# AUTOMATIC BUILDING EXTRACTION FROM UAV-BASED IMAGES AND DSMs USING DEEP LEARNING

Z. Farajzadeh [1]*, M. Saadatseresht [1], F. Alidoost [2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran - (farajzadeh.zahra, msaadat)@ut.ac.ir
[2] Photogrammetry and Geoinformatics, Faculty of Geomatics, Computer Science and Mathematics, Stuttgart University of Applied Sciences (HfT), Germany - fatemeh.alidoost@hft-stuttgart.de

**Commission IV, WG IV/3**

**KEY WORDS:** Building Extraction, Deep Learning, nDSMs, UAV Images, U-Net

## ABSTRACT:

Accurate and automatic building footprint extraction from single UAV images has become essential in many photogrammetry and remote sensing applications such as 3D building modeling, smart city, monitoring, disaster management, and urban planning. In this paper, the capability of U-Net architecture with ResNet as the backbone of the network is investigated to extract the building footprints from UAV-based orthophotos and normalized Digital Surface Models (nDSMs) considering the complementary nature of RGB and height information. The data has been captured from five non-overlapping rural scenes of Yazd province, Iran. After pre-processing, the training and test datasets are prepared to evaluate the performance of U-Net using different hyperparameters and input channels such as RGB (only orthophotos) and RGBD (orthophotos and nDSMs). The experiments highlight the effectiveness of height information to detect and extract the building footprints with significant improvements in precision from 89% to 97% and in recall from 77% to 91%.

## 1. INTRODUCTION

Building detection and footprint extraction from high-resolution images are one of the most challenging tasks in many applications such as urban planning, mapping, 3D building modeling, and change detection analysis (Yuan, 2016; Liu et al., 2019). Challenges in building extraction include but are not limited to the complexity of the shape, size, texture, color, and materials of buildings, and the existence of obstacles such as trees and shadows in RGB images (Sun et al., 2019). Nowadays, Unmanned Aerial Vehicle (UAV)-Photogrammetry offers an affordable, effective and fast approach to real-time acquisition of high-resolution RGB images. The UAV images are a rich source of not only 2D but also 3D information about a scene, thanks to photogrammetric algorithms and software packages. The generated DSMs and DTMs from UAV-based images provide invariant geometric features to localize the boundary of buildings and reduce the complexity of building extraction from non-ground objects including vegetation covers, yards and garages.

There are numerous methods and algorithms for building extraction from RGB images which can categorized into two general methods as conventional and Deep Learning (DL) methods. Manually digitizing of buildings from images is a hard effort and time-consuming task. On the other hand, conventional methods which are based on rules and thresholds on features including edges, shapes, and roof types have to deal with the complexity of a building's appearance in dense man-made structures as well as noise and errors in data (Sun et al., 2019). In the past two decades, machine learning and deep learning algorithms such as Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) have shown promising results in automatic extraction of buildings from remotely sensed data (Bittner et al., 2018; Shi et al., 2019; Sun et al., 2019). In contrast with conventional methods, features in DL are extracted automatically by using convolutional layers and they are remarkably effective in dealing with large amounts of complex data (Chollet, 2018). However, the performance of CNNs depends on the quality of the training data and learning parameters, and therefore, finding the optimum hyperparameters is vital to achieving higher accuracy in building extraction.

This study aims to automatically extract the building's footprints in rural areas. The old or destroyed building structures, inaccurate ground truth, and texture similarities between building roofs and surrounding roads are major challenges of building detection in rural areas. The present paper investigates the capability of the U-Net network (Ronneberger et al., 2015) in building extraction using a combination of single images and height information (nDSMs) and contributes to the literature in three aspects:

- In this paper, we train a well-known CNN on a non-standard and non-benchmark remotely sensed data from rural scenes, and therefore, propose the most important pre-processing steps to prepare the training data with sufficient quality.
- To enhance the performance of building extraction, a weight map and a modified loss function are designed to force the estimator to pay more attention to boundaries of buildings.
- The hyperparameters such as the learning rate, regularization factor, input channels and iterations are optimized for the custom dataset.

