Automatic Detection and Estimation of the Area of Buildings

BRYAN SANCHEZ, University of Twente, The Netherlands

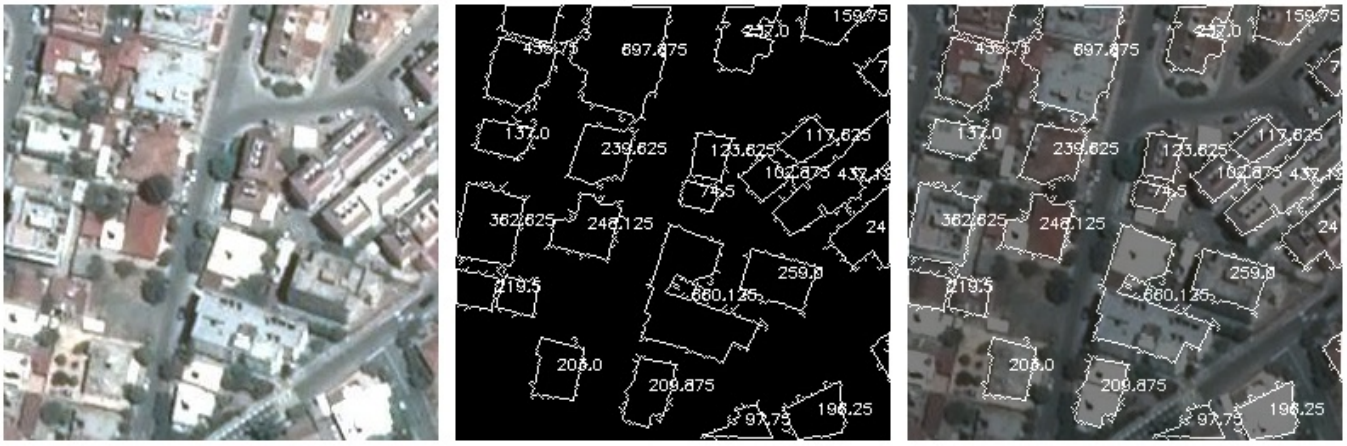

Fig. 1. From left to right: manually annotated image, mask generated by the model and blending mask over the image with area estimation.

This research is an approach to how deep learning networks can be used as a framework to detect different small objects in a satellite image with high precision and estimate the construction area of the segmented buildings. This solution is applied to a problem in the region of Cyprus where the critical concern is the regulation of urban infrastructures and settlements. However, estimating the building area can be a real challenge if such a building has an unusual shape or is too close to another one. The U-Net architecture has been developed for image segmentation and small object recognition, obtaining fast and precise results. Here, we show the improvement in combining this convolutional network for house detection and segmentation with manual annotation. The results indicate that the proposed method can make an improvement in segmentation accuracy by about 10% following an annotation methodology. Furthermore, the mean intersection over union (IoU) score of about 91% validates the promising performance of this model.

Additional Key Words and Phrases: Deep learning, U-Net framework, segmentation model, satellite imagery, convolutional neural network.

1 INTRODUCTION

Over the past decades, people have encountered problems declaring with detail and measuring the exact dimensions of their properties. These issues commonly happen due to inconsistencies when registering the boundaries of the building. This issue is still not solved in many countries since it is time-consuming. The cost is higher for the municipality or government lacking the capability and local expertise to develop such a sophisticated land registration system [14]. The lack of sufficient sources of self-supporting money for local governments may have adverse effects on their ability to function financially as well as on how responsive and accountable they can be to the community [3]. They try to send proper engineers to measure and control the boundaries of the land of such properties, which in most cases, cannot be easily accessed due to geography.

As a consequence of these problems, several issues can be derived and divided into social and environmental problems. Regarding society and how the impact would be within a community, the tax distribution from the government can be affected as this creates significant inequalities. Land taxes have long been recognized as a source of independent income for municipal governments [3]. Social infrastructures benefit from taxes, and if there is a low income, the maintenance of places such as parks or roads cannot be covered. Similarly, this problem impacts governance, incentives for efficient land use, and the types and quantities of public services offered [3].

