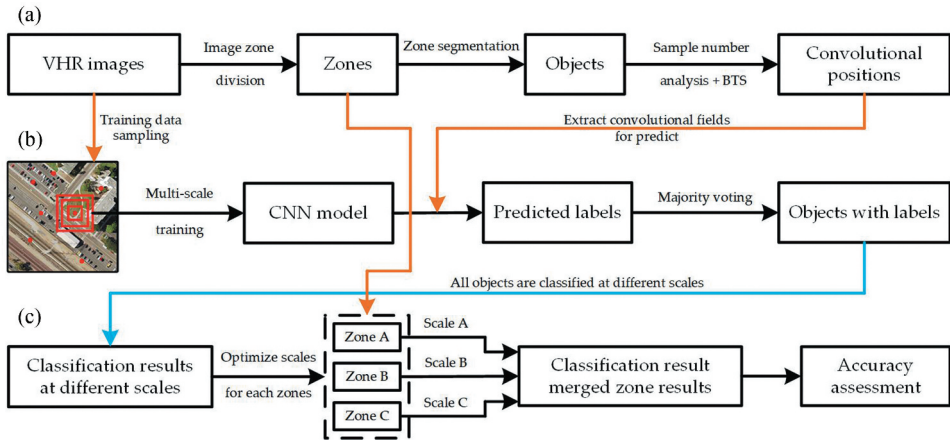


Figure 4. IOCNN workflow for VHR image land cover classification.

selection fusing BTS (Figure 4a), training CNN models and predicting object labels by the CNN models (Figure 4b) with the aid of a majority-voting system, and merging optimal multi-scale results into a final result (Figure 4c).

Zone division and BTS methods are the contributions of the proposed IOCNN. In the pre-processing step, an image is segmented into several zones, each with different land cover types, by using the image stratified zone division method (Xu et al. (2019)). Each zone is subsequently segmented into objects in the form of highly homogeneous sub-image units. The number of convolutional positions for each object is computed based on the geometric properties of the object. BTS is then used to select the convolutional positions from each object. With the generated convolutional positions in the pre-processing phase by BTS, the convolutional positions can then be labelled by the pre-trained CNN and provided multiple labels. Then, the objects can be classified by using the majority-voting scheme for the multiple labels. Finally, the classification results of each zone at their optimized scales are merged into a final classification map.

3.1. Image stratified zone division

Thanks to the inspiration about zone division (Zhou et al. (2018a)), the multi-resolution segmentation (MRS) method is used in zone division and object generation. The VHR image segmentation entailing multiple land cover types requires a suitable scale parameter, which is determined by the size of the smallest objects as studied of interest in the study area (Ming et al. (2015); Ma et al. (2015)). Thus, most objects in the studied areas are over-segmented in terms of its scale parameter. However, there are diverse scale effects in different zones. In other words, it is necessary to avoid to set the scale parameter to accommodate the smallest objects. Therefore, the image stratified zones division method is essential for the image segmentation (Xu et al. (2019); Zhou et al. (2018b)). The image zone division method involves colour space conversion, grey level co-occurrence matrix, and MRS. The flowchart of the image zone division is shown in Figure 5.

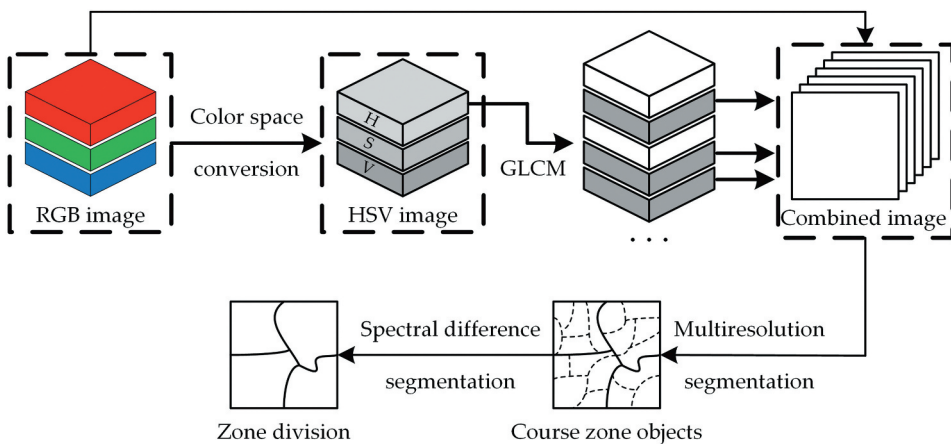


Figure 5. Flowchart of the image zone division.

Remote sensing images can be divided into dominant zones, such as natural, industrial, and residential zones. With the colour space conversion method, an RGB image can be initially converted into an HSV image. The hue band of the HSV image which can differentiate different objects is then used with the grey level co-occurrence matrix (GLCM) to generate texture feature maps. The original RGB image is subsequently stacked with several texture maps optimized by their obvious zone boundaries to form a new image. MRS is applied to the zone division of the stacked image with the optimal scale parameter, appropriate compactness, and shape values. After MRS segmentation, the coarse zone objects can be generated. They are slightly over-segmented for zone-division and under-segmented for ground objects. Spectral difference segmentation is applied to the result of MRS to merge objects with similar spectral and related features. The parameters of the spectral difference segmentation are set as eCognition.

3.1.1. Colour space conversion

The first step of the zone division is to convert the image from the RGB model to the HSV model to improve the coverage information of the GLCM. Conventionally, images are presented using the RGB model in computer vision. RGB, presenting the three primary colours, can directly generate almost all human visible colours, but it has no direct relation with the three attributes of hue, saturation, and value. The HSV model can directly show the relationships between colours and each of the attributes of three bands, and provides better results for image classification compared with the RGB model (Xu et al. (2019); Rabiee, Kashyap, and Rasoul Safavian (1996)). For example, the green objects (vegetation) in the study area usually have different values in the RGB model. Similar colour spectra in RGB may correspond to different objects. After colour space conversion, these green colours in the RGB model can effectively match the similar value of the hue band. The hue band can identify image information from the HSV model. Furthermore, GLCM is necessary to generate all the other required information for the zone division method in this study.

3.1.2. Gray level co-occurrence matrix (GLCM)

Texture is formed by the repeated occurrence of grey level distribution in spatial position. GLCM is a widely used texture analysis approach, with the assumption that spatial distributions of pixels in images contain image texture information (Haralick and Shanmugam (1973)). GLCM is defined as the joint probability distribution of two grey pixels at distance d . The matrix approach can be directly used as a feature to distinguish textures; however, statistical texture features derived from GLCM, including feature parameters, mean, homogeneity, entropy, are generally preferred. GLCM is only used to improve the quality of dividing zone and not used to train the CNN model .

3.1.3. MRS

MRS is used to perform pixel-level bottom-up region-growing segmentation to identify objects and their hierarchical structures (Rabiee, Kashyap, and Rasoul Safavian (1996)). The segmentation follows a minimum heterogeneity principle, i.e. only adjacent pixels similar in the spectra are united and rendered with the same labels. The feature maps (mean,

entropy, or homogeneity) generated by GLCM method from the original RGB images are merged into a new image for zone division. Then, the original image in each zone is subsequently segmented into objects by MRS based on the above divided zones. These objects are then analysed and assigned convolutional positions, as discussed in the following section.

3.2. Binary tree sampling (BTS)

BTS, as a contribution of IOCNN, is applied in this study to generate appropriate convolutional positions based on the idea of binary trees. The BTS method is applied only to the testing phase in the study. Some terminologies need to be clarified before the description about BTS. The fundamental operation of BTS is to clip an object into two new objects. The original object before the clip is called the 'parent object', while the two new objects after the clip are called the 'child objects'. A root object can be a parent object, but not a child object.

3.2.1. Convolutional position number analysis

The number of convolutional positions for a root object depends on its area and shape index. Extremely large or small objects are rare. For extremely small objects, one position is sufficient. As the area increases, the number of positions also increases, although the two are not linearly related. As the internal of large areas tend to be homogeneous, the number of positions should no longer increase when the area reaches a certain level (a user-defined threshold). Thus, the object area can be divided into three intervals: small area, middle area, and large area.

The shape index refers to the smoothness of an image object's border and is defined as the perimeter of the object divided by four times the square root of its area (Lv et al. (2019)). The smoother the border of an object is, the lower its shape index will be. Moreover, the more winding the border of an object is, the higher its shape index will be. Objects with low shape index require few convolutional positions. By contrast, objects with high shape index need more convolutional positions. Similar to how the area is handled, the shape index of an object can be divided into three intervals based on a user-defined threshold.

