CAN A GENERATIVE ADVERSARIAL NETWORK REMOVE THIN CLOUDS IN AERIAL PHOTOGRAPHS?

Toward Improving the Accuracy of Generating Horizontal Building Mask Images for Deep Learning in Urban Planning and Design

KAZUNOSUKE IKENO[1], TOMOHIRO FUKUDA[2] and NOBUYOSHI YABUKI[3]
[1,2,3]*Osaka University*
[1]*ikeno@it.see.eng.osaka-u.ac.jp* [2,3]*{fukuda|yabuki}@see.eng.osaka-u.ac.jp*

Abstract. Information extracted from aerial photographs is widely used in the fields of urban planning and architecture. An effective method for detecting buildings in aerial photographs is to use deep learning to understand the current state of a target region. However, the building mask images used to train the deep learning model must be manually generated in many cases. To overcome this challenge, a method has been proposed for automatically generating mask images by using textured 3D virtual models with aerial photographs. Some aerial photographs include thin clouds, which degrade image quality. In this research, the thin clouds in these aerial photographs are removed by using a generative adversarial network, which leads to improvements in training accuracy. Therefore, the objective of this research is to propose a method for automatically generating building mask images by using 3D virtual models with textured aerial photographs to enable the removable of thin clouds so that the image can be used for deep learning. A model trained on datasets generated by the proposed method was able to detect buildings in aerial photographs with an accuracy of IoU = 0.651.

Keywords. Urban planning and design; Deep learning; Generative Adversarial Network (GAN); Semantic segmentation; Mask image.

1. Introduction

1.1. BACKGROUND

Information extracted from aerial photographs is widely used in urban planning and design. For example, photographs allow for measurement of green coverage rates and sky view factors as well as confirmation of building locations and exteriors. As the use of unmanned aerial vehicle (UAV) technology has become more widespread, aerial photographs have become easier to take. Information that needs to be gathered in real time, such as damage to buildings during a disaster

PROJECTIONS, Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2021, Volume 2, 377-386.

can be grasped using aerial photographs taken by UAVs. To obtain highly accurate information, it is necessary to capture many photographs in a short period of time. An effective method for detecting buildings in aerial photographs is to use artificial intelligence to ascertain the current state of a target region.

Recently, methods using deep learning have been proposed for object detection and segmentation. These methods can quickly and automatically detect target objects in images. It is also possible to detect buildings in aerial photographs by using this method. The accuracy of building detection is greatly influenced by the quantity and features of the dataset used to train the model, and thus it is necessary to adequately train the model for each target area. However, the building mask images used to train the model must be generated manually in many cases. Considerable time is required to generate mask images from aerial photographs for model training because many sets of aerial photographs and mask images are needed to train the model. Photoediting programs such as Adobe Photoshop and GIMP have a function for automatically clipping target objects. However, this function is not effective for generating specific mask images, such as buildings. Therefore, an efficient method to generate mask images is needed.

1.2. PREVIOUS RESEARCH

In recent years, many object detection and segmentation methods that use deep learning have been proposed. By providing training data, features are automatically calculated and objects are detected based on the calculated features. Methods that detect objective areas as rectangles in images by using a convolutional neural network (CNN) such as AlexNet (Krizhevsky et al. 2012) and You Only Look Once (YOLO) (Redmon et al. 2016). Semantic segmentation (Long et al. 2015) classifies each pixel into one of several categories and then segments the objects by their silhouette. A system for automatically calculating green coverage rates and sky factors by semantic segmentation (Cao et al. 2019) has also been developed. A deep CNN-based method for automatically detecting suburban buildings from high-resolution Google Earth images has also been proposed (Zhang et al. 2016) as a building detection method that uses deep learning. In addition, a fused fully convolutional network model has been proposed to perform building segmentation (Bittner et al. 2018). Some research is being conducted with the aim of improving the accuracy of Mask-R-CNN for detecting building footprint boundaries. Furthermore, a method combining Mask-R-CNN with building boundary regularization (Zhao et al. 2018) has been presented, and a method has been proposed for detecting different scales of buildings and segmenting buildings to have accurately segmented edges (Zhou et al. 2019). However, the building mask images for training the model must be generated manually in many cases, which requires considerable time and expense to build.

