4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Automated building extraction from very high-resolution remote sensing data with deep learning approaches

Volkan Dağdelen[1], **Ugur Alganci** *[1], **Elif Sertel** [1]

[1]Istanbul Technical University, Civil Engineering Faculty, Geomatics Engineering Department, İstanbul, Türkiye

Keywords
Remote sensing
Building extraction
Deep learning
Automated detection
Unet Architecture

Abstract

Building extraction from very high-resolution satellite images is an important task due to the various different usage of the extracted information such as population estimation, city planning, and disaster management. Manual extraction of the buildings is a labor-intensive task that is prone to human-induced errors ad mistakes. Index-based and classical machine learning approaches remain insufficient due to diversity in building geometries, changes in reflectance values, and similar properties with other objects. Recently deep learning-based approaches show promising developments and results for this task. Unet architecture is one of the most popular deep learning architectures for building extraction within the scope of semantic segmentation. This study aims to automatically detect buildings by using the Unet architecture. The Unet model was trained twice with the same hyperparameters and Resnet50 backbone on 50 epochs initially with the Massachusetts building detection dataset and secondly with a combination of Massachusetts and Inria datasets to perform a comparative evaluation. According to the independent testing results with data from Massachusetts, Inria, Pleiades and Google Earth, both datasets provided satisfactory IoU scores ranging between 0.71 and 0.89, except for the first dataset testing with Pleiades images that provide a 0.51 IoU score.

1. Introduction

Up to date building information has an important role in several applications such as population estimation, change monitoring, urban planning, smart city applications, map services, and disaster management, thus the extraction of building boundary lines from high-resolution images has always been the main research topic for remote sensing research projects (Xie et al. 2020). The emergence of high spatial resolution satellite images due to current satellite technologies made it possible to extract buildings from these images and to perform useful analyzes by using them in applications such as geographic information systems.

The initial afford in building footprint extraction from satellite images was based on the spectral characteristics and geometric features of the objects on the image, such as spectral information, colors, textures, and geometric shapes, and algorithms were mainly designed as defining thresholds on these parameters to differentiate the buildings from their environment. With the recent advances in hardware and software environments, machine learning-based classification approaches have become popular. The main algorithms for this purpose can be listed as, K-means, support vector machines, random forest, adaptive boosting, and conditional random fields (Liu et al. 2018). The drawbacks of the above-mentioned methods are that they require a high degree of prior knowledge and parameter selection and, accordingly, require a significant amount of time and labor (Liu et al. 2020; Yang et al. 2020).

In recent years, deep learning (DL) networks, especially convolutional neural networks, have been used frequently in remote sensing applications such as classification, change detection, artificial object detection, and extraction (Bakirman et al. 2022; Ekim and Sertel, 2021; Zhang et al. 2020). Over half a decade, the use of DL networks for building detection has frequently been encountered. In the study of Li et al. (2018), Unet and Deeplabv3+ architectures were applied to SpaceNet 2 dataset for the same purpose and their results provided that Unet is more efficient. Bischke et al. (2019) used SegNet architecture and Inria dataset for building segmentation and footprint extraction and they presented high IoU scores for different regions. Another study by Zhang et al. (2019) used Web-net architecture

* **Corresponding Author**

(volkandagdelenn@gmail.com) ORCID ID 0000 – 0001 – 7563 – 8111
*(alganci@itu.edu.tr) ORCID ID 0000-0002-5693-3614
(sertele@itu.edu.tr) ORCID ID 0000-0003-4854-494X

Cite this study

Dağdelen, V, Alganci, U., & Sertel, E. (2022). Automated building extraction from very high-resolution remote sensing data with deep learning approaches. 4th Intercontinental Geoinformation Days (IGD), 127-130, Tabriz, Iran

on Inria and Wuhan University (WHU) datasets and could reach over 0.85 IoU scores.

Inspired by the latest DL-based research for building extraction, this study aims to perform a comparative experiment to evaluate the effects of the training dataset on building extraction using DL-based approaches. Moreover, it presents independent testing to measure the extendability of the trained network in building extraction from different sensor data.

2. Method

In this study, Unet architecture with Resnet 50 backbone was used for building extraction purposes. The Massachusetts and Inria datasets were used for training, while Massachusetts, Inria, Pleiades, and Google Earth data were used for testing.

