DEEP LEARNING BASED BUILDING FOOTPRINT EXTRACTION FROM VERY HIGH RESOLUTION TRUE ORTHOPHOTOS AND NDSM

Mehmet Buyukdemircioglu [1, 2, 4], Recep Can[1, 2], Sultan Kocaman [2, 3 *], Martin Kada [4]

[1] Hacettepe University, Graduate School of Science and Engineering, Ankara, Turkey - mbuyukdemircioglu@hacettepe.edu.tr
[2] Hacettepe University, Department of Geomatics Engineering, Ankara, Turkey – <recepcan><sultankocaman>@hacettepe.edu.tr
[3] ETH Zurich, Institute of Geodesy and Photogrammetry, 8093 Zurich, Switzerland
[4] Technische Universität Berlin, Institute of Geodesy and Geoinformation Science, Berlin, Germany – martin.kada@tu-berlin.de

Commission II, WG II/6

KEY WORDS: Deep Learning, Segmentation, True Orthophoto, nDSM, Building, Footprint.

ABSTRACT:

A challenging aspect of developing deep learning-based models for extracting building footprints from very high resolution (< 0.1 m) aerial imagery is the amount of details contained within the images. The use of convolutional neural networks (CNNs) to tackle semantic image segmentation has been shown to outperform conventional computer vision and machine learning approaches in various applications. Here, we investigated the performances of two different CNN architectures, U-Net and LinkNet by implementing them on various backbones and by using a number of building footprint vectors in a part of Turkey for training. The dataset includes red-green-blue (RGB) true orthophotos and normalized digital surface model (nDSM) data. The performances of the implemented methods were assessed comparatively by using the RGB data only and the RGB + nDSM. The results show that by adding nDSM as the fourth band to the RGB, the accuracy values obtained from the RGB only results were improved by 3.27% and 5.90% expressed in F1-Score and Jaccard (IoU) values, respectively. The highest accuracy reflected by the F1-Score of the validation data was 97.31%, while the F1-Score of the test data that was excluded from the model training was 96.14%. A vectorization process using the GDAL and Douglas-Peucker simplification algorithm was also performed to obtain the building footprints as polygons.

1. INTRODUCTION

Buildings are key structures in cities. Building-related information is used in various applications, such as urban planning, cadastral surveying and registration, natural hazard assessments (Biljecki et al., 2015) and 3D city modelling (Halaa and Kada, 2010). Remotely sensed imagery can be segmented or classified semantically (by pixel level) to provide useful information in land cover mapping, object detection, change detection, and land-use analysis. Semantic image segmentation is an active and challenging research topic. One of the main challenges in this area is the continuously increasing resolutions of remotely sensed imagery. Despite the ability to collect small details, the very high resolution (VHR) causes difficulties in the semantic segmentation process, especially by contributing to greater class imbalance and larger variances between the classes and also within classes (Wang et al., 2016).

In many photogrammetric and remote sensing applications, automatic semantic labelling of urban areas is key to developing and updating a geographic database, monitoring changes in land cover, and extracting information about themes. Computer hardware and sensor technology advancements in recent years have enabled high-resolution samples to be analyzed so that objects such as roof tiles, cars, buildings, and individual branches of trees can be distinguished from each other. A conventional machine learning (ML) classification method typically relies on spectral, spatial, and other handcrafted features for prediction. This structure is sensitive to the expert knowledge of the region, which limits the ability to generalize their findings (Zhao et al., 2021).

Currently, deep convolutional neural networks (DCNNs) are often preferred for semantic image segmentation, whether in remote sensing or the other areas of image analysis (Marmanis et al., 2018). These networks not only classify pixels and determine their content, but also predict the structures of spatial objects. DCNNs can detect, segment or classify a wide range of objects on the ground and predict their spatial extent, including buildings, roads and junctions, trees, or building roof types (Buyukdemircioglu et al., 2021).

Semantic image segmentation usually focuses on two-dimensional (2D) data. However, with the rapid developments in acquiring and analyzing 3D data, it has become possible to use the elevation as a new dimension in addition to the optical information, which are mostly composed of red-green-blue (RGB) band. Various types of elevation information, such as depth maps, Digital Elevation Models (DEMs), or normalized digital surface models (nDSM), etc., are available in different applications. The use of elevation often improves the semantic segmentation results with added 2.5D or 3D information (Qin et al., 2016).

Most building extraction algorithms use only RGB imagery to extract spectral information about buildings (Li et al., 2019). Fusing aerial images with nDSMs could help to overcome some of the limitations (shadows, bad light, clouds, etc.) of aerial

* Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-211-2022 | © Author(s) 2022. CC BY 4.0 License.

images since nDSMs contain the height information of buildings. Our study attempts to develop a framework for the fusion of nDSMs and RGB data to improve the accuracy of building outline (footprint) estimation using VHR (0.1 m GSD) data.

The General Directorate of Land Registry and Cadastre (GDLRC) of Turkey has initiated the acquisition of VHR photogrammetric data in all city centres throughout the country, which encompass more than 11 million buildings. Manual digitization of building footprints by photogrammetry operators using VHR imagery requires a vast amount of time and is costly. Here, we aim to improve the building extraction performance based on the generated workflow through the fusion of data from VHR true orthophotos and nDSMs. The segmented building footprints were converted into vectors as a result. For this purpose, we have created a new building dataset for Izmir, Turkey with true orthophotos, DSMs, digital terrain models (DTMs), and ground truth as building vectors provided by the GDLRC, Turkey.In this study, the performances of two popular deep learning (DL) architectures, i.e., U-Net (Ronneberger et al., 2015) and LinkNet (Chaurasia and Culurciello, 2017), with different backbones were assessed comparatively. The results show that the additional height information improved the overall segmentation quality for building footprint extraction, and provided significant increase in the prediction accuracy.