On the other hand, environmental issues are addressed to potential disasters and hazards such as the spread of a wildfire in a rural zone. The government could not control and prevent these tragedies due to insufficient and inaccurate information. If property rights are created in a way that promotes the wise use of natural resources, external environmental consequences can frequently be internalized [3]. An automated application can help address these environmental and social concerns. The solution presented in this paper relates to analysing imagery from a high-precision satellite and automatically segmenting all buildings.

A convolutional neural network is one of the most significant advancements in the deep learning field related to computer vision [2]. Such a network takes images as input and then assigns learnable values and weight for each pixel to differentiate one from the others. *Semantic segmentation* within images is one main application of these networks. The model can be built using a fully convolutional network for pixel-by-pixel estimation [16] or also adding

TScIT 37, July 8, 2022, Enschede, The Netherlands

deconvolution in an encoder-decoder architecture [5]. Methods proposed by [16] using a fully convolutional network have registered remarkable outcomes for classifying five different items within the same image. The efficiency and accuracy of detecting small objects (buildings, cars, roads, vegetation, lower trees) were greater than four-fifths overall. Another approach using a Dense U-Net framework by the International Society for Photogrammetry and Remote Sensing (ISPRS) achieved excellent results with U-Net. This design was emphasised since it showed that this architecture is appropriate for recognising tiny objects and semantic segmentation [6].

This paper will attempt to describe to what extent the deep learning model can detect houses from satellite imagery and return an estimate of the building area with high efficiency and accuracy.

1.1 Research question

Main question: Can the area of buildings be estimated with high precision given satellite imagery from Cyprus by accurately detecting houses and other constructions?

Research has been performed on deep learning models, which focus on estimating the height of buildings depending on the shadow captured from different angles by satellite imagery [6]. However, there is a minimal study on how to delimit a building's contour and then estimate its area.

Hence, these sub-questions will help answering the previous main research question:

(1) Can the accuracy of detecting houses and other establishments be improved by annotating the dataset in more detail?
(2) How will segmentation models add precision for detecting a building with an unconventional shape or when a house is overlapping?
(3) Is it possible to determine a building area from the mask generated by the model with high confidence?

Most of the buildings in Cyprus have different and unconventional structures. The shape and the neighbourhood arrangements make people build their houses in the best condition possible, taking advantage of every inch. In this scenario, segmenting all structures is challenging if the manual annotation does not adequately delimit their contour with preciseness. Most houses are not detached, which may be a problem differentiating a single one over the rest. Thus, the accuracy and efficiency of the model may be improved by annotating the train set of images with great detail. A generated mask of the image snapshot will highlight all detected buildings and set the remaining pixels to black as they are not of relevance. The area can be obtained by the contour of each structure. Hence, the more detailed the manual annotation is, the better the model classifies the crucial pixels that make the contour of the buildings.

2 RELATED WORK

Table 1 In deep learning, the research proposed by [4] puts some machine learning methods into practice and evaluates their capability and power in landslide detection. They reported that convolutional neural networks do not perform better than other machine learning algorithms like ANN, SVM, and RF. Yet, the best results were obtained by CNN with an input of 16-pixel window size, which is the most efficient in classifying complex object patterns. Tompson et al. [13] employed convolutional training to develop an end-to-end part detector and spatial model for posture estimation with good results by combining FCN and a Markov Random Field. In [1, 2, 16, 18], methods for semantic segmentation show substantial results in doing semantic tasks within an image. Dechesne *et al.* [2] show how semantic segmentation based on Monte Carlo Dropout is measured by its uncertainty so that the model may discard uncertain pixels.

However, this method was applied to buildings where the model can accurately forecast their position with high confidence but not their exact shape. The successful network architecture in [18] combines a convolutional network with up-sampling operators. The model propagates context information to higher resolution layers from the contraction path to the expansion path, which concatenates the corresponding map from the previous path. The trained model in [9] can be taught with a minimum set of images compared to other approaches with the same objective and yet obtain an impressive performance in segmenting biomedical images. On the other hand, the method proposed by [18] uses the modified framework U-Net++, a network built on dense and layered skip connections. The main difference between this technique and the standard implementation is that before fusion, the semantic map between the encoder and decoder feature maps is bridged.