## 2. RELATED WORK

In recent years, there are remarkable studies on building footprint extraction based on deep learning algorithms. Bittner et al. (Bittner et al., 2018) developed fused-FCN4s to fuse three types of remotely-sensed data including RGB and PAN images as well as nDSMs for building semantic segmentation. Their results showed that the fusion of nDSMs with spectral images can

* Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W1-2022-171-2023 | © Author(s) 2023. CC BY 4.0 License.

provide accurate boundaries. Xu et al. (Xu et al., 2018) proposed a ResU-Net model to extract buildings accurately, and guided filters to fine-tune the output of the neural network. Alidoost et al. (Alidoost et al., 2019) employed a multi-scale FC-CNN with the combination of Active Contour Models (ACMs) for boundary extraction of buildings from single aerial images with an accuracy of 68%. To extract buildings from very high-resolution panchromatic satellite images and DSMs, Schuegraf and Bittner (Schuegraf and Bittner, 2019) proposed a hybrid FCN model including two U-Net architectures to extract depth and spectral information and then to fuse the extracted information to detect buildings with accuracy of 97%. Liu et al. (Liu et al., 2019) used two chained CNNs for building footprint extraction and four-channel images composed of UAV-based images and DSMs as the input of the network. Wei et al. (Wei et al., 2019) trained a Multiscale Aggregation Fully Convolution Network (MA-FCN) using aerial images and then applied two post-processing algorithms to refine the output. Yu et al. (Yu et al., 2020) employed a MA-FCN for building footprint extraction using aerial images and corresponding DSMs. For more accurate results, the DSM was classified into buildings and non-buildings and then contour extraction and regularization were applied to extract structured boundaries with the precision of 74%. Recently, Alsabhan and Alotaiby (Alsabhan and Alotaiby, 2022) compared two types of the backbones for U-Net network such as ResNet50 and ResNet152 to extract building footprints with an accuracy of 90%. Their results show that ResNet50 is sufficient for the building extraction task. Also, Buyukdemirciog et.al. (Buyukdemircioglu et al., 2022) trained U-Net and LinkNet architectures using different backbones such as ResNet18, ResNet50 and SeResNet50 to investigate the effect of different architectures on the building extraction. The higher accuracy was obtained by U-Net and ResNet50.

## 3. PROPOSED METHOD

In this paper, a sequential workflow is utilized for automatic extraction of building footprints based on supervised image segmentation techniques (Figure 1). Therefore, the first step of the proposed approach is to prepare train and test datasets as well as annotated data using high-resolution UAV images. Since the datasets are not standard or benchmark data, different pre-processing methods are vital which will be explained in details in the next sub-sections. After data-preparation, a U-Net architecture (Ronneberger et al., 2015) is trained using augmented training dataset. In this step, hyperparameters for training as well as the loss function are optimized to improve the performance of the U-Net on the custom dataset which will be explained in the following sub-sections.

### 3.1 Data Preparation

In this paper, 2D Ortho Image Mosaics (OIMs) are enriched with height information of buildings which is particular useful to have a robust network to detect the building footprints from non-ground and non-building footprints. To aim this, the outlier and errors should be first eliminated from the generated DSMs.

To remove noises from generated DSM, the 3-sigma rule is applied to the points and the points with the height variation higher than 3*SD (i.e. Standard Deviation) are removed as outliers. In the next step, Digital Terrain Models (DTMs) are generated from DSMs using the LAStool (Isenburg, 2014) and LidR tool (Roussel and Auty, 2021). Finally, the nDSM is the difference between DSMs and DTMs.

To generate the annotated data, a 3D vector GIS map is first generated from stereo images, manually. Next, the 3D Vector MAPs (VMAPs) of buildings are selected to prepare the ground truth. The VMAPs are manually corrected for missing buildings by overlaying the OIMs and corresponding DSMs. Then, the VMAPs are rasterized using GDAL (GDAL/OGR contributors, 2022). The results are stored as Rasterized MAPs (RMAPs) in which 0 and 1 values represent non-building and building pixels, respectively. To evaluate the effect of input layers of the CNN on the final segmentation results, RGB orthophos and corresponding nDSMs are combined as a four channel raster image (RGBD). Next, the RGBD images and the corresponding ground truths are cropped to different sizes as 512×512, 1024×1024, and 1536×1536 image tiles with 25% overlap and then resized to 512×512 pixels. Finally, to increase the training datasets, data augmentation techniques like randomly rotating, scaling, and vertical and horizontal flipping are applied to training data.