To overcome this challenge, a method has been proposed to automatically generate mask images of buildings by using VR 3D models for deep learning (Fukuda et al. 2020). By using a 3D virtual model, we can quickly and easily create datasets that include mask images. Given that normal virtual models do not have the realism of a photograph, it is difficult to obtain highly accurate detection

results in the real world even when the image is used for deep learning training. High-precision rendering methods have been developed but they are generally difficult to use because many computers do not have high enough specifications. Using textured 3D virtual models with photographs can overcome this challenge (Ikeno et al. 2020). In addition, photographs may contain obstacles other than the target object. To remove these obstacles in photograph, image generation methods using Generative Adversarial Network (GAN) (Goodfellow et al. 2014) are used.

1.3. OBJECTIVE

The objective of this research is to propose an automatic generation method for horizontal building mask images by using 3D models with textured aerial photographs for deep learning. Specifically, we aim to improve the representation of the VR models by using textured aerial photographs on 3D models. Some aerial photographs include thin clouds, which degrade image quality. The thin clouds on these aerial photographs are removed by using GAN for improving training accuracy. The proposed method can automatically generate mask images by using these 3D models and GAN.

2. Proposed method

Our proposed method automatically generates building mask images and aerial photographs. The generated mask images are used to train the deep learning model for semantic segmentation. The proposed method loads 3D models that include terrain and building objects, classifies by building class and others class, switches between a model with all objects and one with only buildings, and finally generates two upper-view images of the models from multiple viewpoints. The game engine used in this method must be able to import 3D models, classify objects into the two classes, and output images while switching between display and non-display. Aerial photographs that include thin clouds are regenerated as images without thin clouds by using a GAN that can change from an image with one feature to another. This method can generate multiple sets of mask images and aerial photographs without thin clouds from a single 3D model. The flowchart and conceptual diagram are shown in Figures 1 and 2, respectively.

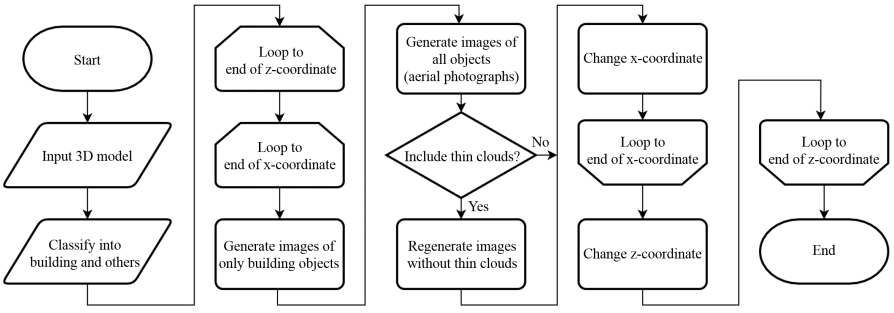

Figure 1. Flowchart of our proposed method.

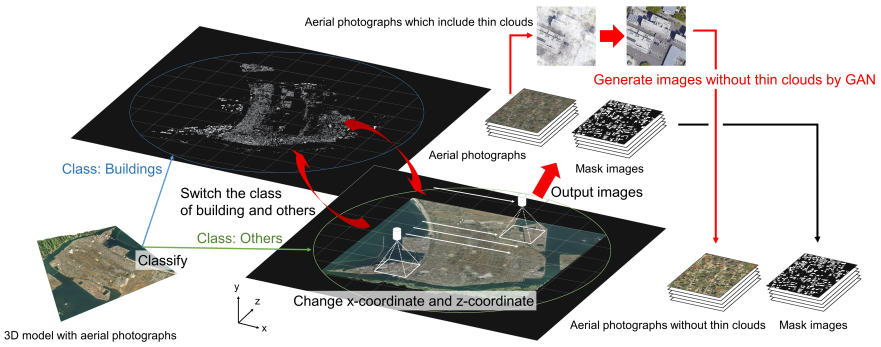

Figure 2. Conceptual diagram of our proposed method.

3. Prototype system

A prototype system was constructed to generate sets of mask images and aerial photographs without thin clouds by our proposed method. The automatic mask-image generation system using an enhanced 3D digital surface model was developed in the game engine Unity (Ikeno et al. 2020). The appearance of the 3D digital elevation model was enhanced by using textured aerial photographs. To build the systems to generate the datasets, Unity was used as a game engine because it can load 3D models.

The 3D models of the target areas were generated by using Autodesk InfraWorks. The building placement was determined according to fundamental geospatial data provided by the Geospatial Information Authority of Japan. The aerial photographs are pasted on objects in the terrain. The generated 3D model is shown in Figure 3.

