Canadian Science Publishing

RESEARCH ARTICLE

A region-based deep learning approach to instance segmentation of aerial orthoimagery for building rooftop extraction[1]

Kyle Gao, Mengge Chen, Sarah Narges Fatholahi, Hongjie He, Hongzhang Xu, José Marcato Junior, Wesley Nunes Gonçalves, Michael A. Chapman, and Jonathan Li

Abstract: Updated building information plays an important role in many fields such as environmental monitoring, disaster assessment, and the creation of base maps for urban planning. High-resolution images captured from Earth observation satellites and airborne platforms provide valuable data that cover large areas at high temporal frequencies. In recent years, deep neural networks have shown great potential in the semantic segmentation of Earth observation images for building detection, significantly exceeding the performance of traditional machine learning methods whenever high-quality training datasets are available. Instance segmentation methods further leverage object detection to focus segmentation onto regions of interest, avoiding certain types of false positives and false negatives when compared to semantic segmentation methods. In this study, we approach building rooftop detection as an instance segmentation problem and propose a region-based deep learning approach to building rooftop extraction based on the Mask Regional Convolutional Neural Network (Mask R-CNN) framework. Our study indicates that searching for suitable hyperparameters results in considerable improvements in deep learning models. We found that hyperparameter optimization could be mandatory in some cases because in our experiments, the baseline Mask R-CNN achieved an unacceptable performance when compared to other methods. Our optimized Mask R-CNN, on the other hand, achieves a precision, recall, and F1-score of 92%, 86.6%, and 89.1%, respectively. Furthermore, we show that by using a region-based instance segmentation model, we can avoid the speckle-like errors sometimes found in semantic segmentation models, resulting in clean and accurate rooftop extraction that is more suited for practical applications.

Key words: building detection, deep learning, Mask R-CNN, aerial orthoimagery, instance segmentation.

Résumé : L'information à jour sur les édifices joue un rôle important dans plusieurs domaines, par exemple la surveillance environnementale, l'évaluation des catastrophes et la création de cartes de base pour la planification urbaine. Les images à haute résolution capturées par les satellites d'observation de la terre et les plateformes aéroportées fournissent des données utiles qui couvrent de vastes zones à des fréquences temporelles élevées.

Received 4 September 2021. Accepted 6 January 2022.

K. Gao, M. Chen, S. Narges Fatholahi, H. He, H. Xu, and J. Li.* Geospatial Sensing and Data Intelligence Lab, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
J. Marcato Junior and W. Nunes Gonçalves. Faculty of Engineering, Architecture, Urbanism and Geography, Federal University of Mato Grosso do Sul, Av. Costa e Silca, Campo Grande, MS 79070-900, Brazil.
M.A. Chapman. Department of Civil Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada.
Corresponding author: Jonathan Li (email: junli@uwaterloo.ca).

*Jonathan Li served as an Associate Editor at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by S. Jabari.

[1] This paper is part of a Special Issue on Advances in Geospatial Mapping and Modeling.

Published at www.cdnsciencepub.com/geomat on 7 February 2022.

Au cours des dernières années, les réseaux de neurones profonds se sont montrés prometteurs dans la segmentation sémantique des images d'observation de la terre pour la détection d'édifices, dépassant largement la performance des méthodes d'apprentissage automatique traditionnelles chaque fois que les ensembles de données de formation de haute qualité sont disponibles. Les méthodes de segmentation d'instance optimisent davantage la détection d'objets pour concentrer la segmentation sur les régions d'intérêt, évitant certains types de faux positifs et de faux négatifs, lorsque comparées aux méthodes de segmentation sémantique. Dans la présente communication, nous abordons la détection des toits d'édifices comme un problème de segmentation d'instance et nous proposons une approche d'apprentissage profond axé sur les régions à l'égard de l'extraction des toits d'édifices fondée sur le cadre du Réseau de neurones à convolution régional avec masque (Mask R-CNN). Notre étude indique que la recherche d'hyperparamètres convenables permet d'améliorer considérablement les modèles d'apprentissage profond. Nous concluons que l'optimisation des hyperparamètres pourrait être obligatoire dans certains cas puisque nos expériences, le Mask R-CNN de base, ont atteint une performance inacceptable lorsque comparé à d'autres méthodes. Notre Mask R-CNN optimisé, par ailleurs, a atteint une précision, un rappel et une note F1 de 92 %, 86.6 % et 89.1 % respectivement. De plus, nous montrons qu'en utilisant un modèle de segmentation d'instance axé sur les régions, nous évitons les erreurs de chatoiement que l'on retrouve parfois dans les modèles de segmentation sémantique, ce qui permet d'obtenir une extraction claire et précise des toits qui convient mieux aux applications pratiques. [Traduit par la Rédaction]

Mots-clés : détection d'édifices, apprentissage profond, Mask R-CNN, orthoimagerie aérienne, segmentation d'instance.

1. Introduction

The extraction of two-dimensional (2D) building footprints from aerial orthoimagery plays an important role in population estimation, disaster response, urban planning, geographic database updates, and other fields (Ghanea et al. 2016). Building footprint extraction involves locating and extracting the footprint areas of buildings from aerial or satellite images. Under the assumption that the rooftop and footprint areas are equal, this is also sometimes called building rooftop extraction. Many complex factors, such as variable scales, complex backgrounds (shadows, vegetation, water, and man-made non-architectural features), roof heterogeneity, and rich topological appearance (Ok et al. 2013), make 2D building extraction from aerial orthoimages a very challenging task. Building footprint extraction is often performed by geographic information system (GIS) specialists on a building-by-building basis. However, this type of manual building extraction is slow and labour-intensive.

In the past decade, several studies have been conducted on automatic building detection and extraction using aerial images. However, with the emergence of deep learning methods, research has largely shifted toward deep neural networks. Broadly speaking, building detection and extraction methods can be classified into classical, deep learning semantic segmentation, and deep learning instance segmentation methods.

There are various types of classical approaches such as energy-function-based, template-based, knowledge-based, and classical machine learning. To develop a morphological architectural index method for automatically detecting buildings from high-resolution remotely sensed images, Huang et al. (2013) and Ok et al. (2012) simulated the spatial relationship between buildings and their shadows using a fuzzy landscape generation method. Belgiu and Drăguț (2014) used classical machine learning and designed a multiresolution segmentation method to detect various-scale variations in buildings. Konstantinidis et al. (2016) designed and optimized an energy function to find building boundaries which was combined with feature-based building detection and provided building footprint extraction. Chen et al. (2018) considered a forwarded edge regularity index and shadow clues as new features of the candidate buildings. These studies used local features with a low-level manual for building extractions. Such methods are difficult to apply in a complex situation with a high diversity of buildings.

In recent years, convolutional neural networks (CNNs) have provided great opportunities for automated detection and extraction of buildings from remote sensing images. The success of CNNs lies in their ability to automatically learn multilayer feature extraction and map the original input into convolutional feature maps. This method can recognize detailed features which may not be recognized by classical object-based classification techniques (Hang and Cai 2020). CNNs are often used to conduct pixel-wise classifications of aerial or satellite remote sensing images (Maggiori et al. 2016). They were trained directly to generate classifications from input images using an end-to-end approach. These self-learning features surpassed and gradually replaced traditional handcrafted features and currently occupy a dominant position in the field of building extraction (Vakalopoulou et al. 2015; Sun et al. 2018).