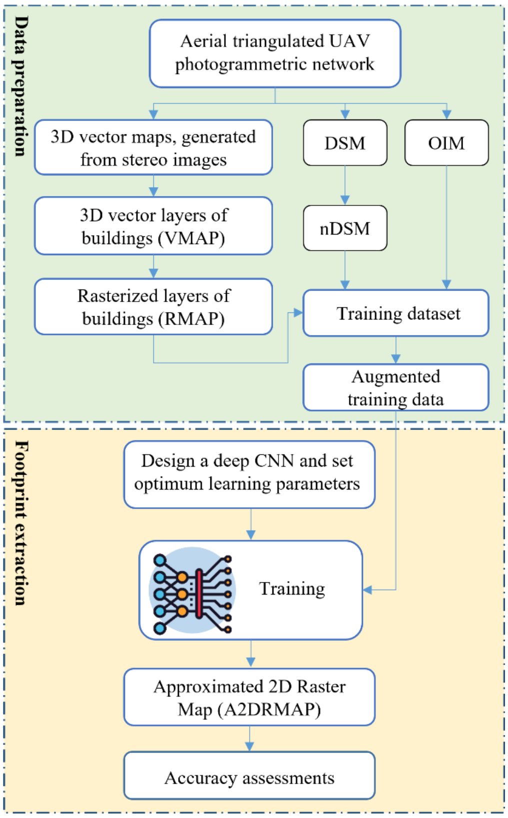
Data preparation
Aerial triangulated UAV photogrammetric network
3D vector maps, generated from stereo images
DSM
OIM
3D vector layers of buildings (VMAP)
nDSM
Rasterized layers of buildings (RMAP)
Training dataset
Augmented training data
Footprint extraction
Design a deep CNN and set optimum learning parameters
Training
Approximated 2D Raster Map (A2DRMAP)
Accuracy assessments

**Figure 1**. The proposed method.

### 3.2 Footprint Extraction

Among many choices for a CNN architecture, the U-Net architecture is selected to train aerial images. U-Net is one of the most common networks for semantic segmentation that was first introduced on biomedical images (Ronneberger et al., 2015). The U-shaped network contains two main parts such as the concatenating path (encoder) and the expanding path (decoder). The concatenating path contains convolution layers and down-sampled feature maps to capture the context information. The expanding path is to up-sample the feature maps to the original resolution. An important part of the U-Net model is the skip-connection layer that connects the concatenation path to the expanding path. It helps the decoder path to learn better-localized information from the encoder (Ronneberger et al., 2015). Moreover, the U-Net-based feature extractor needs to be deep enough to extract complex features from high-resolution aerial images. Therefore, we utilized the ResNet-50 (He et al., 2015) as the encoder (Khoshboresh-Masouleh et al., 2020; Abdollahi and Pradhan, 2021; Dixit et al., 2021). Besides, the weights are initialized based on the pre-trained ImageNet model (Deng et al., 2009). Totally, there are 32,564,253 parameters in the model and 32,516,693 of them are trainable parameters. To train the CNN, a customized loss function is proposed by Equation 1.

$$loss\ function = L_{wbce} + L_{wdice} \tag{1}$$

where $L_{wbce}$ is the weighted binary cross-entropy function (Ronneberger et al., 2015) and $L_{wdice}$ is the weighted dice loss function (Sudre et al., 2017), given by Equations 2 and 3.

$$L_{BCE} = \frac{1}{N}\sum_{i=1}^{N} -(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \tag{2}$$

$$L_{dice} = 1 - 2\frac{\sum_{c\in C}\sum_{x_i\in X} p_c(x_i) r_c(x_i)}{\sum_{c\in C}\sum_{x_i\in X}\left(p_c(x_i) + r_c(x_i)\right) + \varepsilon} \tag{3}$$

where $N$ = number of pixels in one patch
$y_i$ = true labels
$p_i$ = predicted labels
$p$ = softmax prediction of class $c$
$\varepsilon$ = small number to avoid the zero division.

If the predicted class for a pixel is $c$, $r_c$ is equal to 1 and otherwise, it is equal to 0 (Sudre et al., 2017). To force the model to learn the boundaries of the buildings, the weight map is computed by Equation 4.

$$w(x) = w_c(x) + w_0 * \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right) \tag{4}$$

where $w_c$ = weight of the current class
$w_0$ = weight for boundary pixels
$\sigma$ = width for boundary pixels.

The weight maps are generated for the training dataset in the size of 512×512×1 pixels, given by Equation 2. The parameters of $\sigma$ and $w_0$ are experimentally set to 20 and 10, respectively. The weights (e.g. $w_c$) for building and non-building classes are computed by Equation 5.

$$w_c = \frac{n_t}{2 * n_c} \tag{5}$$

where $n_t$ = number of pixels in the training dataset
$n_c$ = number of the pixels belonging to the class $c$.

In this study, $w_c$ for building and non-building classes are 3.5 and 0.5, respectively. To evaluate the performance, the quality measures of precisions (or correctness), recall (or completeness) and F1-score is calculated by Equation 6 using the predicted footprints (A2DRMAP in Figure 1) and the ground truth.