Objects		Types	Classes
Object 1		Building	Building masks
Object 2		Road	Others
Object 3		Water	
…		…	
Object N	Mesh a	Terrain	
	Mesh b	Terrain	
	…		
	Mesh n	Terrain	
	…		

Figure 3. The 3D model created by Infraworks.

To generate thin cloud removal images, we used spatial attention generative adversarial networks (SpA GAN), which use a spatial attention network (Wang et al. 2019) as a generator. The architecture of SpA GAN is shown in Figure 4. The SpA GAN model trained on the open-source RICE dataset was used to generate thin cloud removal images from aerial photographs that include thin

clouds. The color tone of the generated image was corrected to match the original aerial photograph. The specifications of the personal computer used to perform all of these tasks are shown in Table 1.

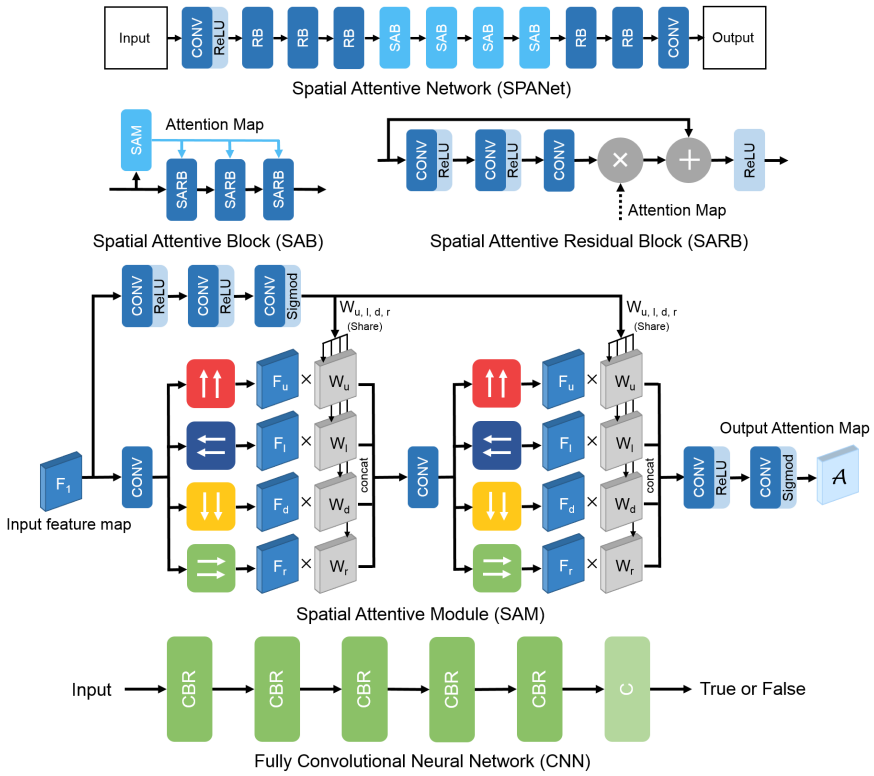

Figure 4. Generator and discriminator (Created by the author with reference to the literature).

Table 1. The specification of PC.

OS	Ubuntu 16.04 LTS
CPU	Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz
GPU	Geforce GTX 1060
RAM	28.0GB

4. Results

4.1. AUTOMATIC GENERATION OF MASK IMAGES

The aerial photographs and mask images automatically generated by our proposed system are shown in Figure 5. The middle column shows automatically generated

mask images generated by the prototype system and the right column shows manually generated mask images as ground truth. The white areas are the building masks. Our prototype system generated 6956 sets in 438 s.

Figure 6 shows the thin-cloud removal results by GAN. The left column shows aerial photographs including thin clouds and the middle column shows images in which the thin clouds were removed by GAN. The time required to generate 192 thin cloud removal images by GAN was 58 s.