Additionally, other approaches that use deep learning [1, 17] with a similar objective were found so that they can be used as a reference for this research. The framework used in [17] has notable performance in semantic segmentation and mapping of the built-up areas in selected regions of Zhengzhou. By combining spatial pyramids through pooling, the model based on Deeplabv3+ showed high accuracy of over 90% on five types of satellite data. In [1], they use a pre-trained network, VGG-16 [10], to predict path loss and then determine the coverage of a wireless communication system in specific areas by analysing satellite imagery.

3 EXPERIMENTAL SETUP

For this research, the graphic card NVIDIA GeForce RTX 2080 Ti was used for training the model. We used the Python library *Segmentation Models* [1] for neural networks intended for image segmentation. This library is based on *Keras* and *TensorFlow* allowing us to use the U-Net architecture and the ResNet backbone. The U-Net architecture was chosen due to the fact that it was designed for semantic segmentation of optical images so that the model can slice different-sized buildings inside the same snapshot [8].

3.1 Model

This paper proposes an pre-trained encoder-decoder convolutional neural network based on the U-Net standard architecture [9]. Implementing the same overlap-tile approach enables the smooth segmentation of huge random images. The missing context can be superposed by mirroring the input pictures to predict pixels in their contour. The method in [9] enables the analysis of large images without limiting their resolution due to the restricted GPU RAM. In this way, the satellite images provided for this research were normalised with this tiling strategy before the training and evaluation process.

[1] https://github.com/qubvel/segmentationmodels

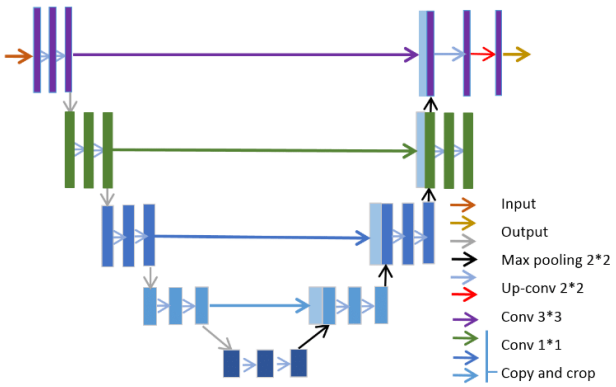

Fig. 2. U-Net architecture of the model used within this research.

This model takes tile images with a width and height of 256 pixels, respectively. The setup of this network contains three image channels since satellite snapshots are in *RGB* format. Figure 3 shows the architecture of the CNN used in this research and the path satellite images go through in the training process. In this model, the expanding path (right side) and the contracting path (left side) both employ kernels to extract context information. On the contracting side, each layer starts to read and store pixels by applying two convolutions using a *kernel* of 3×3 and the same padding [9]. Padded convolutions allow re-scaling input images after disarranging into smaller pieces so that the input and output pictures have identical dimensions after each convolution. In addition, a Rectified Linear Unit follows each layer, and for down-sampling, the network applies max-pooling operations with a pool size of 2×2 [9].

Skip connections are utilised in the U-Net framework to speed up the training process and get the prior data from the respective encoding layers [13]. At the bottom of Figure 3, the coding layer does not require the max-pooling feature since it serves as an informative representation layer that is compressed. In the decoding phase, transpose convolution is necessary for up-sampling, followed by concatenating the feature map extracted from the encoding layer [9]. The network applies another two convolutions with the same parameters and the 3×3 *kernel* proposed for the encoding layer. Finally, each vector with a 128-component feature is mapped to the desired number of classes in the last layer in [9] using a 1 $\times 1 convolution$.

3.2 Backbone

The pre-trained model on the backbone resnet50 was used in this research since training from scratch is time-consuming and requires sophisticated hardware for preparing the model. Pre-weighted weights are included in this backbone for faster and better convergence. It has residual networks of 50 layers. By enabling gradients to flow through this additional short-cut path, ResNet lessens the issue of vanishing gradients [12]. The research in [12] tests the performance of VGG16, ResNet50 and SE-ResNet50. The highest precision and recall are achieved by ResNet50, which also prepares the model in lower epochs.

