5th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Automatic Building Extraction using Kernel-based Deep Learning Approach from VHR Imagery

Tolga Bakirman*[1], Mahmut Oğuz Selbesoğlu [1]

[1]*Yildiz Technical University, Civil Engineering Faculty, Department of Geomatics Engineering, İstanbul, Türkiye*
[2]*Istanbul Technical University, Civil Engineering Faculty, Department of Geomatics Engineering, İstanbul, Türkiye*

Keywords
Deep Learning
Building Extraction
VHR
K-Net

Abstract

Monitoring and analyzing the rapidly changing and growing cities in terms of buildings has become an important demand today. Deep learning approach has been widely used recently in the automatic extraction of buildings, which are important inputs for smart city systems. The recent studies demonstrate that the deep learning approaches greatly improves the accuracy of building extraction from the high-resolution images. The purpose of the study is to investigate the performance of K-Net architecture for building extraction from VHR imagery. In this context, The Wuhan University (WHU) Aerial Building Dataset was used for training, validation and testing. The outcomes of the study demonstrate that the extraction of buildings based on deep learning architectures provides sufficient results with 98.17 % Accuracy, 92.29 % Precision, 91.20 % Recall, 84.74 % IoU and 91.74 % F1-Score.

1. Introduction

Building information is of great importance for urban planning, monitoring engineering structures, and building deformation monitoring. In the last decade, automatic building extraction from high spatial resolution maps has become a very effective way based on recent deep learning approaches. In building extraction studies, there are three widely used platforms which are satellite remote sensing, aerial photogrammetry (Chen et al., 2017), and close-range photogrammetry based on unmanned aerial vehicles (UAV) (Zhuo et al., 2018). These systems have some advantages and disadvantages. Although remote sensing satellites can provide images with spatial panchromatic band resolution up to 0.30 m (Boonpook et al., 2021), it is affected by the orbital period and atmospheric interference. However, the mentioned systems above may have limitations in temporal resolution and cannot respond to emergency monitoring purposes. Furthermore, based on aerial platforms, approximately centimeters high-resolution orthoimages in red, green, blue (RGB) bands (U.S. Department of Interior 2011) can be produced. Widely used recent aerial imagery has many advantages of low flight cost, high accuracy, and real-time monitoring capability.

The uncertainty for extracting buildings caused by the variations in building structure and texture is a challenging factor for conventional methods such as Maximum Likelihood Classification (MLC) and Support Vector Machine (SVM) (Zhong et al. 2018) and object-based classification (Liu and Xia, 2010, Wang et al., 2004). Recently, automatic segmentation of buildings based on deep learning (DL) plays a significant role and provides efficient results especially based on high resolution datasets. The DL approaches can learn complex features depending on the given dataset and classify objects with high accuracy (Li et al., 2017). There have been various architectures throughout the years implemented for building extraction such as U-Net (Guo et al., 2020, Wang and Miao, 2022), DeepLabv3+ (Atik et al., 2022, Li and Dong, 2022), FPN (Sariturk and Seker,

* **Corresponding Author**

*(bakirman@yildiz.edu.tr) ORCID ID 0000-0001-7828-9666
(selbesoğlu@itu.edu.tr) ORCID ID 0000-0002-1132-3978

Cite this study

Bakirman T & Selbesoglu M. O. (2022). Automatic Building Extraction using Kernel-based Deep Learning Approach from VHR Imagery. 5th Intercontinental Geoinformation Days (IGD), page numbers, 160-163, Netra, India

2022), PSPNet (Yuan et al., 2022), U-net++ (Bakirman et al., 2022), etc. In this study, we aim to automatically extract building footprints using recently proposed kernel-based architecture K-Net using high resolution WHU aerial building dataset.

2. Material and methods

In this study, The Wuhan University (WHU) building open access dataset including aerial images was used (Ji et al., 2019). The featured of the dataset is given in Table 1.

