OBJECT-BASED CLASSIFICATION USING MASK R-CNN AND CNN FROM VERY HIGH-RESOLUTION IMAGERY

Batbold Badamdorj[1] and Bolorchuluun Chogsom[1]
[1]National University of Mongolia, Ikh Surguuliin Gudamj-1, Ulaanbaatar, Mongolia,
Email: batbold9909@gmail.com

KEY WORDS: Object-based Image Analysis, Deep learning, Mask R-CNN, CNN

ABSTRACT: With the development of modern technology, the resolution of remote sensing imagery increasing rapidly. As the resolution of remote sensing imagery increases, it becomes more difficult to classify urban land use-land cover types and recognize complex patterns in urban areas using traditional pixel-based methods. Therefore, we combined Object-based Image Analysis and Deep Learning, to accurately classify very high-resolution imagery of urban areas in this study. Object-based Image Analysis is used as the main classifier, while Deep Learning algorithms, Mask R-CNN and CNN, are used as feature extractors. In this study, very high-resolution images were used to classify three urban scenes in Ulaanbaatar, Mongolia. The first scene was used as the training site, whereas the second and the third scenes were used as test sites. As a result, the final classification with ten classes of selected scenes in Ulaanbaatar was created. The overall accuracy of classification result was above 90% in each scene, including 92.69% in the training site, 91.76% in the first test site, and 93.04% in the second test site. Our result shows the combination of Object-based Image Analysis and Deep Learning increases accuracy of classification from very high-resolution imagery of urban areas.

1. INTRODUCTION

For over 40 years since the 1970s, pixel-based analysis has been a dominant *paradigm* in remote sensing (T. Blaschke, 2013) and the principal method and technology for the classification of remotely sensed data, including aerial and satellite imagery. However, since the 2010s, the accuracy of pixel-based image classification methods has begun to decline due to the increasing resolution of remotely sensed data and the complexity of the information it contains. Since pixel-based methods are often used to extract low-level features (H. I. Sibaruddin, 2018), based on per-pixel spectral information, and do not take into account the contextual information, it is not suitable for the classification from very high-resolution imagery, especially in urban areas. For example, man-made objects share similar spectral properties, such as roofs and parking lots with similar construction material, which makes it even harder to accurately classify very high-resolution imagery (W. Zhao, 2016). And rich spatial information presented in very high-resolution images hinders accurate interpretation and classification (D. Tuia, 2009). For another, as the resolution of remote sensing imagery gets finer, small objects are conjointly distributed around target objects which results in a strong confusion (W. Zhao, 2017). In Ulaanbaatar, Mongolia, it is extremely difficult to classify very high-resolution imagery of urban areas using pixel-based methods due to poorly planned residential and industrial areas, high-density buildings, the 'ger district' sprawl (It is a form of the residential district in Mongolian settlements, consist of a group of Mongolian traditional nomadic tents that coexists with urban areas. Ger is also known as yurts.), and diversity of land use-land cover types. Thus, a more efficient method for very high-resolution and high-resolution urban imagery classification needs to be proposed.

In this study, we proposed a combination of the most advanced methods in the field of remote sensing, Object-based Image Analysis (OBIA) and Deep Learning (DL), to accurately classify very high-resolution imagery of urban areas. OBIA is a method of remote sensing, which divides the image into homogeneous regions based on similarities of spatial information, spectral information, and contextual information (E. Guirado, 2021). Object-based methods offer an efficient approach to classify objects with precise boundary information, such as shape, size, and edge from very high-resolution imagery, although it is difficult to recognize or detect complex patterns and objects in urban areas. To reduce this problem, we combined it with deep learning methods, Mask Region-Convolutional Neural Networks (Mask R-CNN) and Convolutional Neural Networks (CNN). Mask R-CNN is a deep learning algorithm for instance segmentation and object detection from an image (X. Zhang, 2018). In other words, it can separate different objects in an image. As well as, CNN is a deep learning algorithm, which is adopted a wide range of aspects in image processing, including image classification, object detection, super-resolution restoration, etc. (Y. Li, 2018).

The main purpose of this study is to combine the most advanced and rapidly developing methods in the field of remote sensing to perform a high accuracy classification from very high-resolution imagery of urban areas, and to introduce the advantages and possibilities of this method, moreover, to develop deep learning models that can be used for classification or object detection in similar urban areas.

The remainder of the paper is organized as follows. Detailed information about the methodology of this study is described in Section 2, and the datasets are introduced in Section 3. Then classification results are discussed in Section 4, followed by the conclusion in Section 5, and acknowledgement in Section 6.