A nine-intersection model is used to calculate the number of convolutional positions in the root object based on the double-limitation strategy, namely, the area and shape index. Then, on the basis of the three intervals defined by the area and the shape index of the object, the convolutional position number can be calculated using a coordinate system corresponding to the sample number analysis theory (Figure 6).

The BTS method clips a parent object into two child objects and recursively clip these child objects until the point number meets the requirements specified by Equation 1, forming a binary tree.

$$S(n) = \begin{cases} P_{s1} & , n = 1 \\ S(\lceil \frac{1}{a} n \rceil) + S(\lfloor \frac{a-1}{a} n \rfloor) & , n > 1 \end{cases} \quad (1)$$

where n is the number of convolutional positions in the root object and a is the area of a child object divided by the area of its parent object. In addition, $S(n)$ is a kernel function for clipping the root object. The BTS method clips an object using a divide-and-conquer

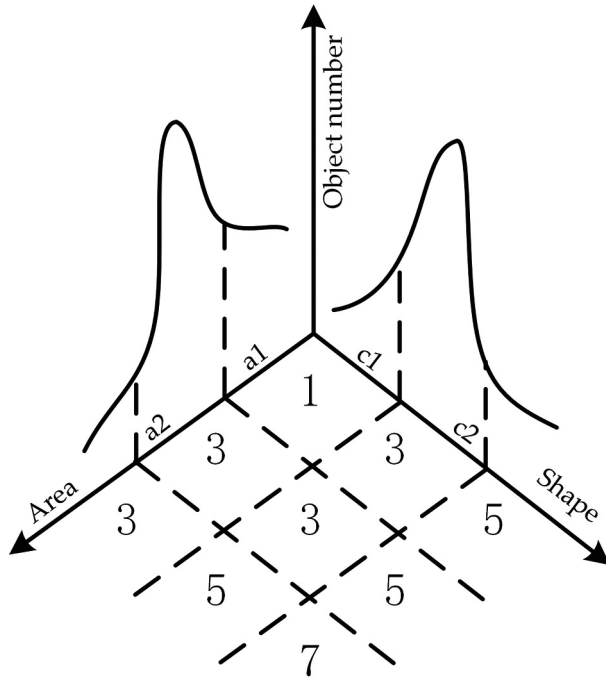


Figure 6. Theory of position number analysis.

strategy. The kernel function $S(n)$ returns the centroid P_{s1} of the child object when n is 1. The centroids of the inseparable child objects are used as the convolutional positions.

3.2.2. Binary tree-based clipping algorithm

The workflow contains two steps. First, a parent object is clipped into two child objects by the minimum bounding rectangle (MBR) and the same step is recursively applied to each child object until the required number of samples is reached. Second, centroids of child objects are extracted and considered as the candidates of the convolutional positions. The

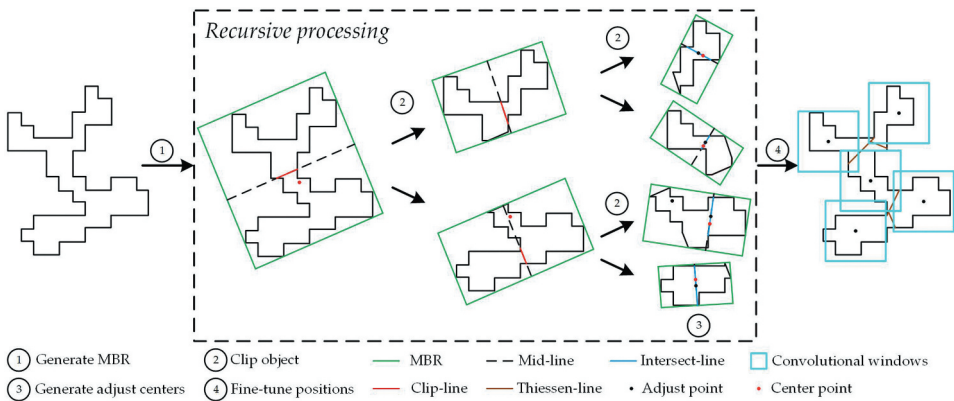


Figure 7. Workflow of the BTS method.

locations of the candidates are then fine-tuned to generate the final positions of the convolutional positions. The flowchart of the BTS method is shown in Figure 7.

Fine-tuning is necessary for two reasons: 1) the candidate positions can be too close to the boundaries of the objects; and 2) in certain cases, the candidates may not be evenly distributed in the objects. To address these issues, the fine-tuning is an iterative process. In each iteration, Voronoi polygons are generated based on the positions of the candidates and the boundaries of the objects. The centroids of the resulting Voronoi polygons are considered as the new candidates. This process is repeated on the candidates for five times.

3.3. Convolutional neural networks (CNNs)

A CNN contains multiple layers including convolutional layers and pooling layers (LeCun, Bengio, and Hinton (2015)). This complex structure enables CNNs to extract deep features from VHR images. Thus, the segmented objects are identified based on the the deep features.

The single CNN model used in the study adopts the same structure as AlexNet with five convolutional layers, three max-pooling layers, and two fully connected layers (Krizhevsky, Sutskever, and Hinton (2012)). Figure 8 shows the detailed structure of the CNN. The training data is reshaped into a 227×227 matrix as the input layer, which is subsequently transformed into five convolutional layers (conv1 to conv5 in Figure 8) with different filters. Similarly, the testing data extracted by BTS is fed into the pre-trained model. The number and size of the filters used for each convolutional layer are shown in Figure 8. All max-pooling layers use the same filter with a window size of 3×3 . The function rectified

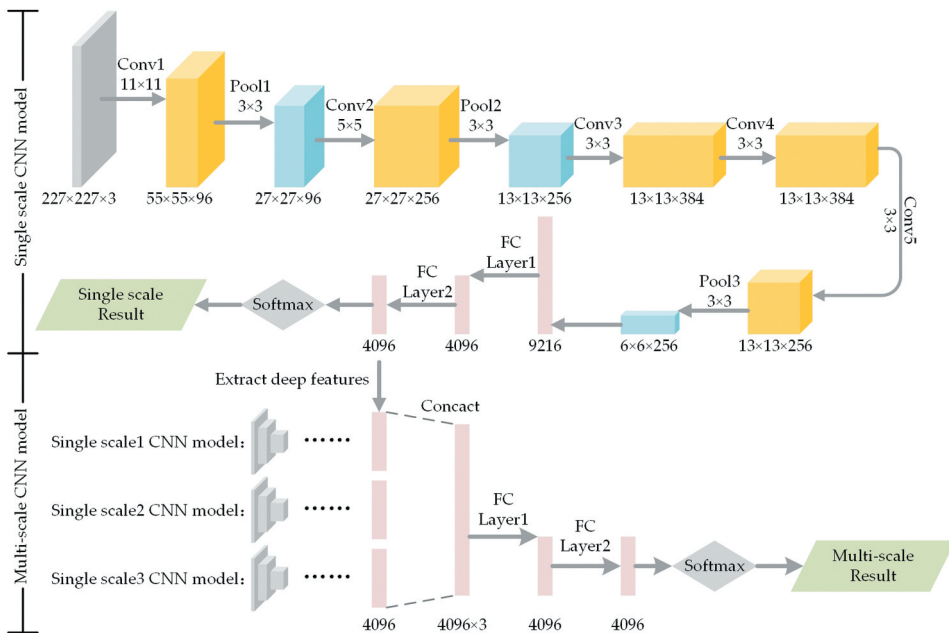


Figure 8. Framework of CNNs in this study.

linear unit is set as an activation function. At the end of this convolutional structure, softmax function is selected as the classifier. As opposed to the single-scale CNNs that are based on single-window sizes, multi-scale CNNs (MCNNs) are based on feature fusion of multiple single-scales. The deep features from two or three single-scales with the highest overall classification accuracy (OA) are combined as multi-scale features. Then, the fully connected layers are retained based on the multi-scale features.

3.4. Accuracy assessment

A confusion matrix is used to assess the accuracy of the VHR image land cover classification. Confusion matrices provide a comprehensive means such as *kappa* coefficient (Mez (1978)), overall accuracy (OA), and *f1*-score (*f1*) to evaluate the performance of land cover classification. The N categories of objects in the classification results represent and validate the proportion of objects correctly matched to the ground truth in an $N \times N$ matrix. OA, denoted as P_0 , is the proportion of correctly classified objects. For a total of n samples, if the number of correctly predicted samples is S , then P_0 is S/n . Here, $PA = TP/(TP + FN)$ and $UA = TP/(TP + FP)$, where PA is product accuracy, UA is user accuracy, TP is the number of positive samples to be correctly classified, FN is the number of positive samples misclassified into be negative, and FP is the number of negative samples misclassified into be positive. The *kappa* is described in Equation 2 and 3.

$$kappa = \frac{P_0 - P_e}{1 - P_e} \quad (2)$$

where P_e is calculated by

$$P_e = \frac{\sum_{i=1}^c a_i b_i}{n \times n} \quad (3)$$

where a_1, a_2, \dots, a_c is the actual number of samples in each category, and b_1, b_2, \dots, b_c is the predicted number of samples for each category (Mez (1978)).

