Semantic Segmentation Uncertainty Assessment of Different U-net Architectures for Extracting Building Footprints

Ehsan Haghighi Gashti [1] , Mahmoud Reza Delavar [2] , Haiyan Guan [3] , Jonathan Li [4]

[1] GIS Department, School of Surveying and Geospatial Eng., College of Engineering, University of Tehran, Tehran, Iran - ehsanhaghighi77@ut.ac.ir
[2] Centre of Excellence in Geomatic Eng. in Disaster Management, and Land Administration in Smart City Lab., School of Surveying and Geospatial Eng., College of Engineering, University of Tehran, Tehran, Iran - mdelavar@ut.ac.ir
[3] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China - guanhaiyan_nj@qq.com
[4] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada - junli@uwaterloo.ca

Keywords: Uncertainty Assessment, Building Footprint, Deep Learning, Semantic Segmentation.

Abstract

Automatic extraction of building footprints from aerial and space imageries has been found ever increasing importance in urban planning, disaster management, and environmental monitoring. However, achieving accurate building footprint extraction poses significant challenges due to diverse building characteristics and their similarities to their background elements. While conventional methods in building footprint extraction have mainly relied on image processing techniques, recent advancements in deep learning, particularly semantic segmentation algorithms like U-Net, have shown promise in addressing these challenges through machine learning. This study explores different depths of the U-Net model for building footprint extraction, aiming to identify the optimum architecture while investigating the semantic uncertainty of the building footprint extraction. Utilizing aerial imagery from cities including Berlin, Paris, Chicago, and Zurich, collected from Google Maps and OpenStreetMap (OSM) data, five U-Net models have been compared with varying depths. In addition, the impact of dataset sizes and learning rates on model performance has been investigated. Results confirmed that the U-Net-32-1024 model achieves the highest intersection over union (IoU), Accuracy, and F1-score. Moreover, increasing the training dataset size leads to significant improvements in model performance with IoU, Accuracy and F1-score reaching their values of 73.73%, 88.65% and 88.53%. However, challenges remain in accurately delineating buildings in dense urban areas. Nonetheless, our findings demonstrated the effectiveness of U-Net models in building footprint extraction.

1. Introduction

Automatic extraction of building footprints from images is particularly valuable for urban planning (Sun, Zhang, Zhao, & Xin, 2018), disaster management (Tian, Cui, & Reinartz, 2013), and environmental monitoring (L. Li, Liang, Weng, & Zhu, 2018). The spatial distribution of buildings plays a crucial role in numerous tasks such as urban settlement monitoring and demographic modelling. However, due to the diverse characteristics of buildings and their similarities and differences to their background elements, the development of accurate building footprints extraction methods presents a significant and challenging research focus, receiving increased attention.

In recent decades, numerous studies on building extraction have relied on conventional image processing techniques such as shadow-based (Chen, Shang, & Wu, 2014), edge-based (Ziaei, Pradhan, & Mansor, 2014), and object-based methods (Norman et al., 2021). For instance, (Y. Dai, Gong, Li, & Feng, 2017) determined building footprints from image-derived point clouds in a two-stage solution including building segmentation and footprint extraction. In the first stage, vegetation points were first extracted using support vector machine (SVM) classifier based on five vegetation indices calculated from colour information. Then the traditional hierarchical stripping classification method was applied to classify and segment individual buildings. However, the primary issue with these algorithms lies in the necessity to craft numerous features for the correct classifier. This could potentially exhaust computational resources, thereby limiting their applicability at a large scale.

Deep learning (DP) techniques have become one of the state-of-the-art solutions for many segmentation problems and been in widespread use in various applications in remote sensing and photogrammetry for object detection, scene classification, and land cover mapping. Deep Convolutional Neural Network (CNN)-based semantic segmentation algorithms such as Fully Convolutional Networks (FCNs) (J. Dai, Li, He, & Sun, 2016), U-Net (Ronneberger, Fischer, & Brox, 2015), ResNet (K. He, Zhang, Ren, & Sun, 2016), and DenseNet (Huang, Liu, Van Der Maaten, & Weinberger, 2017) have been applied extensively to pixel-wise analysis tasks in remote sensing, covering tasks such as road extraction, building detection, urban land use classification, maritime semantic labelling, vehicle detection, damage mapping, weed mapping, and other land cover mapping activities. Several recent studies have utilized semantic segmentation methods specifically for building extraction from remote sensing images.

In this paper we focus on the comparison between different depths of U-Net model and try to identify the best possible architecture for the U-Net model. At first, five models including U-Net-16-512, U-Net-32-512, U-Net-16-1024, U-Net-32-1024 and U-Net-64-1024 were created. Then the dataset was divided into train, validation and test. Two different learning rates (0.001 and 0.0001) were tested and the one with better convergence speed was selected. Then, all the models were trained using half of the training data and the best network was selected. Finally, the selected network was trained again using all the data to see how data quantity affects the network.

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-X-4-2024-141-2024 | © Author(s) 2024. CC BY 4.0 License.