2. METHODOLOGY

The methodology of this study has two stages. In the first stage, deep learning models were developed using Mask R-CNN and CNN, then used for feature extraction from very high-resolution imagery of Ulaanbaatar. Deep learning models of building footprint extraction and car detection from very high-resolution imagery were developed using Mask R-CNN, in order to increase the accuracy of the final classification. The deep learning model of land cover recognition from very high-resolution imagery was developed using CNN. In the second stage, outputs from first stage were used for Object-based Image Analysis, such as object-based classification and rule-based improvement (so-called Knowledge-based improvement). At last, the final classification map with ten classes was created. The methodology workflow is shown in Figure 1.

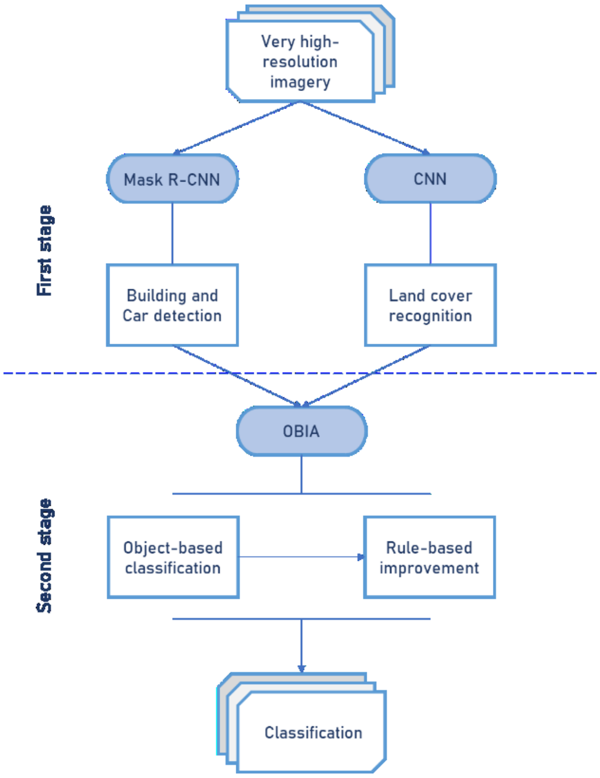


Figure 1. Methodology workflow overview

The building footprint extraction and car detection using Mask R-CNN are described in Section 2.1, including an introduction to Mask R-CNN and a description of the training process of the models. Then, the land cover recognition using CNN is described in Section 2.2, including an introduction to CNN and a description of the training process of the model. Finally, OBIA is outlined in Section 2.3, including an explanation of object-based image classification, segmentation, and rule-based improvement.

2.1 Building Footprint Extraction and Car Detection Using Mask R-CNN

The most difficult elements to classify from very high-resolution imagery of Ulaanbaatar are buildings and cars. The rooftops of buildings in Ulaanbaatar are diverse in shape and structure, and their spectral properties are similar to those of other concrete surfaces, such as roads, parking lots, and playgrounds. As well as, it is difficult to classify cars as separate objects due to their proximity on roads and parking lots. Therefore, we developed deep learning models using Mask R-CNN to detect and extract footprints of these elements with high accuracy, thus increase the accuracy of the final classification. The outputs of these deep learning models are vector layers of detected building and car footprints, which are later used for OBIA.

2.1.1 Mask R-CNN overview: Mask R-CNN is a state-of-the-art algorithm in term of instance segmentation, developed on top of Faster R-CNN (K. He, 2018). While Faster R-CNN has two outputs for each candidate object, a class label and a bounding-box offset, Mask R-CNN is the addition of the third branch that outputs the object mask. The mask output is distinct from class and bounding-box output, requiring the extraction of much finer spatial layout of an object (K. He, 2018). There are two stages of Mask R-CNN. The first stage consists Region Proposal Network (RPN), which runs once per image to give a set of region proposals (S. Ren, 2016). In the second stage, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal (X. Zhang, 2018). The general framework of Mask R-CNN is shown in Figure 2.