Semantic segmentation involves the assignment of category labels to each pixel in an image. For building footprint extraction, this is the assignment of each pixel in an aerial or satellite image as “building” or “background”. Since 2015, fully convolutional networks (FCNs) (Long et al. 2015) have revolutionized semantic segmentation. These models include SegNet (Badrinarayanan et al. 2017), DeconvNet (Huang et al. 2016), U-Net (Ronneberger et al. 2015), and several other variants. Specific to remote sensing, Sherrah (2016) performed semantic segmentation on aerial images, providing a per-pixel classification of buildings, trees, low vegetation, cars, and impervious surfaces. DenseNet-based and attention-based feature maps for building extraction can naturally achieve multiscale monitoring and effectively suppress background interference (Yang et al. 2018). More recently, building extraction from high-resolution aerial images has also been performed using the U-Net and DenseNet models (Erdem and Avdan 2020). Abdollahi et al. (2020) applied the Seg-Unet method, which is a combination of SegNet and U-Net, to extract objects from high-resolution aerial images. The superiority of this method was confirmed using the Massachusetts Building Dataset. The FCN-based fusion of spatial and spectral information can accurately segment complex and small-scale building structures (Schuegraf and Bittner 2019). By applying regularization constraints, Cheng et al. (2016) explicitly enforced the feature representation of training samples before and after rotation. This contributes to the enforcement of rotational invariance in optical remote-sensing images. To solve the problem of diversity and self-similarity among various types of urban buildings, Cheng et al. (2018) proposed a metric learning regularization term for the D-CNN model, resulting in improved building detection. Li et al. (2020) proposed an end-to-end building footprint generation method integrating a CNN and a feature pairwise conditional random field as a graph model to preserve sharp boundaries and fine-grained segmentation. More recently, Cai et al. (2021) conducted a comparative study of deep learning approaches to rooftop detection using their self-developed Waterloo Building Data using aerial orthoimages. Despite these advances, semantic segmentation methods are trained on pixel-wise errors, which can produce false positives by classifying objects with colours similar to buildings, such as pavement and parking lots, as buildings. Likewise, they can produce false negatives by classifying rooftops with colours similar to the background as background.

Building extraction is often formulated as a semantic-segmentation task. Semantic segmentation, when trained on pixel-wise cross-entropy loss, only considers the pixel-wise classification accuracy. However, the research target of building extraction is not whether a pixel is a building or not. As each building is an object, the instance segmentation method is more suitable for building extraction. Instance segmentation first detects objects of interest, then extracts regions of interest around said objects, and finally segments the object from its background. This approach helps avoid certain types of false positives and false negatives found in semantic segmentation methods. Mask Regional Convolutional Neural Network (Mask R-CNN) (He et al. 2017) is the prevalent architecture for this segmentation approach. Mask R-CNN is an extension of the earlier Faster R-CNN model (Ren et al. 2015), which uses a convolutional backbone, a region proposal network (RPN), and a classification head. Faster R-CNN and its predecessors, Fast R-CNN (Girshick 2015) and R-CNN (Girshick et al. 2014), are object detection models which do not perform segmentation. Wen et al. (2019) achieved good building footprint extraction results using this approach, motivating further research in this direction.

With recent advancements in hardware, automated hyperparameter search has achieved outstanding results. A recent literature review by Yu and Zhu (2020) highlighted the importance of hyperparameter search and architecture search, presenting well-known search techniques and introducing automated machine learning (AutoML) toolkits. However, as noted in these studies, most AutoML techniques require vast computing resources; they are either deployed on cloud-computing services or require powerful workstations with many GPUs.

In this study, we performed a hyperparameter/architecture search on Mask R-CNN to produce a fast and accurate building footprint extraction model. The remainder of this paper is organized as follows. Section 2 introduces the overall workflow, dataset, models, and experiments. Section 3 describes and discusses the results. Finally, Section 4 concludes the paper.

2. Methods

As shown in Fig. 1, the overall workflow of this study was divided into three steps: data preprocessing, model training and evaluation, and building segmentation. The aerial images and ground-truth data were preprocessed and split into training, validation, and test sets. An extensive hyperparameter search was performed, and many models were trained and evaluated to determine the optimal hyperparameters. The trained hyperparameter optimized Mask R-CNN model and benchmark models were evaluated on the test dataset, and the test set was used to produce instance segmentation results.

2.1. Datasets and data pre-processing

Aerial orthoimages were acquired from Christchurch City, New Zealand, covering 220 000 buildings of various shapes over an area of 450 km^2. The orthoimages were georeferenced to the New Zealand Transverse Mercator 2000 (NZTM), with a ground sample distance (GSD) of 7.5 cm and RGB bands. As shown in Fig. 2, the research area was divided into training (70%) and testing (30%) areas. Orthoimages were acquired from the Land Information of New Zealand at a flying height of 1600 m (LINZ Data Service 2014). Ji et al. (2018) produced ground-truth images which were vector data containing the individual building masks of the aerial orthoimage. These were converted into binary masks for the deep learning training.

To investigate the effect of spatial resolution and downsampling, we produced two training datasets using the training portion of the orthoimages from Fig 2. In training Dataset 1, the images and associated building masks were tilled into square patches of size 1024 × 1024 with a stride of 500 pixels to create overlapping tiles. These were then filtered such that tiles with incomplete buildings near the tile boundaries or tiles with no buildings were removed. Dataset 1 had the same spatial resolution as the original orthoimages. Because a large patch size is more computationally demanding, a smaller patch size may not include sufficient spatial context; 512 × 512 is the size commonly used in the literature. Using downsampling, we created Dataset 2. The images and associated building masks from Dataset 1

Figure 1. Overall method workflow. [Colour online.]