2.1. Data and preprocessing

The Massachusetts dataset consists of 1500 x 1500 pixel-sized 151 RGB aerial images of Boston with 1m spatial resolution. The footprints of the buildings for this dataset are extracted from OpenStreetMap. The Inria dataset includes 30 cm spatial resolution RGB aerial images with their label data (building – not building). The training and validation data of this dataset covers different cities, which makes the data suitable for such research applications. The test data used in this study includes 10 images from the Massachusetts dataset, 10 images from the Inria dataset. Additionally, 10 images from 50cm spatial resolution Pleiades satellite images and 10 images from Google Earth were used for independent testing. Building footprint labels for Pleiades and Google Earth images were manually generated in the QGIS environment.

This study uses the Google Colab and Kaggle platform to perform the analysis, thus all image sets and the label masks were cropped to create patches with 512 x 512 dimensions in order to minimize the computation bottleneck. After this process, the datasets were augmented to synthetically increase the training data amount, which is proved to improve the learning performance (Roh et al. 2019). Data augmentation parameters applied via "albumentations" toolset are provided in Table 1.

Table 1. Hyperparameter set used to train the Unet Resnet 50 model.

Augmentation Method	Parameters
Horizontal Flip	p: 0.5
Offset and Scale	0.2 scale limit, 0.1 sliding limit, p:1
Gaussian Noise	p: 0.2
Perspective	p: 0.5
CLAHE	p: 1
Sharpness	Selection probability: 0.9, p:1
Blurness	Selection probability: 0.9, blur ratio: 3, p:1
Random Brightness	Selection probability: 0.9, p: 1
Random Gamma	Selection probability: 0.9, p: 1
Random Motion Blur	Selection probability: 0.9, p: 1
Random Contrast	Selection probability: 0.9, p: 1
HSV	Selection probability: 0.9, p: 1

2.2. Model training and validation

Within the scope of the study, the Unet segmentation architecture was used as the basic architecture. The U-Net architecture is a semantic segmentation architecture proposed for biomedical purposes (Ronneberger et al. 2015). The architecture consists of two phases. The first stage is the encoder stage and consists of convolution and max-pooling layers as in classical convolutional neural networks (CNN). In this layer, there are 3x3 convolution layers, followed by corrected linear unit (ReLu) activation functions, followed by max-pooling layers containing two 2x2 sized strides for downsampling. Feature channels are doubled at each downsampling. The second stage is the decoder stage and uses transposed convolutions for precise positioning. Transposed convolutions are used for up-sampling. These convolution layers are 3x3 in size and each convolution layer is followed by the ReLU activation function. This structure is called an end-to-end fully convolutional network, since it has no density layers and only convolutional layers, it accepts any size image as input.

The python programming language and the Pytorch segmentation model library were used to implement the Unet architecture. The hyperparameters used to train Unet Resnet 50 are provided in Table 2.

Table 2. Hyperparameter set used to train the Unet Resnet 50 model.

Parameters	Used Values
Loss Function	Dice
Evaluation Metric	IoU
Optimizer	Adam
Activation Function	Sigmoid
Encoder	Resnet50
Pre Trained Weights	ImageNet
Batch Size	16
# Epochs	50
Learning Rate	0.0001 till 25th epoch, 0.00001 after 25th epoch

This study uses the intersection over Union (IoU) score metric to evaluate the model results. The IoU metric can be calculated as Formulae 1.

$$\text{IoU} = (\text{TP}) / (\text{TP} + \text{FP} + \text{FN}) \qquad (1)$$

Where TP represents true-positive, FP represents false-positive and FN represents false-negative extractions.

3. Results

When the model results are investigated, the highest score was obtained by the model trained with only the Massachusetts dataset and tested with the Massachusetts test set with a 0.8923 IoU score. On the other hand, the model trained with the Massachusetts + Inria dataset provided the highest scores for the remaining test sets (Table 3 and Table 4).

The most critical stage of evaluation is to test the models with data that are not available in both training models. One of the aims of this study is to detect different building types in images obtained from different satellites, with different resolutions, containing different regions. When the test data containing the images of the Pleiades satellite were examined, a 0.5123 IoU score was obtained only in the Massachusetts trained model and a 0.7073 IoU score in the Massachusetts + Inria trained model. When the results on the Google Earth test images were examined, a 0.8393 IoU score was obtained in the model trained with only the Massachusetts dataset, and a 0.8144 IoU score in the model trained with the Massachusetts + Inria dataset.

These results suggest that the model trained with a combination of two datasets provided more generalized performance across different image sources. Especially the improvement in building extraction from Pleiades image points out the advantage of training a model with data from different sensors for extendability requirements.