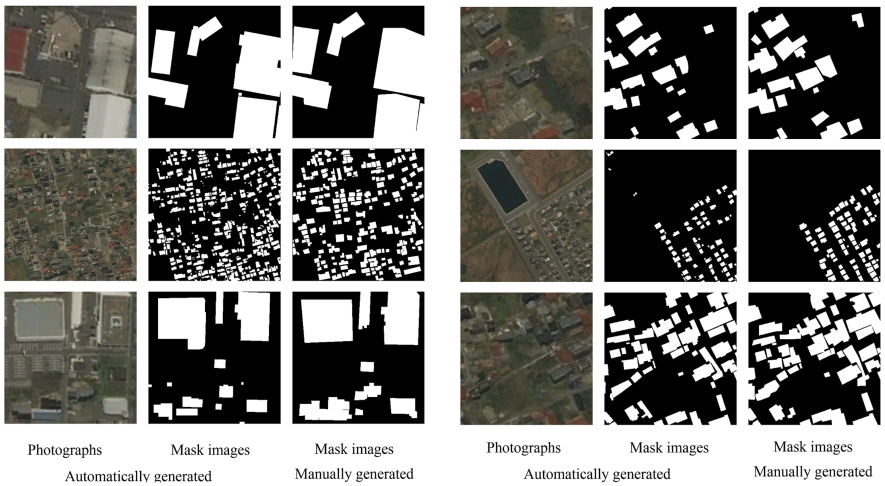

Figure 5. Generated aerial photographs and mask images.

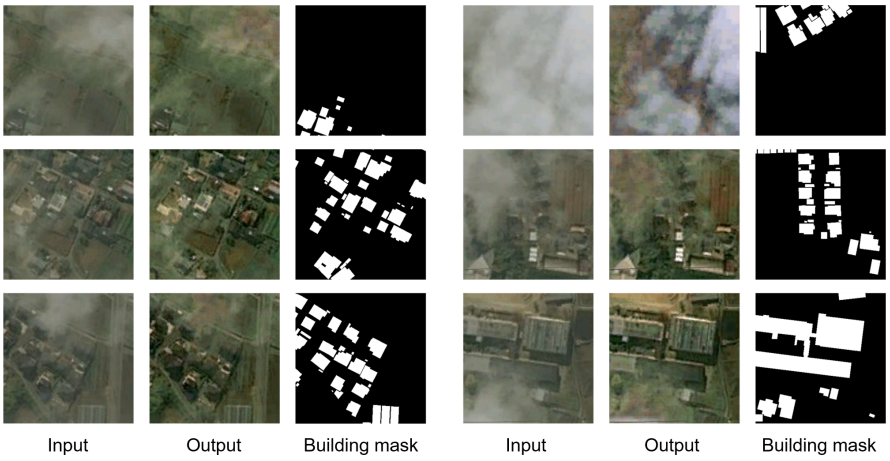

Figure 6. Thin cloud removal results by GAN.

4.2. BUILDING SEGMENTATION BY USING OUR TRAINED MODEL

To verify the accuracy of the generated dataset, the U-Net (Ronneberger et al. 2015) model was trained to detect buildings. U-Net can be used to detect objects accurately in units of pixels for the segmentation of buildings in aerial photographs. Sakaiminato City was chosen as the target area because it has several small buildings and a large proportion of land surface, neither of which are found in existing datasets. Model A is the model trained using unprocessed images and Model B is the model trained using images in which the thin clouds were removed by GAN.

The results for building detection in aerial photographs in Sakaiminato City by the trained model (Model B) are presented in Figure 7. The red areas are the predicted and true areas of buildings. For verification, Intersection over Union (IoU) (Everingham et al. 2015) was used, which is a metric that evaluates how similar predicted areas are to the ground truth. IoU was 0.651 as calculated using the sets of aerial photographs and mask images from the verification class.

