Taylor & Francis
Taylor & Francis Group

Canadian Journal of Remote Sensing
Journal canadien de télédétection

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ujrs20

A Comparative Study of Deep Learning Approaches to Rooftop Detection in Aerial Images

Yuwei Cai, Hongjie He, Ke Yang, Sarah Narges Fatholahi, Lingfei Ma, Linlin Xu & Jonathan Li

To cite this article: Yuwei Cai, Hongjie He, Ke Yang, Sarah Narges Fatholahi, Lingfei Ma, Linlin Xu & Jonathan Li (2021): A Comparative Study of Deep Learning Approaches to Rooftop Detection in Aerial Images, Canadian Journal of Remote Sensing, DOI: 10.1080/07038992.2021.1915756

To link to this article: https://doi.org/10.1080/07038992.2021.1915756

CANADIAN JOURNAL OF REMOTE SENSING
https://doi.org/10.1080/07038992.2021.1915756

A Comparative Study of Deep Learning Approaches to Rooftop Detection in Aerial Images

Une étude comparative des approches d'apprentissage en profondeur pour la détection des toits dans les images aériennes

Yuwei Cai[a], Hongjie He[a], Ke Yang[b], Sarah Narges Fatholahi[a], Lingfei Ma[c], Linlin Xu[b], and Jonathan Li[a,b]

[a]Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, N2L 3G1, Canada; [b]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada; [c]Engineering Research Center of State Financial Security, Ministry of Education, Central University of Finance and Economics, Beijing, 102206, China

ABSTRACT

This paper investigates the deep neural networks for rapid and accurate detection of building rooftops in aerial orthoimages. The networks were trained using the manually labeled rooftop vector data digitized on aerial orthoimagery covering the Kitchener-Waterloo area. The performance of the three deep learning methods, U-Net, Fully Convolutional Network (FCN), and Deeplabv3+ were compared by training, validation, and testing sets in the dataset. Our results demonstrated that DeepLabv3+ achieved 63.8% in Intersection over Union (IoU), 77.8% in mean IoU (mIoU), 74% in precision, and 78% in F_1-score. After improving the performance with focal loss, training loss was greatly cut down and the convergence rate experienced a significant growth. Meanwhile, rooftop detection also achieved higher performance, as Deeplabv3+ reached 93.6% in average pixel accuracy, with 65.4% in IoU, 79.0% in mIoU, 77.6% in precision, and 79.1% in F_1-score. Lastly, in order to evaluate the effects of data volume, by changing data volume from 100% to 75% and 50% in ablation study, it shows that when data volume decreased, the performance of extraction also got worse, with IoU, mIoU, precision, and F_1-score also mostly decreased.

RÉSUMÉ

Cet article étudie les réseaux de neurones profonds pour une détection rapide et précise des toits de bâtiments dans les orthoimages aériennes. Les réseaux ont été formés à l'aide des données vectorielles sur les toits étiquetées manuellement et numérisées sur une orthoimagerie aérienne couvrant la région de Kitchener-Waterloo. Avec les ensembles de formation, de validation et de test dans l'ensemble de données, les performances des trois méthodes d'apprentissage en profondeur, U-Net, Réseau entièrement convolutionnel (FCN) et Deeplabv3 + ont été comparées. Nos résultats ont démontré que DeepLabv3 + a atteint 63,8% en Intersection over Union (IoU), 77,8% en moyenne IoU (mIoU), 74% en précision et 78% en F1-score. La perte focale a été appliquée pour évaluer et améliorer les performances de détection sur les toits. Celle-ci a considérablement réduit la perte d'entraînement et le taux de convergence a connu une croissance significative. Pendant ce temps, la détection sur les toits a également obtenu des performances plus élevées, car Deeplabv3 + an atteint 93,6% en précision de pixel moyenne, avec 65,4% en IoU, 79,0% en mIoU, 77,6% en précision et 79,1% en score F1. Enfin, afin d'évaluer les effets du volume de données, en faisant passer le volume de données de 100% à 75%, 50% dans l'étude d'ablation, il montre que lorsque le volume de données a diminué, les performances d'extraction ont également empiré, avec IoU, mIoU, la précision et le score F1 ont également diminué pour la plupart.

ARTICLE HISTORY
Received 31 October 2020
Accepted 7 April 2021

Introduction

As urban areas have been developing and expanding rapidly in recent years, increasing efforts have been gradually allocated to identify the relevant location information of the buildings (Chen, Papandreou, et al. 2018). For instance, governments should take the responsibility of rational administrating of the city, including city planning, state cadastral inspection, and

CONTACT Jonathan Li junli@uwaterloo.ca

infrastructure development (Cote and Saeedi 2013). In order to achieve these goals, updated information related to rooftops or footprints of buildings should be specified. However, timely mapping and updating of rooftops in urban areas remains a very challenging task.

Aerial images can obtain all buildings images at one shot. High spatial resolution (HSR) aerial images have been the first choice in accurately localizing the buildings within one certain area (Yang et al. 2018). Hence, rooftop extraction from HSR aerial images plays a significant role in urban applications such as urban land use and land cover mapping, population estimation, change detection, urban flood management, and other urban planning issues (Alshehhi et al. 2017; Boogaard et al. 2017; Ha 2017; Shao et al. 2016). HSR aerial images have been used as a major data source for rooftop extraction. However, compared to traditional visual interpretation, advanced methods using machine learning have proved more effective for automated extraction of rooftops from HSR aerial images.

Meanwhile, deep learning methods have demonstrated their high performance in image segmentation in the field of computer vision. Unfortunately, the insufficient extraction and low segmentation accuracy with HSR aerial images always restrict automated extraction of rooftops (Chen, Papandreou, et al. 2018). Both accuracy and efficiency are the main concerns for developing an effective method to extract rooftops from HSR aerial images. Therefore, we aim at testing and comparing several deep learning methods for rooftop extraction in this study to identify potential solutions. The contribution of this paper is threefold: first, it evaluates our new building rooftop dataset that was generated using HSR aerial images with a spatial resolution of 12 cm. Second, it studies three deep learning algorithms, including Fully Convolutional Network (FCN) (Long et al. 2015), U-Net (Ronneberger et al. 2015), and DeepLabv3+ (Chen, Zhu, et al. 2018), and compares them with respect to their accuracy and efficiency in rooftop extraction. Third, it assesses the performance of the three deep learning algorithms by applying focal loss to image segmentation followed by a comparison between focal loss and binary cross entropy. The performance of rooftop extraction with high spatial resolution images can be enhanced by focal loss (Yun et al. 2019).

Related work

Related work in building dataset

Currently, there are several open datasets for building detection with high spatial resolution satellite/aerial images. To compare the classification methods over large areas, Mnih (2013) created building and road classification datasets over Massachusetts, which covered 340 km^2 and 2600 km^2, respectively. An aerial image labeling dataset was constructed by researchers from Inria (Institut national de recherche en informatique et en automatique) in France (Maggiori et al. 2017). Covering 810 km^2 (405 km^2 for training and 405 km^2 for testing), the dataset contains aerial orthorectified color imagery with 30 cm spatial resolution. In this dataset, there are two semantic classes including building and non-building. Ji et al. (2019) constructed an aerial and satellite imagery dataset of building samples covering 450 km^2. It consists of over 220,000 independent buildings, extracted from aerial images with 7.5 cm spatial resolution in Christchurch, New Zealand.

Related work in building rooftop extraction methods

Throughout the process of completing building or rooftop extraction, many approaches have been proposed. These methods start from using machine learning techniques in individual-pixel classification (Mnih 2013), and then turn into higher-level information integration such as shaping features (Maggiori et al. 2015). In details, traditional methods of design features can represent the buildings by extracting them from satellite or aerial images. Color (Sirmacek and Unsalan 2008), spectrum (Zhang 1999; Zhong et al. 2008), length, edge (Ferraioli 2010; Li and Wu 2008), shape (Dunaeva and Kornilov 2017), texture (Awrangjeb et al. 2011; Guo et al. 2016; Tiwari and Pande 2008), shadow (Sirmacek and Unsalan 2008; Chen et al. 2014), height, and semantic (Zhong et al. 2015) have been commonly used to extract buildings from satellite or aerial images. However, these metrics can be influenced by changing the factors such as atmospheric conditions, weather, light, surroundings, density of the buildings, and some other relevant conditions around the buildings in the research area. The initial idea of utilizing designed features can only extract the specific buildings or rooftops in specific research area with specific data (Ji et al. 2019). This does not complete the manual-free process which inspires researchers to find and apply more automatic rooftop segmentation technology for completing the extraction task.

In recent years, deep learning methods have started to gain wide attention, and many networks related to deep learning, especially CNNs, have been used for image segmentation (Maggiori et al. 2017; Maggiori

et al. 2017; Mnih 2013; Volpi and Tuia 2017). CNNs are not only applied for object detection, but also for semantic segmentation to deal with progress fine inference (Ladický et al. 2009). Although the convolutional networks have been employed for years, researchers are still struggling to find new methods to implement the classification tasks in a faster and more accurate way (Chen, Papandreou, et al. 2018). Great progress has been made in semantic segmentation to classify image pixels; however, the challenge is that each pixel need to be labeled with one class (Lu et al. 2019).