The remaining of the paper is organized as follows: In Section 2, a brief overview of the recent work on building extraction with DL is given. In Section 3, the developed methodology is explained and discussed. The datasets used for the investigations and the data pre-processing steps are also explained in this section together with the implementation details. Section 4 summarizes the experimental results for both segmentation and vectorization. The discussion, conclusions and recommendations are presented in Section 5.

2. RELATED WORK

The semantic segmentation of Earth Observation (EO) data has been an active research topic in remote sensing and photogrammetry for many years (Audebert et al., 2016). When analysing EO data on an urban scale, the manual process of extracting building footprints is found extremely time consuming and expensive. The DL method involves a class of techniques in ML, in which models based on multiple layers of processing, such as neural networks (NNs), learn to represent data differently with various abstraction levels (LeCun et al., 2015). Typically, NNs and their weights are trained through supervised learning. Weights and biases are learned by training. For analysing images and detecting and interpreting patterns, convolutional neural networks are mainly used. A CNN is composed of a number of convolutional layers, which consist of filters that perform feature and thereby pattern extraction. Although conventional building extraction methods are still used in many applications, the DL and specifically CNNs have revolutionized this task. The results exhibited in several studies have demonstrated significant improvements, from image orientation to surface reconstruction, scene classification and object detection, as well as object tracking and recognition in image sequences (Heipke and Rottensteiner, 2020). Here, we summarize to most recent studies on DCNNs.

Marmanis et al. (2018) have developed DCNN models that explicitly represent and extract the boundaries between several semantic classes and segment high-resolution aerial images. In their study, a wide range of semantic segmentation architectures, including the use of class boundaries, multi-scale processing, and multi-network ensembles, was analyzed. The DCNN model performed 95.2% F1-score for the “Building” class as best result in ISPRS Vaihingen benchmark dataset. Yi et al. (2019) have developed a novel end-to-end DCNN called DeepResU-Net that effectively performs urban building segmentation at pixel scale from VHR imagery and generates accurate results. When compared to the U-Net, DeepResU-Net increased the F1 score, Kappa coefficient, and overall accuracy (OA) by 3.52%, 4.67%, and 1.72%, respectively. Jiwani et. al. (2021) have proposed a novel approach for extracting building footprints from three-channel RGB satellite imagery by using a modified DeepLabV3+ module with a dilated Res-Net backbone. Through three public benchmark datasets, their method performed state-of-the-art results that produced better-quality visuals regardless of the satellite resolution, scale, and urban density with 92.6%, 96.3% and 83.4% F1-Scores, respectively.Li et al. (2021) combined U-Net, Cascade R-CNN, and Cascade CNN deep learning models for extracting building footprint polygons from VHR aerial imagery. They compared model accuracy with semantic segmentation models on a pixel-by-pixel basis and generated building footprint polygons that are close to the reference data in terms of edges, vertices, and shapes with 92.6% precision, 91.4% recall, and a confidence of 85.1% in the WHU building dataset, respectively. Kada and Kuramin (2021) used PointNet++ and KPConv for classifying building roofs along with other classes from airborne laser scanning (ALS) data with IoU score of 0.948.

Combining different types of data sources and spectral bands is another widely used approach for semantic segmentation tasks with DL. In addition to RGB, the traditional way to provide additional information to a NN is performed through data stacking. This involves feeding the network a four-band input instead of a three-band input, while keeping the rest of the network structure unchanged. FuseNet developed by Hazirbas et al. (2017) fusion approach, by stacking the depth information with the RGB information and training the NN accordingly, the authors argued that it did not fully capitalize the depth information. Sun et al. (2021) combined true orthophotos with 0.25 m resolution and nDSMs for automated building outline extraction in a frame field learning model. By adding the 3D information from the nDSM, the accuracy and regularity were improved. On the composite image test set, the average Intersection over Union (IoU) value was 70%. When compared to 58% IoU value obtained from the RGB images only, it was demonstrated that incorporating the nDSM has improved the IoU by 12%.

By using fully convolutional networks (FCNs), Bittner et al. (2018) were able to combine spectral and height information (RGB, nDSMs and PAN images) from different sources and segment buildings in complex urban areas, thereby automatically creating building masks with full pixel resolution with 85.5% OA. Based on a CNN and recurrent NN (RNN) architecture, Zhao et al. (2021) developed a new approach for building outline extraction in vector format that takes advantage of a CNN for image feature extraction, and a RNN for decoding polygon vertices for generating regularized outlines for buildings. This was accomplished through a workflow that combined traditional feature extraction, semantic segmentation, vectorization, and shape refinement into one end-to-end DL architecture. They have made several improvements following PolyMapper's (Li et al., 2019) work, including improvements to the backbone, detection, and recurrence modules. Another approach developed by Xu et al. (2018) combined DL and guided filtering for extracting urban districts from VHR aerial imagery. Using their proposed method, the segmentation accuracy was improved by 0.43% and 2.94% for the ISPRS Potsdam and Vaihingen datasets, respectively.