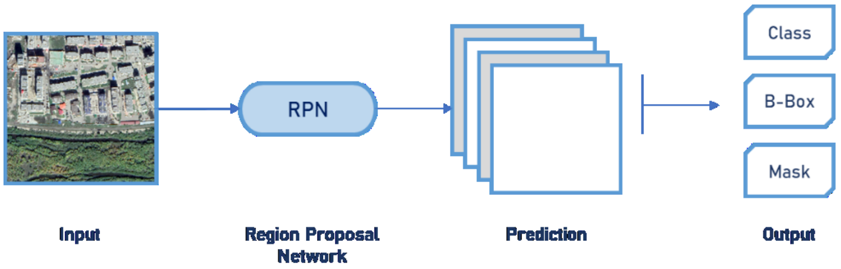


Figure 2. The Mask R-CNN framework

2.1.2 Training the models for building footprint extraction and car detection: Training samples is always the key to deep learning methods (J. Song, 2019), and it should have the proper size. On that account, we set the tile size of training samples to 256 x 256 pixels (about 30 m) for the building footprint extraction model, 56 x 56 (about 3 m) pixels for the car detection model, considering that the average building size in Ulaanbaatar is 30-40 m, and average car size is 3-5 m. We found the optimum learning rate, 3.63078e-05 on building footprint extraction model and 6.30957e-05 on car detection model, which we can train robust models. The learning rate is perhaps the most important hyperparameter (I. Goodfellow, 2016). It controls how quickly the model is adapted to the problem. Smaller learning rates require more training epochs given the smaller changes made to the weights each update, whereas larger learning rates result in rapid changes and require fewer training epochs (J. Brownlee, 2019). Therefore, we set the epoch as 20 and 30 to building footprint extraction and car detection model, respectively.

2.2 Land Cover Recognition Using CNN

The complexity of the spatial and spectral properties of high and very high-resolution imagery makes them difficult to classify by traditional human-dependent classification methods. CNN as the core of deep learning has shown great potential for robust automatic feature extraction and complex object recognition in very high-resolution images (M. A. Ranzato, 2007; Y. Jia, 2014). On that account, the deep learning model of urban land use-land cover recognition from very high-resolution imagery was developed using CNN in this study. In this deep learning model, we pre-defined a total of eight classes, including bare-soil, concrete, ger, grassland, road, shadow, trees, and water. The output of this deep learning model is HeatMap rasters (with floating point pixel values ranging 0 to 1.) of each class defined, which are later used for OBIA.

2.2.1 CNN overview: CNN is a type of deep learning model for processing data that has a grid pattern, such as images, designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns (R. Yamashita, 2018). CNN have become the leading architecture for most image recognition and classification task. CNN architectures come in several variations, however, in general, they consist of convolutional and pooling layers, followed by fully-connected layers (W. Rawat, 2017). The convolutional layers serve as feature extractors, and thus they learn the feature representations of their input images. The pooling layer will then simply perform down sampling along the spatial dimensionality of the given input. Then the fully-connected layers will attempt to produce class scores, to be used for classification (K. O'Shea, 2015). The simple CNN architecture is shown in Figure 3.

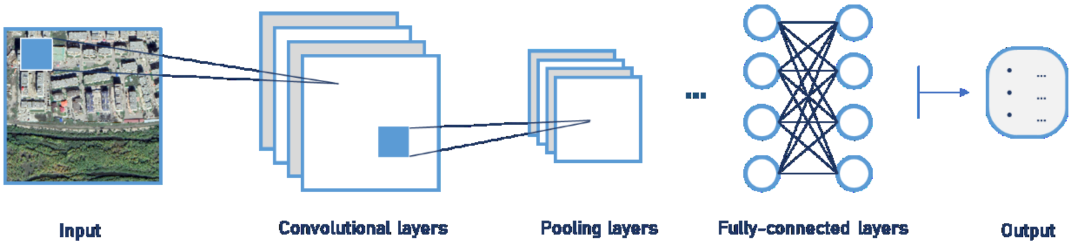


Figure 3. A simple CNN architecture

Learning process of CNN formulated in Equation 1. Where H_i represents the output, W_i denotes the weigh, H_{i-1} indicates the input, f and b stands for activation function and biases.

$$H_i = f(W_i H_{i-1} + b) \qquad (1)$$

2.2.2 Training the model for land cover recognition: The tile size of training samples was set on 64 x 64 pixels, and training samples should be rotated by 30 degrees, in range of 0-330 degrees. The reason why the tile size was 64 x 64 pixels is that some types of land cover do not have a fixed spatial pattern, mostly dependent on spectral properties, which mean it does not need to take the large scale of contextual information. The rotation is necessary to create more training samples, thus we can train more robust model. The optimum learning rate was set to 0.0006, consequently, more training epochs required, we set the epoch as 5000. The depth configuration of the CNN was five layers with the 3 x 3 kernels. As mentioned before, the training data is labeled as eight classes: bare-soil, concrete, ger, grassland, road, shadow, trees, and water.

2.3 Object-based Image Analysis

2.3.1 Object-based Image Classification: As mentioned above, OBIA can accurately classify image objects with precise boundary information using segments. In this stage, object-based image classification is performed using outputs of the deep learning models, vector layers of building and car footprints, and HeatMap rasters of other eight classes, which are obtained in first stage. The vector-based segmentation and the multiresolution segmentation algorithms are used to generate image objects with precise boundaries. First, vector-based segmentation was performed using vector layers of building and car footprints, and each generated segments were classified directly into corresponding classes. Then, with the exception of segments that were classified as buildings and cars, multiresolution segmentation was performed using very high-resolution imagery, and the resulting segments were classified using HeatMap rasters with Membership Function for each class. Membership function allows us to define the relationship between feature values and the degree of membership to a class. The flowchart of object-based image classification using outputs of deep learning models is presented in Figure 4.