The equation to calculate *f1* is shown in Equation 4.

$$f1i = 2 \times \frac{U Ai \times P Ai}{U Ai + P Ai} \quad (4)$$

where i is the i th category, and $f1i$ is the *f1* score of the i th category.

3.5. Comparative methods

The control methods are applied to the same images to evaluate the performance of the land cover classifications of the proposed IOCNN. Five OCNN methods including OCNN₁ (Zhang et al. (2018b)), OCNN₂ (Lv et al. (2018)), OCNN₃ (Zhou et al. (2020)), OCNN₄ (Fu et al. (2018); Lv et al. (2019); Zhang et al. (2020); Chen, Ming, and Xianwei (2019)), OCNN₅ (Tong et al. (2020)) are selected as the comparative methods. In addition, because of the advance of FCNs for semantic segmentation in computer vision in recent years. On this basis, U-Net (Ronneberger, Fischer, and Brox (2015)) and SegNet (Badrinarayanan, Kendall, and Cipolla (2017)) are also selected as the comparative methods in this study.

3.6. Study environment and configuration

The MRS method in eCognition is adopted in this study. The deep learning library TensorFlow is used to run the CNNs on a desktop computer equipped with an NVIDIA RTX2080TI GPU, an Inter(R) Core i7-7820X CPU, and 32 GBs of RAM. The BTS method is performed on the same computer.

For training CNN models, the dropout value is set to 0.5. The learning rate value is 0.01. The number of epochs is set to 60 based on trial and error; that is the model is expected to converge within only 60 epochs. The Gradient Descent Optimizer is used in IOCNN and the comparative methods in the study. The methods are tested at a single scale. Then, an OCNN method and IOCNN are used to make a series of classifications at multiple scales. The scale is selected relatively to the size of ground objects in the study area. Objects vary widely in terms of size. For example, the biggest ground object (an industry building) in our image covers approximately 4100 m². However, the smallest artificial object (tracks and a few cement roads) covers about 50 m² shown in Figure 3. 3. Therefore, we select 25 pixels (about 15 metres) as the side length of the smallest window, and 105 pixels (about 63 metres) as the side length of the largest window. We believe that they can contain enough background information for small and large ground objects. Because there are still a large number of ground objects whose size is between the smallest objects (trucks and cement roads) and largest ground objects (factories and bare soil), we select a series of window sizes, i.e. (25 × 25, 35 × 35, 45 × 45, 55 × 55, 65 × 65, 75 × 75, 85 × 85, 95 × 95, 105 × 105), reshaped into a 227 × 227 matrix as input layer. The same parameters, including learning rate and dropout value, are used for the CNN and MCNN. The epoch number for training the MCNN model is set to 100, which is the only difference in the parameter setting between the single CNN model and the MCNN model. For training FCN models, nine and eight labelled patches with 585 × 585 pixels are used as the training samples of the Sacramento and Auckland images, respectively. The training data are also applied with concepts from the augmentation strategy. The iteration number of training models is 5,000, and the learning rate is 0.001. The accuracy assessment data of the FCNs are the same as OCNNs'.

4. Classification results and analysis

The classification performance of the proposed IOCNN is evaluated on two images. The classification results and the divided zones are analysed for the scale effects. The proposed method is compared with five OCNN methods and two FCN methods. The classification results are evaluated by *OA*, *kappa*, and *f1*. The proposed IOCNN achieves desired land cover classification results. Many classifications are also conducted at different scales, such as single scales and multiple scales (i.e. double and triple scales). The multi-scale results are the feature fusion of single scales. The combinations of multi-scales with the highest accuracies are presented in this paper.

4.1. Single scale classification results

The OAs for each OCNN method at single scales for Sacramento and Auckland are shown in Figure 9, respectively. The IOCNN achieves the highest classification accuracies

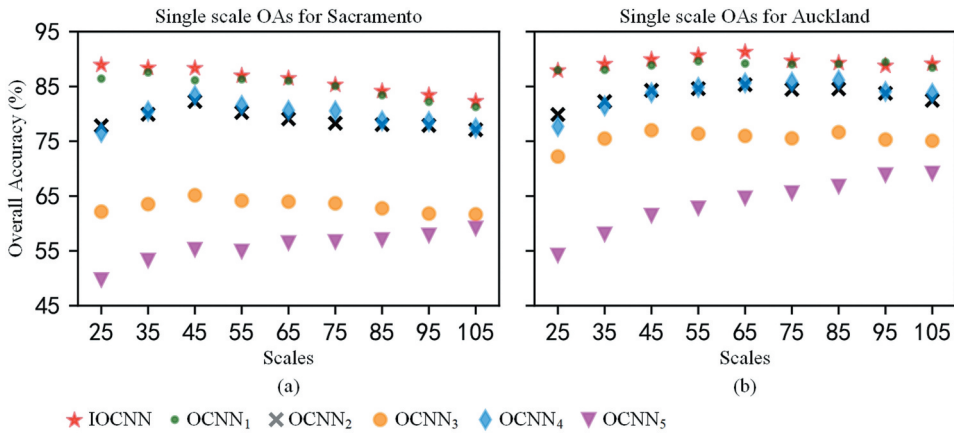


Figure 9. Single scale classification results.

Table 3. F1 of per category for Sacramento.

Results	IOCNN	OCNN ₁	OCNN ₂	OCNN ₃	OCNN ₄	OCNN ₅	U-Net	SegNet
Optimized scales	25	35	45	45	45	105	–	–
OA(%)	88.93*	87.58	82.26	65.16	83.34	59.13	72.60	73.58
kappa	0.8783*	0.8585	0.7978	0.6133	0.8104	0.5338	0.6863	0.6961
Average f1	0.8860*	0.8700	0.8219	0.7133	0.827	0.6036	0.6888	0.6661
Asphalt roads	0.8844*	0.8831	0.8304	0.8028	0.8417	0.6167	0.8046	0.8058
Bare soil	0.9424*	0.9403	0.9191	0.8905	0.9087	0.8129	0.9104	0.9127
Cement roads	0.8101*	0.7906	0.7279	0.5312	0.7409	0.3952	0.4875	0.5573
Factory buildings	0.8968	0.9091*	0.8988	0.7897	0.8812	0.8233	0.7564	0.6168
Residential buildings	0.8703	0.8898*	0.8348	0.6431	0.8585	0.5265	0.7563	0.4290
Shadow	0.9103*	0.8580	0.7659	0.5803	0.7759	0.3547	0.5441	0.8580
Trucks	0.9249*	0.8661	0.8067	0.5234	0.7980	0.4625	0.3149	0.2540
Vegetation	0.8643*	0.8050	0.7220	0.6243	0.7692	0.4319	0.6861	0.7631
Water bodies	0.9566	0.9672*	0.9157	0.9178	0.9350	0.7905	0.8324	0.8224
Wet land	0.8000*	0.791	0.7976	0.8295	0.7609	0.8221	0.7956	0.6421

Table 4. F1 of per category for Auckland.

Results	IOCNN	OCNN ₁	OCNN ₂	OCNN ₃	OCNN ₄	OCNN ₅	U-Net	SegNet
Optimized scales	65	55	65	45	85	105	–	–
OA(%)	91.30*	89.60	85.32	77.01	86.15	69.16	72.39	80.45
kappa	0.8933*	0.8727	0.8198	0.7250	0.8304	0.6306	0.6594	0.7527
Average f1	0.8953*	0.8738	0.8227	0.7892	0.8282	0.6733	0.7032	0.7780
Asphalt roads	0.9293*	0.9197	0.8970	0.8736	0.8990	0.7666	0.8007	0.8489
Cement roads	0.8163*	0.7989	0.6779	0.6069	0.6494	0.2632	0.3326	0.6748
Factory buildings	0.8872*	0.8476	0.8545	0.8064	0.8516	0.7770	0.7059	0.6400
Residential buildings	0.9395*	0.9272	0.8927	0.7759	0.9081	0.7233	0.7238	0.7177
Shadow	0.9068*	0.896	0.7769	0.7506	0.8320	0.5332	0.5861	0.8941
Vegetation	0.8457*	0.7992	0.7056	0.7991	0.7052	0.5539	0.6965	0.8104
Water bodies	0.9670*	0.9670	0.9516	0.9605	0.9223	0.9060	0.9308	0.9242
Wet land	0.8540*	0.8172	0.7955	0.8316	0.8132	0.8670	0.8243	0.7644

compared with other state-of-the-art methods. The IOCNN reaches the best classification result at the optimized scale of 25 for Sacramento (OA: 88.93%, kappa: 0.8783, f1: 0.8860) and the optimized scale of 65 for Auckland (OA: 91.30%, kappa: 0.8933, f1: 0.8953), respectively. OCNN₁, OCNN₂, OCNN₃, OCNN₄, and OCNN₅ get the best classification at

the optimized scales of 35, 45, 45, 45, and 105 for Sacramento, and optimized scales of 55, 65, 45, 85, and 105 for Auckland, respectively.