3. STUDY AREA AND THE METHODOLOGY

3.1 Dataset

Our experiments were performed in a study area with a size of 4.12 km^2 and 13,269 buildings over Selcuk town in Izmir Province, Turkey. The dataset consists of four data types; i.e., true orthophotos (RGB), DSMs and DTMs in raster format with 0.1 m spatial resolution, and building footprint vectors. The data were produced from aerial images taken with 80% forward and 60% sidelap by the GDLRC. The DTMs were produced semi-automatically in a regular grid spacing and all buildings, street furniture, vegetation, etc. were removed. Also, the heights of the objects such as bridges and viaducts were reduced to the terrain level. In order to produce the nDSM, the DTM was subtracted from the DSM. The building footprint vectors, which consists of buildings and structures larger than 10 m^2 were manually delineated from stereo images by photogrammetry operators.

The dataset was split into three parts to be used for training (80%), testing (10%), and validation (10%) tasks with a grid approach. The test grid involved a high variety of buildings (2,185 buildings in total). The validation data was selected randomly using the scikit-learn library. An overview of the study area with building footprints used for training and test is shown in Figure 1. The buildings in the test region (the green square in Figure 1) were not employed in the model training stage.

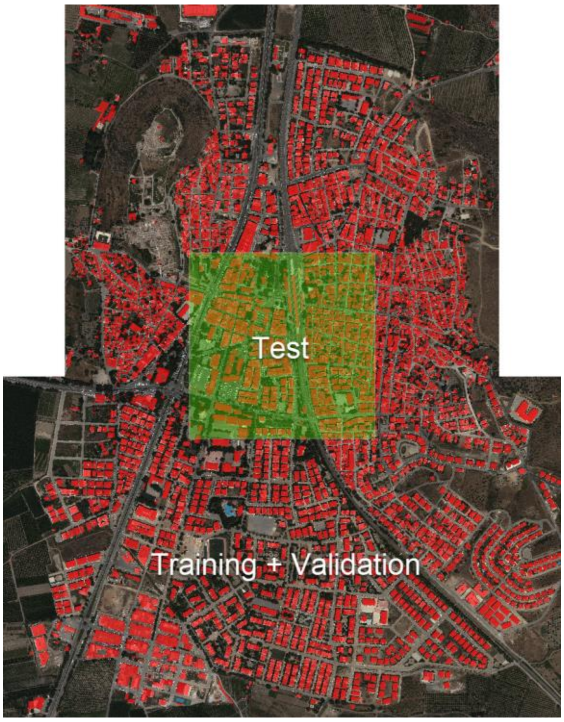

Figure 1. An overview of the study area with the building footprints (red), test area (green), and training and validation data shown on the true orthophoto.

3.2 Data pre-processing

A number of pre-processing steps were applied to produce the input features for the DL method. The manually delineated building footprint vectors were employed to create the ground truth mask. This building footprints were converted to raster format by assigning the pixel values inside the buildings as "1" and outside as "0". The raster data was transformed into non-overlapping tiles with a size of 256 x 256 pixels to be utilized in the DL models, i.e., the U-Net and the LinkNet architectures. The georeferencing information was stored as a Tiff world file (.tfw). A sample tile with the RGB true orthophoto, the building footprint, nDSM and the mask layer is shown in Figure 2.

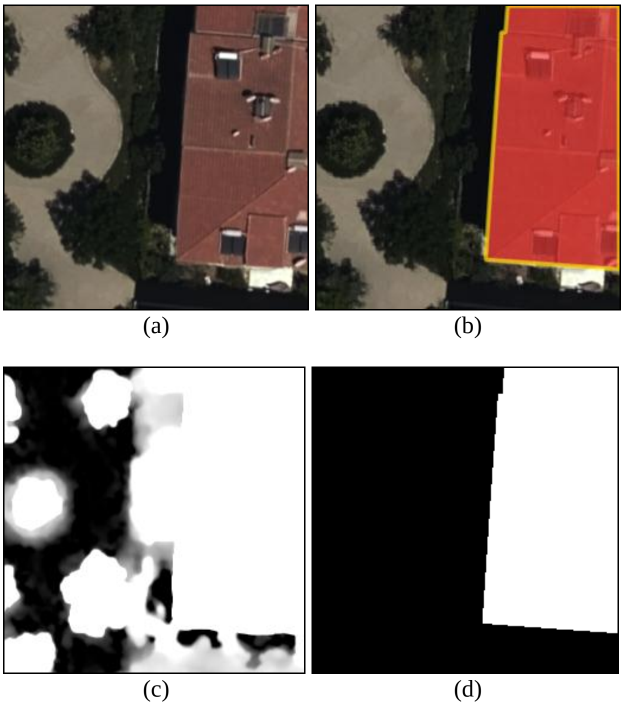

Figure 2. A sample tile from the study area: (a) RGB true orthophoto, (b) building footprint vector, (c), nDSM, and (d) building mask.