In order to evaluate the model, it is necessary to examine and interpret the images as well as the scores. Fig. 1 provides visuals from the testing results of the study. In this figure, green circles represent the better-extracted regions by the model trained with only the Massachusetts dataset and red circles represent the better-extracted regions by the model trained with Massachusetts + Inria dataset. When these visuals are interpreted, it can be asserted that both models provided similar performances for Massachusetts and Inria test sets, on the other hand, there is an obvious gain by training with the Massachusetts + Inria dataset for Pleiades images. For Google Earth test data, the model trained with only the Massachusetts dataset surprisingly provided better extractions.

Another important aspect that should be considered here is that IoU values in the training phase and validation phase converged for the Unet model trained with only the Massachusetts dataset at the 50 epochs, while a 2 percent difference was observed for the model with the combined dataset, which indicates lower learning performance on a more complicated dataset.

Table 3. Performance results of the Unet model trained with only the Massachusetts dataset

Test	#of image	Encoder	Dice Loss	IoU Score
Massachusetts	10	Resnet50	**0.0626**	**0.8923**
Google Earth	10	Resnet50	0.0924	0.8393
Inria	10	Resnet50	0.0846	0.8500
Pleiades	10	Resnet50	0.3247	0.5123

Table 4. Performance results of the Unet model trained with the Massachusetts + Inria dataset

Test	#of image	Encoder	Dice Loss	IoU Score
Massachusetts	10	Resnet50	**0.0645**	**0.8796**
Google Earth	10	Resnet50	0.1044	0.8144
Inria	10	Resnet50	0.0691	0.8742
Pleiades	10	Resnet50	0.1766	0.7073