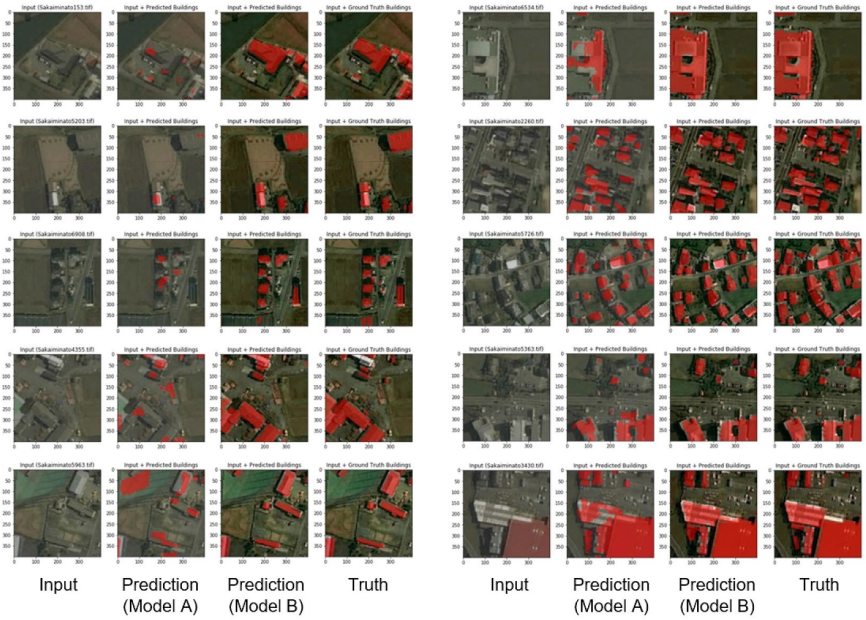

Figure 7. Results of detecting buildings by using model A and B.

5. Discussion

5.1. AUTOMATIC GENERATION OF MASK IMAGES

Our prototype system generated 6956 sets of mask images and aerial photographs without thin clouds in 438 s. The time to generate the mask images was reduced by automatically generating them from 3D models in comparison to the manual

generating method. The mask images generated by our prototype system are nearly the same as those generated manually. This method generates mask images with detailed shapes. However, it was not able to generate mask images of small warehouses. It is, therefore, necessary to prescreen the generated mask images.

5.2. THIN CLOUD REMOVAL BY GAN

The buildings that were covered by the thin cloud cover in images in which thin clouds were removed by GAN are clearly visible. The training accuracy is expected to be improved because the contours of buildings can be clearly recognized when the model is trained. However, when buildings were covered with thick clouds, they remained hidden below the clouds. Areas in which thick clouds have been removed are complemented as the ground surface. This is because the RICE dataset used for training contains many images of the ground surface. Therefore, it is better to remove images in which buildings are completely hidden by thick clouds.

5.3. ACCURACY VERIFICATION

We trained the model on mask images automatically generated by our prototype system and evaluated the accuracy of the trained model for segmenting buildings in aerial photographs of Sakaiminato City. IoU was calculated for accuracy verification by using 1388 test images that were not used for training. The IoU of our trained model (Model B) was 0.651. A comparison of the accuracy of our trained two models (Models A and B) is shown in Table 2. The detection accuracy of Model B, which was trained on the images in which the thin clouds were removed by GAN was improved compared with the detection accuracy of unprocessed Model A. The accuracy, precision, and Recall of Model B were 94.3%, 84.4%, and 74.1%, respectively. Model B is a model with few false positives. In this validation experiment, U-Net was used for comparison with the existing dataset, but the accuracy might be improved by using a more accurate deep learning model.

Figure 8 shows the detection accuracy of each image. The detection accuracy was improved for most of the images, especially for those with a high percentage of buildings in the image. The number of images in which buildings were adequately detected (IoU is over 0.5) was 132 for Model A and 1177 for Model B. The threshold value indicating that the model detected buildings sufficiently well was IoU = 0.5 (Jabbar et al. 2017). This means that it is possible to detect detailed building contours in individual images. The removal of thin clouds by GAN made the building boundaries clearer and the accuracy of the training was improved.

Table 2. IoU of the trained models.

Training datasets	**Objective area**	**IoU**
SpaceNet (Rio de Janeiro)	Rio de Janeiro	0.602
Automatically generated datasets (unprocessed)	Sakaiminato	0.622
Automatically generated datasets (thin cloud removal processed)	Sakaiminato	0.651

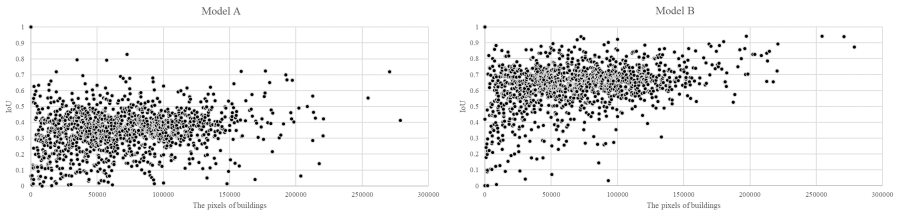

Figure 8. The detection accuracy of each image.