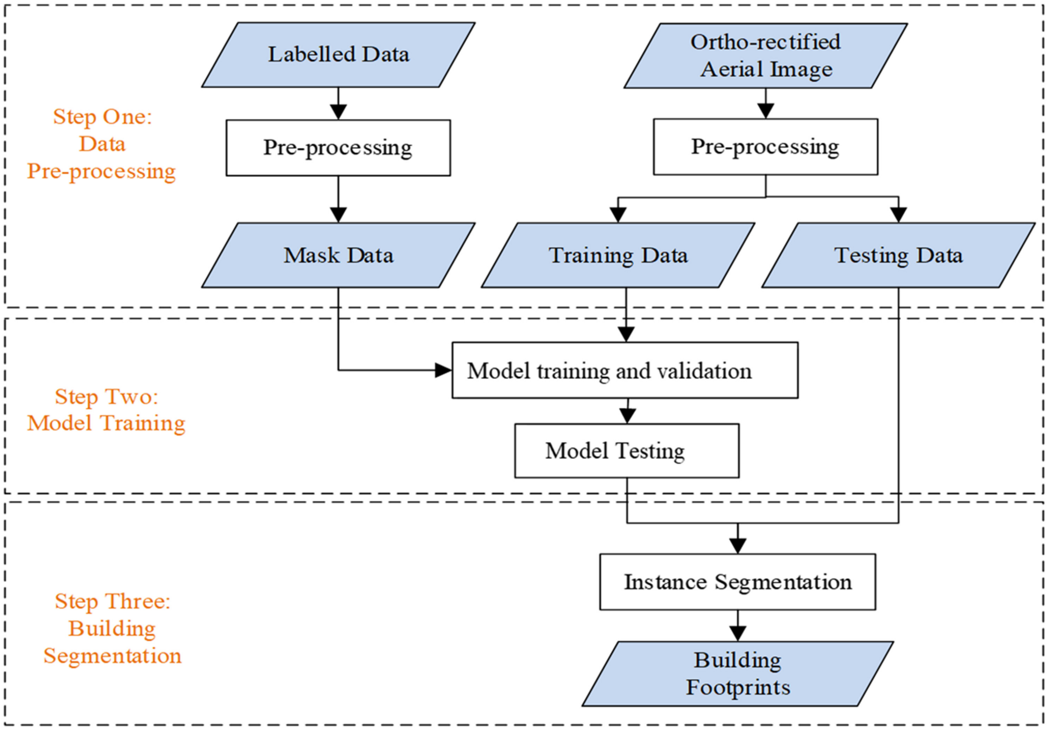
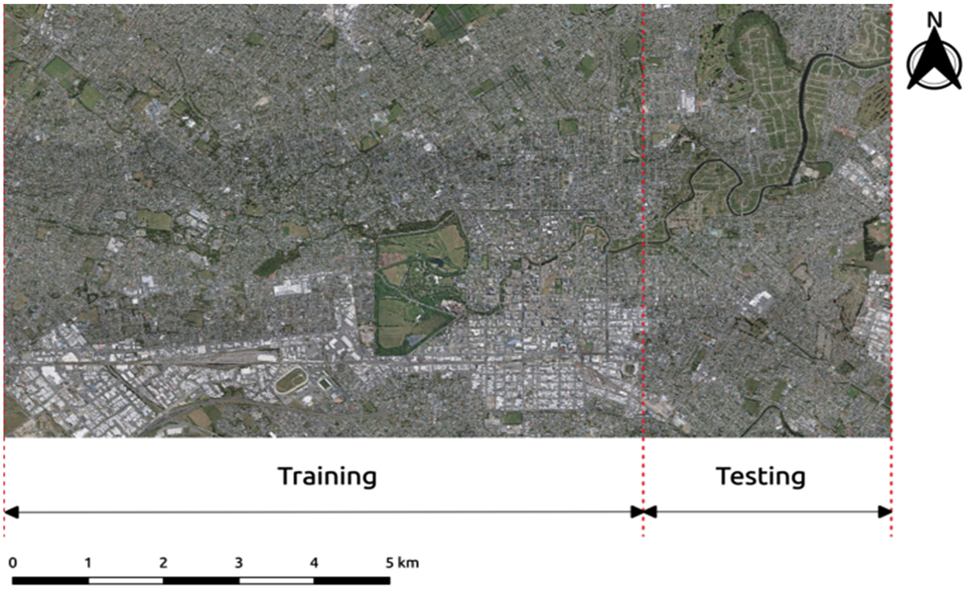


Figure 2. Christchurch, New Zealand, training and testing partitioning. Edited with ArcGIS, image retrieved from LINZ Data Service (2014) under the Creative Commons Attribution 4.0 International License. [Colour online.]



were reduced to 512 × 512 pixels via a 2 × 2 median filter, resulting in images with a spatial resolution of 15 cm. Following Goodfellow et al. (2016), a threshold was applied to the extracted patches to filter out images with no building objects. Band shuffling, in which the red, green, and blue channels in the selected images are shuffled to present images with different combinations of bands to help prevent overfitting (Bei et al. 2018).

The ground-truth building labels were converted into binary mask tiles, with 0 denoting background pixels and 1 denoting building pixels. Occasionally, tiling cropped the building masks across multiple tiles. These incomplete building masks can degrade prediction accuracy; therefore, a two-step solution was implemented to address this problem. The images were tilled using a stride (set to 500 pixels) during the window sliding. This causes the neighbouring tiles to overlap. Finally, the images with incomplete building masks were removed.

2.2. Training and evaluation

2.2.1. Computing environment

The computing environment used in this study included an Intel® CPU i7-9700k, 64 GB RAM, and an NVIDIA GeForce GTX 1080 8 GB GPU. TensorFlow was used to implement the deep learning experiments. The models were trained using stochastic gradient descent with momentum (SGDM), following the original implementation of Mask R-CNN.

2.2.2. Model

Mask R-CNN was the architecture used in this study. Although we made hyperparameter and architecture changes, our model should still be considered as a Mask R-CNN model. In this study we differentiated our model from the default Mask R-CNN model by referring to the default as "baseline Mask R-CNN". Mask R-CNN was innovated on an earlier R-CNN type architecture by adding a mask branch which computed the object mask of a region of interest generated by the RPN. In the context of our building extraction model, the RPN identifies potential buildings and selects the surrounding area as the region of interest. The regions of interest were passed onto the mask branch which in turn identified the building and background pixels. For this study, we optimized the hyperparameters and architecture of the Mask R-CNN which we then compared to the baseline Mask R-CNN and U-Net (Ronneberger et al. 2015). We used two variants of ResNet (He et al. 2016), namely, ResNet-50 with 23.5 million parameters and ResNet-101 with 42.5 million parameters, as convolutional backbones. Figures 3 and 4 show the different components of Mask R-CNN and layers of its original mask branch, respectively. Readers are referred to the original papers for detailed descriptions of the architectures of the respective models.

2.2.3. Performance metrics

It should be noted that Mask R-CNN-based methods formulate the building extraction task as an instance segmentation. In contrast, U-Net-based methods formulate the task as semantic segmentation. Although both methods are compared in this study using precision, recall, and F1-score, for instance segmentation, these three words represent the average precision (AP), average recall (AR), and F1-score at the intersection-over-union (IoU) threshold of 0.5. The IoU is the ratio of the intersection area of the prediction and ground truth to the area of their union; it is used to quantify their mutual overlap. The AP/AR/F1-score at some IoU threshold is calculated for each detection (building in this case) of an IoU with respect to its associated ground truth; the detection is classified as positive if and only if the resulting IoU exceeds the predetermined threshold. Precision (P) is defined as the ratio of true-positive detections to the total number of positive samples. Recall (R) is defined as the ratio of true-positive detections to the total number of detections.

Figure 3. Diagrammatic representation of the different components of Mask R-CNN. [Colour online.]