Section 2 reports on some major previous researches that have been undertaken regarding to the use of U-Net and other models to extract building footprints. In section 3, the dataset that has been used in this research is explained. Section 4 explains the research methodology. Section 5 presents the results of the networks and compares these results to previous research and section 6 concludes the paper and suggests some future research directions.

2. Related work

(Pasquali, Iannelli, & Dell’Acqua, 2019) focused on the architecture of the U-Net to develop a suitable version, capable of competing with the accuracy levels of past SpaceNet competition winners using only one model and one type of data. In their paper U-Nets architectures that had the maximum depth of 512 were studied. It was shown that suitable modifications of the architecture and effective use of data augmentation would lead to a novel network configuration that can be trained in a relatively short time and can achieve comparable performance to the existing state-of-the-art solutions, with simpler processing. (H. He et al., 2022) published a city-scale dataset and then performed an extensive comparative study on the dataset with the existing deep learning methods such as DeepLabV3+, HRNet, FCN and U-Net. (Q. Li, Shi, Huang, & Zhu, 2020) proposed the implementation of feature pairwise conditional random field (FPCRF) as a graph model to preserve sharp boundaries and fine-grained segmentation. Experiments were conducted on four different datasets including Planetscope satellite imagery of the cities of Munich, Paris, Rome, and Zurich; ISPRS benchmark data from the city of Potsdam; Dstl Kaggle dataset; and Inria Aerial Image Labelling data of Austin, Chicago, Kitsap County, Western Tyrol, and Vienna. It was found that the proposed end-to-end building footprint extraction framework with the FPCRF as the graph model can further improve the accuracy of building footprint generation using CNN.(Kaiser et al., 2017) adapted a state-of-the-art CNN architecture for semantic segmentation of buildings and roads in aerial images and compared its performance when using different training data sets, ranging from manually labelled, pixel-accurate ground truth of the same city to automatic training data derived from OpenStreetMap (OSM) data from distant locations. Their results demonstrated that (i) the sheer volume of training data can compensate for lower accuracy (ii)- the large varieties present in very large training sets spanning multiple different cities do improve the classifier ability to generalize to new and unseen locations (iii)- even if high-quality training data is unavailable, the large volume of training data improves classification accuracy and (iv)- large-scale training data allows substitution of the large majority of the manually annotated high-quality data. In order to extract building footprints, (Zhu, Liao, Hu, Mei, & Li, 2020) proposed a solution called Multiple Attending Path neural network (MAP-Net). MAP-Net introduced a multi-parallel path architecture, attention mechanism, and pyramid spatial pooling to address challenges in extracting building footprints. Experimental results showed significant improvements (up to 0.93% F1-score and up to 1.53% IoU score) compared to HRNetv2 model without increasing computational complexity across various datasets. (Bittner, Adam, Cui, Körner, & Reinartz, 2018) proposed Fused-FCN4s architecture to combine spectral and height information from different data sources to generate a full-resolution binary building mask. The proposed network consisted of three parallel networks merged at a late stage to propagate detailed information and produce accurate building outlines. Inputs included RGB, panchromatic, and normalized digital surface model (nDSM) images.

3. Dataset

The image dataset utilized in this study was obtained from Google Maps, while the building masks were acquired from OSM data. It comprises aerial imagery from Berlin, Paris, Chicago, and Zurich. Figure 1 illustrates the study area and Figure 2 displays the aerial images and their corresponding masks for the Berlin region.

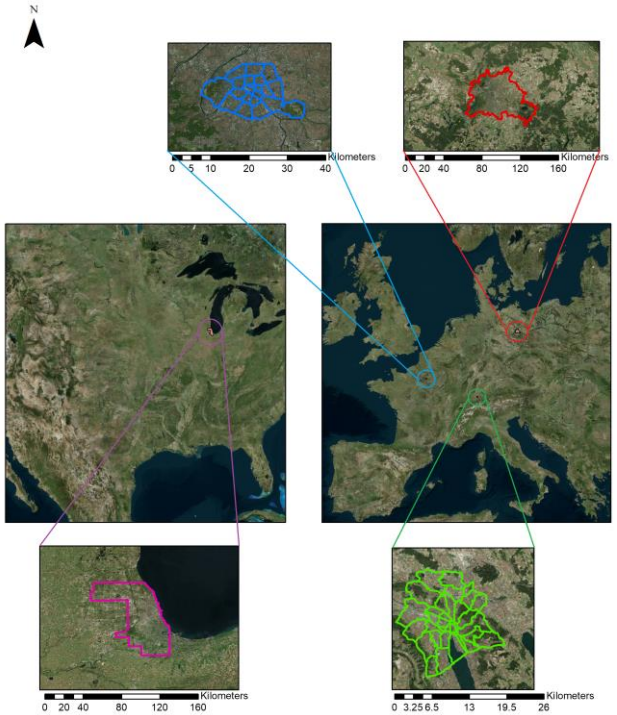

Figure 1. Study areas (Berlin in red, Paris in blue, Zurich in green and Chicago in purple)

(a)

(b)

Figure 2. Image (a) and its corresponding buildings mask (b) for city of Berlin

In Table 1 some metadata of the employed data including data coverage area and ground sample distance (GSD) of the images for each of the study areas are explained.