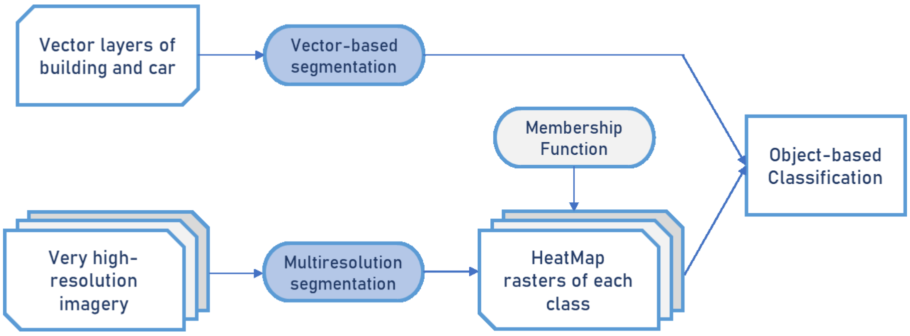

Vector layers of building and car
Vector-based segmentation
Membership Function
Object-based Classification
Very high-resolution imagery
Multiresolution segmentation
HeatMap rasters of each class

Figure 4. The flowchart of object-based image classification using outputs of deep learning models

2.3.2 Segmentation: Image segmentation is fundamental issue in OBIA, as the quality of segmentation results often affects the accuracy of subsequent analysis, such as classification (B. A. Johnson, 2020). It identifies homogeneous and discrete image objects by setting optimal combination of values for three parameters, namely they are scale, shape and compactness. As previously stated, vector-based segmentation and multiresolution segmentation are applied to generate image objects to perform object-based classification in this study. The vector-based segmentation algorithm creates or converts an image object to reflect the content of a vector layer. The multiresolution segmentation algorithm consecutively merges pixels or existing image objects. Essentially, the procedure identifies single image objects of one pixel in size and merges them with their neighbors, based on relative homogeneity criteria. This homogeneity criterion is a combination of spectral and spatial criteria. In multiresolution segmentation, we set the shape parameter, shape parameter and compactness parameter to 20, 0.1 and 0.5, respectively. There are no unique values for these parameters, and their values always depends on object of interest (E. Guirado, 2021). These parameters are not needed for vector-based segmentation, since it uses vector layers.

2.3.3 Rule-based Improvement: In object-based classification, it is possible to improve the classification results and accuracy by using certain criteria during and at the end of the classification, which is one of the advantages that distinguish it from pixel-based classification methods. In addition to the spatial characteristics of the segments, the criteria can be based on the spectral properties of the image pixels contained in the segments. Classification maps that have improved based on rules (or criteria) are the final results of this classification.

3. DATASETS

In this study, very high-resolution satellite images were used to classify three urban scenes in Ulaanbaatar, Mongolia. The first scene was used as the training site, while the second and third scenes were used as test sites. The training site was selected in a location that included all the urban and geographical elements of Ulaanbaatar. And test sites were selected in locations that similar to the training site and include all the urban and geographical elements. The area of each scene is about 120 ha, and it has three spectral bands with a spatial resolution of 0.12 m, radiometric resolution of 8 bits. It contains typical residential and commercial buildings with different heights and sizes, ger districts, and other urban and geographical elements, such as water, grassland, trees, roads, parking lots, cars, concrete, and bare soil. Each scene was acquired by Maxar Technologies in September 2019 and extracted from Google Earth in March 2021. Scenes used in this study and their location is shown in Figure 5.

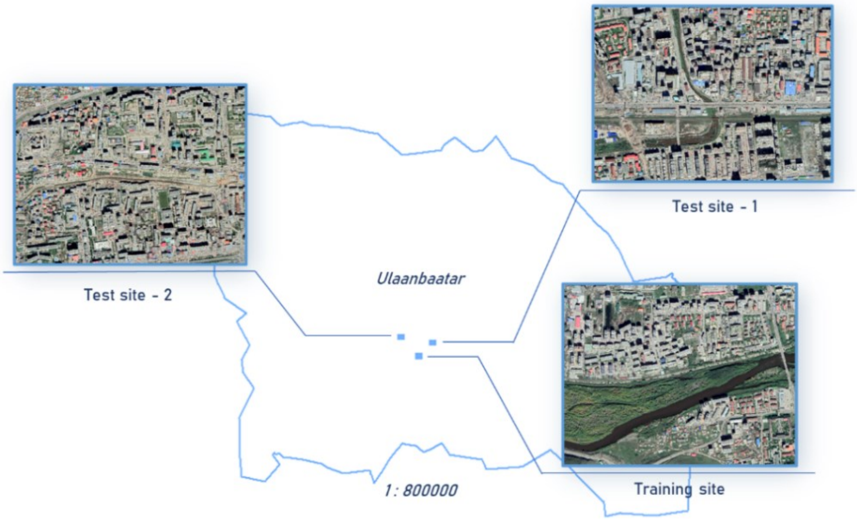


Figure 5. Scenes used in this study

4. RESULTS

This section presents the classification results of the combination of OBIA and deep learning methods. The results of building footprint extraction and car detection using Mask R-CNN, the results of land cover recognition using CNN, and the results of final classification are described in Section 4.1, Section 4.2, and Section 4.3, respectively.