3.3 Application of the DL Methods

In order to perform a viable comparison and to obtain best possible accuracy with the data, we performed several experiments using different DL architectures and backbones. In addition, we used two different the input data (RGB and RGB-nDSM). For this purpose, the U-Net and LinkNet, which are well known for their success in image segmentation, were combined with different backbones (ResNet-18, ResNet-50 and SeResNet-18) and trained separately. The model training step was performed from scratch using the study dataset, thus no pre-trained weights were used. There are a few parameters that should be tweaked as part of the learning process, including initial learning rates, batch size, number of epochs, and other factors, such as the loss function, optimization algorithm, metrics, and data augmentation. The model training parameters for the U-Net and the LinkNet architectures are summarized in Table 1.

Parameter	**U-Net**	**LinkNet**
Backbone	ResNet-18, ResNet50, SeResNet-18	
Weight Initialization	Pre-trained	
Learning Rate	0.001 (Default)	
Optimizer	Adam	
Metrics	F1-Score	
Loss Function	BCE-Dice Loss	
Number of Epochs	100	
Data Augmentation	None	
Activation Function	Sigmoid	
Batch Size	16	
Input Size	256 x 256 x 3 (True Ortho only) 256 x 256 x 4 (True Ortho + nDSM)	

Table 1. Model training parameters

In this study, the experiments were conducted using the Adam optimizer, which iteratively updates the weights of networks. A big difference between various optimizers is the implementation of the learning rate and the update frequency of the parameters (weights). The learning rate of Adam is determined by the parameter that is maintained within a network. These rates are separately adjusted throughout the learning process. The training was executed with a default learning rate of 0.001, and sigmoid function was used for activation in classification layer.

In the model training process, different numbers of epochs were investigated to obtain the optimum value for training. A training epoch represents a cycle through the entire training dataset. Since the highest accuracies were obtained with up to 100 epochs in many models and there was no further improvement at larger values, a fixed-value of 100 epochs were selected for training. After each epoch, the model was evaluated using the validation data. When the accuracy tends to decrease in the validation dataset, i.e., the begin of the rise of the loss, then the training process could be forced to stop earlier. As the models were not overfitted during training, the early stopping function was disabled.

Batch size is another important parameter that specifies the number of samples (images) which will be propagated along the network. Different batch sizes (4, 8, 16) were used during the model training, and a batch size of 16 was chosen optimal since it provided the highest performance. F1-score was used as accuracy metric and the Binary cross entropy (BCE)-Dice was used as loss function. The F1-score is the harmonic mean of precision and recall that gives an accurate measure of incorrect classifications. The BCE-Dice loss is typically used for segmentation. Using both approaches allows some diversity in the loss while still benefiting from the stability of BCE.

Using the data augmentation techniques, a model could be prevented from overfitting by modifying the images, such as changing the pixel values, or applying geometric transformations, such as flipping, scaling, and rotation. As part of this study, no data augmentation was used. Another common solution is to use pre-trained model weights that have been trained with large datasets and fine-tuning. As part of this study, neither pre-trained weights nor fine tuning was used, and the model was trained from scratch using the generated data set. The implementation and training of the models were performed using the Tensorflow 2.5 in Python 3.8 environment on a workstation with 32GB RAM and GeForce GTX 1080 GPU. An overview of the DL framework is given in Figure 3.

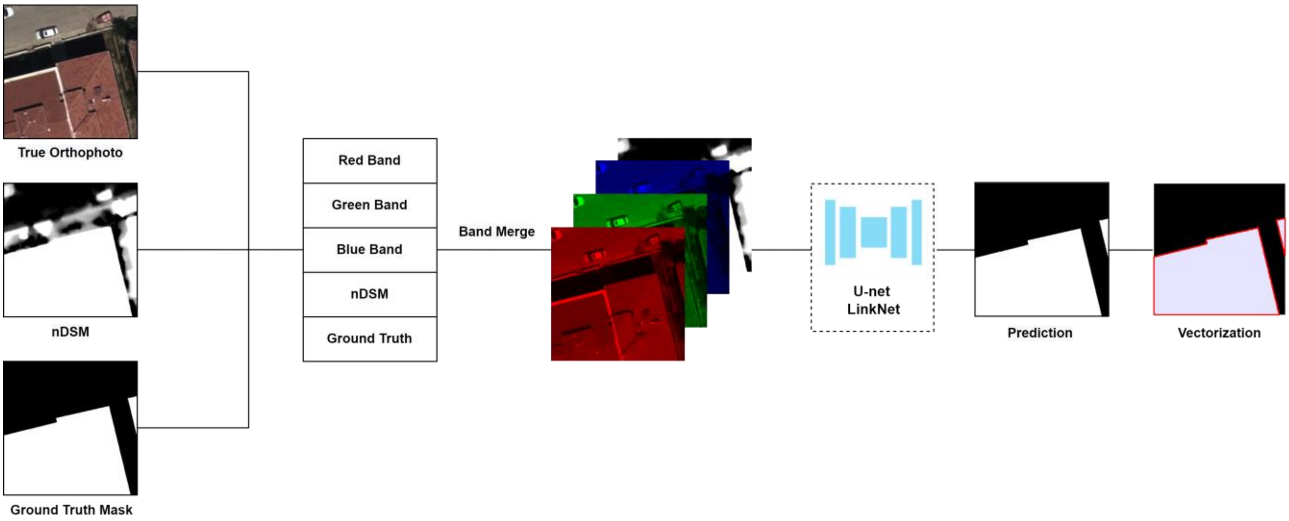

Figure 3. An overview of the workflow for building extraction and vectorization with the help of DL methods.