Table 1. Dataset features

Image Size	512 × 512 pixels
Number of images	8189 natural image tiles
Overlap	no overlaps
Resolution	0.30 meter spatial resolution
Raw resolution	0.075 meter original data.
Number of labels	187,000 independent buildings.

Original vector data provided by land in-formation service which covers rural, residential, cultural and industrial areas of the area. The labels were improved with manually editing. A sample image tile with corresponding labels is given in Figure 1.

Figure 1. sample image tile with labels

In the study, WHU dataset separated as training, validation and test sets consisting of 4736, 1036, and 2416 image tiles, respectively.

K-Net architecture (Zhang et al., 2021) is built on a collection of convolutional kernels that have been randomly initialized and may be applied to panoptic, semantic, and instance segmentation. The semantic kernels use convolutions to produce the corresponding segmentation predictions. Globally, the kernels are dynamically modified to enhance their capacity for better discrimination. The bipartite matching approach is used to recognize objects that create a one-to-one mapping between kernels and instances in an image. K-Net architecture can be implemented for semantic segmentation, instance segmentation and panoptic segmentation. In this study, we exploited the K-Net architecture for semantic segmentation of building footprints from VHR aerial imagery. The general structure of the K-Net architecture can be seen in Figure 2.

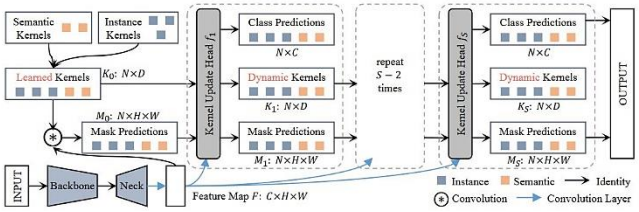

Figure 2. The general structure of K-Net (Zhang et al., 2021)

3. Results and discussion

In this study, K-Net architecture is trained on a workstation equipped with 11th Gen Intel(R) Core(TM) i9-11900 @ 2.50GHz processor and NVIDIA Quadro RTX 5000 16 GB graphic processor unit. The architecture is implemented with the MMSegmentation library for PyTorch in the Python environment. The hyperparameters used in the training are given in Table 2. We have also used pretrained weights from ADE20K dataset.

Table 2. Hyperparameters for K-Net training

Number of Images (Training)	4736
Number of Images (Validation)	1036
Number of Images (Testing)	2416
Backbone	Swin Transformer
Image Size	512 x 512
Iterations	5000
Loss Function	Cross Entropy Loss
Optimizer	AdamW
Learning Rate	0.00006
Weight Decay	0.0005
Batch Size	2
Augmentation	Random Flip, Photo Metric Distortion

We report accuracy, precision, recall, intersection over union (IoU) and F1-score in order to evaluate the results. The accuracy metrics were calculated on pixel level based on True-Positive (TP), True-Negative (TN),

False-Positive (FP) and False Negative (FN). Accuracy metrics is ratio of correctly predicted pixels to total count of all predicted pixels which is shown in Equation 1. In the case of class imbalance between the target class and background, accuracy metric may provide misleading results. In order to overcome this issue, the IoU, also known as the Jaccard Index, was calculated by the ratio of overlap area between the prediction and the ground truth divided by the area of union between the prediction and the ground truth. Precision is the ratio of pixels predicted as buildings to all pixels predicted as buildings. As can be seen in Equation 2, the precision is highly affected by FP pixels. On the other hand, recall is the ratio of pixels predicted as buildings to all pixels that are labeled as buildings in ground truth (Equation 3). Similarly, the recall metric is heavily dependent on FN pixels which are classified as background instead of building. Naturally, there is a trade-off between precision and recall. Therefore, F1-score, also known as Dice coefficient, provides a more balanced metric through harmonic mean of precision and recall which was calculated by Equation 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 - Score = \frac{Precision\ x\ Recall}{Precision + Recall}$$

The evaluation results that were calculated with the test set of the WHU dataset are given in Table 3. All accuracy metrics are above 90% except IoU. However, considering that IoU metric is quite essential for semantic segmentation studies, more experiments should be conducted to obtain a more efficient solution.