To improve the accuracy of semantic segmentation on the whole image, Long et al. (2015) applied FCNs for transferring pre-trained classifier weights to fuze various layer representations to train end-to-end and pixel-to-pixel. The fully connected layers of CNNs were then replaced by convolutional and de-convolutional layers (Pan et al. 2019) to construct the pixel-based encoder–decoder architectures (Long et al. 2015). The operations of end-to-end and pixel-to-pixel both simplify and accelerate the learning speed and reasoning. After FCNs, increased CNN architectures were proposed to improve the image segmentation performance (Pan et al. 2019). In 2015, U-Net was proposed to modify and extend the architecture of FCNs by replacing the pooling operators with upsampling operators, and concatenation of feature maps of encoder and decoder (Pan et al. 2019). U-Net is a training strategy that performs better than others in biomedical segmentation (Ronneberger et al. 2015). With concatenation architecture and by taking adequate use of both low-level features and high-level features, Ronneberger et al. (2015) obtained more accurate segmentation results. Consequently, U-Net has been proved as a first-class convolutional network with high efficiency. After that, DeepLabv1 (Chen et al. 2014), DeepLabv2 (Chen, Zhu, et al. 2018), and DeepLabv3 (Chen, Papandreou, et al. 2018) were proposed to alleviate the information loss due to the pooling operations (Pan et al. 2019). The atrous convolution which is also called dilated convolution (Russakovsky et al. 2015) was proposed to both increase receptive field size and protect high resolution of the feature maps (Giusti et al. 2013; Holschneider et al. 1990; Papandreou et al. 2015; Sermanet et al. 2014; Pan et al. 2019). By updating the atrous convolutions with upsampled filters, DeepLabv3 can simultaneously extract the dense feature maps and capture long range context to achieve the best performance. Based on the architecture of DeepLabv3, in 2018, DeepLabv3+ was developed after adding depth wise separable convolution in atrous convolution and encoder-decoder architecture (Badrinarayanan et al. 2017; Ronneberger et al. 2015). The structure of encoder-decoder in DeepLabv3+ can identify sharp object boundaries for semantic segmentation. Besides, DeepLabv3+ also employed a new Deep Convolutional Neural Network (DCNN) to its architecture (Chen, Papandreou, et al. 2018). According to previous studies, U-Net and Deeplabv3+ always achieve better performance than FCNs (Hui et al. 2019; Ji et al. 2019; Li et al. 2018; Pan et al. 2019).

Related work in class imbalances and focal loss

One-stage object detector and two-stage detector are two methods usually used in object detection. When using one-stage detectors, there are $10^4 \sim 10^5$ locations in one image, but just a few locations are included in objects that need to be detected. This phenomenon usually leads to two problems: one is training efficiency and the other is overwhelmed training process and degenerate models (Lin et al. 2014). However, data-level and algorithm-level methods can be chosen to alleviate these problems (Dong et al. 2019; Zhao et al. 2019). In data-level methods, the imbalance problem can be alleviated by re-sampling the training data (Kong et al. 2017), while in algorithm-level methods, the algorithmic behavior can be modified by making changes in weight for each class, especially minority classes (Zhao et al. 2019). Lin et al. (2014) chose focal loss as an algorithm-level method to improve the performance of object detection. By stressing sparse hard examples and preventing considerable easy negatives by overwhelming the detection methods during training, Lin et al. (2014) concluded that class imbalances can be dealt with as the focal loss in one-stage object detector and its training accuracy cannot be lower than results with two-stage object detectors. Therefore, to improve the performances, focal loss have been presented to take place of the binary cross entropy (BCE) loss.

Nie et al. (2019) applied focal loss in U-Net and detected that it can improve the performance of prostate's base and apex parts with medical image segmentation. Focal loss improved the Dice Similarity Coefficient (DSC) of U-Net for more than 2% in medical image segmentation (Nie et al. 2019). Ma et al. (2020) applied focal loss in U-Net for road segmentation. Compared with binary cross entropy, focal loss raised the IoU of experimental results by 3% (Ma et al. 2020). In most of the recent studies, focal loss

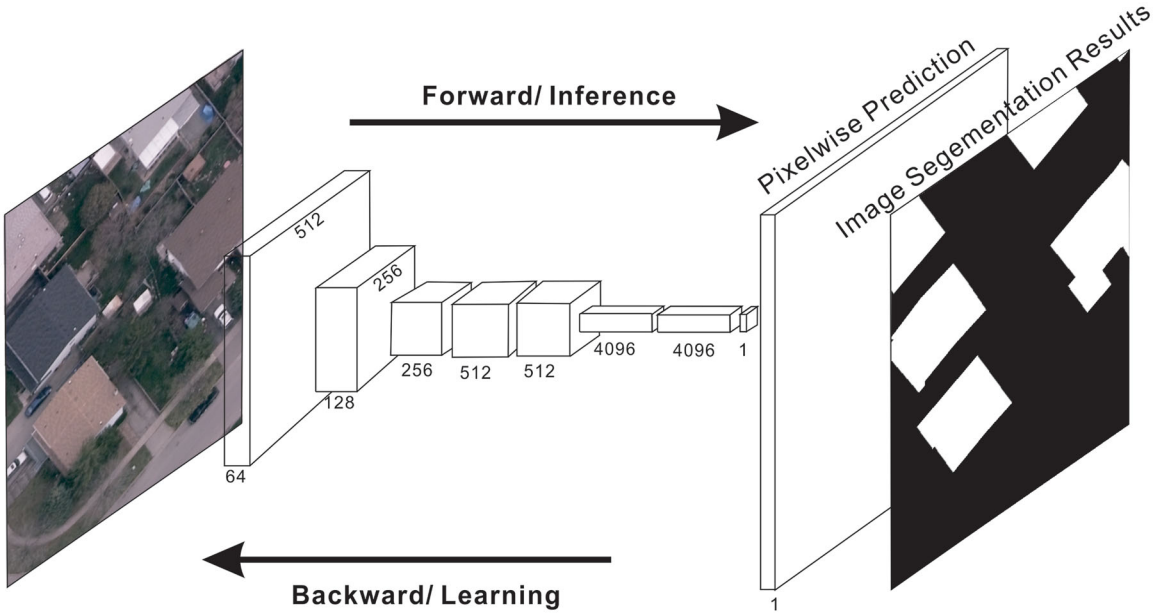

Figure 1. FCN architecture modified by Long et al. (2015).

has always enhanced the performance of image segmentation, especially by increasing IoU and mIoU (Doi and Iwasaki 2018; Ma et al. 2020; Nie et al. 2019; Ye et al. 2019; Zhao et al. 2019).

Deep learning methods

FCN in building segmentation

FCN can be accurately trained to make dense predictions, and can train semantic segmentation in an end-to-end, pixel-to-pixel manner (Long et al. 2015). Semantic segmentation is a process that matches each object's pixel with a label. It first downsizes the input image step by step, and then up-samples output size at the very end to make it similar to the input. It also combines the output from shallow and deep layers together to enhance the accuracy of results. Figure 1 shows the modified architecture of FCN.

Contrary to convolution, which is a process of decreasing output size, FCNs go through a deconvolution process to expand the size of output. For the sake of making output and input images equal in size, up-sampling occurs before getting the output label map. When going to the deeper layers, deep features are discovered, however, location information can be lost which makes the output image rough. To solve this problem, FCNs combine the feature hierarchies by fuzing multiple layers (Abramson et al. 2006). Based on fixed bottom-up grouping, this fuzing operation combines features through layers to construct a nonlinear local-to-global representation, and to produce more accurate prediction results (Shelhamer et al. 2017). In this paper, FCN-8s were used to detect building rooftops in high spatial resolution aerial images.

U-Net in building segmentation

U-Net is developed based on FCNs. According to Ronneberger et al. (2015), U-Net performs better than others in biomedical segmentation. Following the typical convolutional network architecture, U-Net network's architecture contains both contracting and expansive paths. The contracting path is a down-sampling step which adds the number of feature channels twice before the step. By completing two consecutive 3×3 convolution (black arrows) and 2×2 max pooling (red arrows), more features can be extracted, and the size of features can be decreased in the down-sampling process. In contrast, expansive path is an up-sampling step which changes the number of feature channels into half. In order to recover the size of segmentation map, two consecutive 3×3 Conv (black arrows) and 2×2 Up-conv (green arrows) is added to the U-Net architecture. Critically, they add more feature channels after applying unpadded convolutions to rectify linear unit (ReLU), and max pooling operation in network architecture. Although up-sampling step obtains more advanced features, it also experiences location information loss.