4.1 Results of Building Footprint Extraction and Car Detection

For each selected very high-resolution imagery scene in Ulaanbaatar, building footprints and cars were extracted and detected using deep learning models. Examples of these deep learning model outputs are shown in Figure 6, Figure 7.

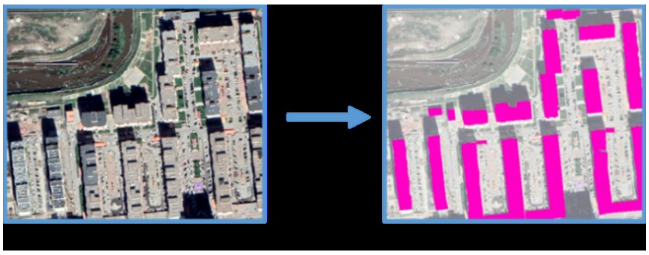


Figure 6. Example of building footprint extraction model output; (a) input image, (b) extracted building footprints

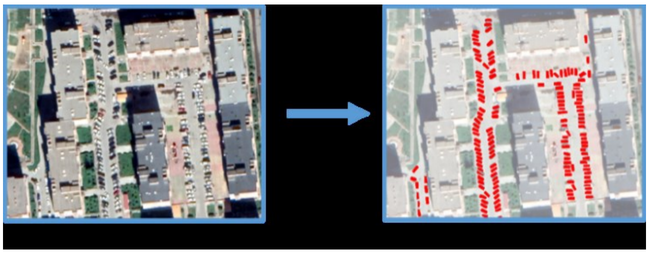


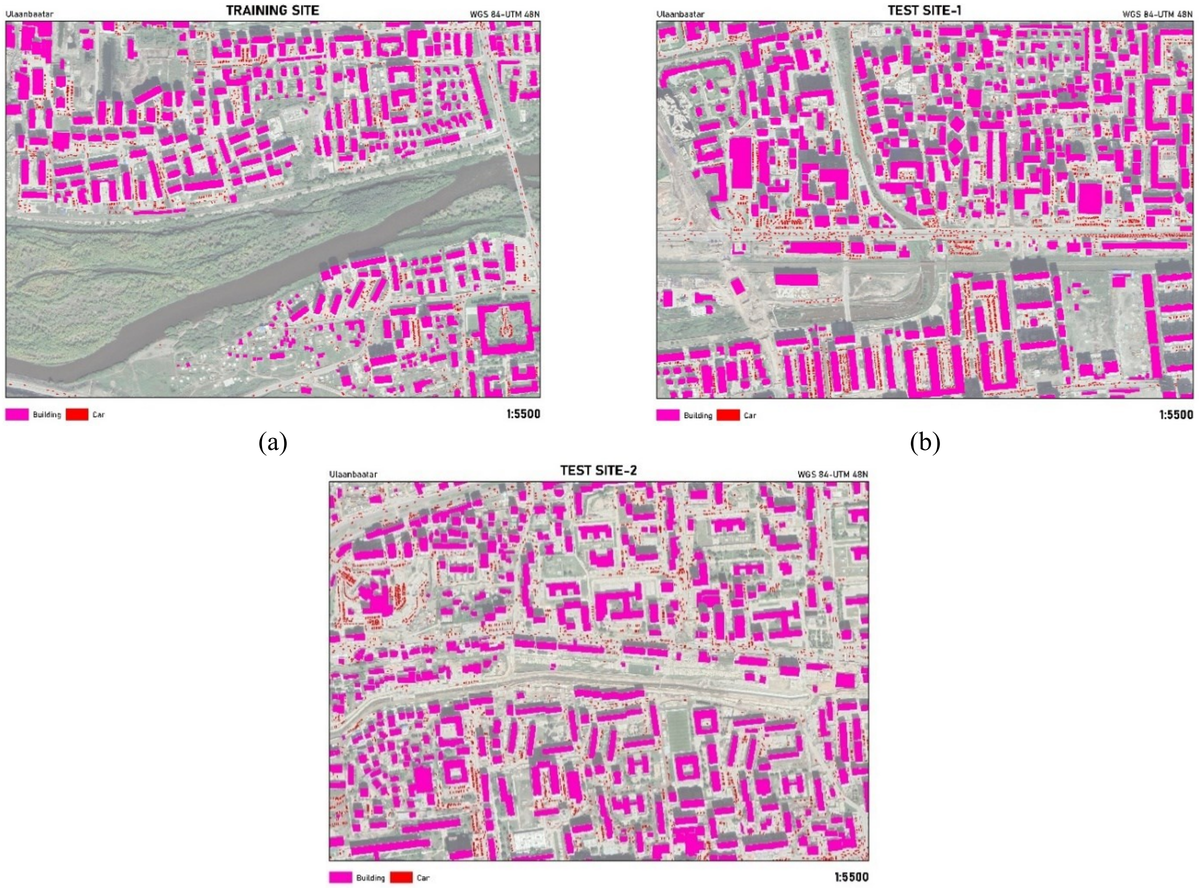
Figure 7. Example of car detection model output; (a) input image, (b) detected cars