City	Data coverage area (Square kilometre)	GSD (Cm)
Chicago	4.1×10^9	11.1
Paris	6.3×10^9	9.8
Zurich	3.5×10^9	10.1
Berlin	1.28×10^9	9.1

Table 1. Image dataset for the study areas

The same data has been used in (Kaiser et al., 2017) and therefore we have compared the results from their FCN model with our U-Net model. As with any opensource data like OSM there are bound to be errors in the dataset. They tackled this problem by introducing 4 hypotheses which are: (i)-The sheer volume of training data can possibly compensate for the lower accuracy (if used with an appropriate, robust learning method). (ii)- The large variety present in very large training sets (e.g., spanning multiple different cities) could potentially improve the classifier’s ability to generalise to new, unseen locations. (iii)- Even if high-quality training data is available, the large volume of additional training data could potentially improve the classification. (iv)- If low-accuracy, large-scale training data helps, then it may also allow one to substitute a large portion of the manually annotated high-quality data. Given that we are using the same dataset as they, we include these hypotheses in our study as well. The images in the dataset were divided into 512 by 512 tiles. This size was selected through experiments to ensure sufficient geographical context within each tile without having excessive load on the graphics processing unit (GPU). In Table 2 the number of tiles for each of the cities of the study areas is shown.

City	Number of 512 by 512 tiles
Chicago	13470
Paris	22500
Zurich	10920
Berlin	4000

Table 2. Number of tiles for each city

4. Methodology

The conceptual model of the study is illustrated in Figure 3. In part A, first we acquired the data for cities of Berlin, Paris, Chicago and Zurich. In part B, we tested three different ratios for dividing the data to training, validation and test data and two different learning rates to find the best possible ratios and learning rates. In part C, we developed five different U-Net models with various depths and train them using half the data. Then, we selected the best model and then trained it with all the data in part D. Finally in part E, we evaluated the final model.

4.1 Different employed U-Net architectures

In the field of automatic buildings detection, various models have been used such as convolutional neural networks (CNN), fully convolutional networks (FCN), U-Net and Region-based Convolutional Neural Network (RCNN) (Girshick, 2015) where, RCNN is a model used primarily for object detection tasks. The main advantage of RCNN is its ability to localize objects accurately by generating region proposals and then classifying and refining those proposals. However, RCNN has limitations in terms of speed and efficiency due to its multi-step processes. This research focuses on the U-Net model. U-Net is a CNN architecture designed for semantic segmentation of images. The U-Net model has a structure that collects information through an encoder and then retrieves spatial information through a decoder. In addition, U-Net uses skip connections that transfer information directly between layers, helping to preserve fine details during scaling operations and preventing possible data loss. U-Net also shows acceptable performance with a small amount of training data which makes it a suitable option for building detection tasks in aerial images. Therefore, in this research we have focused on the U-Net architecture to extract building footprints. The overall architecture of U-Net model is illustrated in Figure 4. The architectute for U-Net in the original paper (Ronneberger et al., 2015) started at depth level 64 and then continued to bridge layer 1024. After that, the decoding path went back to depth 64 and then the binary output was obtained4. We added the layer 16 and 32 to the original architecture, and the final depth was set to 1024 in our implemented architecture, while depth of 512 could also be used as a bridge layer. If layer 512 was to be considered as the bridge layer, then green path should be

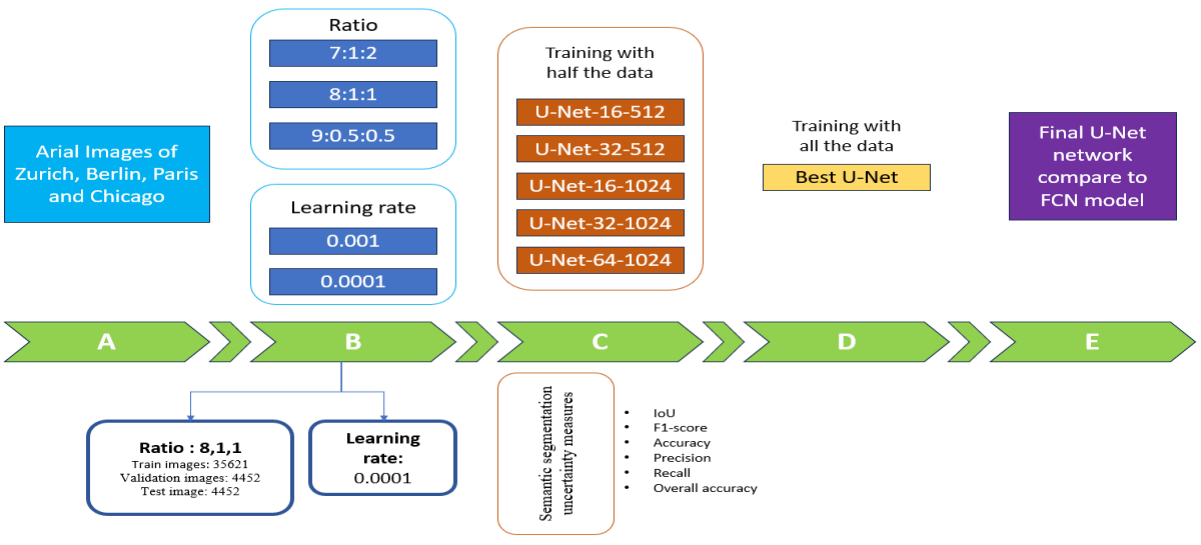

Figure 3. Conceptual model of the study of different U-Net a and their comparison with different models