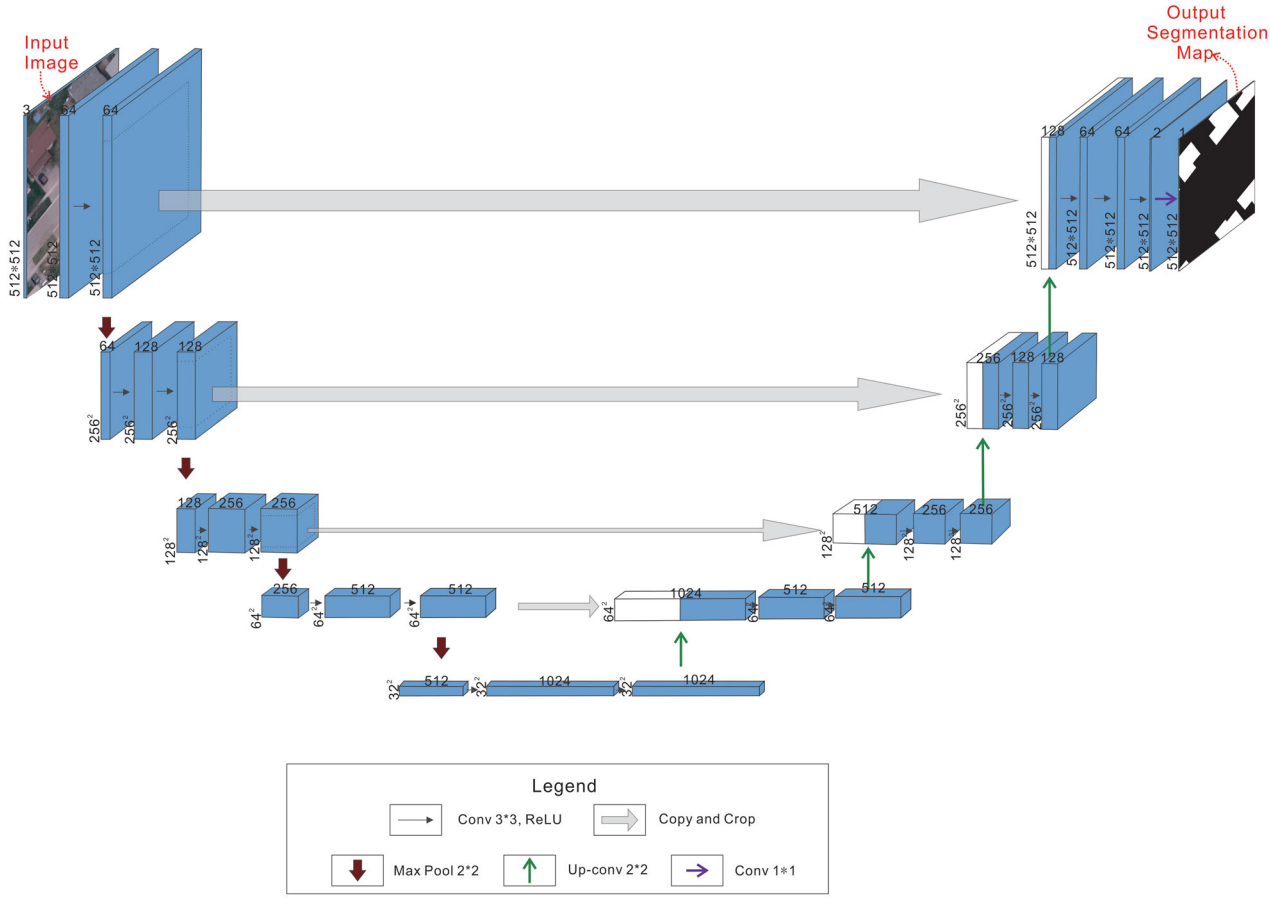

Figure 2. U-Net architecture modified from Ronneberger et al. (2015).

The concatenation of feature maps provides the expansion path with more location information from the contraction path (grey arrows). At the end, 1×1 convolutional filter (purple arrows) changes the size of feature map from 2 to 1 due to the only two classes, which are the cell and membrane of the output feature map. Figure 2 represents the architecture of U-Net network proposed by Ronneberger et al. (2015).

DeepLabv3+ in building segmentation

DeepLabv3 contains three special parts in the proposed module including atrous convolution, multi-grid methods, and Atrous Spatial Pyramid Pooling (ASPP). When rate = 1, it is the standard convolution. By adjusting rate to 6, 12, and 18, the filter's field-of-view was modified. In atrous convolution, the relevant parameters can be calculated by Equation 1 as follow:

$$y[i] = \sum_k x[i + r * k]\omega[k] \quad (1)$$

where i is the location; y is the output of each location i; r is the atrous rate and ω is the filter. Atrous convolution was applied to the input feature map x, where the atrous rate r corresponds to the stride, with which the input signal was sampled. This process also means that by inserting r-1 'zeros' between two consecutive filter values along each spatial dimension, upsampled filters were convolved in the input feature map x.

When going deeper with atrous convolution, according to the multi-grid methods, different atrous rates were applied within block 4 to block 7 in the architecture of DeepLabv3. However, because of the state-of-art neural networks (Chollet 2017; Krizhevsky et al. 2012; Ladický et al. 2009; Simonyan and Zisserman 2015; Szegedy et al. 2015) and restricted GPU memory, as the resolution of output feature maps is sometimes 1/8 or 1/4 times of the input images, it is difficult to extract the output features. The output features from each branch are then concatenated and passed through another 1×1 convolution, which has 256 filters and Batch Normalization(BN), before the final 1×1 convolution, which generates the final logits (Chen, Zhu, et al. 2018) (Figure 3).

Compared with DeepLabv3, DeepLabv3+ adds depth-wise separable convolution in atrous convolution encoder-decoder architecture, which can recover the location information. Modified Aligned Xception, and

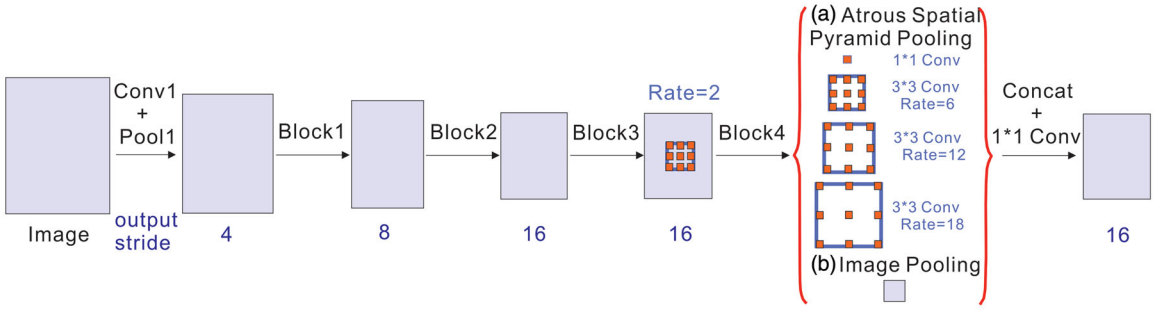

Figure 3. DeepLabv3 Architecture, modified from Chen, Zhu, et al. (2018).

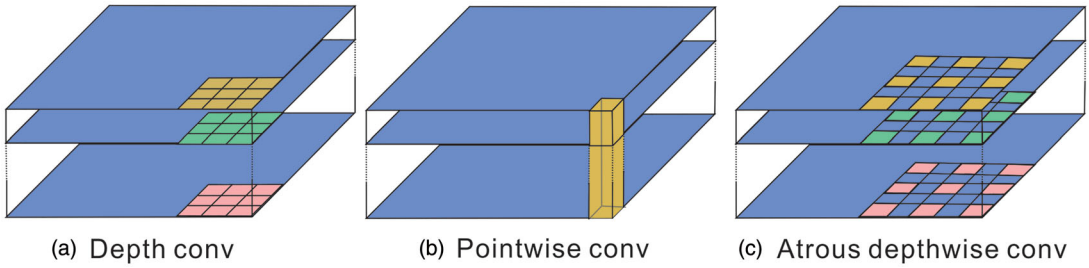

Figure 4. 3 * 3 Depthwise Separable Convolution, modified from Chen, Papandreou, et al. (2018).

Atrous Separable Convolution are developed to promote a faster and stronger network. By decomposing a standard convolution into a depth-wise convolution (Figure 4a), and a point-wise convolution (Figure 4b), Atrous Separable Convolution has drastically contracted computation complexity. Furthermore, in order to achieve similar or even better performance, and decrease the computation complexity at the same time, atrous convolution which also exists in DeepLabv3, is used to support depth-wise convolution in DeepLabv3+ (Chen, Zhu, et al. 2018) (Figure 4c).

The encoder process is similar in DeepLabv3 and DeepLabv3+. When extracting various classes in the images, the output stride is 16, as the spatial resolution of feature maps of the final output is usually 1/16 of input image spatial resolution. Since that spatial resolution is not enough for completing semantic segmentation, the striding in the last few blocks was replaced by atrous convolution to fulfill output stride of 16 or 8. In the encoder part in DeepLabv3 and DeepLabv3+, atrous spatial pyramid pooling module is expanded so as to have access to multi-scale convolutional features (Chen, Zhu, et al. 2018). While in the decoder part, 1×1 convolution low-level features were concatenated with the upsampled output features of the encoder part. After concatenation, several 3×3 convolutions then were applied to refine those concatenation features with another upsampling. The ratio of 4 is adopted in each upsampling layer to decode the feature maps and generate pixel-wise labeling results (Chen, Zhu, et al. 2018) (Figure 5).

Focal loss

Binary cross entropy (BCE) is widely used in many deep learning-based images segmentation methods such as FCN-8s, U-Net, and DeepLabv3+. BCE can be calculated as follow:

$$\text{Cross Entropy} = -\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{P} d_{nkwh} log p \qquad (2)$$

where N, K, and P are the mini-batch size, number of categories, and the number of pixels, respectively. p is the model's estimated probability for the class with certain label, d is a one-hot vector and $d_k = 1$ when k is a true classification.

Researchers have shown that BCE loss cannot avoid the curse from class imbalance in dataset, which impede the further improvement of segmentation accuracy. He et al. (2016) first proposed focal loss, which has been used in many deep learning models, especially in object detection tasks. Focal loss can be calculated as Equation (3). Users have shown its power to leverage the problem from class imbalances with many experiments (Doi and Iwasaki 2018).

$$\text{Focal Loss} = -\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{P} d_{nkwh}(1-p)^{\gamma} log p \qquad (3)$$

where N, K, and P are the mini-batch size, number of categories, and number of pixels, respectively; d is a one-hot vector; and $d_k = 1$, when k is a true classification. The term $(1-p)^{\gamma}$ is a modulating factor, which controls cross entropy loss. When $\gamma = 0$, focal