6. Conclusion

We improved the appearance of the 3D model by using aerial photographs as textures. We also improved the training accuracy of the deep learning model by removing thin clouds in the aerial photographs by using GAN. The prototype system using the proposed method in this study can automatically generate training datasets for deep learning, including aerial photographs and mask images, in a short time compared with methods requiring the manual generation of mask images. We believe that this method can be used not only for buildings in aerial photographs but also for other types of objects. It is expected that this method could be used to automatically generate supervised datasets for objects such as roads and rivers in aerial photographs as well as buildings seen from street level. In addition, we believe that the mask images generated by this method can be used not only as training data for deep learning but also for visualization to understand cities. The conclusions of the present study are summarized below.

- Our prototype system can generate sets of aerial photographs in which thin clouds are removed by GAN as well as mask images from a 3D model.
- Aerial photographs before and after the removal of thin clouds by GAN were compared.
- The model-trained datasets generated by our prototype system can detect buildings in aerial photographs with an accuracy of IoU = 0.651.

Future work will aim to train a deep learning model by using the datasets generated by our prototype system and evaluate the accuracy of the trained model.

Acknowledgement

This research has been partly supported by JSPS KAKENHI Grant Number JP19K12681.

References

"RICE_DATASET" : 2019. Available from <https://github.com/BUPTLdy/RICE_DATASET > (accessed 19th September 2020).

"SpaceNet" : 2019. Available from <https://explore.digitalglobe.com/SpaceNet-Thank-You.html> (accessed 9th September 2019).

"Photoshop" : 2020. Available from <https://www.adobe.com/products/photoshopfamily.html > (accessed 13th September 2020).

"GNU Image Manipulation Program (GIMP)" : 2020. Available from <https://www.gimp.org/> (accessed 13th September 2020).
"Site of basic map information" : 2020. Available from <https://www.gsi.go.jp/kiban/> (accessed 10th September 2019).
"SpA-GAN_for_cloud_removal" : 2020. Available from <https://github.com/Penn000/SpA-GAN_for_cloud_removal> (accessed 19th September 2020).
Bittner, K., Adam, F., Cui, S., Körner, M. and Reinartz, P.: 2018, Building Footprint Extraction From VHR Remote Sensing Images Combined With Normalized DSMs Using Fused Fully Convolutional Networks, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **11**, 2615–2629.
Cao, R., Fukuda, T. and Yabuki, N.: 2019, Quantifying Visual Environment by Semantic Segmentation Using Deep Learning, *Proceedings of the 24th International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA 2019)*, 623–632.
Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J. and Zisserman, A.: 2015, The PASCAL Visual Object Classes Challenge: A Retrospective, *International Journal of Computer Vision*, **111**, 98–136.
Fukuda, T., Novak, M., Fujii, H. and Pencreach, Y.: 2020, Virtual reality rendering methods for training deep learning, analysing landscapes, and preventing virtual reality sickness, *International Journal of Architectural Computing*, **16th September 2020**, https://doi.org/10.1177/1478077120957544.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: 2014, Generative adversarial nets, *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, 2672–2680.
Ikeno, K., Fukuda, T. and Yabuki, N.: 2020, Automatic Generation of Horizontal Building Mask Images by Using a 3D Model with Aerial Photographs for Deep Learning, *Proceedings of eCAADe 2020*, 271–278.
Jabbar, A., Farrawell, L., Fountain, J. and Chalup, S. K.: 2017, Training Deep Neural Networks for Detecting Drinking Glasses Using Synthetic Images, *Neural Information Processing*, 354–363.
Krizhevsky, A., Sutskever, I. and Hinton, G. E.: 2012, ImageNet classification with deep convolutional neural networks, *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS 2012)*, 1097–1105.
Long, J., Shelhamer, E. and Darrell, T.: 2015, Fully Convolutional Networks for Semantic Segmentation, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.
Redmon, J., Divvala, S., Girshick, R. and Farhadi, A.: 2016, You Only Look Once: Unified, Real-Time Object Detection, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.
Ronneberger, O., Fischer, P. and Brox, T.: 2015, Unet: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.
Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q. and Lau, R.: 2019, Spatial Attentive Single-Image Deraining with a High Quality Real Rain Dataset, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12262–12271.
Zhang, Q., Wang, Y., Liu, Q., Liu, X. and Wang, W.: 2016, CNN based suburban building detection using monocular high resolution Google Earth images, *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 661–664.
Zhao, K., Kang, J., Jung, J. and Sohn, G.: 2018, Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization, *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 247–251.
Zhou, K., Chen, Y., Smal, I. and Lindenbergh, R.: 2019, Building segmentation from Airborne VHR Images Using mask R-CNN, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 155–161.