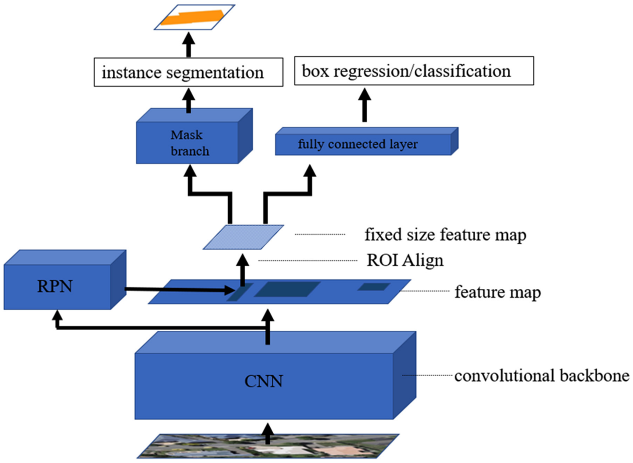
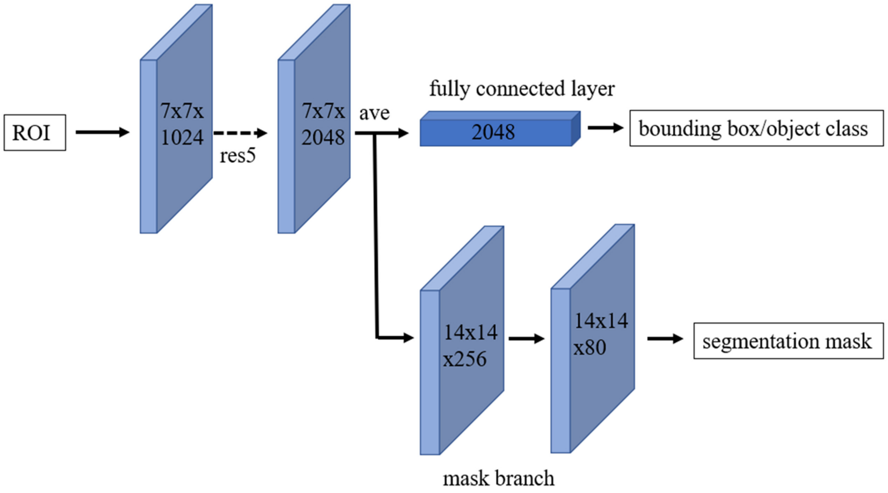


Figure 4. Example mask branch and bounding box branch of Mask R-CNN. [Colour online.]

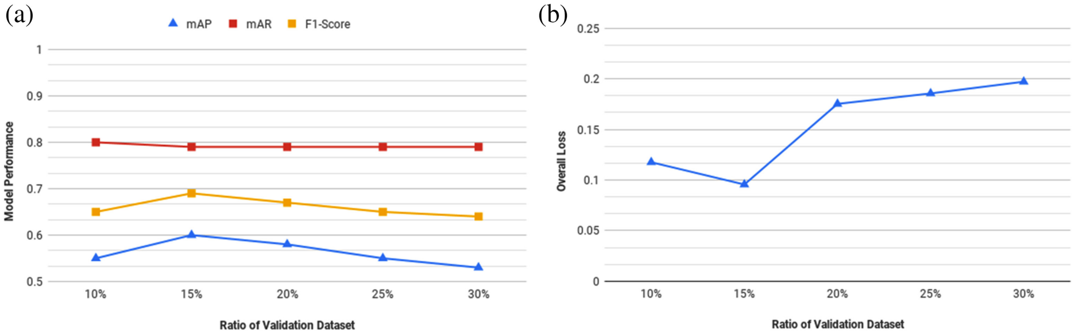


The F1-score is given by the harmonic mean of precision and recall: $\text{F1} = \frac{2(P \times R)}{P + R}$. Precision, recall, and F1-score were calculated according to the detection threshold for IoU. This procedure was repeated for each object class and then averaged. For building extraction, this step was not applied because we only had a single type of object. For comparison purposes, we used instance segmentation metrics to evaluate both U-Net and Mask R-CNN. The mean average precision/recall (mAP/mAR) averaged the respective metrics over different IoU thresholds.

2.3. Building extraction results

Four hyperparameters/architecture optimization experiments were conducted to produce an optimized Mask R-CNN. These were the dataset partitioning ratio, mini-mask size, learning rate, backbone type, and initialization choice. The models were trained for 40

Figure 5. (*a*) Performance metrics for different ratios of training set to validation set. (*b*) Difference of training loss to validation loss for different validation–training ratios. [Colour online.]



epochs on Dataset 2 during the search phase, and compared using the validation set(s) associated with Dataset 2.

The optimized Mask R-CNN was trained for 200 epochs for each dataset. The detailed results of AP and AR at different IoU thresholds are tabulated and presented. We found the most well-trained model based on the detailed performance metrics across the datasets and training epochs. The model was evaluated using a test set. The results were benchmarked against the baseline Mask R-CNN and U-Net trained for the same number of epochs on the same dataset. The results of the optimized Mask R-CNN are presented in detail in the following section.

3. Results

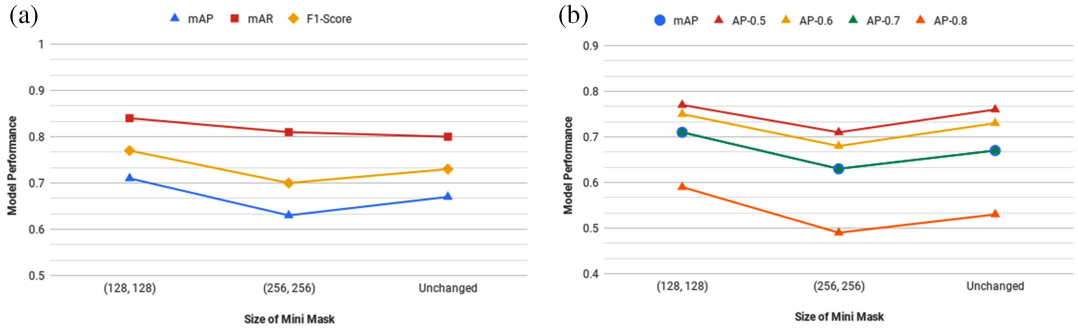
In this section, we present the detailed results of the proposed framework which was conducted as part of the thesis work of Chen (2019). First, the results for hyperparameters/architectures are presented. Then, we report the building rooftop detection results of our optimized Mask R-CNN across different epochs when trained on the two datasets, followed by a comparison with the baseline Mask R-CNN and the popular U-Net.

3.1. Hyperparameter and architecture search

To investigate the optimal partitioning of the validation set into the training set, five ratios (10%, 15%, 20%, 25%, and 30%) were used. The performance measures of mean average precision (mAP), mean average recall (mAR), and F1-score were used to benchmark a Mask R-CNN model trained on Dataset 2 for each of these ratios. As shown in Fig. 5*a*, when using 15% validation data, the F1-score and mAP reached their highest values. With an increase in the proportion of the validation dataset, the mAP value decreased significantly. Except for the 10% validation-training ratio, the other four ratios had the same mAR scores. Figure 5*b* shows the difference between the training loss and validation loss across the five split ratios, which was minimized at a 15% validation-training split. Therefore, a ratio of 15% was chosen as the optimal split ratio for the validation set to the training set for this dataset and model architecture.

The mini-mask size of the mask branch was typically the same as that of the region of interest. A small region of the original image usually contains a single object of interest generated by the RPN. To determine whether adjusting the mini-mask size would significantly affect model performance, we experimented with mini-mask sizes of 128 × 128 pixels, 256 × 256 pixels, and default mask size of 56 × 56 pixels. The performance metrics mAP, mAR, and F1-Score were used to compare the models using the three different

Figure 6. (*a*) Performance metrics with respect to mini-mask size. (*b*) Average precision for different IoU thresholds with respect to mini-mask size. [Colour online.]



sizes, as shown in Fig. 6. The mini-mask (128 × 128 pixels) model was superior to the other two models in terms of all three metrics.