4 ANNOTATION AND TRAINING

The data set for training and evaluating the model generated with a CNN were provided from the Pleiades satellite by the European Space Agency. The resolution of the high spatial-resolution satellite imagery is 0.5m, where the dimensions of each pixel are $0.5 \times 0.5 m^2$. Snapshots from dense cities and small villages are remarkable large for using them to train the model. As mentioned before, for minimizing overhead and making the most of the GPU memory, 256×256×3 are the corresponding dimensions for all tile pictures.

A dataset of 266 cropped images was manually annotated using a tool for labelling photos. Label-Studio is an open-source tool for labelling data types such as text, images, audio and time series. Figure 3 exhibits the environment for manually segmenting houses for this project. Each building has to be labelled for semantic segmentation by shaping its contour. In this way, the predictions of pixels from the border region could be as precise as possible. Although seven labels were proposed to detect small objects (agriculture, coastal, forest, pasture, road, urban, water), only the label for houses and buildings was used throughout the annotation process.

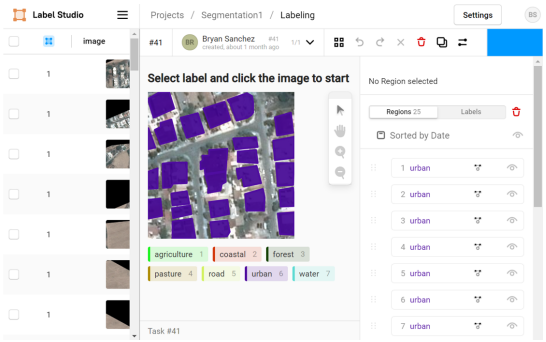

Fig. 3. U-Net architecture of the model used within this research.

For the first training, one hundred images were chosen as the first set of annotations from the region of Nicosia, Cyprus. The goal was to label houses separately (as much as possible) and include all the most minor establishments. Nevertheless, the resolution of some pictures made the separate labelling difficult. Most houses have one or more little backyard cabins or structures that cannot be differentiated effortlessly from other similar buildings. Consequently, a considerable amount of tiny structures were merged into one whole facility. The chosen pictures for the second set of annotations have houses next to each other and buildings with unusual shapes. These images help boost the model performance, giving the ability to detect and segment that kind of establishment. The final methodology for building labelling was to increment the number of train data and correct the previous annotation. Output images with model-generated annotations showed inconsistency. A technique based on active learning was used during this phase to choose images with masks labelled with low confidence based on previous knowledge. Therefore, those images were encountered within the entire dataset and manually annotated to increment the accuracy and improve the model.

5 RESULTS

5.0.1 Unmethodical annotations . The first training aimed to evaluate the model's performance in identifying and segmenting houses and other facilities. Manually labelled snapshots from the region of Nicosia in Cyprus were the training dataset for the CNN model using U-Net as the backbone. After training, the model generated masks for more than five thousand photographs from the same region.

Fig. 4. Image 6144_2560 with its model-generated mask.

Results from the initial version have been encouraging. Having tested the remaining twenty per cent of the dataset, the model reported more than 86% accuracy. Nevertheless, some masks generated by this prototype were not correctly identified and segmented. Figure 4 shows a picture from the dataset that was model-generated. It can be noticed that most houses have some noise in their mask. In addition, Figure 5 displays recognised white covers of homes on the right side that are not even there.

Fig. 5. Image 17664_256 with its model–generated mask.

5.1 Targeted annotations

The main objective for the second set of annotations was to improve the accuracy in detecting the different types of buildings with unusual construction. The methodology was to analyse arbitrary photographs from the previous result that contain such houses and manually annotate as precise as possible. Figure 6 shows the improvements in this procedure. The segmentation of each building was more detailed and with no significant errors because ground truth information about that specific picture now exists. This figure shows almost square shapes in most houses and buildings.

Fig. 6. Image 6144_2560 with its model-generated mask.

The updated model version displayed an improved mask for Figure 7. In this picture, more detailed houses can be appreciated. The model has segmented all its buildings with high confidence. Despite the fact that there is no significant noise in every mask, it still cannot segment the precise shape of the building. Besides, the incorrect-masked structure that the model segmented in the previous version remains, but with a more detailed contour.

Fig. 7. Image 17664_256 with its model-generated mask.