The total count of buildings and cars extracted and detected from very high-resolution imagery of each scene using the deep learning models and their average confidence score are shown in Table 1.

Table 1. The total count of extracted and detected features using the deep learning model and their confidence score

		Training site	**Test site-1**	**Test-2**
Buildings	Total count	374	389	439
	Confidence score	96.47%	96.09%	96.26%
Cars	Total count	1458	2907	2725
	Confidence score	98.15%	97.92%	97.60%

Results of deep learning models of building footprint extraction and car detection in the selected scenes are presented in Figure 8.



(a)

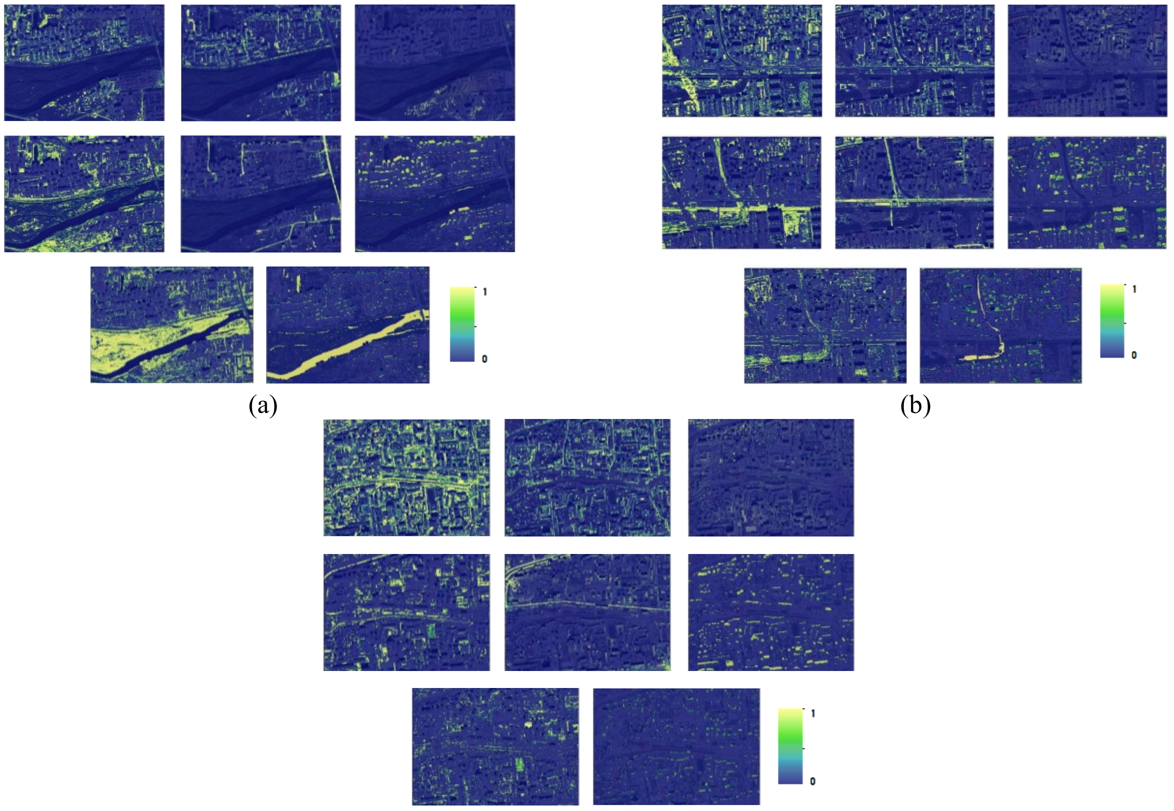
(b)

(c)

Figure 8. Results of deep learning models of building footprint extraction and car detection; (a) training site, (b) test site-1, (c) test site-2

4.2 Results of Land Cover Recognition

As a result of the deep learning model of land cover recognition, HeatMap rasters of land cover type classes were created in each scene and shown in Figure 9. As indicated above, the pixels of the HeatMap rasters have floating point values between 0 and 1. The value closer to zero denotes the likelihood of the pixel or object belonging to the class is very low. On the contrary, the higher it gets in the range between 0 and 1, it is more likely to belong to the class.