4. RESULTS

In this Section, results obtained from the two data sources, i.e., true orthophotos (RGB only) and the nDSM as additional information to the RGB, are provided in the following subheadings. The vectorization results are also presented and discussed.

4.1 True Orthophoto Results

The results obtained from models trained with the RGB images only are provided in Table 2. In terms of the F1-score and loss values, U-Net + SeResNet-18 achieved the highest accuracy with 0.987 and 0.025, respectively. The LinkNet + ResNet-18 provided the accuracy on the validation data with an F1-Score of 0.949. As can be seen from the Table, although the differences obtained from the models are very small, it appears that U-Net + ResNet-18 provided the highest F1-Score and Jaccard score on the test data with 0.928 and 0.867, respectively. Based on the visual analysis of the segmentation results, the U-Net and LinkNet have similar quality. However, the segmented output of LinkNet generally comprises unorganized and less homogeneous predictions compared to U-Net results. In addition, U-Net has proven success in separating structures in small areas. Trees covering the building roofs caused incorrect predictions in almost all architectures. The U-Net performed better in predicting small gaps between or in the middle of some roofs. In general, shadows were the greatest challenge for segmentation quality in all models. Shaded areas that belong to the building class are often mis-classified as non-buildings. A large portion of the lower performance in both networks could be attributed to the lack of pre-trained weights, which could be investigated in future studies. The results obtained from the different models in a part of the study area are given in Figure 4 for visual assessment of the predictions.

Model	Best Epoch Result	F1-Score	Loss	Validation F1-Score	Validation Loss	Test F1-Score	Test Jaccard (IoU) Score
U-Net + ResNet-18	78	0.981	0.037	0.949	0.157	**0.929**	**0.867**
U-Net + ResNet-50	96	0.986	0.027	0.949	0.174	0.897	0.814
U-Net + SeResNet-18	83	**0.987**	**0.025**	0.947	0.184	0.918	0.849
LinkNet + ResNet-18	94	0.984	0.030	**0.950**	0.163	0.924	0.858
LinkNet + ResNet-50	67	0.975	0.050	0.947	**0.154**	0.927	0.865
LinkNet + SeResNet-18	88	0.986	0.026	0.945	0.189	0.908	0.832

Table 2. The performance results obtained from the true orthophotos (image only results).

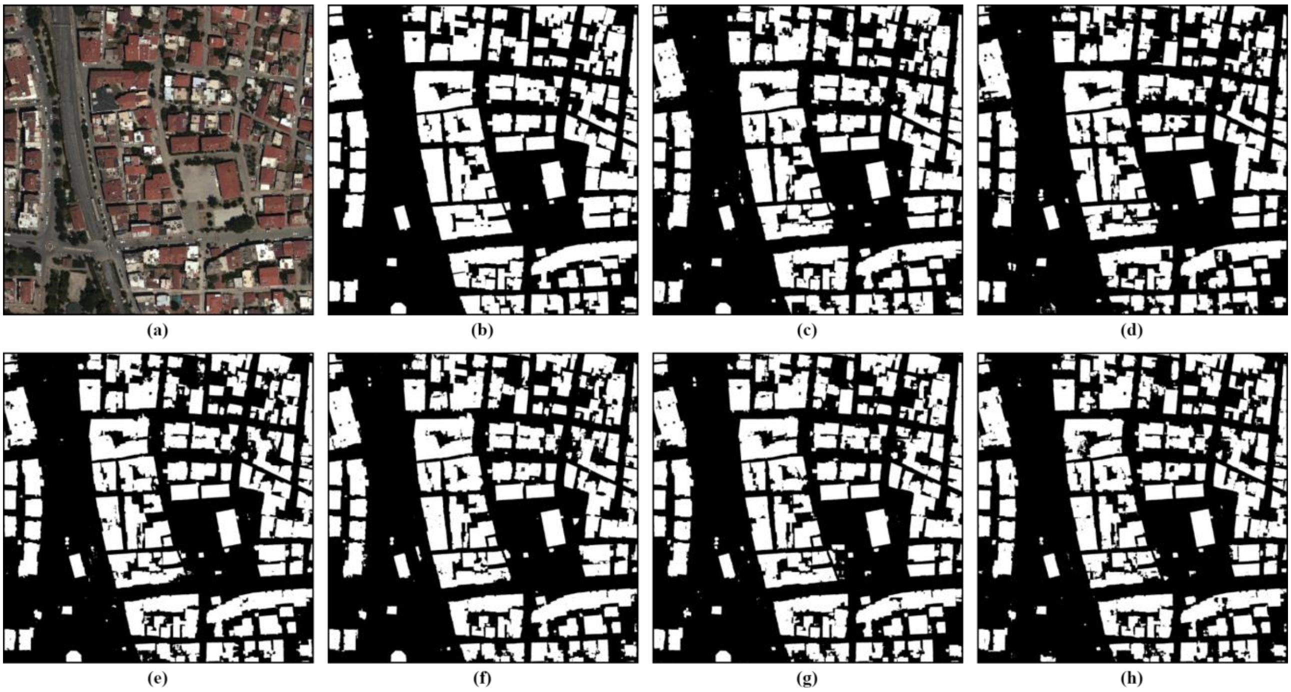

Figure 4. An overview of the RGB only predictions for the test area: (a) True orthophoto, (b) Ground truth, (c) U-Net+ResNet-18, (d) U-Net+ResNet-50, (e) U-Net+SeResNet-18, (f) LinkNet+ResNet-18, (g) LinkNet+ResNet-50 and (h) LinkNet+SeResnet-18