The performance increase owing to the adjustment of the mini-mask size could be caused by the size of the buildings in the input images and the resulting region of interest. Upscaling the mini-mask could result in more detailed segmentation boundaries. However, an excessively large change in the mask size can introduce various geometric errors. The mini-mask size of 256 × 256 pixels exhibited the worst performance. When comparing the AP at an IoU threshold of 0.5, a mini-mask size of 128 × 128 pixels showed a similar performance to the default 56 × 56 pixel mini-mask. However, at a higher IoU threshold, the AP of the mini-mask size of 128 × 128 pixels exceeded that of the default mini-mask size, indicating that the predictions at a mini-mask size of 128 × 128 pixels had a higher percentage overlap with the ground truth. Considering the AP at different IoUs, we selected 128 × 128 pixels as the mini-mask size for our model.

Using the SGDM, four learning rates (0.005, 0.001, 0.0005, and 0.0001) were tested as part of the hyperparameter search. As shown in Fig. 7, the learning rates of 0.005 and 0.001 exhibited the best performance; these two learning rates showed similar mAP values. However, the learning rate of 0.001 showed a higher mAR than the other learning rates, which was also reflected in the F1-score. Consequently, we chose to train our models at a learning rate of 0.001.

A more complex and larger backbone results in a neural network with a greater representational power. However, training a larger backbone is often difficult. This is particularly apparent for smaller datasets and simpler tasks, in which a larger model trained on a large number of epochs can easily overfit. Moreover, it is well known that transfer learning is an effective technique; pre-trained models can save significant training time and often achieve better performance. In our study, we were able to procure pre-trained weights for ResNet-50 trained on ImageNet and ResNet-101 trained on COCO. For our architecture search, we tested ResNet50 (ImageNet) and ResNet-101 (COCO) as Mask R-CNN backbones. To investigate the effects of transfer learning, we also compared the aforementioned ResNet-50 initialized with ImageNet weights and ResNet-101 initialized with COCO weights with a random initialization ResNet-101. As shown in Fig. 8, the model using the ResNet-101 (COCO) backbone exhibited better performance across the three metrics, while the training time was not significantly longer. Therefore, ResNet-101 initialized with COCO pretrained weights was chosen as the backbone of the optimized Mask R-CNN model.

Table 1 presents a summary of the hyperparameter/architectural search. We used a validation-training set ratio of 15%, SGDM, and a learning rate of 0.001 to train our building

Figure 7. Performance metrics for different learning rates. [Colour online.]

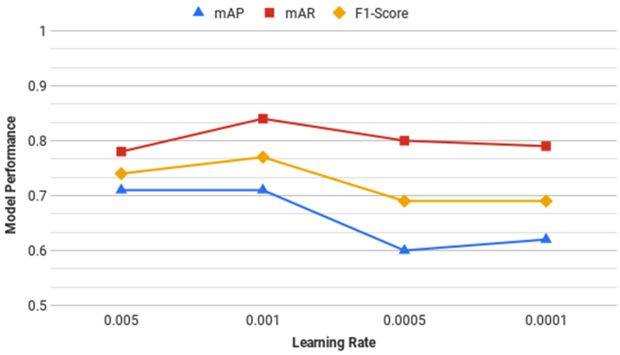


Figure 8. Performance metrics for ResNet-50 (ImageNet) and ResNet-101 (COCO), with a ResNet-101 random initialization backbone. [Colour online.]

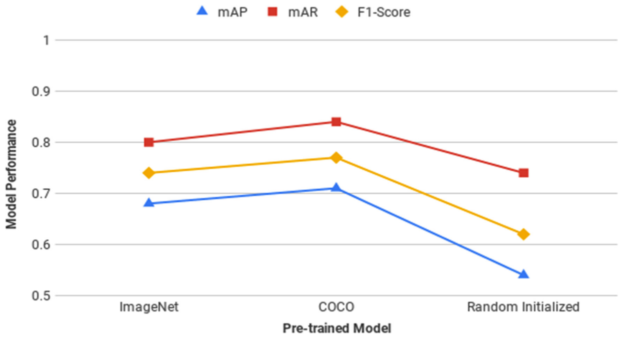


Table 1. Summary of the hyperparameter/architecture search.

Hyperparameters/architecture	Best value
Ratio of the validation dataset	15%
Mini-mask size	(128, 128)
Learning rate	0.001
Backbone	ResNet-101
Model initialization	ResNet-101 (COCO pretrained)

extraction model, which was a Mask R-CNN with a mini-mask size of (128, 128), with a ResNet-101 backbone pretrained on the COCO dataset.

3.2. Model training and results

In the experiment, we used Mask R-CNN trained on two datasets with different spatial resolutions. Figure 9 shows the loss functions of the training and validation sets over 200 epochs. By examining the training and validation losses in Fig. 6, we noticed signs of overfitting at approximately 200 epochs for Dataset 2 and 140 epochs for Dataset 1. Therefore, the training was stopped at 200 epochs. The detailed AP and AR scores at different IoU thresholds for different epochs are presented in Tables 2 and 3.

Figure 9. Training and validation loss on Datasets 1 and 2 over 200 epochs. [Colour online.]

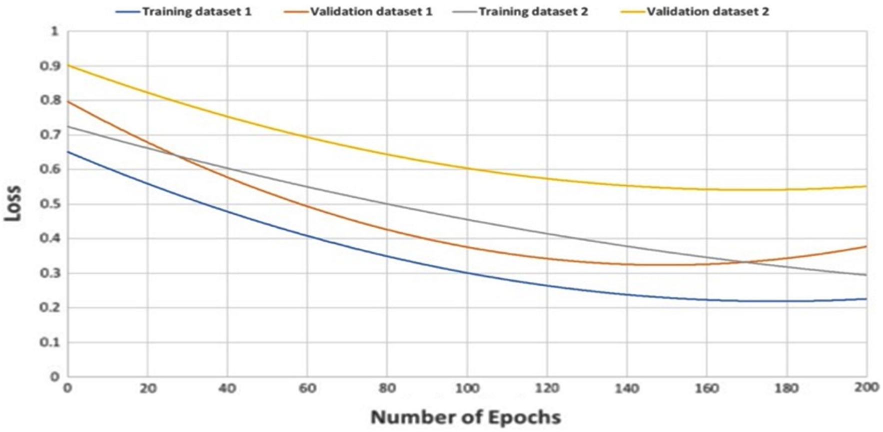


Table 2. Average precision and average recall at different IoU thresholds for Dataset 1 (%).

Epoch	$AP_{0.5}$	$AP_{0.6}$	$AP_{0.7}$	$AP_{0.8}$	mAP	mF1-score
10	89.2	85.3	79.0	61.1	78.6	67.0
40	89.5	86.5	81.4	67.7	81.3	77.4
120	92.0	88.4	83.1	68.5	83.0	**80.5**
200	**94.2**	**90.6**	**85.5**	**71.3**	**85.4**	80.4
	$AR_{0.5}$	$AR_{0.6}$	$AR_{0.7}$	$AR_{0.8}$	mAR	
10	66.1	63.8	58.6	45.3	58.3	
40	81.3	78.5	74.0	61.5	73.8	
120	**86.6**	**83.3**	**78.2**	**64.5**	**78.1**	
200	83.8	80.6	76.0	63.4	75.9	

Table 3. Average precision and average recall at different IoU thresholds for Dataset 2 (%).