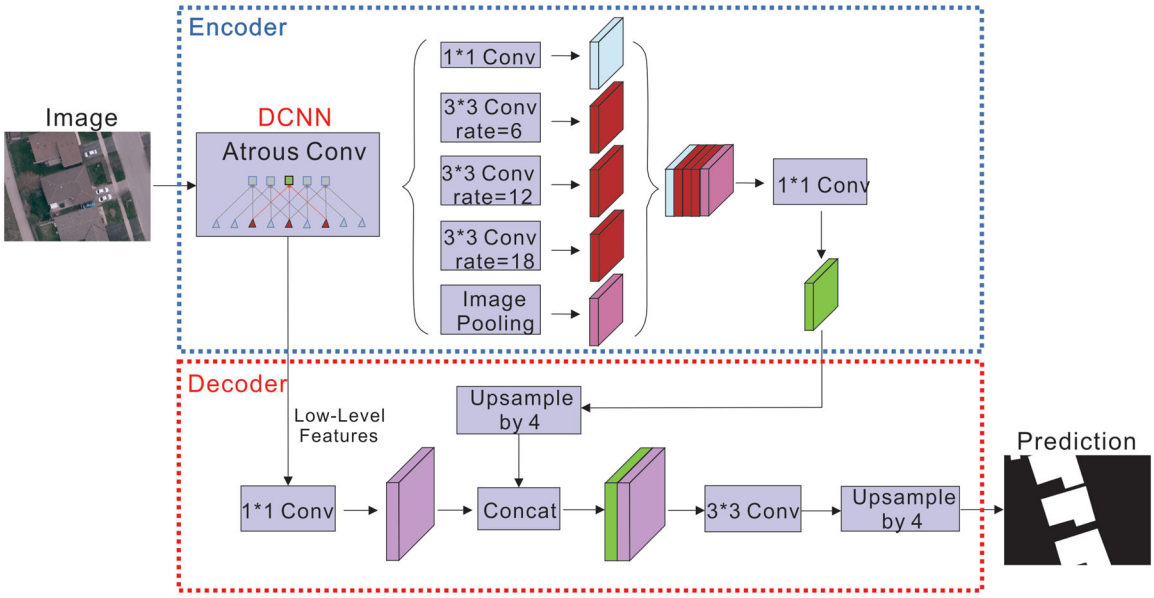

Figure 5. DeepLabv3+ Architecture, modified from Chen, Zhu, et al. (2018).

loss is equivalent to cross entropy loss. When γ increases, the differences between loss value of easy and hard classes also increase.

In this research, class imbalances mean the number of pixels (or the area) of non-rooftops which take up much more than that of rooftops in the aerial images. Since class imbalances also exist in our dataset, the performance of focal loss was tested to achieve a better segmentation result.

Metrics

In order to compare those methods mentioned at the beginning, we incorporate several commonly used metrics to evaluate their accuracy and efficiency.

Accuracy assessment

We use average accuracy, IoU, mIoU, precision, recall, and F1-score (Bischke et al. 2019; Chen et al. 2019; Khalel and El-Saban 2018; Maggiori et al. 2017; Tan et al. 2020) as the most used metrics in image segmentation to evaluate and compare deep learning based image segmentation methods mentioned above. The average accuracy can be expressed as follows:

$$\text{Average accuracy } (\%) : \text{PA} = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij}} \tag{4}$$

where k is the number of categories for classification; i, j represent different classification; p_{ii} indicate pixels that are correctly classified in the image; and p_{ij} denote pixels that are wrongly classified in the image.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{5}$$

$$\text{mIoU} = \frac{\left[\frac{TP}{TP+FP+FN} + \frac{TN}{TN+FP+FN}\right]}{2} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1} = \frac{2TP}{2TP + \text{FP} + FN} \tag{9}$$

where TP is true positive that is the overlap between target mask and the prediction output; FP is false positive and is an error in output while in the test result mistakenly indicates presence of a condition; FN is false negative and is an error while in the test result mistakenly identifies it as no presence of a condition; TN is true negative and is an example in which the model predicted the negative class correctly in the output.

Efficiency evaluation

Test Rate: Test rate is the image that has been tested in 1 s during the testing process. The testing rate can be calculated as follow:

$$\text{TR} = \text{TI}/\text{t} \tag{10}$$

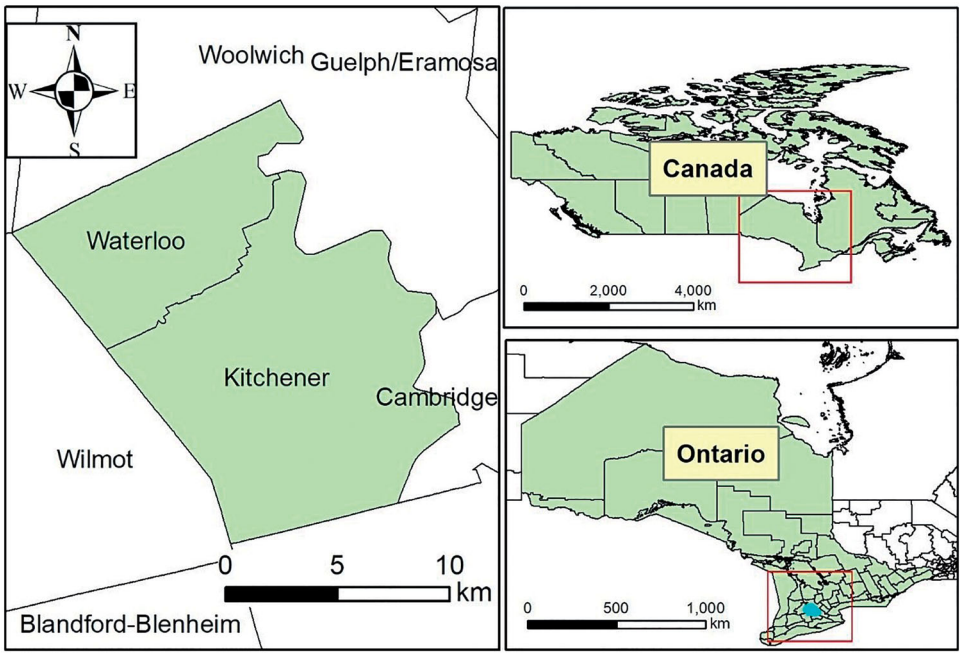

Figure 6. Dataset covering the Kitchener-Waterloo area (the KW area).

where TI is the number of images that has been test, and t is the total time cost for testing. FPS has similar meaning. Thus, we use FPS in the paper to denote the testing rate.

Dataset

Aerial imagery

As for the area covered by the aerial imagery dataset, Waterloo is a city lying in the south-east of Ontario, Canada. It is the smallest among three cities in the Regional Municipality of Waterloo, and the City of Kitchener is adjacent to it. As shown in Figure 6[1], Kitchener and Waterloo are often jointly referred to the "Kitchener–Waterloo (KW) area".

The dataset area incorporates rural and urban areas including residential, industrial, commercial, and education-functional buildings. These buildings have various shapes, size, colors, heights, and numbers covering a land area of 205.83 km^2. According to census data from Statistics Canada, the population of KW area increased from 302,143 in 2006 to 317,933 in 2011, and 338,208 in 2016 (Statistics Canada 2017). It means that an increasing number of buildings are required in the area, which motivated us to choose this area as the dataset region, and 2014 as the time of the most recent dataset.

[1]Main map is projected into Universal Transverse Mercator (UTM), Zone 17N coordinates and horizontal datum is the North American Datum of 1983 (NAD 83). Two sub maps are provided with their original Geographic Coordinate System WGS 1984.

The HSR aerial orthoimages mainly containing the Kitchener-Waterloo area were provided by the Geospatial Center of the University of Waterloo. The Geospatial Center at the University of Waterloo scanned the images at 600 dpi and used ArcMap 9.2 for georeferencing process. Using paper indexes, they found the geographical location of the images and tagged each image with latitude and longitude coordinates. Street intersections, river bends and unique terrain patterns on farmland were the main and common used features for orienting the images. These aerial images covering the Region of Waterloo were acquired using an UltraCam D camera in 2014 and were then calibrated and orthorectified. With the spatial resolution of 12 cm, such orthoimages can be used to recognize and delineate the boundary of the rooftops. These geo-referenced tiles are available in the Universal Transverse Mercator (UTM), Zone 17 N coordinates and horizontal datum is the North American Datum of 1983 (NAD 83).

Dataset labeling

There are usually two requirements for an image dataset that can be used to effectively train a high-performance deep learning model: (1) a large quantity of data covering a large geographic area, and (2) HSR images to see rooftops clearly (Chen, Papandreou, et al. 2018).

The 12 cm resolution aerial orthoimages cover a land area of 205.83 km^2 in the KW area with almost 150,000 individual buildings. All building rooftops

Figure 7. Manually edited polygons examples of labeled results of rooftops in the dataset of the KW area (blue lines in the first row: boundary of the rooftops, blue shaded areas in second row: labeled rooftops).

were manually labeled based on building footprint data from Statistics Canada in ArcGIS shapefiles. However, most of those building footprints were not accurate enough, and some buildings had been omitted from the shapefiles. Therefore, the roof-by-roof manual correction was conducted for those missing rooftop labels. All the rooftops in the KW area were checked and modified with ArcGIS10.6 in order to guarantee the high accurate annotations of rooftops. Moreover, multiple persons worked on checking the annotations at least three times to control the labeling error within 3 pixels to minimize the manual mistakes. Given that some buildings were partly shaded or even covered by trees, their rooftop shapes were manually delineated. Through all these processes, all polygons of labeled rooftops were created, revised, and validated within the dataset of the KW area (Figure 7).

By utilizing the 'polygon to raster' tool in the ArcToolbox, the edited polygons can be successfully transferred into raster images. Subsequently, the raster results of the images can be obtained by the raster calculator in map algebra (to fill the null value in converted results), and 'copy raster' tool (to compress data) in ArcGIS software (Figure 8).

As the data were labeled into two classes of roof and non-roof, the annotated images were then cut from the resolution of 8350×8350 pixels into 512×512 pixels. No data area in the boundary of each image were padded with 127. No data area in the boundary of each labeled image were also padded with 0. The purpose of cutting the images into 512×512 pixels is that, not only the efficiency of calculation can be improved, but also the sample size can be augmented. In that way, the generalization performance of the model can be improved. Finally, these images were separated into three sets: training, validation, and testing sets.

Experiments

Dataset

In order to construct a new dataset and test the accuracy of labeling work, building rooftops on HSR aerial images were manually labeled based on rooftop shapefiles from Statistics Canada. Then, the labeled images were transferred into binary images. By slicing the binary images and dividing them into three sets, training set, validation set and testing set, the dataset was constructed. In detail, around 1/6 data from the dataset was extracted to test the accuracy of the labeling work.