Among the eight methods, the IOCNN and OCNN₁ methods produce the desirable classification results. The classification results of OCNN₃ and OCNN₅ are the worst as shown in the Figure 9. The classification results for each category of these OCNN methods at their optimized scales for the two studied images are shown in Tables 3 and 4, respectively. In addition, the details of the two FCN methods (U-Net and SegNet) are also shown in the two Tables. The highest *OA*, *kappa*, average *f1* of each method and *f1* for each category are marked with stars and in bold.

The classification results of factory buildings, residential buildings, and water bodies of OCNN₁ are more effective than IOCNN in Sacramento. The *f1*s of the three categories between IOCNN and OCNN₁ are comparable. Nevertheless, the highest *f1* of almost all categories in Sacramento and all categories in Auckland are all achieved by using IOCNN. IOCNN not only performs better than other state-of-the-art OCNNs, but also works more effectively than U-Net and SegNet. Even though the U-Net and SegNet made a progress in semantic segmentation in computer vision, they cannot work effectively in VHR image classification, which is determined by the attributes of VHR remote sensing images with large area.

4.2. Multi-scale classification results

Multi-scale classifications are conducted based on the single scale CNN models. IOCNN and OCNN₁ achieve the desirable single scale classification results so that only these two methods are only selected in the multi-scale study. The scale selection for scale combination is based on the single-scale classification accuracies on the study areas. We selected the top six scales in accuracy for combining scales. The top six scales in Sacramento are 25, 35, 45, 55, 65, and 75 shown in Figure 9 in the manuscript. The top six scales in Auckland are 35, 45, 55, 65, 75, and 105. Based on the classification results of the two methods at single scales, the scale combinations of 25–35, 35–45, 45–55, 55–65, 65–75, 25–35–45, 35–45–55, and 55–65–75 are conducted in Sacramento, and the scale combinations of 34–45, 45–55, 55–65, 65–75, 75–105, 34–45–55, 35–45–65, 45–55–65, and 65–75–105 are conducted in Auckland. The multi-scale classification results are shown in Figure 10.

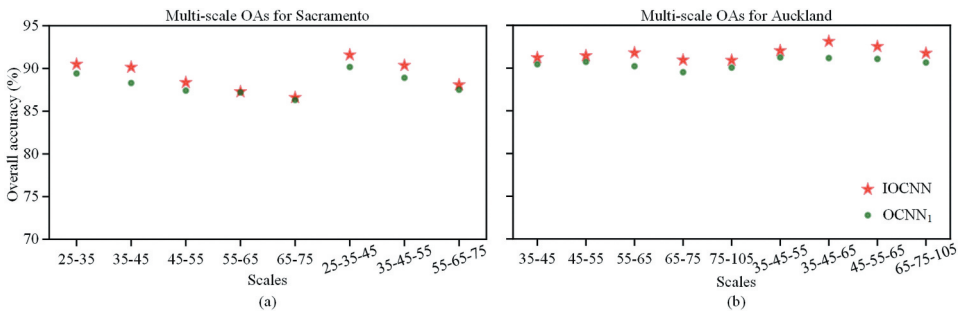


Figure 10. Multi-scale classification results of IOCNN and OCNN₁.

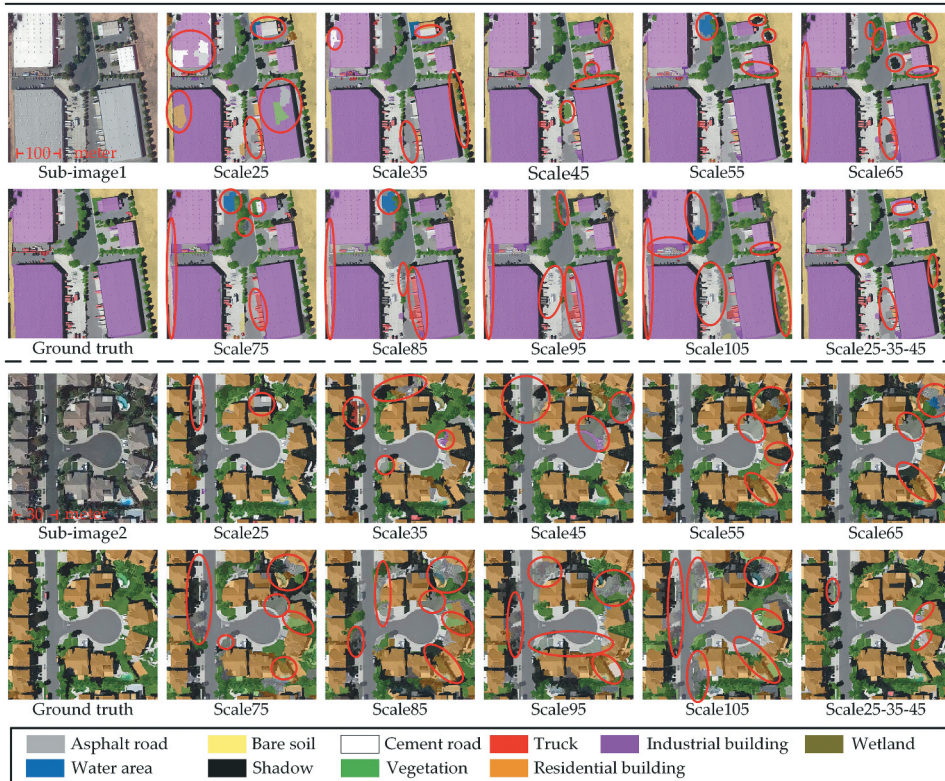


Figure 11. Scale effects in IOCNN classification results.

The multi-scale classification results show that IOCNN performs better than that of $OCNN_1$ at all scales, demonstrating the effectiveness of IOCNN. In addition, the OAs of triple scale results are often higher than that of double scale. The optimized scale combinations of IOCNN (OA : 91.59%, $kappa$: 0.9041, $f1$: 0.9108) and $OCNN_1$ (OA : 90.17%, $kappa$: 0.888, $f1$: 0.8959) are the same triple scale 25–35–45 for Sacramento. In addition, the optimized scale combinations of IOCNN (OA : 93.13%, $kappa$: 0.9156, $f1$: 0.914) and $OCNN_1$ (OA : 91.25%, $kappa$: 0.8926, $f1$: 0.8942) are triple scale 35–45–65 and 35–45–55 for Auckland, respectively.

The scale effects inherent to $OCNN$ are illustrated in Figure 11. The classification for the single scales perform poorly relative to the classification results for the multiple scales. Moreover, typical object-level salt-and-pepper errors are found at the small scales (i.e. 25×25 and 35×35 scales), and the classification results of small objects are better than those of large objects. Conversely, object-level salt-and-pepper errors are markedly reduced in the ground object areas at the large scales (i.e. 95×95 and 105×105 scales). However, objects with small areas or narrow widths are easily misclassified as their neighbour objects. The multi-scale methods are proposed to address the scale effect issue in single scales.

The multi-scale method combines multiple single-scale features and retrains the fully connected layers in the CNN model to obtain the multi-scale features. This technique could considerably improve the classification results. Obviously, the salt and pepper errors are basically solved by the multi-scale method from the Figure 11.