Epoch	$AP_{0.5}$	$AP_{0.6}$	$AP_{0.7}$	$AP_{0.8}$	mAP	mF1-score
10	74.2	695	62.3	45.4	62.8	69.8
40	81.7	77.8	71.7	56.6	72.0	76.7
120	**85.2**	80.9	73.6	55.3	73.8	76.8
200	84.8	**81.6**	**76.2**	**62.7**	**76.3**	**80.1**
	$AR_{0.5}$	$AR_{0.6}$	$AR_{0.7}$	$AR_{0.8}$	mAR	
10	92.7	86.8	77.9	56.6	78.5	
40	93.2	88.8	81.8	64.6	82.1	
120	92.4	87.8	79.9	60.0	80.0	
200	**93.7**	**90.2**	**84.2**	**69.3**	**84.4**	

Figure 10. Regional-level results: (*a*) aerial image, (*b*) ground truth, (*c*) segmentation by model trained on Dataset 1, and (*d*) segmentation by model trained on Dataset 2. [Colour online.]

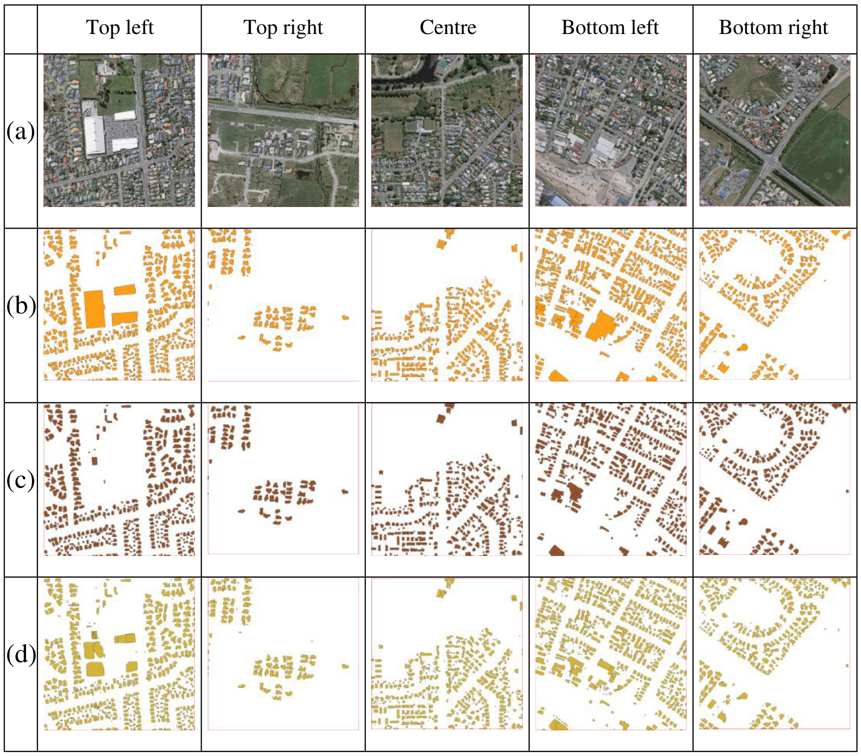


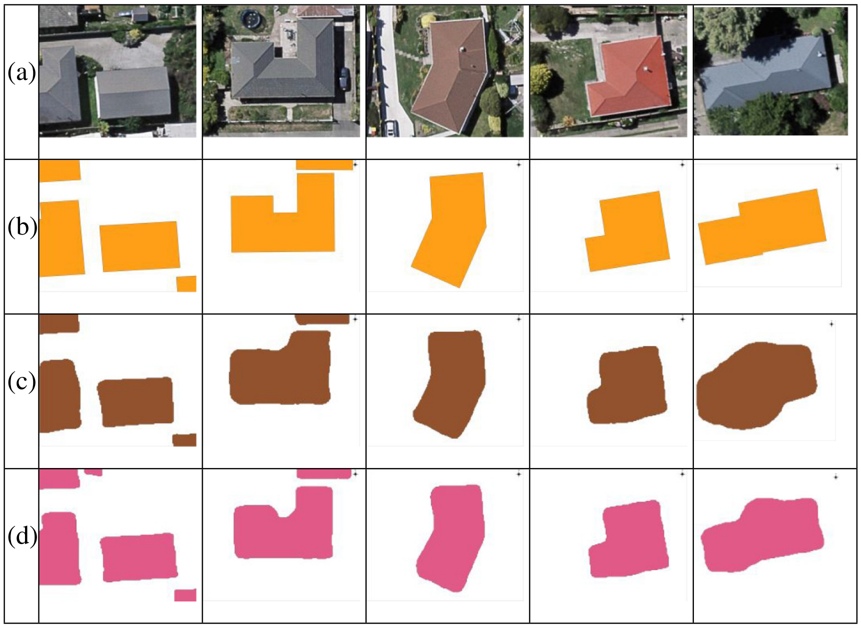
Figure 10 shows visual results of the models trained on Datasets 1 and 2 and compares them with the original images and ground truths. The detection results at the regional level included five research areas: upper left, upper right, middle left, lower left, and lower right. Each image represented a square area of 576 m^2. Across all regions, the models trained on both datasets performed well visually. We did notice some over-segmentation of large buildings in the upper left and lower left regions.

Figure 11 shows the detection results for typical single-house-level samples, including a challenging scenario of occlusion by trees and shadows. Models trained on both datasets were able to extract building footprints for houses of different shapes, sizes, and orientations; neither model struggled with tree- and shadow-occlusion scenarios. However, the segmentation results struggled to achieve sharp edges and corners of the ground truth. This is particularly apparent in tree and shadow occlusion scenarios.

We used AP and AR at various IoU thresholds, as well as mAP and mean F1-score (mF1) for quantitative analysis. The detailed results for the models trained on Datasets 1 and 2 and evaluated on their respective validation sets are presented in Tables 2 and 3, respectively. The model trained on Dataset 1 achieved the highest mF1-score of 80.5% at epoch 120. The model trained on Dataset 2 achieved the highest mF1-score of 80.1% at epoch 200. These two scores are sufficiently similar to conclude that the two models trained for their respective numbers of epochs have similar performances.

When trained on Dataset 1, we observed that the model achieved its highest AP at an IoU threshold of 0.5 at epoch 200, whereas it achieved its highest AR at epoch 120. This was expected because a higher IoU threshold requires a much larger and more accurate overlap

Figure 11. Representative samples of segmentation results at the single-house level: (*a*) aerial image, (*b*) ground truth, (*c*) models using training Dataset 1, and (*d*) models using training Dataset 2. [Colour online.]



between the segmentation result and ground truth before accepting a detection as positive. We observed signs of overfitting in Dataset 1 at approximately 140 epochs. It is likely that the AR at different IoU thresholds at 200 epochs was lower than that at 120 epochs owing to overfitting. It is interesting to note that overfitting in this case manifested itself on the AR scores and not the AP scores. This indicates that the main overfitting error was the relative increase in false negatives across all IoU thresholds, that is, not detecting a building in the validation set where there was one, which in turn decreased the AR.