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F1 &= \frac{2TP}{2TP + FN + FP} \\ IoU &= \frac{TP}{TP + FN + FP} \end{aligned} \tag{6}$$

where TP = True-positive
FP = False-positive
FN = False-negative

*TP* refers to the pixels that are truly classified. *FP* refers to the pixels that are not classified as true. *FN* refers to the pixels that are wrongly classified as true.

## 4. EXPERIMENTS AND RESULTS

The main dataset consists of aerial UAV photogrammetric images from five rural areas in Yazd city, Iran. The RGB images are captured by a Phantom 4 Pro UAV with a GSD of 5 cm over Aghda, Zardin, Haftabad, Kahduiyeh, and Karimabad, as shown in Figure 2. The first four villages including 4720 buildings are considered for training and validation datasets which are divided by a 70% to 30% ratio, while the last village (e.g. Karimabad) including 917 buildings is a test data to evaluate the trained model. Besides, DSM and DTM are generated with a GSD of 10 cm. The image tiles are pre-processed and copped to a size of 512×512 pixels with four channels as RGB and corresponding nDSM, as shown in Figure 3. After applying data augmentation, the training dataset is increased from 4333 to 9976 images.

To optimize the hyperparameters for training, several tests are designed using U-Net and RGBD data, as shown in Table 1. According to the precision and recall results on validation data, hyperparameters such as the Learning Rate (LR), L2 regularization, number of epochs, and batch size are set to 0.0001, 0.01, 50, and 4, respectively. In addition, the Adam algorithm is employed as the optimizer with the parameters of beta 1 of 0.9 and beta 2 of 0.999. The L2 regularization is also applied to prevent overfitting. The optimized learning settings are reported in Table 2. The proposed model is trained using Tensorflow Keras Framework on Nvidia Geforce RTX 2080 Ti.

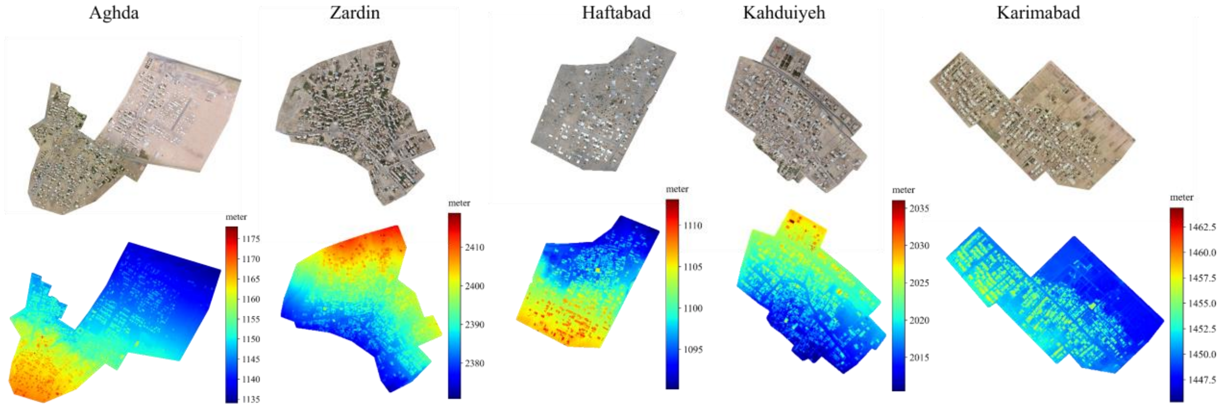

**Figure 2**. Overview of training and test datasets.

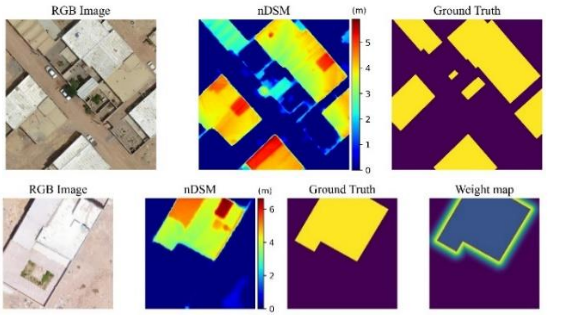

**Figure 3**. Test (the first row) and train (the second row) data.

| Parameters | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| LR | 0.01 | 1e-4 | 1e-4 | 1e-4 | 1e-4 | 1e-4 |
| L2 | - | - | 1e-2 | - | 1e-2 | 1e-2 |
| Early Stop | - | - | - | yes | yes | yes |
| Epoch | 50 | 50 | 50 | 6 | 24 | 16 |
| Precision (%) | 98 | 93 | 91 | 83 | 91 | **94** |
| Recall (%) | 18 | 59 | 72 | 82 | 89 | **90** |
| IoU (%) | 18 | 56 | 67 | 70 | 81 | **85** |