4.2 True Orthophoto + nDSM Results

The results obtained from the true orthophoto and the nDSM are presented in Table 3 for the implemented architectures. When compared with the results given in Table 2, the use of height information led to a significant increase in F1 and Jaccard scores. In terms of the F1-score and loss and validation F1-Scores, the U-Net with ResNet-50 backbone achieved the best results with 0.987, 0.023 and 0.973 respectively. Linknet + ResNet-50 achieved the best F1-Score and Jaccard score on the test data with 0.961 and 0.926, respectively. The validation F1-Score was improved by 2.3% when the nDSM data was combined with the RGB information. In addition, the validation loss was decreased from 0.154 to 0.073. When the F1-Score and the Jaccard scores of the test data are considered, both scores were improved significantly over the RGB only results by 3.27% and 5.90%, respectively. The visual inspection also shows that all models performed better with the height information added by providing smoother, homogenous, and structured outputs (e.g., see Figure 5). The boundary representation is also less fuzzy when the height information is added. It was also observed that the U-Net provided a higher segmentation quality with complex buildings.

Model	Best Epoch Result	F1-Score	Loss	Validation F1-Score	Validation Loss	Test F1-Score	Test Jaccard (IoU) Score
U-Net + ResNet-18	93	0.982	0.034	0.972	**0.073**	0.958	0.919
U-Net + ResNet-50	95	**0.987**	**0.023**	**0.973**	0.086	0.960	0.924
U-Net + SeResNet-18	84	0.983	0.032	0.972	0.080	0.958	0.920
LinkNet + ResNet-18	93	0.985	0.028	0.972	0.086	0.958	0.919
LinkNet + ResNet-50	89	0.986	0.027	0.973	0.079	**0.961**	**0.926**
LinkNet + SeResNet-18	95	0.987	0.025	0.971	0.093	0.960	0.925

Table 3. The prediction performance results of the true orthophoto + nDSM.

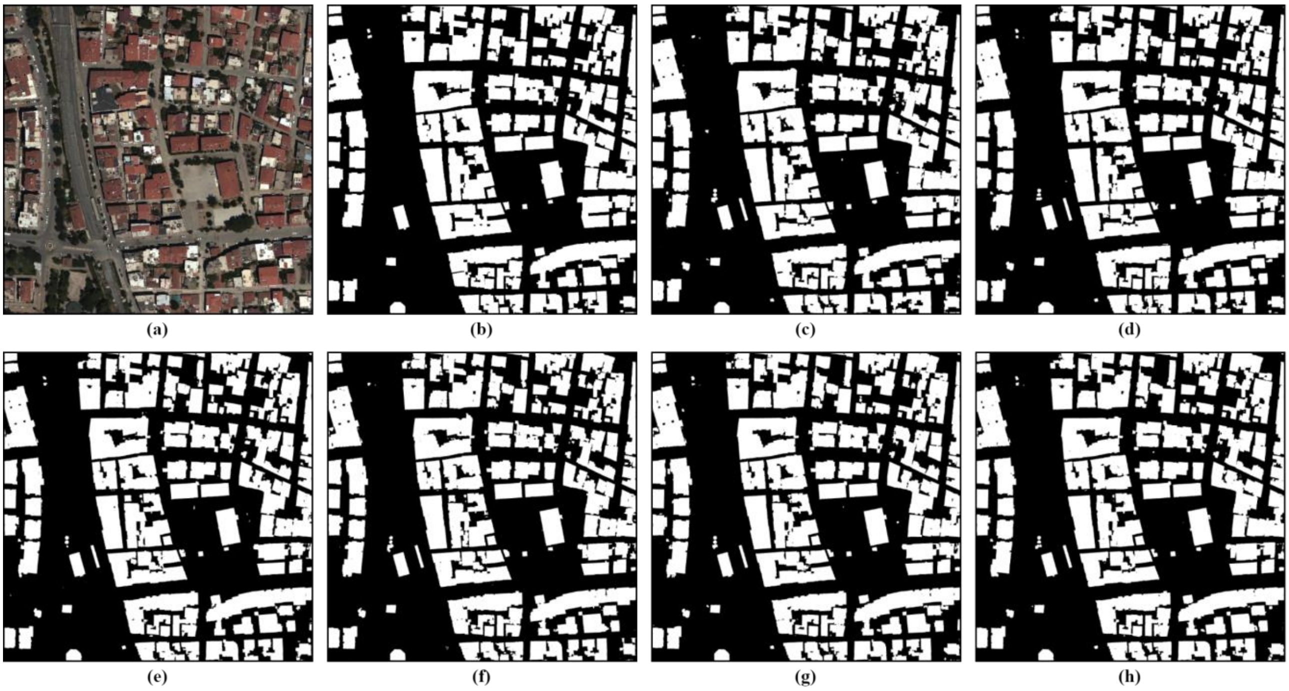

Figure 5. An overview of the RGB + nDSM predictions for the test area: (a) True orthophoto, (b) Ground truth, (c) U-Net+ResNet-18, (d) U-Net+ResNet-50, (e) U-Net+SeResNet-18, (f) LinkNet+ResNet-18, (g) LinkNet+ResNet-50 and (h) LinkNet+SeResnet-18

4.3 Vectorization

The vectorization process was performed using the GDAL (2021) library. The simplification of vector data was performed by applying the Douglas-Peucker (Visvalingam and Whyatt, 1990) line simplification algorithm to the vector data generated after the vectorization. As part of the Douglas-Peucker algorithm, it was necessary to define a tolerance value, which refers to the distance between the initial and the output geometries. The value was determined iteratively to minimize the average area changes of the geometries in the vector data. A part of the generated building footprints in vector format from the test area is shown in Figure 6. The tolerance values used for the prediction results of the different architectures and the average change values obtained from the study area are presented in Table 4.