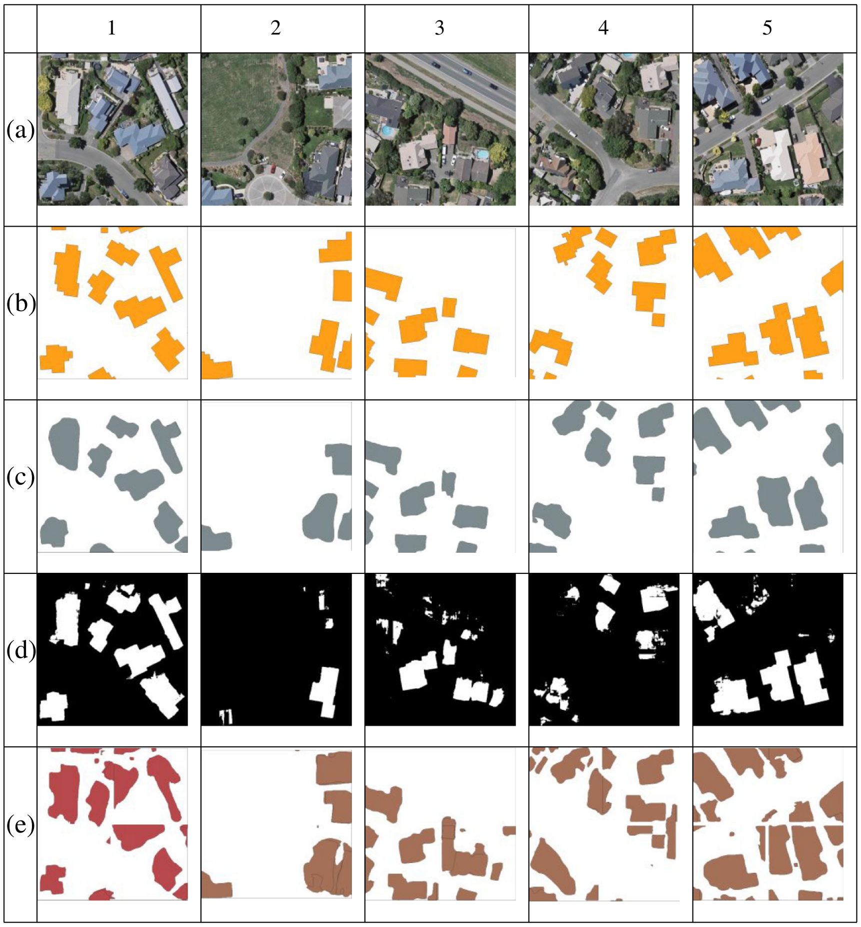
The model trained on Dataset 2 did not exhibit strong overfitting at 200 epochs because we terminated the training as soon as we began to observe signs of overfitting. This model achieved its highest mF1-score of 80.1%, which was comparable to that of the model trained on Dataset 1. The model achieved the highest AR across all the IoU thresholds at 200 epochs. This finding is consistent with our belief that the model did not overfit. The model achieved its highest AP across all IoU thresholds at 200 epochs except for the lowest threshold of 0.5, which was achieved at 120 epochs. The model's $AP_{0.5}$ dropped by 0.4% from epoch 120 to epoch 200. This may indicate that, from epochs 120 to 200, the model learned to produce both more true positives and false positives when the IoU threshold was low. These false positives were not reflected in the AR scores and were easily filtered out by demanding a higher overlap between the segmentation results and ground truth using higher IoU thresholds. Therefore, this problem does not manifest when a higher IoU threshold is used.

Table 4 summarizes the training results. An IoU threshold of 0.5 is typically used in instance segmentation tasks, for which we report the AP, AR, and F1-score. The model achieved the highest validation scores when trained on Dataset 1 for 120 epochs. The training time for Dataset 1 was 256 min, nearly twice as long as that of Dataset 2 (145.52 min). Considering the final F1-scores, we selected the model trained on Dataset 1 for 120 epochs as the building detection model.

Table 4. Summary of results of the performance metrics at IoU thresholds of 0.5 (%).

	$AP_{0.5}$	$AR_{0.5}$	F1-score$_{0.5}$
Dataset 1 at epoch 120	92.0	86.6	**89.1**
Dataset 1 at epoch 200	**94.2**	83.8	88.6
Dataset 2 at epoch 120	85.2	92.4	88.7
Dataset 2 at epoch 200	84.8	**93.7**	89.0

Figure 12. Five building detection scenes: (*a*) input image, (*b*) ground truth, (*c*) optimized Mask R-CNN, (*d*) U-Net, and (*e*) baseline Mask R-CNN. [Colour online.]



3.3. Comparison to benchmark models

The optimized Mask R-CNN and two benchmark models (U-Net (Ronneberger et al. 2015) and baseline Mask R-CNN (He et al. 2017)) were evaluated on the test set after being trained for 120 epochs on training Dataset 1. Figure 12 compares the results from the three models for five representative scenes. The baseline Mask R-CNN, which we used as the benchmark,

Table 5. Performance metrics comparison on test set at IoU thresholds of 0.5 (%).

Method	Precision	Recall	F1-score
Baseline Mask R-CNN	30.0	88.7	44.7
U-Net	91.8	77.1	83.8
Optimized Mask R-CNN	92.0	86.6	89.1

had a mini-mask size of (56,56) and used ResNet-101 as the backbone, but was randomly initialized as opposed to pretrained on COCO. The same learning rate was used across all models for a fair comparison. We compared our optimized model to the baseline to quantify the effects of our hyperparameter/architecture search. We selected U-Net to represent the semantic segmentation approach for building rooftop extraction. Although U-Net is not state-of-the-art, it is a foundational model commonly used in benchmarks and can be used as a basis for cross-benchmark comparisons.

The proposed model properly identified all buildings in the image; however, the detection results lacked the sharp edges and corners of the ground-truth masks. U-Net performed well in producing sharp and defined boundaries but tended to oversegment and produce speckle-like errors. This was likely because Mask R-CNN is an instance segmentation method that considers individual objects, whereas U-Net is a semantic segmentation method designed to approach the task on a per-pixel basis. The baseline Mask R-CNN exhibited the worst performance with both inconsistent segmentation boundaries and a high rate of false positives.

Table 5 lists the performance metrics at the IoU threshold of 0.5. The optimized Mask R-CNN achieved the best precision, recall and F1-score. In particular, its recall was significantly higher than that of U-Net, which indicates that Mask R-CNN produces much fewer false negative results; the optimized Mask R-CNN tended to correctly identify buildings where there should be one, whereas U-Net sometimes missed detection. This can be observed in the second and fifth columns of Fig. 12. The baseline Mask R-CNN produced surprisingly low AP scores, indicating a high rate of false positives, which can be seen in all five scenes in Fig. 12. The baseline Mask R-CNN's mask branch seemed to associate colour with the rooftop, therefore falsely identifying the pavement as a building. The main roads did not suffer from this issue because the RPN did not identify roads as regions of interest. However, the poorly trained mask branch resulted in false positives on driveways near houses which were inside the regions of interest.

Overall, we noted that the optimized Mask R-CNN significantly outperformed the baseline Mask R-CNN, which indicates the importance of hyperparameters and architecture tuning. Furthermore, the optimized Mask R-CNN outperformed U-Net by producing fewer false negatives and reducing speckle-like errors. However, we noticed a tendency of our model to produce overly rounded boundaries. The ground-truth building boundaries tend to have sharp corners with straight edges, which our model struggled to reproduce. At very high resolutions, these boundary errors were less significant; however, at lower resolutions, they were no longer negligible. We suggest investigating boundary regularization and super-resolution methods before applying our model to lower-resolution aerial or satellite imagery.