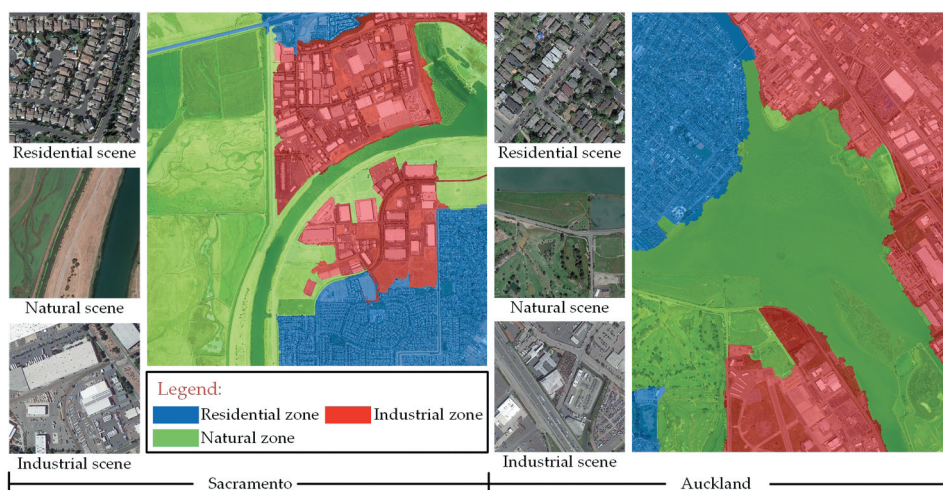


Figure 12. Zone division results.

4.3. Dividing zone results

The contributions of dividing zones involve data pre- and post-processing. The number of geographical objects is heavily reduced by dividing zones, which simultaneously improves the efficiency of remote sensing image classification. In addition, a higher classification *OA* can be obtained by combining the classification results of each zone at their optimized scales. Zone division results are shown in Figure 12. Both images are divided into three zones with coarse boundaries. In fact, the zones have no any labels. To emphasize the dividing results, these zones match the residential, industrial, and natural zones in the study areas.

Each zone can be re-segmented into objects with optimized segmentation parameters based on the land covers in each zone. Thus, the segmentation result cannot be affected by the smallest objects in images. The segmentation effect is improved by dividing zones, which simultaneously reduces the number of geographical objects. Furthermore, the classification efficiency is improved by using this strategy. Figure 13 shows the number of objects in dividing zones and none-dividing zones.

The number of objects with dividing zones is reduced to 57.5% of that in none-dividing zones for Sacramento. In addition, the number of objects with dividing zones in Auckland is reduced by 26.4% compared with objects with none dividing zones. However, the scale effects during classification still exists and varies across different zones. Therefore, each image is initially divided into smaller zones, and each zone is separately processed by IOCNN. The results of all zones are eventually merged to improve the *OA* of the image classification. The optimized scales of residential, natural, and industrial zones in Sacramento are 25–35–45, 55–65–75, and 25–35–45, respectively. In addition, the optimized scales of residential, natural, and industrial zones in Auckland are 35–45–65, 25–45–65, 65–75–105, respectively. By combing results of these zones, new highest values for Sacramento (*OA*: 91.65%, *kappa*: 0.9048, and *f1*: 0.9138) and Auckland (*OA*: 93.49%, *kappa*: 0.9201, and *f1*: 0.9204) are achieved.

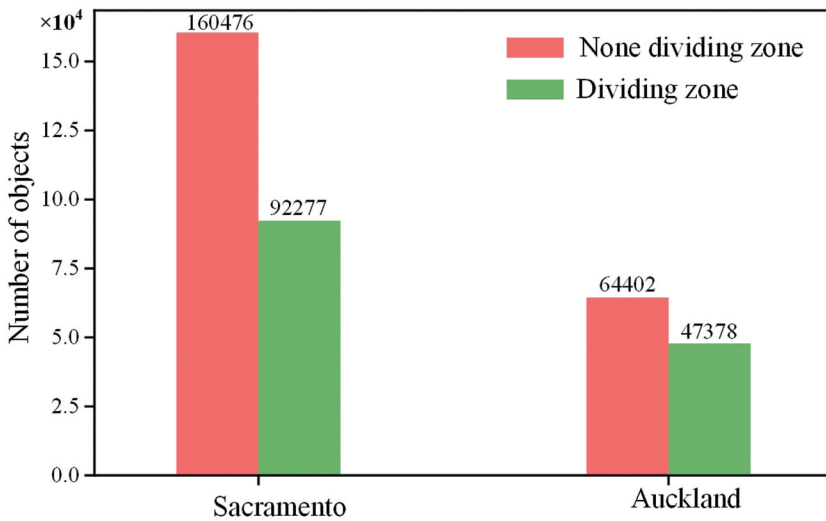


Figure 13. Number of objects in dividing zones and none dividing zones.

5. Discussion

5.1. CNNs versus FCNs

Convolutional neural networks (CNNs) and fully convolutional networks (FCNs) are the two key directions of deep learning in image classification, semantic segmentation, and image fusion. In the conventional computer vision, CNNs are commonly used to label images. FCNs often segment images and then mark pixels semantics. FCNs perform well in semantic segmentation of natural or life images, but not in remote sensing image classification. Figure 14 shows the disadvantages of U-Net and SegNet compared with IOCNN in image classifications. The FCNs show great potential for remote sensing image

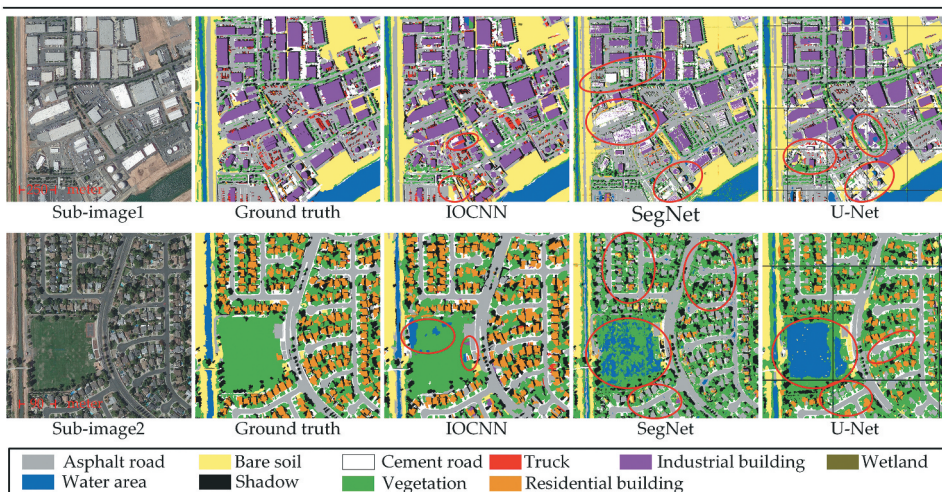


Figure 14. Local comparison of FCNs and IOCNN.

classification. However, FCNs cannot overcome a classical issue in remote sensing, where objects of the same land cover type show different spectrums and objects of different land cover types show the similar spectrums. Obviously, most of the misclassifications in [Figure 14](#) are due to this issue which is not common in computer vision. Given that CNNs have large respective field that covers surroundings. As The context information can be learned by CNNs, CNNs can better handle this problem. Thus, IOCNN is more suitable for remote sensing classification comparing with FCNs.

5.2. Dividing zone for pre- and post-processing and scale effect

The zone division method is a highlight of this study for pre- and post-processing. The segmentation parameters are defined by the smallest objects in the image due to the complexity of land covers, leading to an over-segmentation of the other categories. Therefore, the overall efficiency of OCNNs is decreased by the tremendous objects generated by the segmentation. Most categories are only located in the corresponding functional zones and form local clusters (for example in the two study areas, residential buildings are mainly located in residential zones and factory buildings mainly located in industrial zones). Thus, the classification parameters can be optimized for each divided zone. Additionally, the classification result with optimized scales of different zones can be merged on the whole image. In this situation, the number of over segmented objects is dramatically reduced.

The scale effect is a tricky question in OBIA. This effect is a kind of limiting effect that exists objectively and is expressed by scales. The logic behind this effect does not regard unconditional reasoning and the infinite extension of scales. Nonetheless, macroscopic movements can be inferred, and macroscopic laws can be replaced with microscopic experimental results. This scenario is an important philosophical root for many theoretical paradoxes. The scale effect in remote sensing image classification means that the classification results may vary greatly across different classification scales, particularly in terms of an object size in an image. Objects vary widely in terms of sizes. For example, the biggest artificial object (an industry building) in the studied image covers approximately 4100 m². However, the smallest artificial object (tracks and a few cement roads) covers about 50 m². If only the segmented objects (parts of actual objects) are considered, the smallest segmented objects will cover 1 pixel (the inevitable product of the MRS algorithm), while the biggest segmented objects will cover more than 60,000 pixels in the study images. Different objects also vary in terms of optimized scales. Given this condition, a single scale cannot be applied to achieve the desired classification performance. Meanwhile, multi-scale methods can be applied to extract object features. However, these methods lead to a compromise between different optimized scales. According to Tobler's First Law of Geography, geographical objects or attributes are related to one another in spatial distribution. In urban or near-urban areas, the clustering and regular distributions are obvious. The study areas in our research contain clustering and regular distributions. The proposed method is also conducted to achieve zone-level self-adaptative scale based on the dividing and conquering strategy. This work addresses the scale effects, which is a fundamental issue overlooked by many remote sensing studies. The proposed dividing zone method conforms with Tobler's First Law of Geography and the idea of the scale effects.