Configurations of deep learning models

Similar configurations were set in all three deep learning models. Learning rate was set to 10^{-4} and the Adam optimizer was used. Batch size was set to 5 and epochs set to 100.

For each deep learning model, we use two different loss function to show the superiority of focal loss. We also used different training dataset volume (100%, 75%, 50%) to test the quality of our dataset. As a result, we have $3 * 2 * 3$ experiments in this work. As mentioned in section 3.5, we evaluated those models using pixel accuracy, IoU, mIoU, precision, recall, F_1 score and testing rate. In this work, all models were trained on a GPU of GeForce RTX 2080ti, a CPU of Intel(R) Core (TM) i9-9900X and CUDA 10.2.

Figure 8. RGB color aerial orthoimages (top row) and their corresponding reference data (bottom row: labeled building rooftops).

Subsequently, in order to achieve the best performance and to obtain apparent comparison results between the binary cross entropy and the focal loss, the value of α and γ were set to be 0.25 and 2.0, respectively. As these two settings were tested to help achieve the best performance compared with other settings (Doi and Iwasaki 2018; Lin et al. 2014; Ye et al. 2019; Yun et al. 2019; Zhao et al. 2019).

Results and discussion

Results

First, almost one-sixth data from the dataset were chosen to test the accuracy of the labeling work. All chosen data were cut into 10,404 images which contains 6,069 labeled images for training set, 1,156 labeled images for validation set, and 3,179 labeled images for testing set. In the training process, FCN and U-Net have no non-trainable parameters, and only DeepLabv3+ has non-trainable parameters. Meanwhile, FCN owns the highest number of total parameters among the three, while the U-Net has the lowest (Table 1).

As for the training and validation accuracy of the three deep learning algorithms, the training and validation accuracy of FCN has always been a little bit lower than that of U-Net and DeepLabv3+. When the epoch is approaching 100, the training and validation accuracy of FCN approaches 99.2%, while the training and validation accuracy of U-Net and DeepLabv3+ approaches 99.8%.

As shown in Figure 9a, the training and validation accuracy of U-Net and DeepLabv3+ is always higher than that of FCN. When epoch is approaching 100, the training and validation accuracy of the three deep learning methods all approach 1. Turning to the loss function of the three deep learning algorithms, the loss in the training and validation process with the FCN has always been higher than that of U-Net, and DeepLabv3+. When epoch is approaching 100, the loss of FCN reaches 2.05% in the training and validation process, while the loss of U-Net, and DeepLabv3+ approaches almost 0.58% on average in the training and validation process. Finally, the accuracy of U-Net and DeepLabv3+ is much lower than the FCN in the training and validation process (Figure 9b).

Table 1. Parameters of three deep neural network.

Deep neural networks	Trainable parameters	Non-trainable parameters	Total parameters
FCN	49,789,326	0	49,789,326
U-Net	31,032,837	0	31,032,837
Deeplabv3+	41,050,273	202800	41,253,073

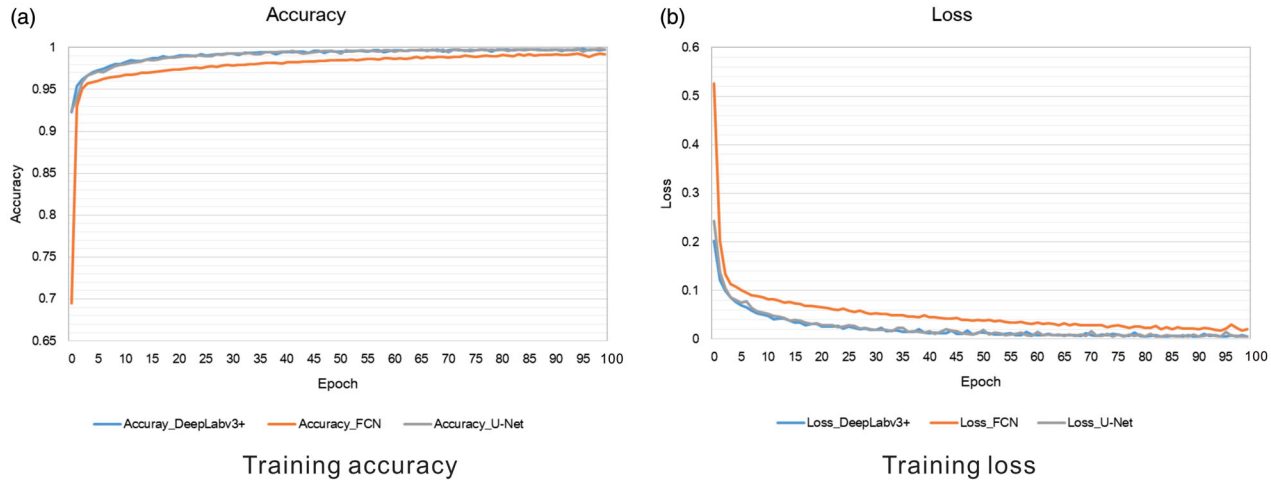

Figure 9. Training accuracy and loss of three deep learning algorithms.

With output images, the predicted results are relatively more accurate in results with the U-Net and DeepLabv3+ than results with the FCN. Most of the predicted polygons accurately covered the polygons of the ground truth in the results of U-Net and DeepLabv3+, and the wrong classified results are much less in U-Net and DeepLabv3+ than those in U-Net (Figure 10).

In the close-ups, the performance of the results with U-Net and DeepLabv3+ are better than that of FCN. In the output segmentation results, mistakes mostly appear in classifying the boundary of the rooftops and roads (Figure 11).

From the results, although DeepLabv3+ owns the most complex architecture, and the lowest training and testing rates, it shows the highest accuracy in the rooftop segmentation with high spatial resolution images. U-Net has almost similar average accuracy as the DeepLabv3+, but it works much slower than the DeepLabv3+. As for FCN, the training and testing rate is high, however, the training and validation accuracy, and the average accuracy are all the lowest, and the loss in the training and validation process is the highest. mIoU is the average of IoUs for building masks (foreground) and background. Due to imbalanced classification, mIoU appears to be higher than IoU for all three algorithms. Precision and recall, represent correctness and completeness (Shu 2014). The percentage of correct-classified building masks to all predicted building masks equals correctness, and the percentage of correct-classified building masks to all ground truth building masks equals completeness. Therefore, these metrics can reveal the performance of the algorithms when processing comparison. According to the metrics, IoU, mIoU, precision, recall, and F_1-score in the DeepLabv3+ show the best performance compared with the other two algorithms. Therefore, compared with FCN-8s and U-Net, DeepLabv3+ should be the first choice in rooftop segmentation with high spatial resolution images (Table 2).

After focal loss was applied to all three deep learning algorithms, Deeplabv3+, U-Net, and FCN-8s, great differences can be detected between focal loss and binary cross entropy in the training accuracy and loss (Figure 12). When focal loss is applied to deep learning methods, the training loss is greatly reduced, while the training accuracy is just slightly affected. The training loss was especially minimized in Deeplabv3+, which was almost reduced by 0.1 points. Besides, the application of focal loss extremely accelerates the converging rate of training loss curve when epoch increases.

By changing the training data volume by using original data to using 75% and 50% of the original data to complete the training, the influence from the training data volume has been analyzed. The ablation study shows that when training data size increases, the training accuracy rises, while the training loss decreases. This phenomenon is obviously detected in all three deep learning methods (Figure 12).

We also calculated the relevant metrics to analyze the performance of deep learning algorithms. Firstly,