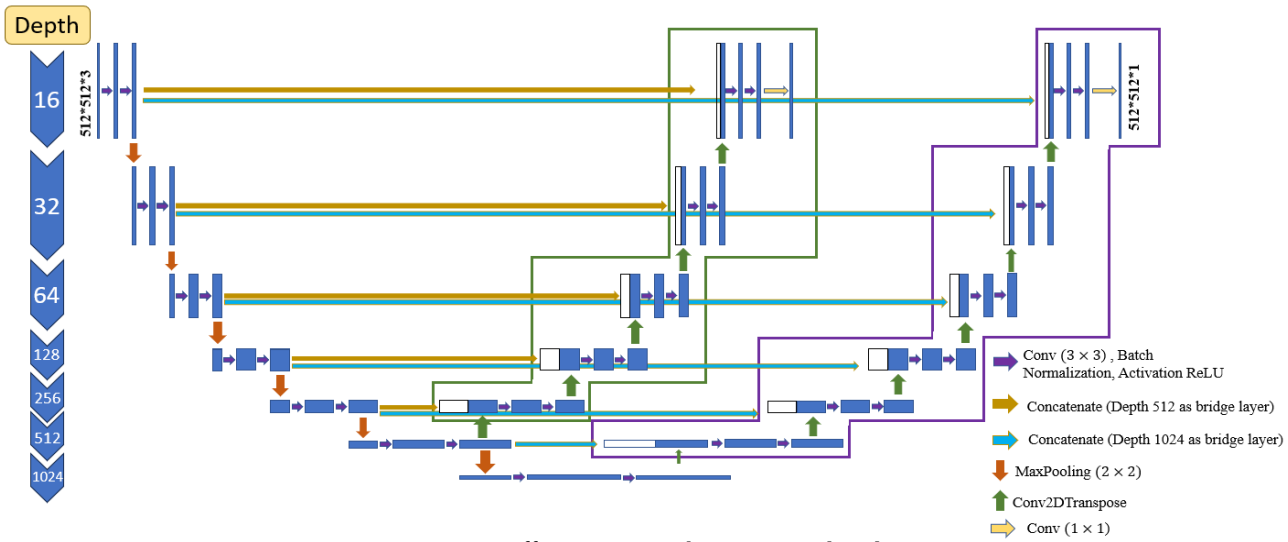

Figure 4. Different U-Net architectures employed

followed. If layer 1024 was to be considered as the bridge layer, then purple path should be followed. So with all these changes, five new models with different depths were created. In Table 3 The created models and their start and bridge layers are shown.

As shown in Figure 4, in the encoding path, at each depth, the network begins by taking the input and passing it through a 3 by 3 convolutional layer (conv (3×3)). following this, Rectified Linear Unit (ReLU) activation function is applied. This process is repeated once more to reach the final output before applying a 2 by 2 Max pooling operation, which reduces the image dimensions by half. Then, the number of convolutional layers is doubled, allowing for more features to be extracted in deeper layers. Upon reaching either the depth of 512 or 1024, the decoding path initiates. Initially, a Conv2DTranspose layer is applied to the starting layer, doubling its dimensions but halving the number of filters. Subsequently, the resulting layer is concatenated with the output from the encoding path, a pivotal step facilitating the retrieval of localized information crucial for accurate semantic segmentation in U-Net. These steps are repeated until the image reaches its original dimensions. Finally, a 1 by 1 convolutional layer is employed to generate a binary output, extracting building footprints.

Model	Start layer	Bridge layer
U-Net-16-512	16	512
U-Net-32-512	32	512
U-Net-16-1024	16	1024
U-Net-32-1024	32	1024
U-Net-64-1024	64	1024

Table 3. Employed models and their depths

4.2 Semantic segmentation uncertainty assessment

For the building footprint extraction semantic uncertainty assessment, the employed measures calculated from the error matrix are explained below.

TP represents the true positive, which indicates the correct prediction of the positive class identifying that the real value on the ground is building and the model has recognized the building correctly. FP refers to a false positive that occurs when the model predicts a negative class as positive. It means that the real value on the ground is not building, but the model has recognized it as building.

FN stands for false negative, where the model classifies the positive class into the negative class. In other words, the real value on the ground was the building, however, the model did not recognize the building. TN stands for the true negative, where the model correctly predicted the negative class at the output where the true value on the ground was no building and the model correctly predicted no building.