5.3. Convolutional positions for object classification

Parts of objects can be determined and extracted by the BTS. These parts can also be evenly distributed within the objects. Thus, the detailed inner and surrounding information can also be derived by the OCNN model. From the data perspective, the convolutional data extracted by BTS can be reduced greatly. For example, other convolutional selection methods may generate positions located on the boundary of objects. Given this condition, two objects of different categories may present themselves in the same convolutional window, a scenario that will cause huge interference to the classification results. The BTS method can be used to overcome the issue. From the feature perspective, the extracted features from the convolutional window generated by the BTS can be classified much easier. We select three methods as comparative methods on the BTS. Method one (M1), two (M2), and three (M3) are derived from OCNN₁, OCNN₂, and OCNN₄. The convolutional patches extracted by OCNN₃ and OCNN₅ are generated based on the entire study image, not based on the object, so that we do not take them as the comparative methods on the BTS. Figure 15 shows the examples of objects with their corresponding convolutional positions and convolutional windows based on BTS and comparative methods. The

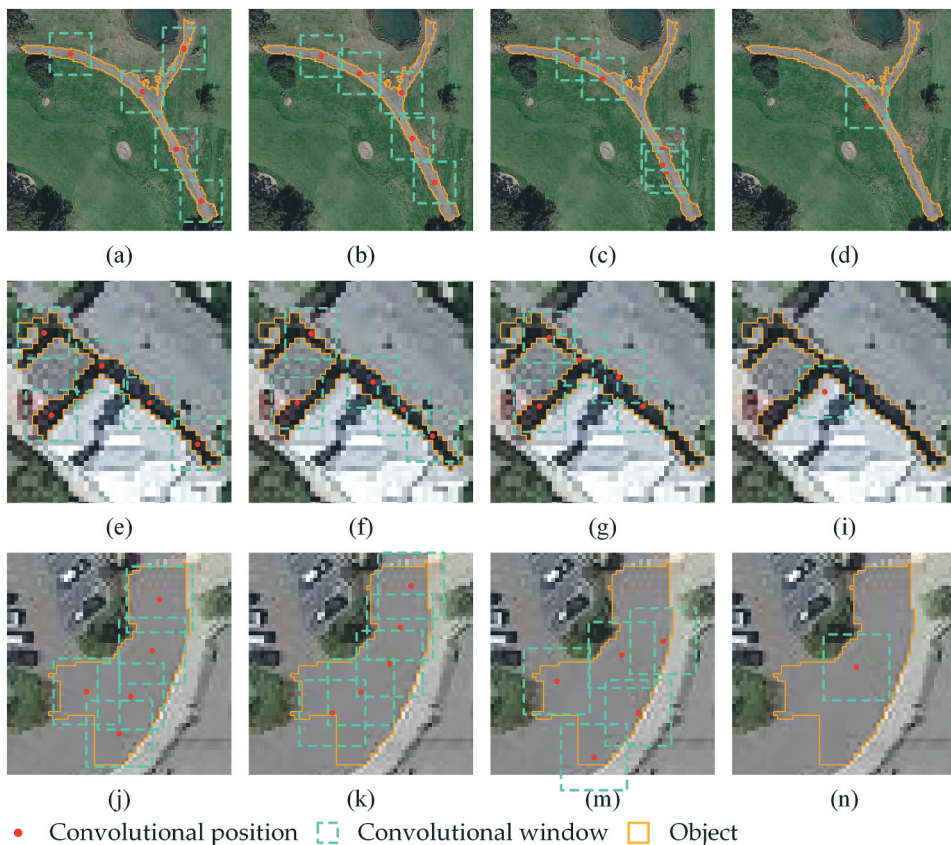


Figure 15. Results of comparative methods on the BTS. (a), (e), and (j) are the results of the BTS; (b), (f), and (k) are the results of M1; (c), (g), and (m) are the results of M2; (d), (i), and (n) are the results of M3.

convolutional positions are evenly distributed within objects, including those with highly irregular shapes, thus demonstrating the effectiveness and robustness of the sampling algorithm.

5.4. Multi-scale deep features for land cover classification

The features used for land cover classifications are abstractions observed from VHR images. The abstract features hidden in the spatial structures or patterns of images are defined as single features when only single-extracting convolutional windows are used. However, this method is inadequate to express the deep features. The visualization of inadequate expressions leads to scale effects, as shown in [Figure 11](#), in which object-level salt-and-pepper errors exist in the classification results at the small single scales, and fuzzy boundaries are presented in the classification results at the large single scales. The issues can be addressed using the proposed method with multi-scale features. Instead of simple combinations of classification results at the different single scales, the features extracted from fully connected layers are combined in the multi-scale method. As shown in [Figure 10](#), all of the highest classification accuracies are achieved in IOCNN by using multi-scale features. Therefore, multi-scale CNNs are more effective than single CNNs for feature representation.

5.5. IOCNN for VHR image land cover classification

The IOCNN method is designed by combining zone division, the BTS method, multi-scale feature characterization, and the OCNN classification approach. IOCNN, as it inherits the capability of OCNNs in mining spatial information, structures, and patterns, can analyse the variations in deep features across different zones, use optimized segmentation scale parameters for each zone, derive convolutional positions from the shapes of objects, and label the objects according to their multi-scale deep features.

Each divided zone has its distinctive domain requiring unique scale parameters, and the result is the varying scale effects across different zones. However, a uniform scale parameter cannot effectively meet the requirements of complex scale effect issues for the different zones in an image. Therefore, the object segmentation needs to use adaptive scale parameters for each zone. Zone division not only plays an important role in the object segmentation in the image's pre-processing phase but also helps merge the optimized classification result for each zone in the post-processing phase. For example, the best classification result in Sacramento with an *OA* of 91.65% is achieved by merging the residential zone classification result at the scale of 25–35–45, the natural zone classification result at the scale of 55–65–75, and the industrial zone classification result at the scale of 25–35–45. In terms of the distribution of convolutional positions, the BTS method contributed effectively to the representation of multi-scale convolutional deep features. The classification of highly complex objects benefits from the optimization of convolutional positions generated by BTS.

Objects that are slightly over-segmented by MRS have high homogeneity. Theoretically, the spatial relationships and neighbourhood patterns of objects represent important information in the classification when deriving convolutional positions. On the one hand, the internal deep features of the objects should be considered. On the other hand, several other

neighbourhood deep features of the objects should be identified. The results of this study indicate that the convolutional positions generated by the BTS method can utilize both the internal and neighbourhood deep features of the objects. Only the objects with high shape indexes are sensitive to their convolutional positions. Although not all categories in the study areas could be classified correctly using the positions generated by BTS, the method constantly generates good classification results in most categories. The benchmark for generating convolutional positions are the different stratified zones. By combining these divided zones with the BTS method, IOCNN can reach a high classification accuracy.

5.6. Future research

The proposed IOCNN method attains notably high *OA*, *kappa*, and *f1* for VHR image land cover classification. Currently, precise zone boundaries cannot be extracted by the method, and this scenario may cause incorrectly divided zones. The efficiency of the BTS algorithm is reportedly low. Theoretically, the BTS undergoing five recursive local fine-tuning requires lots of running time. The local fine-tuning of convolutional positions is also I/O-intensive and requires a high computational resource. Moreover, multi-scale convolutional deep features comprise various categories in different zones. Each category cannot be extracted at its optimal scales, so that feature representations cannot easily label objects at their optimized scales.

6. Conclusion

Accurately labelling all categories for land cover classification is challenging because of the complexity of objects and the diversity of the zones. A classification method that considers object complexity and regional heterogeneity is therefore needed. Moreover, geographic phenomena are inherently hierarchical and stratified. Stratified processing is suitable for VHR image land cover classification with multiple zones. To achieve this objective, the IOCNN method is proposed in this study including a novel zone division and BTS method for VHR image classification.

Four contributions are made in this study:

- (1) Dividing zones are firstly used for classifying VHR images with multiple zone structures. Introducing zone division into OCNN prevents the over-segmentation of each zone and improves the efficiency of the algorithm.
- (2) Novel binary tree sampling (BTS) for generating convolutional positions of Object-based CNNs is used. The BTS based on zone division can provide a highly reasonable and suitable labels of objects. The high classification accuracy achieved in this work proves the effectiveness of the BTS. The convolutional positions in the study areas further confirm the results. Additionally, the classification results prove that *OAs* can be substantially improved by the IOCNN.
- (3) The proposed method achieves combinations of the optimized scales at different zones in an image. The combination of multi-scale results of different zones can improve the classification performance. The idea of 'divide and conquer' is theoretically consistent with the stratified processing theory in geography.