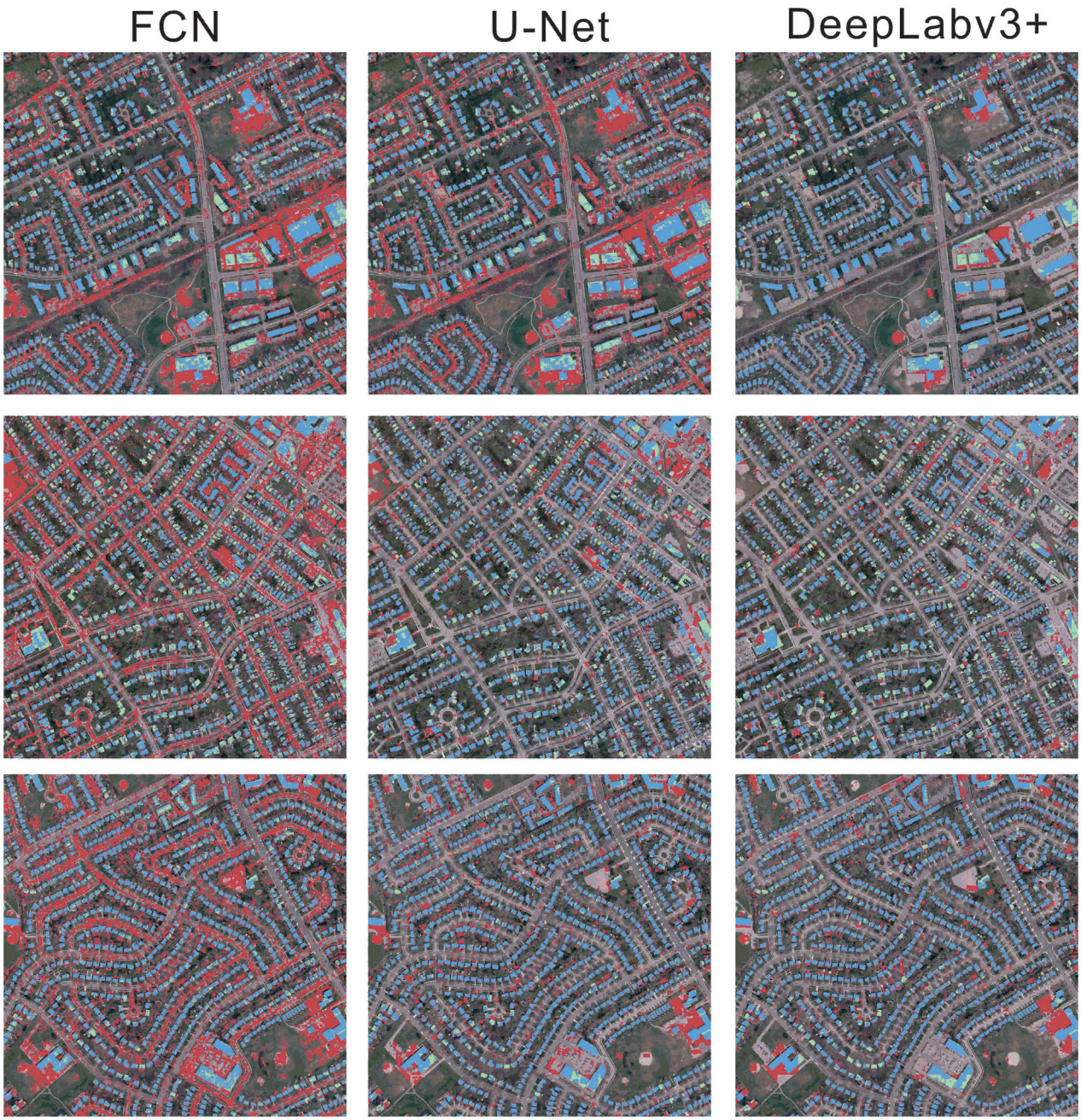

Figure 10. Examples of segmentation results using FCN, U-Net, and DeepLabv3+, respectively (blue: predicted, green: ground truth, red: wrongly classified).

considering the data volumes are equal, Deeplabv3+ keeps showing the best performance after applying the focal loss, reaching 93.6% in average accuracy, and all highest performance in IoU, mIoU, precision, and F_1-score. It means that using the same data size and loss function, Deeplabv3+ can still preform the best among others. Moreover, according to the results, when data size increases, the performance usually becomes greater, except for U-Net. When applying focal loss in U-Net, 75% of training data achieves the best performance in terms of all accuracy metrics. The reason behind the performance is unexpected and we will explore this in detail in our future work.

Lastly, focal loss always enhances the performance of the algorithms. When data volume is constant, the deep learning methods can achieve higher performance with assistance of the focal loss. For example, with 100% data volume, the average accuracy of Deeplabv3+ trained with focal loss achieves 93.6%, higher than original Deeplabv3+ with BCE by 0.7 points, while IoU and mIoU of Deeplabv3+ trained with focal loss achieves 65.4%, 79.0%, respectively,

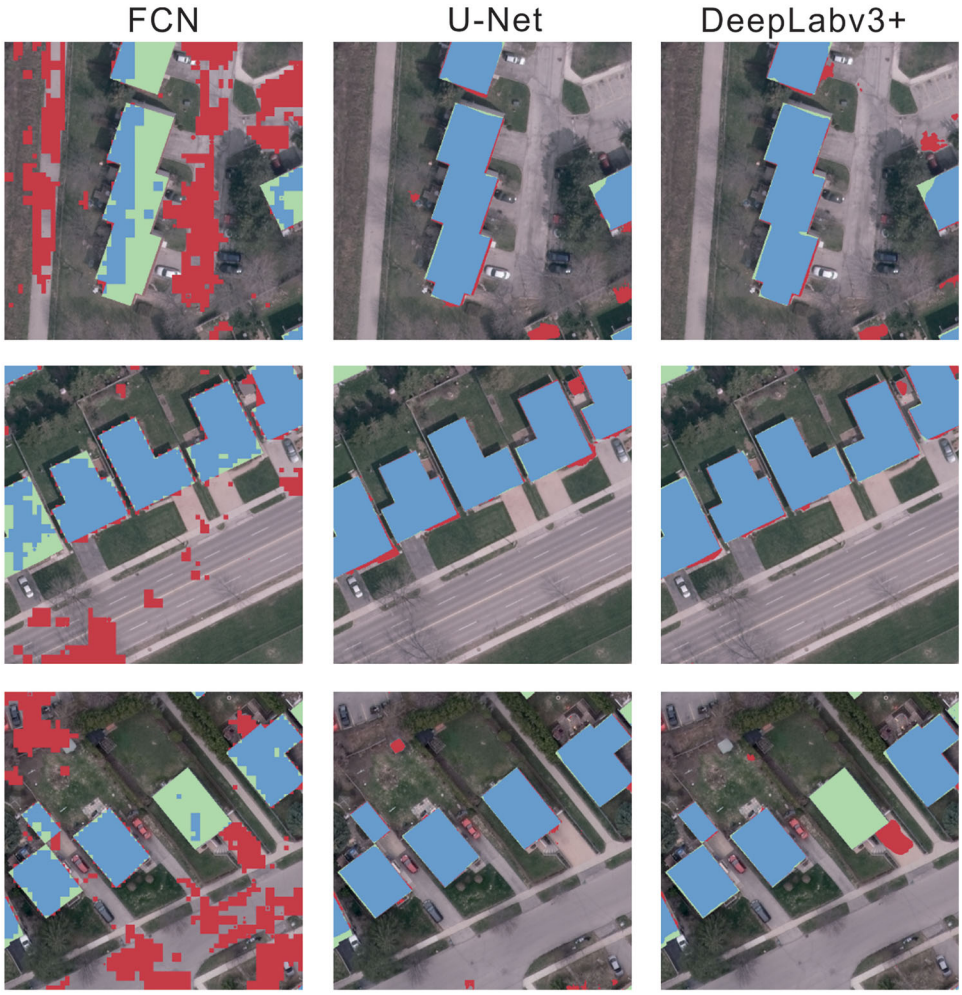

Figure 11. Close-ups of the segmentation results by using FCN, U-Net and DeepLabv3+, respectively (blue: predicted, green: ground truth, red: wrongly classified).

Table 2. Evaluation of the performance of three methods (%).

	FCN-8s	U-Net	DeepLabv3+
Average Accuracy	78.3	91.9	**92.9**
Testing Rate (FPS)*	**19.4**	14.5	16.9
IoU	35.2	61.1	**63.8**
mIoU	55.3	75.9	**77.8**
Precision	39.2	69.5	**73.7**
Recall	78.0	**83.5**	82.5
F1-score	52.1	75.8	**77.9**

Note: FPS: frames per second. The bolded figures refer to the highest value in one evaluation factor.

higher than the results of Deeplabv3+ with BCE by almost 2 points (Table 3).

Discussion

The reason for inaccurate output results has been discussed. Although the average accuracy of the three algorithms is high, many obvious mis-extractions of the rooftops can be still found in the output results. It is quite difficult for the algorithms to identify the exact location features of the rooftops when part of the rooftop is covered by tree or its shadows. In this case, inaccurate rooftop segmentation can decrease the average accuracy. Also, it is obvious that the output results of FCN do not have great quality as the U-Net and the Deeplabv3+. The contours of rooftops in the FCN output results are zigzag, while the contours of rooftops in the output results of U-Net and Deeplabv3+ are smoothed curves or polygonal lines. The reason for this may come from the differences in the expansive path in the U-Net and Deeplabv3+. Additionally, when trees shade the buildings, all experimental results of the three methods extract inaccurate shading rooftops, especially in the results of FCN. Although all the shading parts of the rooftops are not extracted, the shading does have some impacts on the accuracy of the outputs (Figure 13).

Moreover, there are many wrong classified results in the parking lots. The reason may come from the large-scale impervious surfaces that can be easily misclassified into rooftops. Both the rooftops and parking