The semantic segmentation uncertainty measures for the building footprint extraction that were used in this study are Accuracy, Intersection over Union (IoU), Overall accuracy, Precision, Recall and F1-Score. The formulas for these measures are as follow (H. He et al., 2022):

$$Accuracy = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

$$IoU = \frac{TP}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (2)$$

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \quad (3)$$

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \quad (4)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

$$Overall\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

4.3 Dividing the data for training, validation and testing

The dataset comprises a total of 50,890 tiles, which needed to be divided into three groups for training, validation, and testing purposes. The initial phase of the study involved determining the most effective approach to divide the data into these groups. Initially, tiles without buildings were excluded from the dataset to alleviate computational burdens on the GPU, resulting in a reduction to 44,525 tiles. Subsequently, the data was divided into three sets using ratios of 7:1:2, 8:1:1, and 9:0.5:0.5 for training, validation, and testing, respectively. These ratios were

selected based on precedents established in previous studies such as (Schuegraf & Bittner, 2019) and (Bittner, Cui, & Reinartz, 2017). Table 4 represents the number of train, validation and test data for each ratio employed in this study.

Ratio	Train	Validation	Test
7:1:2	31168	4452	8905
8:1:1	35621	4452	4452
9:0.5:0.5	40073	2226	2226

Table 4. Number of train, validation and test data for each employed ratio

To determine the optimum ratio, we selected the U-Net-32-512 model as the test network due to its relatively shorter training time compared to other models and promising results reported by (Pasquali et al., 2019). To expedite the process, only one-third of the data for each ratio was used as input for model training. Figure 5 illustrates the train loss of the model.

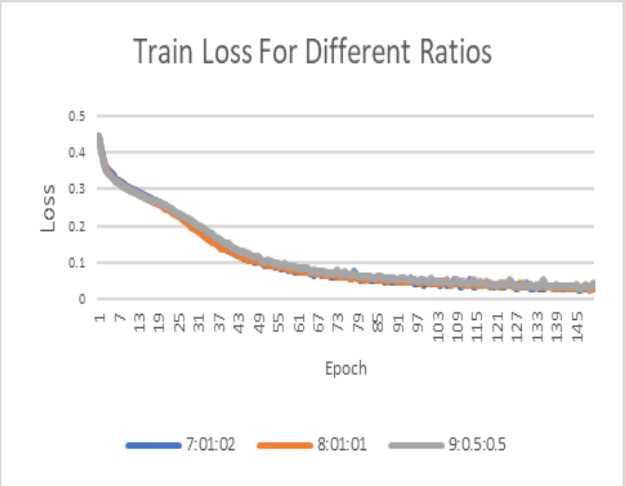

Figure 5. Train loss of model with different ratios for the employed data

In Table 5, the resulted IoU and the train time are shown.

Ratio	Total train time (h)	IoU for test data (%)
7:1:2	25.6	66.57
8:1:1	30.5	67.84
9:0.5:0.5	30.6	67.86

Table 5. IoU for different ratios

Although the models converged in all of the undertaken scenarios, it was observed that IoU increased by more than 1% when transitioning from the 7:1:2 ratio to the 8:1:1 ratio. However, there was not a substantial increase in IoU from the 8:1:1 ratio to the 9:0.5:0.5 ratio. In addition, the 9:0.5:0.5 ratio would result in a very small dataset for testing and validation. Hence, the 8:1:1 ratio was selected as the most suitable choice for dividing the dataset.

4.4 Selection of the learning rate

The learning rate is a critical hyperparameter in deep learning that significantly influences the training process, convergence behaviour, and overall performance of the model. Proper tuning and selection of the learning rate are essential for achieving optimum results in training deep neural networks. Many different types of learning rates have been used in a number of studies. In (Woo, Park, Lee, & Kweon, 2018), the learning rates started from 0.1 and dropped every 30 epochs. In (Xie, Girshick, Dollár, Tu, & He, 2017) the learning rates started from 0.1 and then divided by 10 at the 150th epoch and then again divided by 10 at the 225th epoch. In (H. He et al., 2022) the learning rate was set to 0.0001 for the whole training time. In our study we selected two different learning rates including 0.001 and 0.0001, respectively. Then we trained the U-Net-32-512 model with both of these learning rates. The resulted accuracy and loss are shown in Figures 6 and 7.

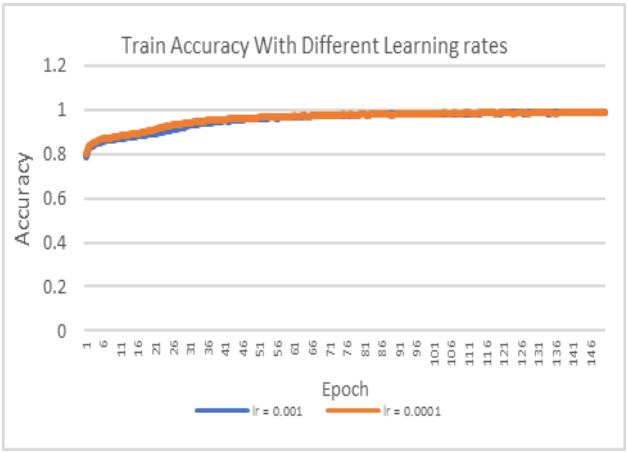

Figure 6. Accuracy for different learning rates (lr)

Figure 7. Loss for different learning rates

It was seen in Figures 6 and 7 that using both the learning rates, the model would converge and reach a good accuracy value. However, the convergence speed was higher with learning rate set to 0.0001. So, it was selected as the learning rate to train all the models.