- (4) The classification *OA*, *kappa*, and *f1* of the proposed IOCNN method are higher than other state-of-the-art methods.

In conclusion, the IOCNN method is robust, effective, and useful for the land cover classification of VHR images with multiple zones. However, the coarse boundaries of the divided zones, the efficiency of BTS, and the optimized scales for certain objects need to be addressed in future studies.

Data availability statement

The data that support the findings of this study are openly available in figshare at <https://figshare.com/s/99c1fbcc7d36c15c3779>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China with grant numbers 42090012, 41771452 and 41771454.

ORCID

Xianwei Lv  <http://orcid.org/0000-0002-1574-8500>

Zhenfeng Shao  <http://orcid.org/0000-0003-4587-6826>

References

- Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. "Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.2016.2644615.
- Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing. ISPRS J Photogramm Remote Sens." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. doi:10.1016/j.isprsjprs.2009.06.004.
- Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, et al. 2014. "Geographic Object-based Image Analysis—towards a New Paradigm." *ISPRS Journal of Photogrammetry and Remote Sensing* 87: 180–191. doi:10.1016/j.isprsjprs.2013.09.014.
- Cai, Y., K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, and L. Zhan. 2018. "A High-performance and In-season Classification System of Field-level Crop Types Using Time-series Landsat Data and A Machine Learning Approach." *Remote Sensing of Environment* 210: 35–47. doi:10.1016/j.rse.2018.02.045.
- Chen, B., B. Huang, and X. Bing. 2017. "Multi-source Remotely Sensed Data Fusion for Improving Land Cover Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 124: 27–39. doi:10.1016/j.isprsjprs.2016.12.008.
- Chen, G., Q. Weng, G. J. Hay, and H. Yinan. 2018. "Geographic Object-based Image Analysis (GEOBIA): Emerging Trends and Future Opportunities." *GIScience & Remote Sensing* 55 (2): 159–182. doi:10.1080/15481603.2018.1426092.

- Chen, G., H. Yinan, A. De Santis, L. Guosheng, R. Cobb, and R. K. Meentemeyer. 2017. "Assessing the Impact of Emerging Forest Disease on Wildfire Using Landsat and KOMPSAT-2 Data." *Remote Sensing of Environment* 195: 218–229. doi:10.1016/j.rse.2017.04.005.
- Chen, Y., D. Ming, and L. Xianwei. 2019. "Superpixel Based Land Cover Classification of VHR Satellite Image Combining Multi-scale CNN and Scale Parameter Estimation." *Earth Science Informatics* 12 (3): 341–363. doi:10.1007/s12145-019-00383-2.
- Chen, Y., H. Jiang, L. Chunyang, X. Jia, and P. Ghamisi. 2016. "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks." *IEEE Transactions on Geoscience and Remote Sensing* 54 (10): 6232–6251. doi:10.1109/TGRS.2016.2584107.
- Chen, Y., X. Zhao, and X. Jia. 2015. "Spectral–spatial Classification of Hyperspectral Data Based on Deep Belief Network." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (6): 2381–2392. doi:10.1109/JSTARS.2015.2388577.
- Fu, P., Y. Xie, Q. Weng, S. Myint, K. Meacham-Hensold, and C. Bernacchi. 2019. "A Physical Model-based Method for Retrieving Urban Land Surface Temperatures under Cloudy Conditions." *Remote Sensing of Environment* 230: 111191. doi:10.1016/j.rse.2019.05.010.
- Fu, T., M. Lei, L. Manchun, and B. A. Johnson. 2018. "Using Convolutional Neural Network to Identify Irregular Segmentation Objects from Very High-resolution Remote Sensing Imagery." *Journal of Applied Remote Sensing* 12 (2): 025010. doi:10.1117/1.JRS.12.025010.
- Haralick, R. M., and K. Shanmugam, Its' Hak Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3 (6): 610–621. doi:10.1109/TSMC.1973.4309314.
- He, Y., G. Chen, A. De Santis, D. A. Roberts, Y. Zhou, and R. K. Meentemeyer. 2019a. "A Disturbance Weighting Analysis Model (DWAM) for Mapping Wildfire Burn Severity in the Presence of Forest Disease." *Remote Sensing of Environment* 221: 108–121. doi:10.1016/j.rse.2018.11.015.
- He, Y., G. Chen, C. Potter, and R. K. Meentemeyer. 2019b. "Integrating Multi-sensor Remote Sensing and Species Distribution Modeling to Map the Spread of Emerging Forest Disease and Tree Mortality." *Remote Sensing of Environment* 231: 111238. doi:10.1016/j.rse.2019.111238.
- Heydari, S. S., and G. Mountrakis. 2019. "Meta-analysis of Deep Neural Networks in Remote Sensing: A Comparative Study of Mono-temporal Classification to Support Vector Machines." *ISPRS Journal of Photogrammetry and Remote Sensing* 152: 192–210. doi:10.1016/j.isprsjprs.2019.04.016.
- Huang, B., B. Zhao, and Y. Song. 2018. "Urban Land-use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery." *Remote Sensing of Environment* 214: 73–86. doi:10.1016/j.rse.2018.04.050.
- Hughes, L. H., M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu. 2018. "Identifying Corresponding Patches in SAR and Optical Images with a Pseudo-siamese CNN." *IEEE Geoscience and Remote Sensing Letters* 15 (5): 784–788. doi:10.1109/LGRS.2018.2799232.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." *In Advances in Neural Information Processing Systems* 2012 (25): 1097–1105.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *nature* 521 (7553): 436–444. doi:10.1038/nature14539.
- Liu, P., H. Zhang, and K. B. Eom. 2016. "Active Deep Learning for Classification of Hyperspectral Images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10 (2): 712–724. doi:10.1109/JSTARS.2016.2598859.
- Liu, T., and A. Abd-Elrahman. 2018. "Deep Convolutional Neural Network Training Enrichment Using Multi-view Object-based Analysis of Unmanned Aerial Systems Imagery for Wetlands Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 139: 154–170. doi:10.1016/j.isprsjprs.2018.03.006.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440. Boston, MA, USA.
- Lv, X., D. Ming, Y. Chen, and M. Wang. 2019. "Very High Resolution Remote Sensing Image Classification with SEEDS-CNN and Scale Effect Analysis for Superpixel CNN Classification." *International Journal of Remote Sensing* 40 (2): 506–531. doi:10.1080/01431161.2018.1513666.