5. CONCLUSIONS AND FUTURE WORK

The main objective of this study was to assess the performances of two different DL architectures, i.e., the U-Net and the LinkNet, implemented on various backbones for building roof segmentation from VHR aerial true orthophotos and nDSM. The results were provided for two different scenarios, such as RGB only and RGB + nDSM. The models were trained using building footprints manually delineated by photogrammetry operators. The data was provided by the GDLRC of Turkey.

The results showed that by adding the nDSM as the fourth band to the RGB data, the accuracy increased significantly. By using RGB only, the best F1-score and Jaccard (IoU) score were obtained from the U-Net + ResNet-18 model with values of 0.929 and 0.867, respectively. With the addition of nDSM, the highest accuracy results were achieved from the LinkNet + ResNet-50 model with F1-score and Jaccard score values of 0.961 and 0.926, respectively. These results indicate an accuracy improvement of 3.2% and 5.9% for F1-score and Jaccard on the test data, respectively.

Based on the visual inspection, it was observed that false predictions were caused by the roofs covered by trees, areas in the shade, and areas between close buildings. We plan to increase the amount of training data by adding buildings from different provinces and training further DL architectures. The hyperparameters can also be analyzed in more detail in future studies. By using our approach, we can effectively generate and update building footprints and reduce the manual efforts carried out by the mapping agencies. The codes and model training log of the study are available under this GitHub page: https://github.com/buyukdemircioglu/building_footprint_extraction.

Model	Data	Tolerance Value	Mean Difference
U-Net + ResNet-18	True Orthophoto	0.25	0.014 m²
U-Net + ResNet-50	True Orthophoto	0.20	0.006 m²
U-Net + SeResNet-18	True Orthophoto	0.30	0.004 m²
LinkNet + ResNet-18	True Orthophoto	0.25	0.008 m²
LinkNet + ResNet-50	True Orthophoto	0.30	0.006 m²
LinkNet + SeResNet-18	True Orthophoto	0.20	0.009 m²
U-Net + ResNet-18	True Ortho + nDSM	0.25	0.015 m²
U-Net + ResNet-50	True Ortho + nDSM	0.25	0.046 m²
U-Net + SeResNet-18	True Ortho + nDSM	0.20	0.009 m²
LinkNet + ResNet-18	True Ortho + nDSM	0.25	0.011 m²
LinkNet + ResNet-50	True Ortho + nDSM	0.25	0.015 m²
LinkNet + SeResNet-18	True Ortho + nDSM	0.15	0.025 m²

Table 4. Tolerance values used in the vectorization process and the mean difference values obtained from the predictions of the DL methods.

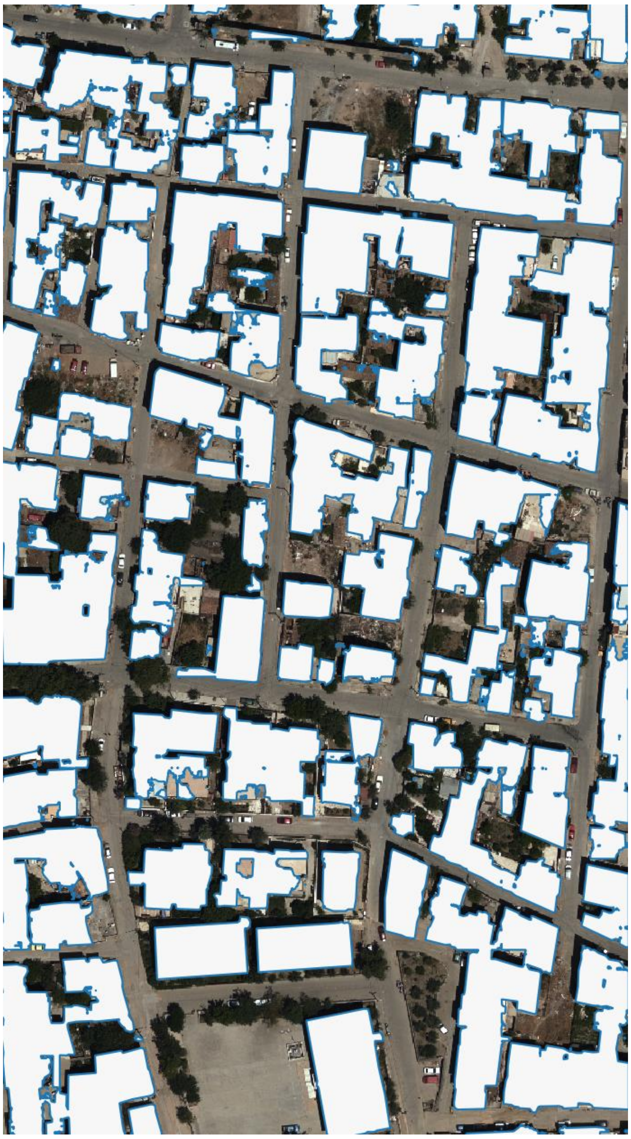

Figure 6. A close view of generated building footprints from the test area.