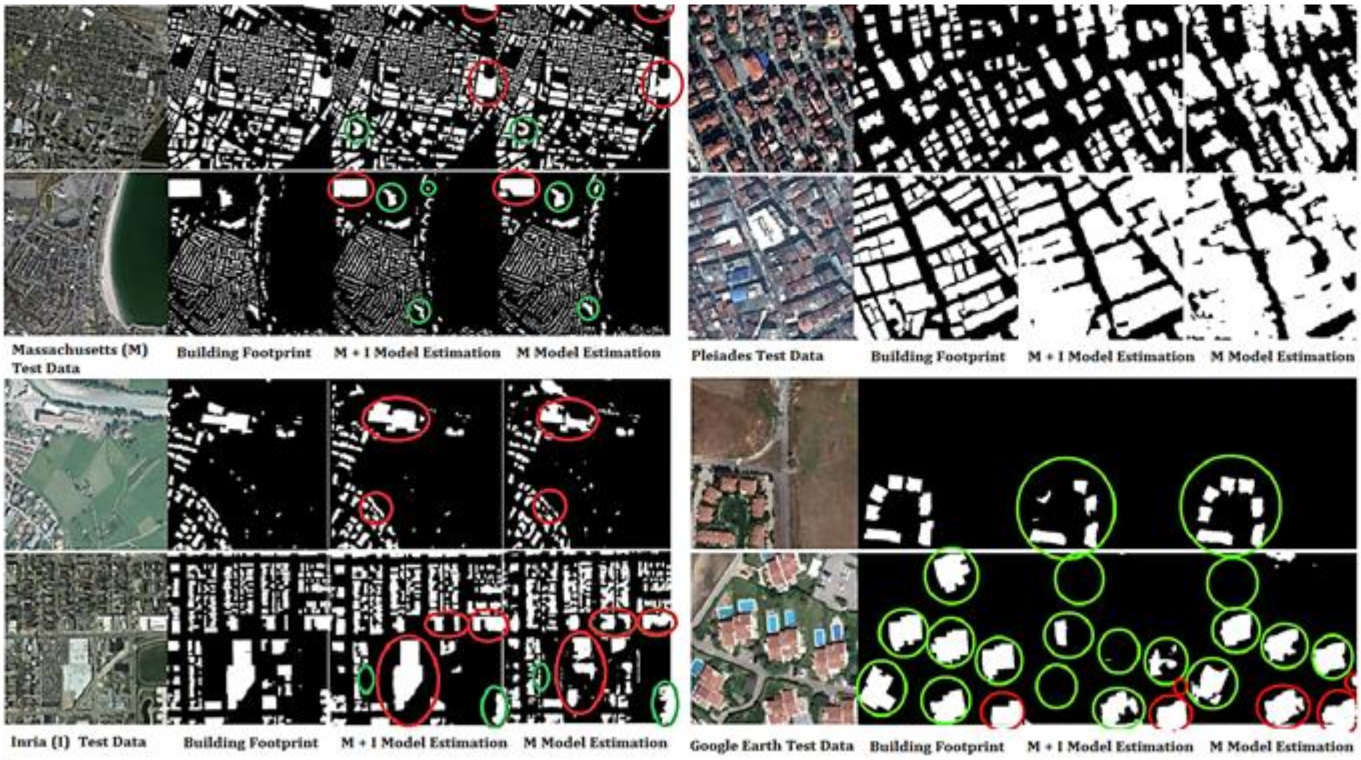

Figure 1. Sample visuals from test results. Green circles represent the better-extracted regions by the model trained with only the Massachusetts dataset and red circles represent the better-extracted regions by the model trained with Massachusetts + Inria dataset

4. Discussion

When the metric scores and test images were examined as a result of all tests, it was found that the model trained with Massachusetts + Inria dataset provided improved performance with the same hyperparameters and architecture. As a result of this study, it is observed that improved results are obtained by increasing the training data. In addition, it may be possible to obtain better results by using different segmentation architectures and encoders. The increase in the number of parameters in different encoders and the extraction of different features can affect the model performance.

It is obvious that there is a benefit in diversifying data sources, but it is also worth mentioning that this increase may not always have positive effects, as in the example of Google Earth images. This situation points out the importance of the test data set in examining the model results and that its diversity is of great importance for the interpretation and accuracy assessment of the model results.

5. Conclusion

This study evaluated the effects of the training dataset on the building extraction from very high-resolution remote sensing data by performing two experiments with different training data setups on the same Unet architecture. Trained models were evaluated across different test data setup two of which is completely independent and sourced by different sensors. Results provided that the model trained by a combination of two datasets provided comparatively higher performance and more importantly the accuracy levels seemed more stable across different test data. This finding points to the expendability of the model with combined training to be used in building extraction from multisensory datasets.

References

Bakirman, T., Komurcu, I., & Sertel, E. (2022). Comparative analysis of deep learning-based building extraction methods with the new VHR Istanbul dataset, Experts Systems with Applications, 202, 117346, https://doi.org/10.1016/j.eswa.2022.117346.

Bischke, B., Helber, P., Folz, J., Borth, D., & Dengel, A. (2019). Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks. AI4SocialGood. International Conference on Learning Representations (ICLR-2019) May 6-9 New Orleans Louisiana United States arXiv 2019.

Ekim, B., & Sertel, E. (2021). A Multi-Task Deep Learning Framework for Building Footprint Segmentation, International Geoscience and Remote Sensing Symposium (IGARSS-2021), 11-16 July, Brussels, Belgium.

Liu, Y., Zhou, J., Qi, W., Li, X., Gross, L., Shao, Q., Zhao, Z., Fan, X., & Li, Z. (2020). ARC-Net: An Efficient Network for Building Extraction from High-Resolution Aerial Images, IEEE Access, 8, 154997-155010. https://doi.org/10.1109/ACCESS.2020.3015701

Li, W., He, C., Fang, J., & Fu, H. (2018). Semantic Segmentation Based Building Extraction Method Using Multi-source GIS Map Datasets and Satellite Imagery. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 233 – 236. https://doi.org/10.1109/CVPRW.2018.00043.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 234-241.

Xie, Y., Zhu, J., Cao, Y., Feng, D., Hu, M., Li, W., Zhang, Y., & Fu, L. (2020). Refined Extraction of Building Outlines from High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1842-1855. https://doi.org/10.1109/JSTARS.2020.2991391.

Yang, G., Zhang, Q., & Zhang, G. (2020). EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images. Remote Sensing, 12 (13), 2161. https://doi.org/10.3390/rs12132161

Zhang, Y., Gong, W., Sun, J., & Li, W. (2019). Web-Net: A Novel Nest Networks with Ultra-Hierarchical Sampling for Building Extraction from Aerial Imageries. Remote Sensing. 11, 1897, https://doi.org/10.3390/rs11161897.

Zhang, Y., Li, W., Gong, W., Wang, Z., & Sun, J. (2020). An Improved Boundary-Aware Perceptual Loss for Building Extraction from VHR Images. Remote Sensing. 12. 1195. https://doi.org/10.3390/rs12071195.