(a)

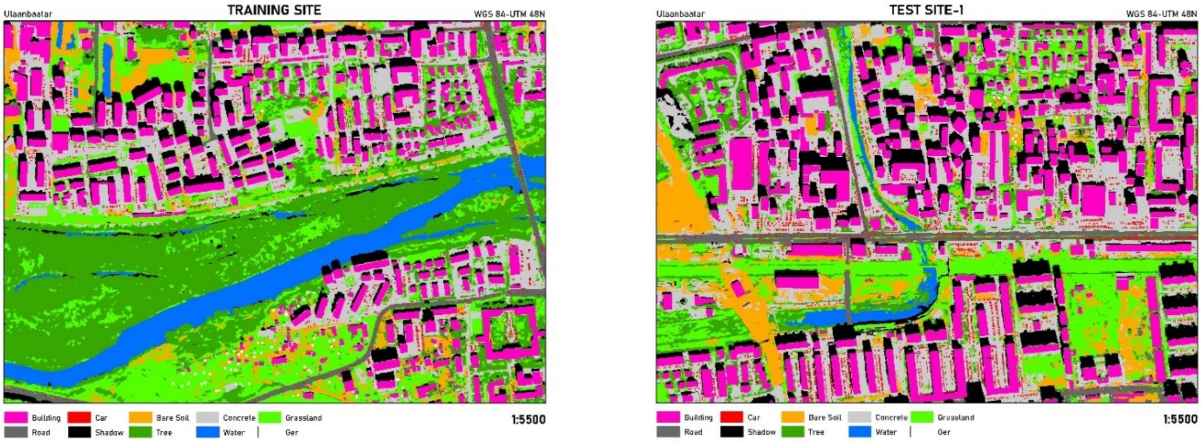
(b)

(c)

Figure 9. Results of the deep learning model of land cover recognition; (a) training site, (b) test site-1, (c) test site-2

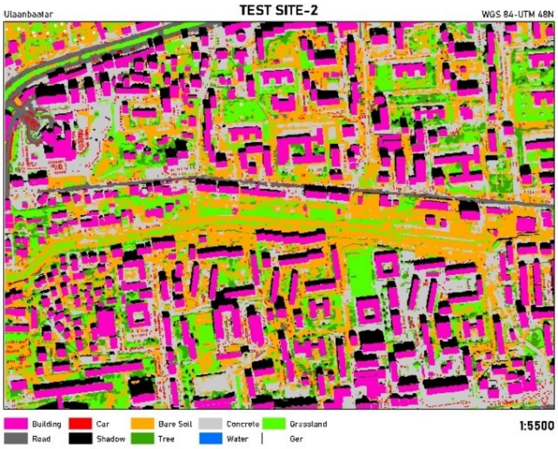
4.3 Final Classification Result

The final classification with a total of ten classes is created by combination of OBIA and deep learning methods in selected scenes in Ulaanbaatar. Classes are buildings, cars, bare soil, concrete, grassland, road, shadow, trees, water, and ger. The final classifications in selected scenes are presented in Figure 10.



(a)

(b)



(c)

Figure 10. Results of the final classification; (a) training site, (b) test site-1, (c) test site-2

The overall accuracy of classification results was above 90% in each scene, including 92.69% in the training site, 91.76% in the first test site, and 93.04% in the second test site. The class with the lowest accuracy was road (90.61%) in the training site, concrete (87.88%) in the first test site, and road (85.65%) in the second test site. The road class has lowest accuracy in two scenes, the reason for this confusion is possibly that the road has similar spectral properties to concrete. In addition, the class with the highest accuracy was bare soil (96.57%) in the training site, and 'ger' in the first and second test site, with 100% accuracy. The reason why the 'ger' class has 100% accuracy in two scenes is because of its shape, size and spectral properties. Since, the ger has fixed circular shape and size, and has a white color, it is easy to detect and classify using OBIA and deep learning. The final classification accuracies of each selected scene are shown in Table 2.