- Lv, X., D. Ming, L. Tingting, K. Zhou, M. Wang, and H. Bao. 2018. "A New Method for Region-based Majority Voting CNNs for Very High Resolution Image Classification." *Remote Sensing* 10 (12): 1946. doi:10.3390/rs10121946.
- Ma, L., L. Cheng, L. Manchun, Y. Liu, and M. Xiaoxue. 2015. "Training Set Size, Scale, and Features in Geographic Object-Based Image Analysis of Very High Resolution Unmanned Aerial Vehicle Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 102: 14–27. doi:10.1016/j.isprsjprs.2014.12.026.
- Ma, L., Y. Liu, X. Zhang, Y. Yuanxin, G. Yin, and B. A. Johnson. 2019. "Deep Learning in Remote Sensing Applications: A Meta-analysis and Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 152: 166–177. doi:10.1016/j.isprsjprs.2019.04.015.
- Ma, L., L. Manchun, M. Xiaoxue, L. Cheng, P. Du, and Y. Liu. 2017. "A Review of Supervised Object-based Land-cover Image Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 130: 277–293. doi:10.1016/j.isprsjprs.2017.06.001.
- Ma, L., F. Tengyu, and L. Manchun. 2018. "Active Learning for Object-based Image Classification Using Predefined Training Objects." *International Journal of Remote Sensing* 39 (9): 2746–2765.
- Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2016. "Convolutional Neural Networks for Large-scale Remote-sensing Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 55 (2): 645–657. doi:10.1109/TGRS.2016.2612821.
- Marcos, D., M. Volpi, B. Kellenberger, and D. Tuia. 2018. "Land Cover Mapping at Very High Resolution with Rotation Equivariant CNNs: Towards Small yet Accurate Models." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 96–107. doi:10.1016/j.isprsjprs.2018.01.021.
- Merkle, N., W. Luo, S. Auer, R. Müller, and R. Urtasun. 2017. "Exploiting Deep Matching and SAR Data for the Geo-localization Accuracy Improvement of Optical Satellite Images." *Remote Sensing* 9 (6): 586. doi:10.3390/rs9060586.
- Mez, C. E. 1978. "Basic Principles of ROC Analysis." In *Seminars in nuclear medicine*, 8, 283–298. WB Saunders. doi:10.1016/S0001-2998(78)80014-2
- Ming, D., L. Jonathan, J. Wang, and M. Zhang. 2015. "Scale Parameter Selection by Spatial Statistics for GeOBIA: Using Mean-shift Based Multi-scale Segmentation as an Example." *ISPRS Journal of Photogrammetry and Remote Sensing* 106: 28–41.
- Mui, A., H. Yuhong, and Q. Weng. 2015. "An Object-based Approach to Delineate Wetlands across Landscapes of Varied Disturbance with High Spatial Resolution Satellite Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 109: 30–46. doi:10.1016/j.isprsjprs.2015.08.005.
- Pan, B., Z. Shi, and X. Xia. 2018. "MugNet: Deep Learning for Hyperspectral Image Classification Using Limited Samples." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 108–119. doi:10.1016/j.isprsjprs.2017.11.003.
- Rabiee, H. R., R. L. Kashyap, and S. Rasoul Safavian. 1996. "Multiresolution Segmentation-based Image Coding with Hierarchical Data Structures." In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 4*, 1870–1873. IEEE, Atlanta, GA, USA.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-net: Convolutional Networks for Biomedical Image Segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, 234–241, Munich, Germany. Springer.
- Shao, Z., and J. Cai. 2018. "Remote Sensing Image Fusion with Deep Convolutional Neural Network." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (5): 1656–1669. doi:10.1109/JSTARS.2018.2805923.
- Shao, Z., F. Huyan, L. Deren, O. Altan, and T. Cheng. 2019a. "Remote Sensing Monitoring of Multi-scale Watersheds Impermeability for Urban Hydrological Evaluation." *Remote Sensing of Environment* 232: 111338. doi:10.1016/j.rse.2019.111338.
- Shao, Z., Y. Pan, C. Diao, and J. Cai. 2019b. "Cloud Detection in Remote Sensing Images Based on Multiscale Features-convolutional Neural Network." *IEEE Transactions on Geoscience and Remote Sensing* 57 (6): 4062–4076. doi:10.1109/TGRS.2018.2889677.

- Shao, Z., P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai. 2020a. "BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction from High-Resolution Remote Sensing Images." *Remote Sensing* 12 (6): 1050. doi:10.3390/rs12061050.
- Shao, Z., L. Wang, Z. Wang, W. Du, and W. Wu. 2019c. "Saliency-aware Convolution Neural Network for Ship Detection in Surveillance Video." *IEEE Transactions on Circuits and Systems for Video Technology* 30 (3): 781–794. doi:10.1109/TCSVT.2019.2897980.
- Shao, Z., W. Zhou, X. Deng, M. Zhang, and Q. Cheng. 2020b. "Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 318–328. doi:10.1109/JSTARS.2019.2961634.
- Song, H., Q. Liu, G. Wang, R. Hang, and B. Huang. 2018. "Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (3): 821–829.
- Sun, G., H. Huang, Q. Weng, A. Zhang, X. Jia, J. Ren, L. Sun, and X. Chen. 2019. "Combinational Shadow Index for Building Shadow Extraction in Urban Areas from Sentinel-2a Msi Imagery." *International Journal of Applied Earth Observation and Geoinformation* 78: 53–65. doi:10.1016/j.jag.2019.01.012.
- Tong, X.-Y., G.-S. Xia, L. Qikai, H. Shen, L. Shengyang, S. You, and L. Zhang. 2020. "Land-cover Classification with High-resolution Remote Sensing Images Using Transferable Deep Models." *Remote Sensing of Environment* 237: 111322. doi:10.1016/j.rse.2019.111322.
- Wurm, M., T. Stark, X. X. Zhu, M. Weigand, and T. Hannes. 2019. "Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 150: 59–69. doi:10.1016/j.isprsjprs.2019.02.006.
- Xu, L., D. Ming, W. Zhou, H. Bao, Y. Chen, and X. Ling. 2019. "Farmland Extraction from High Spatial Resolution Remote Sensing Images Based on Stratified Scale Pre-estimation." *Remote Sensing* 11 (2): 108. doi:10.3390/rs11020108.
- Zhang, C., I. Sargent, X. Pan, A. Gardiner, J. Hare, and P. M. Atkinson. 2018a. "VPRS-based Regional Decision Fusion of CNN and MRF Classifications for Very Fine Resolution Remotely Sensed Images." *IEEE Transactions on Geoscience and Remote Sensing* 56 (8): 4507–4521. doi:10.1109/TGRS.2018.2822783.
- Zhang, C., I. Sargent, X. Pan, L. Huapeng, A. Gardiner, J. Hare, and P. M. Atkinson. 2018b. "An Object-based Convolutional Neural Network (OCNN) for Urban Land Use Classification." *Remote Sensing of Environment* 216: 57–70. doi:10.1016/j.rse.2018.06.034.
- Zhang, C., I. Sargent, X. Pan, L. Huapeng, A. Gardiner, J. Hare, and P. M. Atkinson. 2019. "Joint Deep Learning for Land Cover and Land Use Classification." *Remote Sensing of Environment* 221: 173–187. doi:10.1016/j.rse.2018.11.014.
- Zhang, C., P. Yue, D. Tapete, B. Shangguan, M. Wang, and W. Zhaoyan. 2020. "A Multi-level Context-guided Classification Method with Object-based Convolutional Neural Network for Land Cover Classification Using Very High Resolution Remote Sensing Images." *International Journal of Applied Earth Observation and Geoinformation* 88: 102086. doi:10.1016/j.jag.2020.102086.
- Zhang, X., G. Chen, W. Wang, Q. Wang, and F. Dai. 2017. "Object-based Land-cover Supervised Classification for Very-high-resolution UAV Images Using Stacked Denoising Autoencoders." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10 (7): 3373–3385. doi:10.1109/JSTARS.2017.2672736.
- Zhao, W., and S. Du. 2016. "Learning Multiscale and Deep Representations for Classifying Remotely Sensed Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 113: 155–165. doi:10.1016/j.isprsjprs.2016.01.004.
- Zhao, W., S. Du, and W. J. Emery. 2017. "Object-based Convolutional Neural Network for High-resolution Imagery Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10 (7): 3386–3396. doi:10.1109/JSTARS.2017.2680324.
- Zhao, W., Z. Guo, J. Yue, X. Zhang, and L. Luo. 2015. "On Combining Multiscale Deep Learning Features for the Classification of Hyperspectral Remote Sensing Imagery." *International Journal of Remote Sensing* 36 (13): 3368–3379. doi:10.1080/2150704X.2015.1062157.

- Zheng, Z., Y. Zhong, M. Ailong, and L. Zhang. 2020. "FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(8): 5612-5626.
- Zhong, Y., X. Han, and L. Zhang. 2018. "Multi-class Geospatial Object Detection Based on a Position-sensitive Balancing Framework for High Spatial Resolution Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 138: 281–294. doi:10.1016/j.isprsjprs.2018.02.014.
- Zhou, W., D. Ming, Z. Hong, and L. Xianwei. 2018a. "Scene Division Based Stratified Object Oriented Remote Sensing Image Classification." In *2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, 1–5. IEEE, Xi'an, China.
- Zhou, W., D. Ming, L. Xianwei, K. Zhou, H. Bao, and Z. Hong. 2020. "SO-CNN Based Urban Functional Zone Fine Division with VHR Remote Sensing Image." *Remote Sensing of Environment* 236: 111458. doi:10.1016/j.rse.2019.111458.
- Zhou, W., D. Ming, L. Xu, H. Bao, and M. Wang. 2018b. "Stratified Object-oriented Image Classification Based on Remote Sensing Image Scene Division." *Journal of Spectroscopy* 2018: 1–11. doi:10.1155/2018/3918954.
- Zhu, L., Y. Chen, P. Ghamisi, and J. A. Benediktsson. 2018. "Generative Adversarial Networks for Hyperspectral Image Classification." *IEEE Transactions on Geoscience and Remote Sensing* 56 (9): 5046–5063. doi:10.1109/TGRS.2018.2805286.
- Zou, Q., N. Lihao, T. Zhang, and Q. Wang. 2015. "Deep Learning Based Feature Selection for Remote Sensing Scene Classification." *IEEE Geoscience and Remote Sensing Letters* 12 (11): 2321–2325. doi:10.1109/LGRS.2015.2475299.