5.2 Boosting performance

After examining the results from the second training, the model needed another round of annotations. Around one hundred pictures were chosen with caution and segmented in great detail. In addition, previous manually annotated snapshots were modified according to a new comparison between the actual image and the resulting image with masks generated by the model. Figure 8 displays two images with masks in three colours. Black colour pixels denote no detected building, and white colour pixels for building segmented with assurance. These two colours are the same as previous results. Red colour pixels indicate structures detected with less confidence.

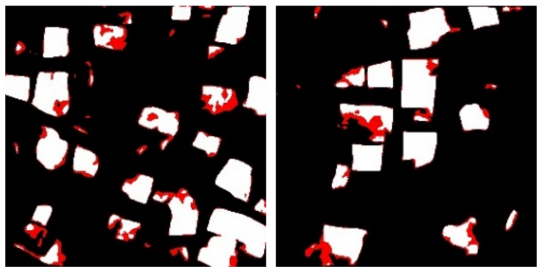

Fig. 8. Image 1792_2048 and 94725632.

The confidence-based learning methodology attempts to determine the correctness and confidence of the model's knowledge. The model uses previous annotations and generated masks to guess pixels and label them as a building, but with low confidence. This technique selected the complex (error-prone) images with red-marked pixels for further annotation. The labelling methodology for this final phase aimed to include red pixels that should form a shape. This new red form should be included in the white mask by manually modifying the previous labels or creating new ones if the picture does not have any. As previous results, the model-generated mask from all inputs of the dataset was significantly improved due to its revised annotation. Figure 9 displays two pictures of the same snapshot containing facilities with distinctive construction. At the top of the photograph, the significant infrastructure is not segmented accurately. The model was boosted with some building annotation with the same features as this one.

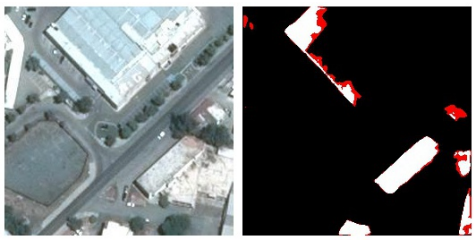

Fig. 9. Image 1792_4608 with its model-generated mask.

5.3 Area estimation

Given the encouraging results of the segmentation model, the area estimation of all detected buildings was no problem. The methodology proposed in this research describes using the white mask contour from all houses. Figure 10 displays three images with a box denoting the border of the structure. In addition, the area estimation of all places is shown next to the frame.

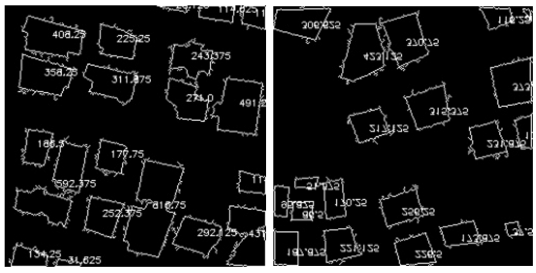

Fig. 10. Image 9472_256 and 28161536 marking borders and areas.

The border of the segmented house is used to calculate the area of the building. For doing this, the following consideration must be considered before reporting the result. The output area should be scaled according to the dimensions of each pixel, $0.5{\times}0.5m^2$. Figure 11 denotes a photograph displaying each house's border and the construction area's estimation.

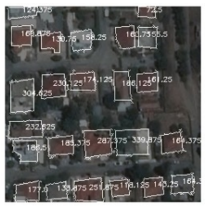

Fig. 11. Image 2560_768 superposed with borders and areas.

5.4 Evaluation

The model's performance was measured using Intersection over Union, the standard metric for segmenting pictures. The IoU measure, defined as the intersection size divided by the union of the two areas, determines how closely an item in an image matches up with the predicted region and the ground-truth region [7]. The model implemented in this approach uses *SoftMax* as a loss operation to address the image segmentation problem. This function seeks to maximize overall accuracy. For instance, if the model incorrectly labelled pixels as background and not the object, the IoU metric can penalize the model and give a low score. Compared to pixel-wise accuracy, IoU is a considerably better indicator of success in segmentation assignments, especially when the input data is significantly and thinly distributed [15].