**Table 1**. Tests on validation data to find the optimum hyperparameters.

| Settings | Values |
|---|---|
| Network | U-Net |
| Backbone | ResNet50 |
| Initial weight | ImageNet |
| Learning rate | 0.0001 |
| Optimizer | Adam |
| Regularization (L2) | 0.01 |
| Batch size | 4 |
| Epoch | 50 |
| Input size | (512,512,4) RGBD, (512,512,3) RGB |

**Table 2**. The selected learning parameters.

To select an appropriate loss function and evaluate the effect of the dice loss function, one network is trained using a weighted cross-entropy loss function with and without the dice component. As shown in Figure 4, the accuracy of prediction on the validation data are significantly improved by adding the dice loss function.

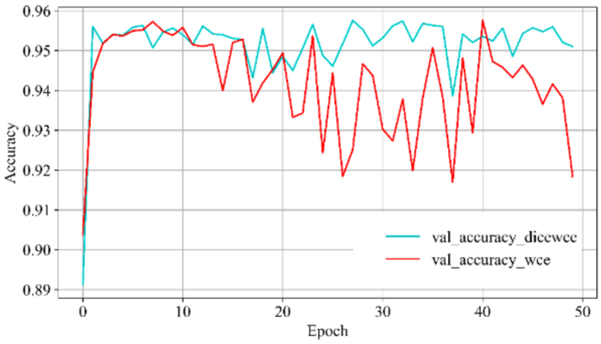

**Figure 4**. Accuracy comparison between two models trained using cross entropy and dice loss functions.

The color, textural, and structural similarities between buildings and other 3D objects make it difficult to distinguish buildings when attempting to extract building footprints using only spectral features, and building shadows cause color differences and confuse the CNN model. On the other hand, if the height information is only employed in the input layer, any elevated object could be extracted as a building. To retain the advantages of both spectral and height data, RGB images and nDSMs are concatenated. Besides, the building edges are extracted more accurate and sharper in nDSMs. Therefore, two strategies are designed to investigate the effect of height information on the footprint extraction as the RGB model to use only RGB images to train the model and the RGBD model to fuse the spectral and height data as the input layer for the network (Figure 5).