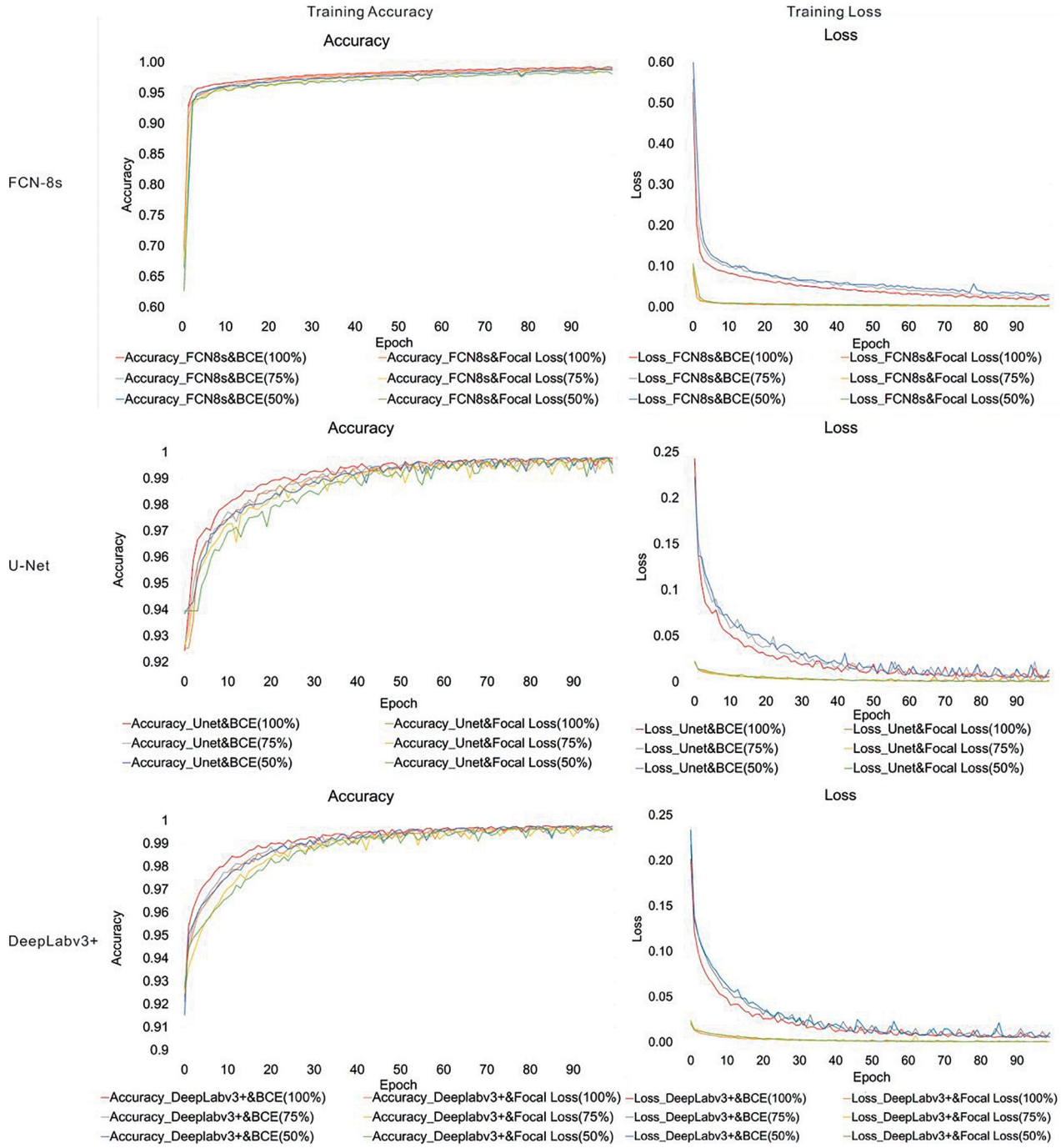

Figure 12. Application of focal loss to three deep learning algorithms (The percentages behind focal loss are percentages of training data to total data in the experiments). Examples of outputs with tree shading rooftops and non-shading rooftops.

lots are impervious surfaces with similar color and other characteristics, and when extracting rooftops in the aerial images, the parking lots can sometimes be misidentified as rooftops. According to output images in Figure 14, it is obvious that the proportion of wrongly classified parking lots decreases from the results of FCN to U-Net, and then to DeepLabv3+. It can be predicted that deepLabv3+ achieves the high performance in producing the output results, because of its accurate extraction in identifying the parking lots. In fact, by not extracting parking lots as rooftops, DeepLabv3+ can achieve high accuracy and high value in all assessment parameters (Figure 14).

Therefore, deep learning methods that can accomplish the rooftop extraction with rooftops even shading by trees need further study. Besides, correctly classifying parking lots and rooftops with the deep learning methods may also be one key point in rapid extraction from aerial images. The method with excellent performance can effectively complete rooftop

Table 3. Evaluation of the performance of focal loss in three algorithms (%).

		Average accuracy	IoU	mIoU	Precision	Recall	F_1-score	Testing rate (FPS)
FCN-8s BCE loss	100% *Data*	78.3	35.2	55.3	39.2	78.0	52.1	19.4
	75% *Data*	75.4	31.3	51.8	35.2	74.3	47.7	**20.5**
	50% *Data*	64.3	25.4	42.4	27.1	80.4	40.5	20.4
FCN-8s focal loss	100% *Data*	84.7	42.4	62.6	49.6	74.6	59.6	20.4
	75% *Data*	82.8	37.3	59.1	45.5	67.5	54.3	20.4
	50% *Data*	73.9	29.4	50.1	33.3	71.8	45.5	20.3
U-Net BCE loss	100% *Data*	91.9	61.1	75.9	69.5	**83.5**	75.8	14.5
	75% *Data*	89.3	53.9	70.8	60.8	82.6	70.0	14.7
	50% *Data*	78.9	38.4	57.0	40.7	86.9	55.5	14.9
U-Net focal loss	100% *Data*	91.9	59.0	74.9	71.9	76.6	74.2	14.9
	75% *Data*	92.1	60.2	75.7	72.3	78.3	75.2	14.8
	50% *Data*	90.2	55.0	71.9	64.1	79.5	71.0	15.0
DeepLabv3+ BCE loss	100% *Data*	92.9	63.8	77.8	73.7	82.5	77.9	16.9
	75% *Data*	92.7	62.3	77.0	74.0	79.7	76.8	16.9
	50% *Data*	91.1	51.6	70.9	74.5	62.7	68.1	17.5
DeepLabv3 + focal loss	100% *Data*	**93.6**	**65.4**	**79.0**	**77.6**	80.6	**79.1**	17.5
	75% *Data*	92.7	54.7	73.3	89.1	58.7	90.7	17.6
	50% *Data*	92.0	54.7	72.9	78.8	64.1	70.7	17.4

**Note*: FPS: frames per second. *Data* here refers to training and validation dataset used in training stage. The bolded figures refer to the highest value in one evaluation factor.

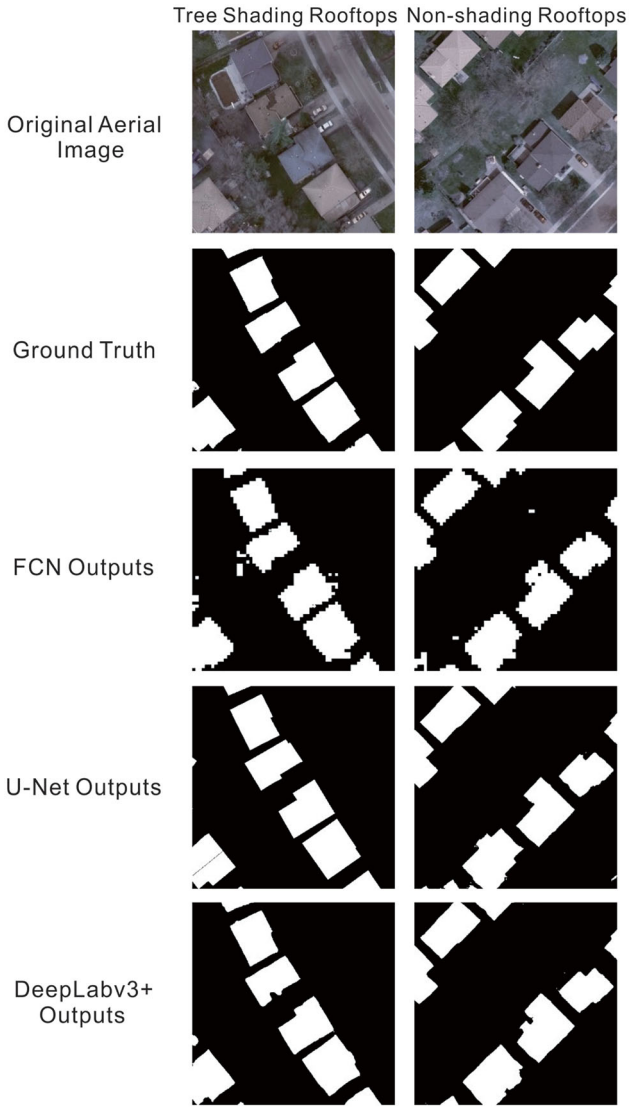

Figure 13. Examples of outputs with tree shading rooftops and non-shading rooftops

segmentation in one region, which can assist the government in urban practical applications such as building construction characteristics, population estimation, change detection, urban flooding prevention, and other urban planning issues.

Lastly, when using average accuracy, IoU, mIoU, precision, recall and F_1-score to evaluate the performance of rooftop extraction, the value of recall fluctuated. When average accuracy, IoU, mIoU, precision and F_1-score all increased, the value of recall irregularly increased or decreased. Therefore, the value of recall may not be a reasonable metric for assessing the performance of rooftop extraction when using deep learning methods. According to Equation (8), reasons leading to weird value in recall may come from high value in FN, which means in the test result there are many spots of buildings in the images mistakenly identified as non-building. That implied while extraction accuracy increased from FCN, to U-Net and to Deeplabv3+, TP increased while FN decreased.

Conclusions

In this paper, a deep learning approach was proposed to complete the rapid extraction of building rooftops from aerial imagery. Firstly, the building rooftops on HSR aerial images were manually labeled in the Kitchener-Waterloo area, and the high-quality dataset was constructed for building rooftop extraction. Almost one-sixth of the data from the dataset proved the labeling work's accuracy. The accuracy approached nearly 100%, and the loss approached to 0 when epoch approached 100.