Table 1 describes the average scores evaluating the model's performance and accuracy during this project. The three model versions were merged into a final version. As discussed in previous subsections, version-1 and version-2 are the ones that used 80 and 147 annotated images, respectively. The precision for making white pixels from the first version is 72.25%. Version-3 contains the data from the last two adding confidence-based techniques and denotes an accuracy for both classes (black and white pixels) are 98.26% and 92.16%, respectively. The final results indicate an improvement in the white class over the initial version. In addition, Appendix A.1 and A.2 includes the function graphs of IoU corresponding to these two models' performance.

Table 1. Model's average score in IoU and accuracy.

Metric / Model	First version	Final version
Mean pixel–wise accuracy	85.21	**95.21**
Mean IoU	78.67	**91.55**

Results from similar methods intended for semantic segmentation within images are compared to the model used in this project. These researches use a pre-trained U-Net model with different backbones to detect dense buildings and other small objects such as forests, farms, and wastelands. Table 2 lists their performance regarding the metric IoU. Method [11] proposes a U-Net model with five categories for labelling buildings in the centre of China. The Bayesian U-Net model presented in the method [2] used a MonteCarlo dropout to improve the accuracy and obtain high-quality scores on four datasets[2]. The research in [16] obtained in validation a high mean IoU score

[2]Vaihingen, Toulouse, Massachussets and INRIA datasets

given the resolution of 1 meter implementing a residual multi-scale model with the dataset of the Dachangshan island, China. However, Table 2 only shows the models' performance using different datasets for training and testing. Therefore, their mIoU cannot be compared in great detail with our implementation since this model was only tested with Nicosia's dataset due to the brief research period.

Table 2. Comparison of the results of the independent tests.

Model	mIoU	Dataset	Resolution
U-Net 5 categories [11]	72.00	Centre of China	0.8m
Bayesian U-Net [2]	72.59	Four datasets	0.6m
Residual multi-scale network [16]	91.00	Dachangshan	1m
Our model	**91.55**	Nicosia	**0.5m**

6 DISCUSSION

The following conclusion can be drawn from these results. First of all, convolutional neural networks seem to perform the best for image segmentation. This performance is possible because of the method's proposed architecture for biomedical image segmentation [9]. As stated in the previous subsection, the model can segment buildings with high confidence. The result from the final training phase reported accuracy in detecting houses and other types of infrastructures equal to 94%. Before annotating the first set of images, the model reported an average of 78% using the metric IoU. Since this research is about object detection and image segmentation, the typical metric used for these approaches is IoU. Therefore, the first sub-question is answered positively since the segmentation model uses annotation as its ground truth. This model improved the generated annotation after the second and third rounds of annotations. It succeeded because it implemented targeted annotating and boosted its performance by adding more pixels to the previous manual annotations.

For the second sub-question, it was determined that selecting images that only contain buildings with unusual shapes. Massive infrastructures or houses that are in a row next to each other were of relevance. The model improved its performance by avoiding structures similar to those annotated before and taking buildings with these characteristics. It can detect these buildings and generate a mask almost identical to the one provided in the input. Figure 12 shows three images. The left one is the actual image, the picture in the middle corresponds to the given annotation, and the one in the right denotes the mask image generated by the model. Both mask images are almost identical, which outlines and answers the second sub-question about how this model will add precision for detecting a building with an unconventional shape or when a house is overlapping. Appendix B.1 illustrates more similar images.

Fig. 12. Comparison between their actual and generated mask.

Nevertheless, there are structures with modern architecture where the model's execution cannot segment them with high precision. Appendix B.2 lists four snapshots in which the model did not perform accurately, each including irregular and inaccurate white masks. Architectural shapes vary due to local cultures, climates, terrains, and even politics or economics, making it difficult to extract such buildings accurately [11]. Dividing the buildings into different categories and labelling them can improve accuracy. The method proposed in [11] states having four or five classes boosts the precision and increases the IoU. These types include dense and messy buildings, small constructions with a fuzzy boundary, neatly organized houses and extensive facilities with a clear border.

Based on these results, the last sub-question argues about the buildings' area estimation accuracy. Blending images with the actual image are displayed in Figure 13. The area calculated depends on how precise the model-generated mask is. Although, after removing the covers' inside, the model generated noise around the edge. Another problem where the model does not perform with precision is shown in Figure 13. This issue arises when two or more segmented white boxes collide or overlap. The model removed the inner part of the boxes with this attribute and merged them. Consequently, the estimated areas do not correspond to the expected calculation since now it reports only one estimation for a combined construction surface. This problem can be seen at the top of Figure 13.