Table 3. Performance of the K-Net DL method

Accuracy	98.17 %
Precision	92.29 %
Recall	91.20 %
IoU	84.74 %
F1-Score	91.74 %

Figure 3 shows prediction examples from the test dataset. The visual inspections show that even though the used architecture can extract general structure of the buildings, it fails to preserve edge details of the buildings. Figure 3 also reveals that the DL method is more successful on larger buildings. It can be seen that the morphological features are lost especially in small sized buildings. On the other hand, the architecture is able to detect both residential and industrial buildings.

4. Conclusion

Monitoring of rapidly changing and growing cities today can be carried out by advanced remote sensing technologies based on recent deep learning approaches. It has been seen that automatic building extraction with recent kernel-based deep learning approach used in the study produces sufficient results. However, the architecture still fails to predict building edges efficiently which may require post-processing. Future studies, it is planned to perform more experiments and analyze the region with different patterns using different deep-learning techniques.

Figure 3. Prediction examples from the dataset. Top Row: Test Image, Middle Row: Ground Truth, Last Row: Predictions

Acknowledgment

The authors would like to acknowledge the Group of Photogrammetry and Computer Vision (GPCV) at Wuhan University for providing the WHU building dataset.

References

Atik, S. O., Atik, M. E. & Ipbuker, C. J. J. O. A. R. S. (2022). Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. 16**,** 024510.

Bakirman, T., Komurcu, I. & SERTEL, E. (2022). Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Systems with Applications,* 202**,** 117346.

BOONPOOK, W., TAN, Y. & XU, B. 2021. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *International Journal of Remote Sensing,* 42**,** 1-19.

CHEN, K., FU, K., GAO, X., YAN, M., SUN, X. & ZHANG, H. Building extraction from remote sensing images with deep learning in a supervised manner. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017. IEEE, 1672-1675.

GUO, M., LIU, H., XU, Y. & HUANG, Y. J. R. S. 2020. Building extraction based on U-Net with an attention block and multiple losses. 12**,** 1400.

JI, S., WEI, S. & LU, M. 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing,* 57**,** 574-586.

LI, Y., HE, B., LONG, T. & BAI, X. Evaluation the performance of fully convolutional networks for building extraction compared with shallow models. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017. IEEE, 850-853.

LI, Z. & DONG, J. J. R. S. 2022. A Framework Integrating DeeplabV3+, Transfer Learning, Active Learning, and Incremental Learning for Mapping Building Footprints. 14**,** 4738.

LIU, D. & XIA, F. J. R. S. L. 2010. Assessing object-based classification: advantages and limitations. 1**,** 187-194.

SARITURK, B. & SEKER, D. Z. 2022. Comparison of Residual and Dense Neural Network Approaches for Building Extraction from High-Resolution Aerial Images. *Advances in Space Research*.

WANG, H. & MIAO, F. J. E. J. O. R. S. 2022. Building extraction from remote sensing images using deep residual U-Net. 55**,** 71-85.

WANG, L., SOUSA, W. & GONG, P. 2004. Integration of object-based and pixel-based classification for mapping mangroves with IKONOS imagery. *International journal of remote sensing,* 25**,** 5655-5668.

YUAN, W., WANG, J. & XU, W. J. R. S. 2022. Shift Pooling PSPNet: Rethinking PSPNet for Building Extraction in Remote Sensing Images from Entire Local Feature Pooling. 14**,** 4889.

ZHANG, W., PANG, J., CHEN, K. & LOY, C. C. K-net: Towards unified image segmentation. Advances in Neural Information Processing Systems, 2021. 10326-10338.

ZHUO, X., FRAUNDORFER, F., KURZ, F. & REINARTZ, P. J. R. S. 2018. Optimization of OpenStreetMap building footprints based on semantic information of oblique UAV images. 10**,** 624.