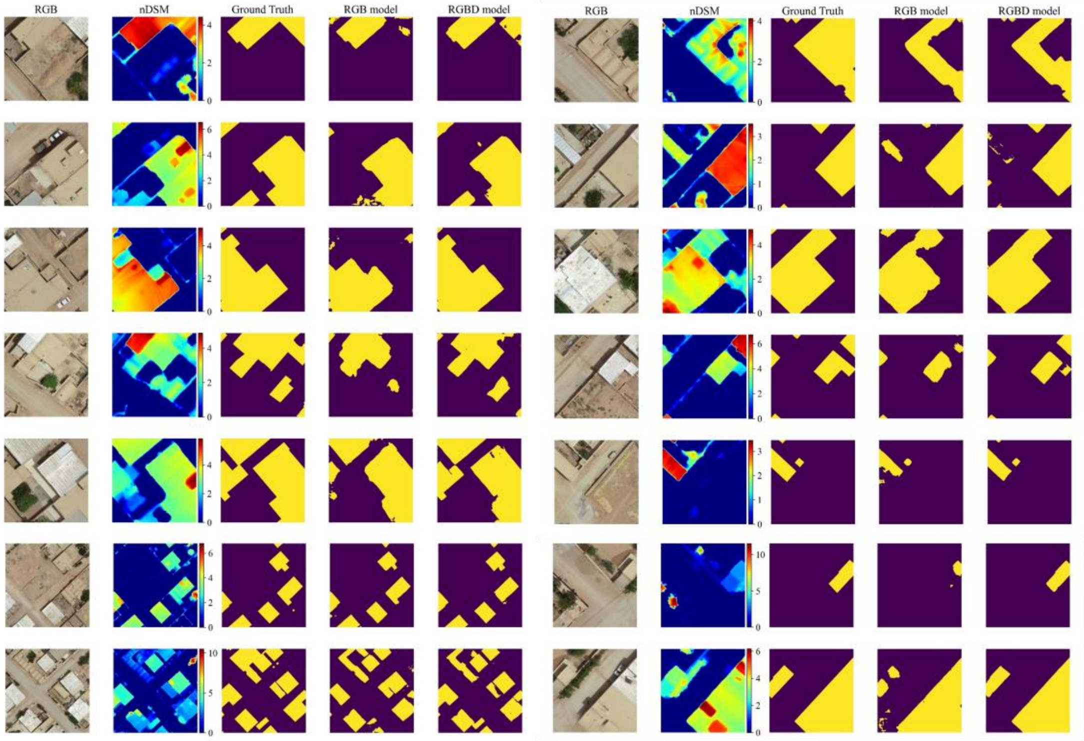

**Figure 5**. The results of RGB and RGBD models on the test samples.

https://doi.org/10.5194/isprs-annals-X-4-W1-2022-171-2023 | © Author(s) 2023. CC BY 4.0 License.

As shown in Table 3, the average of precision and recall measures using RGBD model are about 97% and 91% respectively, while these values are about 89% and 77% for the RGB model.

| Metric | RGB model | RGBD model |
|---|---|---|
| Loss | 0.0765 | **0.0346** |
| Accuracy (%) | 95 | **98** |
| Precision (%) | 89 | **97** |
| Recall (%) | 77 | **91** |
| IoU (%) | 70 | **88** |

**Table 3**. The results of footprint extraction on the test dataset.

The experimental results in Figure 5 show that the trained model based on RGB image and nDSMs can reliably extract the building footprints in different areas and non-building objects such as tress, roads as well as closed areas including walls are correctly detected as background values. The error map is shown in Figures 6 and 7 to present the differences between the ground truth and the predicted map using the RGBD model for two different areas. According to the error maps, the green areas are the locations of buildings in the ortho images that are not in the ground truth due to the digitization errors, and therefore, they were detected by the netwek, correctly. On the other hand, the red areas (e.g. FN pixels) are incorrectly detected as the buildings which are mostly in the boundaries of the buildings. As a conclusion, all large buildings are extracted completely. Compared to similar studies (Khoshboresh-Masouleh et al., 2020; Xu et al., 2018) the building footprints are extracted more accurately in the present study. However, the model has some problems with the prediction of footprints for small buildings.

## 5. CONCLUSION

The aim of this study was to evaluate the performance of the U-Net for building footprint extraction based on the combination of spectral (e.g. RGB orthophotos) and geometrical features (e.g. nDSMs). Several test scenarios are designed to first optimize the hyperparameters in the training step and the best model is then trained using a custom dataset. The qualitative and quantitative assessments indicate that quite promising results with significant high completeness and correctness rates are obtained for the RGBD model. Since the nDSM data provides a rich source of geometrical information, the height values of building roofs can be embedded into the model using RGB images for learning CNN, and therefore, extracting more accurate and distinguishable features to improve accuracy building footprints. The buildings footprints have been extracted with the precision of 97% and recall of 91% for Karimabad village. In order to improve the results of boundries, more investigations on deeper networks such as Deeplab-v3(Chen et al., 2017) is suggusted for the future work.

## ACKNOWLEDGEMENTS

The authors would like to thank the laboratory of Close Range Photogrammetry of School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran for acquiring the data used in this study and thank Parsa Dajkhosh master student of Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran for preparing the training dataset.