4.5 Training the models

The hardware configuration for training the models comprised a Nvidia RTX 4070ti GPU, alongside an Intel Core i9 13700k Central Processing Unit (CPU). Although the system boasted 128 Gigabytes of DDR4 RAM, the neural networks were designed to access data directly from the hard drive, limiting RAM usage to only 10 Gigabytes at peak. Binary Cross Entropy was employed as the loss function, and Adam as the optimizer for all the models. The models were trained for 150 epochs, with a batch size of 8. However, for the U-Net-64-1024 model,

the batch size had to be reduced to 6 due to insufficient GPU capacity.

5. Results

First the models were trained using just half the data to check how well the U-Net models perform with a limited data. Table 6 represents the results of the first training.

Model	Recall (%)	Overall Accuracy (%)	Precision (%)	F1-Score (%)	IoU (%)	Accuracy (%)
U-Net-16-512	85.61	68.74	87.10	81.45	67.48	87.78
U-Net-32-512	88.26	69.19	88.54	81.88	66.57	87.91
U-Net-16-1024	89.28	69.37	86.17	81.90	68.63	88.25
U-Net-32-1024	87.65	69.33	86.50	81.94	68.99	88.31
U-Net-64-1024	89.88	69.24	86.84	81.91	67.74	87.95

Table 6. The results of different models employed

Due to the fact that the U-Net-32-1024 had the highest IoU, accuracy and F1-score among all the models, it was selected as the best model. U-Net-32-512 had the highest Precision and U-Net-64-1024, which was the model identical to the original model of U-Net, had the highest Recall. Precision can be defined as the number of true positive predictions divided by the total number of positive predictions made by the model. Having a high Precision is desirable in applications that have a cost for false positive. It can be concluded that the U-Net-32-512 model classified lower pixels as buildings but with a better accuracy than that of all the other models.

Recall can be defined as the number of true positive predictions divided by the total number of actual positive instances in the dataset. Having a high Recall is desirable in applications where the cost of false positive is low. The U-Net-64-1024 had the highest Recall which means that the model predicted the biggest number of building pixels.

Another parameter which is important in this study is the training time of the models which is represented in Table 7.

Model	Time per epoch (s)	Total train time (h)
U-Net-16-512	740	30.8
U-Net-32-512	1535	64
U-Net-16-1024	802	33.4
U-Net-32-1024	1720	71.7
U-Net-64-1024	3983	166

Table 7. Time of training of the models

It was seen that even though U-Net-64-1024 had the longest training time among all the models, it was not the best model by the evaluated measures.

After the initial training, the U-Net-32-1024 was then trained using all the images in the dataset. The model was evaluated again using the measures and the results are as shown in Table 8.

Model	Recall (%)	Overall Accuracy (%)	Precision (%)	F1-Score (%)	IoU (%)	Accuracy (%)
U-Net-32-1024	89.08	74.25	87.99	88.53	73.73	88.65

Table 8. Evaluated measures (train with entire dataset)

All the measures indicated better performance by the increase in training data size especially the value of IoU which increased by more than 4% from 68.99% to 73.73%. F1-score increased from 81.94% to 88.53% and the accuracy increased from 88.31% to 88.65%.

In Figures 8. and 9., some examples of model outputs are visualised.

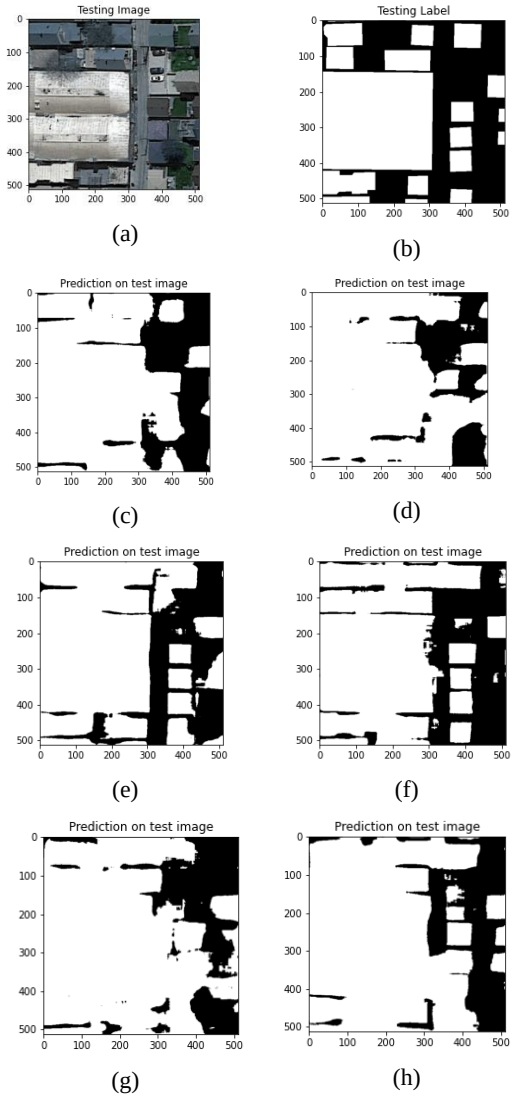

Figure 8. Example of building footprint extraction results in the densely populated areas in Berlin: (a) Test image, (b) Test label, (c) U-Net-16-512, (d) U-Net-32-512, (e) U-Net-16-1024, (f) U-Net-32-1024, (g) U-Net-64-1024, (h) U-Net-32-1024 (trained with all the data)