ACKNOWLEDGEMENTS

The authors would like to thank General Directorate of Land Registry and Cadastre of Turkey for providing the data used in this study. Mehmet Buyukdemircioglu acknowledges the support of the Scientific and Technological Research Council of Turkey (TUBITAK) 2214-A International Doctoral Research Fellowship Programme, 2224-A Grant Program for Participation in Scientific Meetings Abroad and Higher Education Council of Turkey (YÖK) for 100/2000 Ph.D. research scholarship. The paper is part of the Ph.D. thesis research of the Mehmet Buyukdemircioglu.

REFERENCES

Audebert, N., Saux, B. L., & Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *In Asian conference on computer vision (pp. 180-196).* Springer, Cham. http://doi.org/10.1007/978-3-319-54181-5_12

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S. and Çöltekin, A., 2015. Applications of 3D city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4), pp.2842-2889. https://doi.org/10.3390/ijgi4042842

Bittner, K., Adam, F., Cui, S., Körner, M., & 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(8), 2615-2629. https://doi.org/10.1109/JSTARS.2018.2849363

Buyukdemircioglu, M., Can, R., and Kocaman, S., 2021. Deep Learning Based Roof Type Classification Using Very High Resolution Aerial Imagery, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B3-2021, 55–60, https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-55-2021, 2021.

Chaurasia, A., & Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. *In 2017 IEEE Visual Communications and Image Processing (VCIP)* (pp. 1-4). IEEE. https://doi.org/10.1109/VCIP.2017.8305148

Geospatial Data Abstraction Library (GDAL), 2021. https://gdal.org (accessed on 01.04.2022).

Haala, N., & Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 570-580. https://doi.org/10.1016/j.isprsjprs.2010.09.006

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D., 2017. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. *Computer Vision – ACCV 2016*, volume 10111, pages 213–228. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science. http://doi.org/10.1007/978-3-319-54181-5_14

Heipke, C., & Rottensteiner, F., 2020. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spatial Information Science*, 23(1), 10-19. https://doi.org/10.1080/10095020.2020.1718003

Jiwani, A., Ganguly, S., Ding, C., Zhou, N., & Chan, D. M., 2021. A Semantic Segmentation Network for Urban-Scale Building Footprint Extraction Using RGB Satellite Imagery. *arXiv preprint.* https://arxiv.org/abs/2104.01263

Kada, M. and Kuramin, D., 2021. ALS Point Cloud Classification Using Pointnet++ And Kpconv with Prior Knowledge, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVI-4/W4-2021, 91–96, https://doi.org/10.5194/isprs-archives-XLVI-4-W4-2021-91-2021, 2021.

LeCun, Y., Bengio, Y. & Hinton, G., 2015 Deep learning. *Nature* 521, 436–444. https://doi.org/10.1038/nature14539

Li, Z., Wegner, J. D., & Lucchi, A., 2019. Topological map extraction from overhead images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1715-1724). https://doi.org/10.1109/ICCV.2019.00180

Li, Z.; Xin, Q.; Sun, Y.; Cao, M., 2021. A Deep Learning-Based Framework for Automated Extraction of Building Footprint Polygons from Very High-Resolution Aerial Imagery. *Remote Sensing*, 13, 3630. https://doi.org/10.3390/rs13183630

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., & Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 158-172. https://doi.org/10.1016/j.isprsjprs.2017.11.009

Qin, R., Tian, J., and Reinartz, P., 2016. 3D change detection – Approaches and applications. *ISPRS Journal of Photogrammetry and Remote Sensing,* 122:41–56. https://doi.org/10.1016/j.isprsjprs.2016.09.013

Ronneberger, O., Fischer, P., & Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Sun, X., Zhao, W., Maretto, R. V., and Persello, C., 2021. Building Outline Extraction from Aerial Imagery and Digital Surface Model with A Frame Field Learning Framework, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2021, 487–493, https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-487-2021

Visvalingam, M., & Whyatt, J. D., 1990. The Douglas-Peucker algorithm for line simplification: re-evaluation through visualization. *In Computer Graphics Forum (Vol. 9, No. 3, pp. 213-225). Oxford, UK: Blackwell Publishing Ltd.* https://doi.org/10.1111/j.1467-8659.1990.tb00398.x

Wang, Q., Lin, J., and Yuan, Y., 2016. Salient Band Selection for Hyperspectral Image Classification via Manifold Ranking. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1279–1289. https://doi.org/10.1109/TNNLS.2015.2477537

Yi, Y.; Zhang, Z.; Zhang, W.; Zhang, C.; Li, W.; Zhao, T., 2019. Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network. *Remote Sensing*. 2019; 11(15):1774. https://doi.org/10.3390/rs11151774

Zhao, W., Persello, C., & Stein, A., 2021. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 119-131. https://doi.org/10.1016/j.isprsjprs.2021.02.014

Xu Y, Wu L, Xie Z, Chen Z., 2018. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sensing*. 2018; 10(1):144. https://doi.org/10.3390/rs10010144