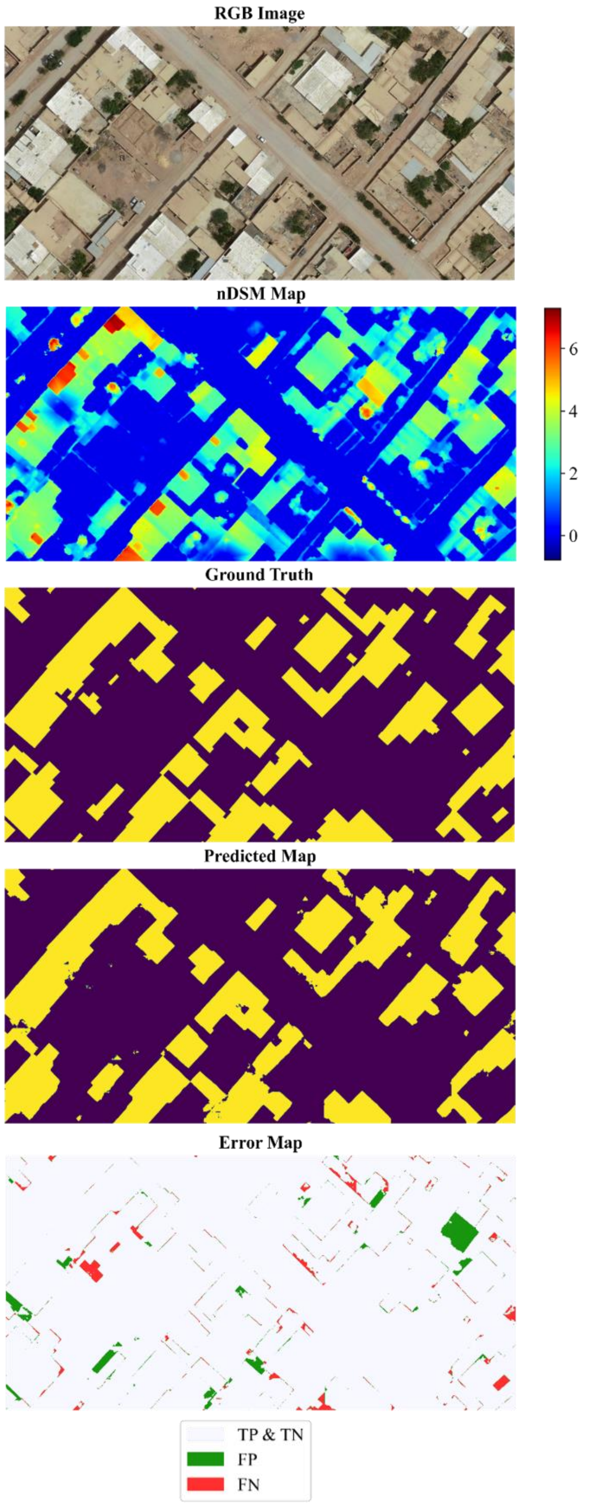

**Figure 6**. The error map between the ground truth and the predicted map using the RGBD model for the sample area 1.

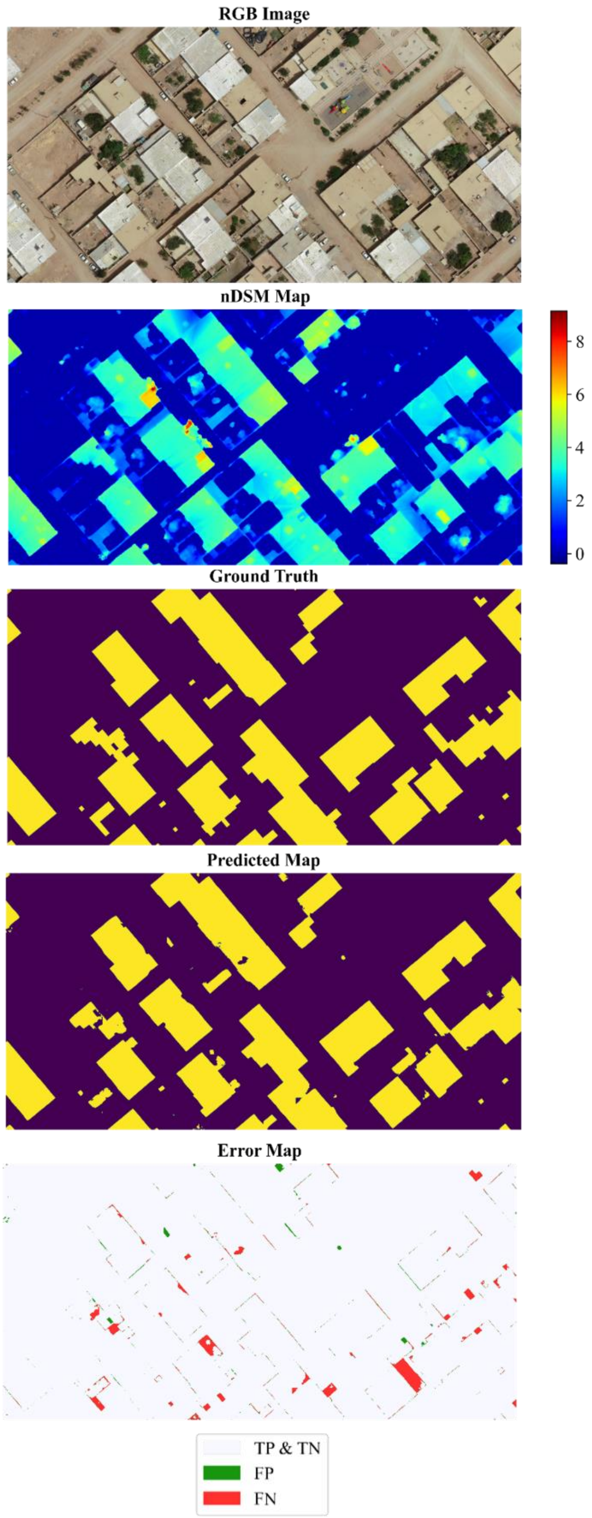

**Figure 7**. The error map between the ground truth and the predicted map using the RGBD model for the sample area 2.

## REFERENCES

Abdollahi, A., Pradhan, B., 2021. Integrating semantic edges and segmentation information for building extraction from aerial images using UNet. *Machine Learning with Applications* 6, 100194. doi.org/10.1016/j.mlwa.2021.100194