It was found that in places that buildings are far apart, the U-Net model performs surprisingly well as shown in Figure 9. However, in regions where buildings are densely located, the model has a problem in distinguishing buildings from other objects like roads or cars as shown in Figure 8. The same data was used in the study that was conducted by Kaiser, Wegner et al. (2017). In their study they developed a FCN model to extract buildings and roads simultaneously.

In Table 9., F1-score measure for the FCN model is compared to our U-Net models. F1-score for buildings using the FCN model at average was 82.74%, however, we were able to achive F1-score of 81.94% with just half the data. Furthermore, when

the network used all the dataset, we were able to achieve F1-score of 88.53% and surpass the FCN model.

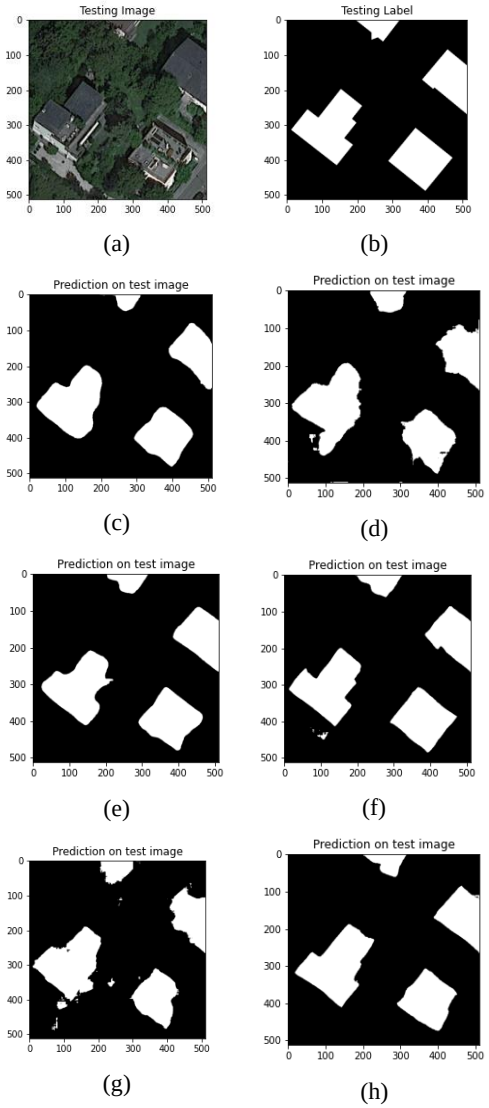

Figure 9. Example of building footprint extraction results in a sparsely populated areas of Zurich: (a) Test image, (b) Test label, (c) U-Net-16-512, (d) U-Net-32-512, (e) U-Net-16-1024, (f) U-Net-32-1024, (g) U-Net-64-1024, (h) U-Net-32-1024 (trained with all the data)

Model	F1-score (%)
FCN	82.74
U-Net-32-1024 (trained with half the data)	81.94
U-Net-32-1024 (trained with all the data)	88.53

Table 9. F1-score for U-Net model compared to FCN

6. Conclusion and Future Directions

Deep learning networks have revolutionized many fields including the extraction of building footprints from aerial images. The importance of deep learning networks in this task stems from their ability to automatically learn complex patterns and features from large amounts of data, making them highly effective in extracting precise and accurate building footprints. The scalability of deep learning approaches is another significant advantage. Deep learning networks can be trained on vast amounts of data, allowing them to learn complex patterns and generalize well to unseen regions or datasets. Having mentioned these, one of the most important questions is, what is the best model for extracting building footprints? Numerous models such as FCN, HRNet, and DeepLab have been designed and tested in multiple instances. The U-Net model has emerged as a cornerstone in various image segmentation tasks. Its importance lies in its unique architecture and design, which enable highly accurate and precise segmentation results, particularly in scenarios where detailed delineation of object boundaries is critical. A key advantage of the U-Net model is its ability to handle limited annotated data effectively. This is particularly important in tasks like building footprint extraction, where obtaining large annotated datasets can be challenging and costly. Furthermore, the U-Net model is highly versatile and adaptable to different domains and modalities. This versatility underscores the broad applicability and importance of the U-Net model across diverse fields and applications. Therefore, the focus of this research was on the U-Net model.