Table 2. Final classification accuracies of selected scenes (in percentage)

Class	**Training site**	**Test site-1**	**Test site-2**
Buildings	94.64	91.84	98.85
Cars	91.39	96.99	99.34
Bare soil	96.57	92.27	89.12
Concrete	96.34	87.88	98.60
Grassland	91.24	92.91	97.29
Road	90.61	88.74	85.65
Shadow	94.89	92.91	96.85
Trees	91.53	90.41	88.18
Water	93.15	94.01	-
Ger	91.78	100.00	100.00
Overall accuracy (OA)	**92.69**	**91.76**	**93.04**

5. CONCLUSION

In this study, we proposed a new method to classify very high-resolution imagery by combining Object-based Image Analysis and Deep Learning, and tested the method in three similar very high-resolution imagery scenes in Ulaanbaatar, Mongolia. OBIA methods, object-based classification and rule-based improvement are used as the primary classifier, whereas Deep Learning algorithms, Mask R-CNN and CNN are used as feature extractors in the classification. We trained and developed the deep learning models of building footprint extraction and car detection using Mask R-CNN. As well as, using CNN, we trained and developed the deep learning model of land cover recognition. Then outputs of these deep learning models were used for Object-based Image Analysis to perform final classification. The overall accuracy of classification using this method was over 90% in each scene, while the classification accuracy was 55%-60% in the pixel-based method. From this contrast, we can notice that the combination of Object-based Image Analysis and Deep Learning is completely outperforming the traditional pixel-based method in the task of classifying very high-resolution imagery of urban areas.

The results of classification indicate that the combination of Object-based Image Analysis and Deep Learning method is highly effective and suitable for classification from very high-resolution imagery of complex urban areas. In addition, these deep learning models can be used in similar urban areas and scenes, as the classification results of test sites were as high as the training site classification result. Furthermore, these deep learning models can be separately used depending on the purpose and case of study or activity.

6. ACKNOWLEDGEMENT

This study and the research behind it would not have been possible without the exceptional support of Trimble Geospatial. Special thanks to Trimble Geospatial for licensing the eCognition Developer 10.0 software used in this study.

REFERENCES

- B. A. Johnson, L. Ma, 2020. Image Segmentation and Object-Based Image Analysis for Environmental Monitoring: Recent Area of Interest, Researchers' Views on the Future Priorities. Remote Sensing, 12, pp. 1772.
- D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, W. J. Emery, 2009. Active learning methods for remote sensing image classification. IEEE Trans. Geoscience Remote Sensing, 47 (7), pp. 2218–2232.
- E. Guirado, J. Blanco-Sacristán, E. Rodriguez-Caballero, S. Tabik, D. Alcaraz-Segura et al., 2021. Mask R-CNN and OBIA Fusion Improves the Segmentation of Scattered Vegetation in Very High-Resolution Optical Sensors. Sensors, 21 (1), pp. 320.
- H. I. Sibaruddin, H. Z. M. Shafri, B. Pradhan, N. A. Haron, 2018. Comparison of pixel-based and object-based image classification techniques in extracting information from UAV imagery data. In: IOP Conference Series: Earth and Environmental Science, 169.
- I. Goodfellow, J. Bengio, A. Courville, 2016. Deep Learning. pp. 429.
- J. Brownlee, 2019. Understand the Impact of Learning Rate on Neural Network Performance, Retrieved March 21, 2021, from https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/
- J. Song, S. Gao, Y. Zhu, C. Ma, 2019. A survey of remote sensing image classification based on CNNs. Big Earth Data, 3 (3), pp. 232-254.
- K. He, G. Gkioxari, P. Dollár, R. Girshick, 2018. Mask R-CNN. arXiv:1703.06870.
- K. O'Shea, R. Nash, 2015. An Introduction to Convolutional Neural Networks. arXiv:1511.08458v2.
- M. A. Ranzato, F. J. Huang, Y. L. Boureau, Y. LeCun, 2007. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: IEEE Conference on Computer Vision Pattern Recognition. pp. 1–8.
- R. Yamashita, M. Nishio, R. K. Gian Do, K. Togashi, 2018. Convolutional neural networks: an overview and application in radiology. Insight Into Imaging, 9, pp. 611-629
- S. Ren, K. He, R. Girshick, J. Sun, 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497.
- Thomas Blaschke, Geoffrey J. Hay, Maggi Kelly et al., 2014. Geographic Object-Based Image Analysis – Towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing, 87, pp. 180-191.
- W. Rawat, Z. Wang, 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation, 29, pp. 2352-2449.
- W. Zhao, S. Du, 2016. Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. IEEE Trans. Geoscience Remote Sensing, 54 (8), pp. 4544–4554.
- W. Zhao, S. Du, W. J. Emery, 2017. Object-Based Convolutional Neural Network for High-Resolution Imagery Classification. IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, 10 (7), pp. 3386-3396.
- X. Zhang, 2018. Simple Understanding of Mask RCNN, Retrieved March 21, 2021, from https://alittlepain833.medium.com/simple-understanding-of-mask-rcnn-134b5b330e95

- Y. Jia et al., 2014. Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia, pp. 675–678.
- Y. Li, H. Zhang, X. Xue, 2018. Deep learning for remote sensing image classification. WIREs Data Mining and Knowledge Discovery, 8 (6).