Alidoost, F., Arefi, H., Tombari, F., 2019. Building outline extraction from aerial images using convolutional neural networks. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-4/W18, 57–61. doi.org/10.5194/isprs-archives-XLII-4-W18-57-2019

Alsabhan, W., Alotaiby, T., 2022. Automatic building extraction on satellite images using Unet and ResNet50. *Computational Intelligence and Neuroscience* 2022. doi.org/10.1155/2022/5008854

Bittner, K., Adam, F., Cui, S., Korner, M., Reinartz, P., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11, 2615–2629. doi.org/10.1109/JSTARS.2018.2849363

Buyukdemircioglu, M., Can, R., Kocaman, S., Kada, M., 2022. Deep learning based building footprint extraction from very high resolution true orthophotos and NDSM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-2–2022, 211–218. doi.org/10.5194/isprs-annals-V-2-2022-211-2022

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587. doi.org/10.48550/arXiv.1706.05587

Chollet, F., 2018: *Deep learning with Python*. Manning Publications Co, Shelter Island, New York.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition. (CVPR 2009)*, doi.org/10.1109/CVPR.2009.5206848

Dixit, M., Chaurasia, K., Mishra, V.K., 2021. Automatic building extraction from high-resolution satellite images using deep learning techniques, in: *Int. Conf. on Paradigms of Computing, Communication and Data Sciences, Algorithms for Intelligent Systems*. doi.org/10.1007/978-981-15-7533-4_61

GDAL/OGR contributors, 2022. GDAL/OGR geospatial data abstraction software library. https://gdal.org (1 June 2022)

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. in: *CVPR2016*. Las Vegas, NV, USA, pp. 1–9. https://doi.org/10.1109/CVPR.2016.90

Isenburg, M., 2014. LAStools: efficient LiDAR processing software. https://rapidlasso.com/lastools (1 June 2022)

Khoshboresh-Masouleh, M., Alidoost, F., Arefi, H., 2020. Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors. *J. Appl. Rem. Sens.* 14, 1. doi.org/10.1117/1.JRS.14.034503

Liu, W., Yang, M., Xie, M., Guo, Z., Li, E., Zhang, L., Pei, T., Wang, D., 2019. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. *Remote Sens.* 11, 2912. doi.org/10.3390/rs11242912

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. in: *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*. doi.org/10.1007/978-3-319-24574-4_28

Roussel, J.-R., Auty, D., Boissieu, F.D., Sánchez, A., et al., 2021. Airborne LiDAR Data Manipulation and Visualization for Forestry Applications. https://github.com/r-lidar/lidR (1 June 2022)

Schuegraf, P., Bittner, K., 2019. Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN. *ISPRS Int. J. Geo-Inf.* 8(4), 191. doi.org/10.3390/ijgi8040191

Shi, Y., Li, Q., Zhu, X.X., 2019. Building footprint generation using improved generative adversarial networks. *IEEE Geosci. Remote Sensing Lett*. 16, 603–607. doi.org/10.1109/LGRS.2018.2878486

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support . DLMIA ML-CDS 2017*, vol 10553. Springer, Cham. doi.org/10.1007/978-3-319-67558-9_28

Sun, G., Huang, H., Zhang, A., Li, F., Zhao, H., Fu, H., 2019. Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images. *Remote Sens*. 11(3), 227. doi.org/10.3390/rs11030227

Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing* 58(3), 2178–2189. doi.org/10.1109/TGRS.2019.2954461

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 Sens*. 10(1), 144. doi.org/10.3390/rs10010144

Yu, D., Wei, S., Liu, J., Ji, S., 2020. Advanced approach for automatic reconstruction of 3d buildings from aerial images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 541–546, doi.org/10.5194/isprs-archives-XLIII-B2-2020-541-2020

Yuan, J., 2016. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. arXiv:1602.06564. doi.org/10.48550/arXiv.1602.06564

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X., Yan, G., 2016. An easy-to-use airborne lidar data filtering method based on cloth simulation. *Remote Sens.* 8(6), 501. doi.org/10.3390/rs8060501