4. Concluding remarks

Manual extraction of the building footprint area is costly and time-consuming. In recent years, deep learning methods have achieved impressive results for the automatic extraction of building footprints from aerial images. In particular, instance segmentation models have shown promise in avoiding certain types of errors found in more commonly used semantic segmentation models. In this study, we developed a hyperparameter/architecture and created an optimized Mask R-CNN model for instance segmentation of aerial orthoimagery for building rooftop extraction. Compared with the other two studied building rooftop extraction models, our method achieved a better performance, with a precision of 92%, recall of 86.6%, and F1-score of 89.1%. We note that by using Mask R-CNN and formulating the building extraction problem as instance segmentation, we avoided the speckle-like errors often found in semantic segmentation methods. The results also showed that optimizing the hyperparameters of an older model is a simple and effective way of achieving near-state-of-the-art results. This is especially important because the baseline Mask R-CNN achieved significantly worse results, which further emphasizes the importance of the hyperparameter search. Overall, Mask-RCNN-type architectures and instance segmentation of building footprints show great potential and warrant further research. In future research, we seek to address the issue of rounded edges and corners in our instance segmentation results to further improve the building extraction accuracy via boundary regularization. Combining boundary regularization and super-resolution would allow us to extend the model to lower-resolution aerial and satellite images in future research, thereby significantly improving the adaptability of our model.

References

Abdollahi, A., Pradhan, B., and Alamri, A.M. 2020. An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images. Geocarto Int. doi:10.1080/10106049.2020.1856199.

Badrinarayanan, V., Kendall, A., and Cipolla, R. 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12): 2481–2495. doi:10.1109/TPAMI.2016.2644615. PMID:28060704.

Bei, Y., Damian, A., Hu, S., Menon, S., Ravi, N., and Rudin, C. 2018. New techniques for preserving global structure and denoising with low information loss in single-image super-resolution. Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 987–994. doi:10.1109/CVPRW.2018.00132.

Belgiu, M., and Drăguț, L. 2014. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. ISPRS J. Photogramm. Remote Sens. **96**: 67–75. doi:10.1016/j.isprsjprs.2014.07.002. PMID:25284960.

Cai, Y., He, H., Yang, K., Fatholahi, S.N., Ma, L., Xu, L., and Li, J. 2021. A comparative study of deep learning approaches to rooftop detection in aerial images. Can. J. Remote Sens. **47**(3): 413–431. doi:10.1080/07038992.2021.1915756.

Chen, M. 2019. Building detection from very high resolution remotely sensed imagery using deep. Neural Networks. M.Sc. Thesis, University of Waterloo. UWSpace. Available from http://hdl.handle.net/10012/14593.

Chen, R., Li, X., and Li, J. 2018. Object-based features for house detection from RGB high-resolution images. Remote Sens. **10**(3): 451. doi:10.3390/rs10030451.

Cheng, G., Zhou, P., and Han, J. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. **54**(12): 7405–7415. doi:10.1109/TGRS.2016.2601622.

Cheng, G., Yang, C., Yao, X., Guo, L., and Han, J. 2018. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans. Geosci. Remote Sens. **56**(5): 2811–2821. doi:10.1109/TGRS.2017.2783902.

Erdem, F., and Avdan, U. 2020. Comparison of different U-Net models for building extraction from high-resolution aerial imagery. Int. J. Environ. Geoinformatics, **7**(3): 221–227. doi:10.30897/ijegeo.684951.

Ghanea, M., Moallem, P., and Momeni, M. 2016. Building extraction from high-resolution satellite images in urban areas: Recent methods and strategies against significant challenges. Int. J. Remote Sens. **37**(21): 5234–5248. doi:10.1080/01431161.2016.1230287

Girshick, R. 2015. Fast R-CNN. In Proc. IEEE International Conference on Computer Vision. pp. 1440–1448. doi:10.1109/ICCV.2015.169.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587. doi:10.1109/CVPR.2014.81.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep learning. MIT Press. Available from https://mitpress.mit.edu/books/deep-learning.

Hang, L., and Cai, G.Y. 2020. CNN based detection of building roofs from high resolution satellite images. ISPRS Arch. **42**(3/W10): 187–192. doi:10.5194/isprs-archives-xlii-3-w10-187-2020.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. doi:10.1109/CVPR.2016.90.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. 2017. Mask R-CNN. In Proc. IEEE International Conference on Computer Vision. pp. 2980–2988. doi:10.1109/ICCV.2017.322.

Huang, X., Zhang, L., and Zhu, T. 2013. Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**(1): 105–115. doi:10.1109/JSTARS.2013.2252423.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., and Pan, C. 2016. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proc. IGARSS 2016. pp. 1835–1838. doi:10.1109/IGARSS.2016.7729471.

Ji, S., Wei, S., and Lu, M. 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. **57**(1): 574–586. doi:10.1109/TGRS.2018.2858817.

Konstantinidis, D., Stathaki, T., Argyriou, V., and Grammalidis, N. 2016. Building detection using enhanced HOG–LBP features and region refinement processes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(3): 888–905. doi:10.1109/JSTARS.2016.2602439.

Li, Q., Shi, Y., Huang, X., and Zhu, X.X. 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). IEEE Trans. Geosci. Remote Sens. **58**(11): 7502–7519. doi:10.1109/TGRS.2020.2973720.

LINZ Data Service. 2014. Christchurch post-earthquake 0.1 m urban aerial photos (24 February 2011). Available from https://data.linz.govt.nz/layer/51932-christchurch-post-earthquake-01m-urban-aerial-photos-24-february-2011/.

Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440. doi:10.1109/CVPR.2015.7298965.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. **55**(2): 645–657. doi:10.1109/TGRS.2016.2612821.

Ok, A.O. 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS J. Photogramm. Remote Sens. **86**: 21–40. doi:10.1016/j.isprsjprs.2013.09.004.

Ok, A.O., Senaras, C., and Yuksel, B. 2012. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. IEEE Trans. Geosci. Remote Sens. **51**(3): 1701–1717. doi:10.1109/TGRS.2012.2207123.

Ren, S., He, K., Girshick, R., and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In Proc. 28th Int. Conf. Neural Information Processing Systems, vol. 1. pp. 91–99.

Ronneberger, O., Fischer, P., and Brox, T. 2015. U-net: convolutional networks for biomedical image segmentation. In Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

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

Sherrah, J. 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. Available from https://arxiv.org/abs/1606.02585.

Sun, Y., Zhang, X., Zhao, X., and 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 Sens. **10**(9): 1459. doi:10.3390/rs10091459.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., and Paragios, N. 2015. Building detection in very high-resolution multispectral data with deep learning features. *In* Proc. IGARSS 2015. pp. 1873–1876. doi:10.1109/IGARSS.2015.7326158.

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., and Wang, P. 2019. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. Sensors, **19**(2): 333. doi:10.3390/s19020333.

Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., and Xu, Y. 2018. Building extraction in very high resolution imagery by dense-attention networks. Remote Sens. **10**(11): 1768. doi:10.3390/rs10111768.

Yu, T., and Zhu, H. 2020. Hyper-parameter optimization: a review of algorithms and applications. Available from https://arxiv.org/abs/2003.05689.