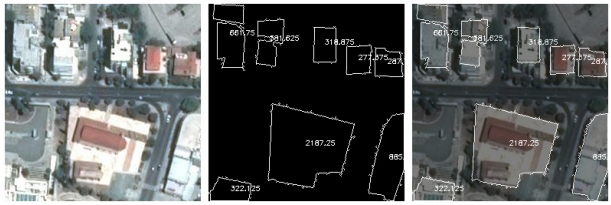

Fig. 13. Image 8448_768 with its estimated area.

After discussing all sub-questions, buildings within satellite imagery can be detected with high precision by a CNN. This approach can be made by annotating structures in more detail, specifically those with unconventional shapes. Besides, homes with overlapping are relevant since annotating them can boost the neural network's performance. Finally, the area estimation will depend on the accuracy of the segmentation model. Hence the accuracy will be as high as the model. These results outline the central question in this project.

7 CONCLUSIONS AND FUTURE WORK

This paper discusses the automated detection of buildings, powered by a convolutional neural network. The U-Net architecture achieves excellent execution in segmenting small objects in the region of Nicosia in Cyprus. Because of the methodology in the annotation process and the label correction, the model improved and outperformed its previous versions.

The results for building detection and area estimation are promising. Nevertheless, this segmentation model can be boosted by combining

it with other techniques to reduce incorrect labelling for unconventional structures and implement a method for area overlapping. As for the dataset, it would be worth observing how the model performs in other regions of Cyprus with different data distributions.

ACKNOWLEDGMENTS

I would like to thank the Secretariat of Higher Education, Science, Technology and Innovation (SENESCYT) in Ecuador, who provided direct funding support for this project by a scholarship. The completion of this research could not have been possible without the expertise of Senior Research Fellow Indrajit Kalita and Dr. Andreas Kamilaris, a beloved thesis adviser. Finally, the author is very grateful to his parents and family; without none of you this would have been possible.

REFERENCES

[1] Omar Ahmadien, Hasan F Ates, Tuncer Baykas, and Bahadir K Gunturk. 2020. Predicting path loss distribution of an area from satellite images using deep learning. *IEEE Access* 8 (2020), 64982–64991.
[2] Clément Dechesne, Pierre Lassalle, and Sébastien Lefèvre. 2021. Bayesian U-Net: Estimating uncertainty in semantic segmentation of earth observation images. *Remote Sensing* 13, 19 (2021), 3836.
[3] Klaus W Deininger et al. 2003. *Land policies for growth and poverty reduction*. World Bank Publications.
[4] Omid Ghorbanzadeh, Thomas Blaschke, Khalil Gholamnia, Sansar Raj Meena, Dirk Tiede, and Jagannath Aryal. 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing* 11, 2 (2019), 196.
[5] Maurice Herlihy. 1993. A methodology for implementing highly concurrent data objects. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 15, 5 (1993), 745–770.
[6] Jishnu Mukhoti and Yarin Gal. 2018. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709* (2018).
[7] Md Atiqur Rahman and Yang Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*. Springer, 234–244.
[8] Abdelatif Rajji, Abdessamad Najine, Amina Wafik, and Amroumoussa Benmoussa. 2022. Building height estimation from high resolution satellite images. *International Journal of Innovation and Applied Studies* 35, 2 (2022), 268–281.
[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[11] Shuting Sun, Lin Mu, Lizhe Wang, Peng Liu, Xiaolei Liu, and Yuwei Zhang. 2021. Semantic segmentation for buildings of large intra-class variation in remote sensing images with O-GAN. *Remote Sensing* 13, 3 (2021), 475.
[12] Dhananjay Theckedath and RR Sedamkar. 2020. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. *SN Computer Science* 1, 2 (2020), 1–7.
[13] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems* 27 (2014).
[14] Camilla Toulmin. 2009. Securing land and property rights in sub-Saharan Africa: The role of local institutions. *Land Use Policy* 26, 1 (2009), 10–19. https://doi.org/10.1016/j.landusepol.2008.07.006 Formalisation of Land Rights in the South.
[15] Floris van Beers, Arvid Lindström, Emmanuel Okafor, and Marco A Wiering. 2019. Deep Neural Networks with Intersection over Union Loss for Binary Image Segmentation.. In *ICPRAM*. 438–445.
[16] Chengyi Wang and Lianfa Li. 2020. Multi-scale residual deep network for semantic segmentation of buildings with regularizer of shape representation. *Remote Sensing* 12, 18 (2020), 2932.
[17] Haibo Wang, Xueshuang Gong, Bingbing Wang, Chao Deng, and Qiong Cao. 2021. Urban development analysis using built-up area maps based on multiple high-resolution satellite data. *International Journal of Applied Earth Observation and Geoinformation* 103 (2021), 102500. https://doi.org/10.1016/j.jag.2021.102500
[18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 3–11.