This study delved into the architecture of the U-Net model, undertaking a comparison of various U-Net models to address some questions and challenges like (i)-what is the best U-Net architecture for building extraction in arial images? (ii)-what effect does the size of data on the semantic segmentation uncertainty of model? and (iii)- does U-Net model outperform previous models such as FCN?

The dataset used consisted of aerial images acquired from Google Maps, with corresponding building footprint masks obtained from OSM data. Initially, the data was divided into three groups including training, validation, and testing using three different ratios (7:2:1, 8:1:1 and 9:0.5:0.5). Subsequently, the U-Net-32-512 model was trained with each dataset, revealing the 8:1:1 ratio as the most effective one. Following this, two distinct learning rates, one equals to 0.001 and the other equals to 0.0001, were experimented and the value of 0.0001 was selected as the optimum learning rate for all the models. Five different models, featuring diverse starting and bridge layers, were developed and initially trained using only half of the data.

The top-performed model, U-Net-32-1024, was then selected and further trained with the entire dataset. Remarkably, even with half the data, this model exhibited notable performance with F1-score of 81.94%, nearly matching the accuracy of the FCN model with F1-score of 82.74% for building footprints. When trained on the complete dataset, it significantly outperformed the FCN model and achieved F1-score of 88.53%. However, challenges pertaining to hardware limitations were encountered, necessitating a reduction in batch size to maintain training efficiency.

Furthermore, it was observed that while the U-Net model excelled when buildings were widely spaced, like in rural areas, its performance deteriorated in dense urban areas. This suggests that in such environments, alternative networks like ResNet or DenseNet may yield superior results. It should be noted that this research focused on the semantic segmentation uncertainty assessment of U-Net model. Thus, for future research, the geometric uncertainty can be evaluated. In addition, investigating U-Net performance compared to models such as ResNet can be investigated in future research.

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-X-4-2024-141-2024 | © Author(s) 2024. CC BY 4.0 License.

References

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.

Bittner, K., Cui, S., & Reinartz, P. 2017. Building extraction from remote sensing data using fully convolutional networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, 481-486.

Chen, D., Shang, S., & Wu, C. 2014. Shadow-based Building Detection and Segmentation in High-resolution Remote Sensing Image. *J. Multim., 9*(1), 181-188.

Dai, J., Li, Y., He, K., & Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems, 29*.

Dai, Y., Gong, J., Li, Y., & Feng, Q. 2017. Building segmentation and outline extraction from UAV image-derived point clouds by a line growing algorithm. *International Journal of Digital Earth, 10*(11), 1077-1097.

Girshick, R. 2015. Fast r-cnn. *Paper presented at the Proceedings of the IEEE international conference on computer vision.*

He, H., Jiang, Z., Gao, K., Narges Fatholahi, S., Tan, W., Hu, B., . . . Li, J. 2022. Waterloo building dataset: A city-scale vector building dataset for mapping building footprints using aerial orthoimagery. *Geomatica, 75*(3), 99-115.

He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep residual learning for image recognition. *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.*

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017. Densely connected convolutional networks. *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.*

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., & Schindler, K. 2017. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing, 55*(11), 6054-6068.

Li, L., Liang, J., Weng, M., & Zhu, H. 2018. A multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote Sensing, 10*(9), 1350.

Li, Q., Shi, Y., Huang, X., & Zhu, X. X. 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). *IEEE Transactions on Geoscience and Remote Sensing, 58*(11), 7502-7519.

Norman, M., Shahar, H. M., Mohamad, Z., Rahim, A., Mohd, F. A., & Shafri, H. Z. M. 2021. Urban building detection using object-based image analysis (OBIA) and machine learning (ML) algorithms. *Paper presented at the IOP Conference Series: Earth and Environmental Science.*

Pasquali, G., Iannelli, G. C., & Dell’Acqua, F. 2019. Building footprint extraction from multispectral, spaceborne earth observation datasets using a structurally optimized U-Net convolutional neural network. *Remote Sensing, 11*(23), 2803.

Ronneberger, O., Fischer, P., & Brox, T. 2015. *U-net:* Convolutional networks for biomedical image segmentation. *Paper presented at the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.*

Schuegraf, P., & Bittner, K. 2019. Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN. *ISPRS International Journal of Geo-Information, 8*(4), 191.

Sun, Y., Zhang, X., Zhao, X., & Xin, Q. 2018. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. *Remote Sensing, 10*(9), 1459.

Tian, J., Cui, S., & Reinartz, P. 2013. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing, 52*(1), 406-417.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. 2018. *Cbam:* Convolutional block attention module. *Paper presented at the Proceedings of the European conference on computer vision (ECCV).*

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. 2017. Aggregated residual transformations for deep neural networks. *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.*

Zhu, Q., Liao, C., Hu, H., Mei, X., & Li, H. 2020. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing, 59*(7), 6169-6181.

Ziaei, Z., Pradhan, B., & Mansor, S. B. 2014. A rule-based parameter aided with object-based classification approach for extraction of building and roads from WorldView-2 images. *Geocarto International, 29*(5), 554-569.

https://doi.org/10.5194/isprs-annals-X-4-2024-141-2024