A FUNCTION GRAPHS

A.1 First Version

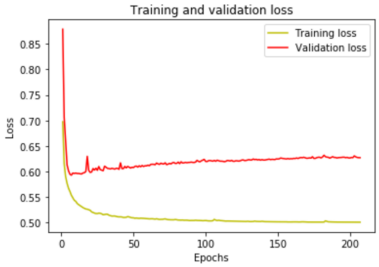

Fig. 14. Training and validation loss graph.

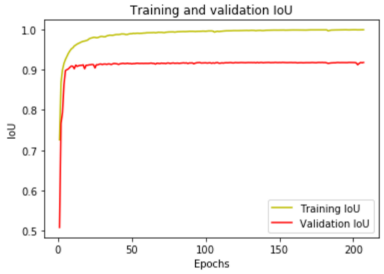

Fig. 15. Intersection over Union graph.

A.2 Final version (version-3)

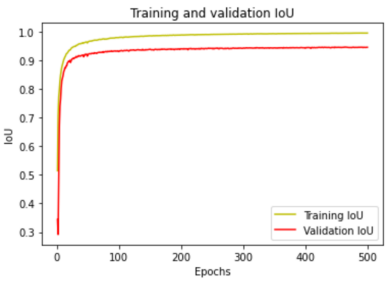

Fig. 16. Training and validation loss graph.

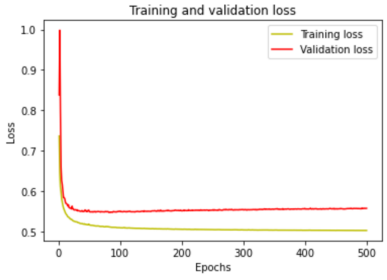

Fig. 17. Intersection over Union graph.

B SEGMENTED PICTURES

B.1 Images with high precision

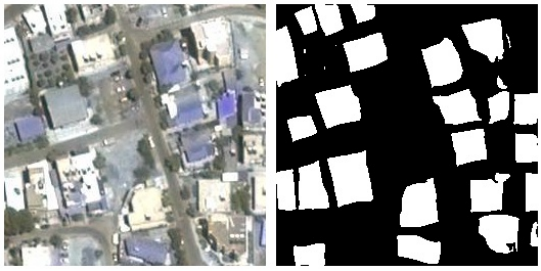

Fig. 18. Image 4608_3328 with a model-generated mask

Fig. 19. Image 7680_4096 with a model-generated mask

Fig. 20. Image 14848_0 with a model-generated mask

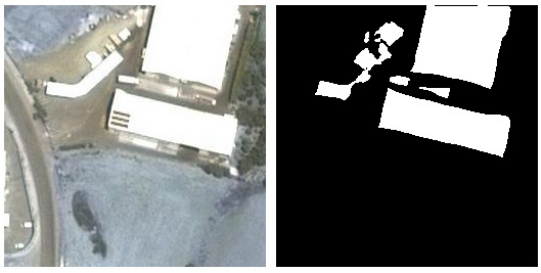

Fig. 21. Image 25344_7168 with a model-generated mask

B.2 Images with low precision

Fig. 22. Image 18944_4096 with a model-generated mask

Fig. 23. Image 16384_1536 with a model-generated mask

Fig. 24. Image 11264_7680 with a model-generated mask

Fig. 25. Image 4352_4608 with a model-generated mask