Secondly, the performance of the three common deep learning algorithms including FCN, U-Net, and

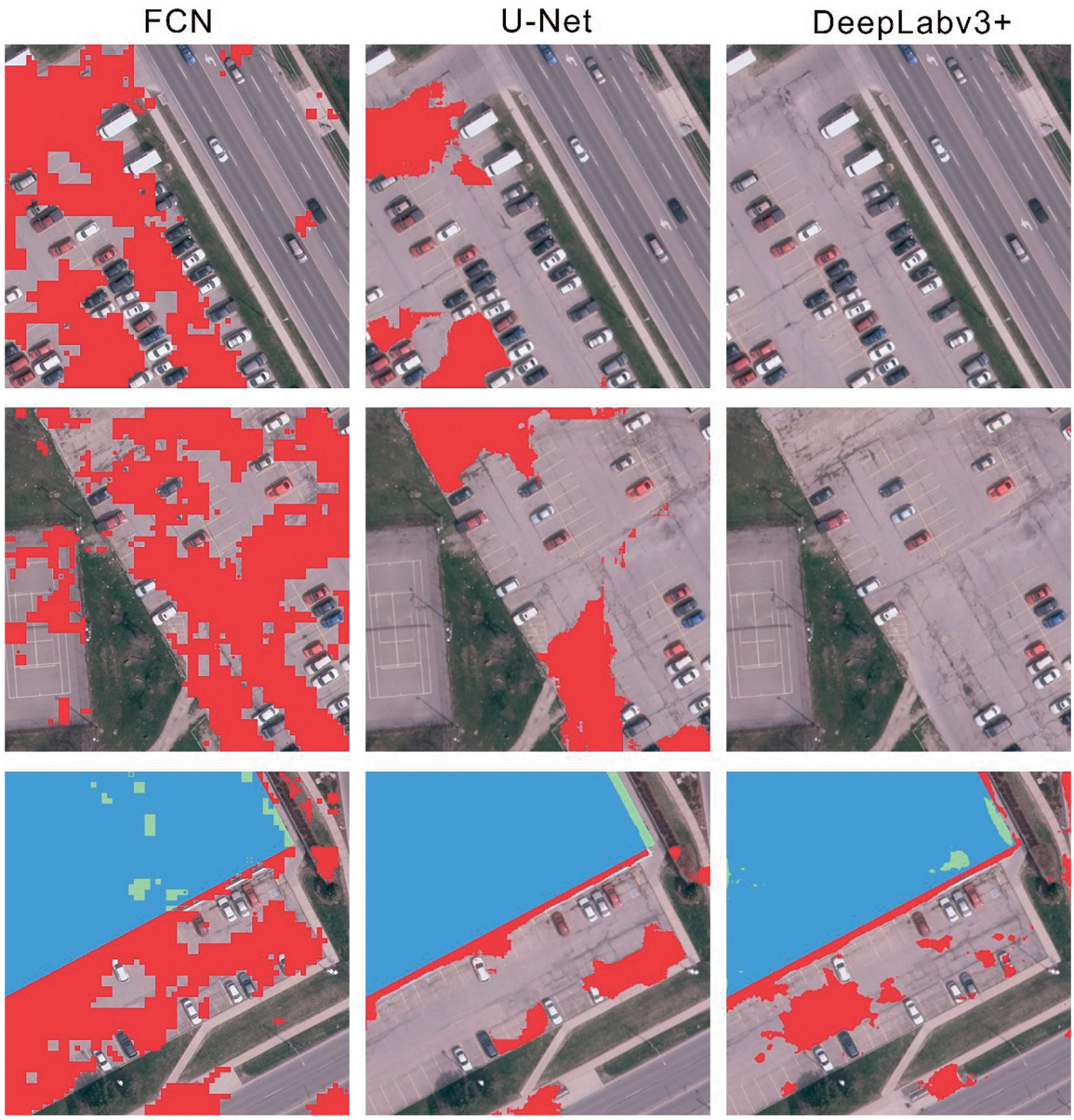

Figure 14. Close-ups of parking lots wrongly classified into rooftops in segmentation results by using FCN, U-Net and DeepLabv3+, respectively (blue: predicted, green: ground truth, red: wrongly classified).

Deeplabv3+ were tested and compared. The results indicated that the DeepLabv3+ owns the highest accuracy, while the FCN has the lowest. As for the U-Net, although it owns similar average accuracy as the DeepLabv3+, its working rate is much lower than that of the DeepLabv3+. The higher accuracy is the result of sophisticated deeper architecture design among different methods. The higher efficiency of DeepLabv3+ comparison to U-Net also come from the architecture design. If rooftops are extremely considerable in one dataset, or the dataset's scale is extensive, then the U-Net can be a more time-consuming choice than the DeepLabv3+. After comparing the performance of the algorithms, it was found that DeepLabv3+ achieved the greatest performance, with 63.8% IoU, 77.8% mIoU, 74% precision, and 78% F_1-score. As a result, the accuracy of labeled work meets expectations among the three selected deep learning methods, and that DeepLabv3+ achieves the highest accuracy in building rooftop extraction from HSR aerial orthoimages.

Lastly, focal loss can deal with the class imbalance problem, and enhance the performance of deep learning methods in extracting rooftops with high resolution aerial images. After applying the focal loss to algorithms, the training loss can be greatly cut down, and reducing rate can undergo a significant growth. Meanwhile, the performance of rooftop extraction can become better, as the average accuracy, IoU, mIoU, precision, and F_1-score greatly increase after applying

focal loss. In the ablation study, when data volume decreases, the parameters such as IoU, mIoU, precision, and F_1-score mostly decline, and the performance of extraction gets worse.

ORCID

Jonathan Li http://orcid.org/0000-0001-7899-0049

References

Abramson, N., Braverman, D.J., and Sebestyen, G.S. 2006. "Pattern recognition and machine learning." *Publications of the American Statistical Association*, Vol. 103(No. 4): pp. 886–887.

Alshehhi, R., Marpu, P.R., Woon, W.L., and Dalla Mura, M. 2017. "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 130: pp. 139–149. doi:10.1016/j.isprsjprs.2017.05.002.

Awrangjeb, M., Zhang, C., and Fraser, C.S. 2011. "Improved building detection using texture information." *ISPRS Archives*, Vol. 38: pp. 143–148.

Badrinarayanan, V., Kendall, A., and Cipolla, R. 2017. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39(No. 12): pp. 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Bischke, B., Helber, P., Folz, J., Borth, D., and Dengel, A. 2019. "Multi-task learning for segmentation of building footprints with deep neural networks". In Proceedings of 2019 IEEE International Conference on Image Processing (ICIP), pp. 1480–1484.

Boogaard, F., Vojinovic, Z., Chen, Y. C., Kluck, J., and Lin, T. P. 2017. "High resolution decision maps for urban planning: A combined analysis of urban flooding and thermal stress potential in Asia and Europe." *Proceedings of MATEC Web of Conferences*, Vol. 103: pp. 04012.

Chen, D., Shang, S., and Wu, C. 2014. "Shadow-based building detection and segmentation in high-resolution remote sensing image." *Journal of Multimedia*, Vol. 9(No. 1): pp. 181–188. doi:10.4304/jmm.9.1.181-188.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. 2018. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40(No. 4): pp. 834–848. doi:10.1109/TPAMI.2017.2699184.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., and., and Adam, H. 2018. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of ECCV*, pp. 801–818.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., and Waslander, S.L. 2019. "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 147: pp. 42–55. doi:10.1016/j.isprsjprs.2018.11.011.

Chollet, F. 2017. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of CVPR*, pp. 1251–1258.

Cote, M., and Saeedi, P. 2013. "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51(No. 1): pp. 313–328. doi:10.1109/TGRS.2012.2200689.

Doi, K., and Iwasaki, A. 2018. "The effect of focal loss in semantic segmentation of high resolution aerial image." In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 6919–6922. doi:10.1109/IGARSS.2018.8519409.

Dong, Q., Gong, S., and Zhu, X. 2019. "Imbalanced deep learning by minority class incremental rectification." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41(No. 6): pp. 1367–1381. doi:10.1109/TPAMI.2018.2832629.

Dunaeva, A.V.E., and Kornilov, F.A. 2017. "Specific shape building detection from aerial imagery in infrared range." *Computer Science, Engineering and Control*, Vol. 6(No. 3): pp. 84–100.

Ferraioli, G. 2010. "Multichannel InSAR building edge detection." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48(No. 3): pp. 1224–1231. doi:10.1109/TGRS.2009.2029338.

Giusti, A., Cireşan, D.C., Masci, J., Gambardella, L.M., and Schmidhuber, J. 2013. "Fast image scanning with deep max-pooling convolutional neural networks." *Proceedings of IEEE International Conference on Image Processing*, pp. 4034–4038.

Guo, Z., Shao, X., Xu, Y., Miyazaki, H., Ohira, W., and Shibasaki, R. 2016. "Identification of village building via Google Earth images and supervised machine learning methods." *Remote Sensing*, Vol. 8(No. 4): pp. 271. doi:10.3390/rs8040271.

Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. 1990. "A real-time algorithm for signal analysis with the help of the wavelet transform." In *Wavelets Chapter: Inverse problems and theoretical imaging*, edited by J.M. Combes, A. Grossmann, and P. Tchamitchian, pp. 286–297. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-75988-8_28.

Ji, S., Wei, S., and Lu, M. 2019. "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57(No. 1): pp. 574–586. doi:10.1109/TGRS.2018.2858817.

He, K., Zhang, X., Ren, S., and Sun, J. 2016. "Deep residual learning for image recognition." *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Hui, J., Du, M., Ye, X., Qin, Q., and Sui, J. 2019. "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 5): pp. 786–790. doi:10.1109/LGRS.2018.2880986.

Khalel, A., and El-Saban, M. 2018. "Automatic pixel-wise object labeling for aerial imagery using stacked U-Nets." *arXiv Preprint arXiv*

Kong, D., Tang, J., Zhu, Z., Cheng, J., and Zhao, Y. 2017. "De-biased dart ensemble model for personalized

recommendation." In Proceedings of 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 553–558.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. 2012. "Imagenet classification with deep convolutional neural networks." *Proceedings of NeurIPS*, pp. 1097–1105.

Ladický, L.U., Russell, C., Kohli, P., and Torr, P.H. 2009. "Associative hierarchical CRFs for object class image segmentation." *Proceedings of ICCV*, pp. 739–746.

Li, X., Yao, X., and Fang, Y. 2018. "Building-a-nets: robust building extraction from high-resolution remote sensing images with adversarial networks." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 11(No. 10): pp. 3680–3687. doi:10.1109/JSTARS.2018.2865187.

Li, Y., and., and Wu, H. 2008. "Adaptive building edge detection by combining LiDAR data and aerial images." *ISPRS Archives*, Vol. 37: pp. 197–202.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., and Zitnick, C.L. 2014. "Microsoft coco: Common objects in context." In *Proceedings of ECCV*, pp. 740–755.

Long, J., Shelhamer, E., and Darrell, T. 2015. "Fully convolutional networks for semantic segmentation." *Proceedings of CVPR*, pp. 3431–3440.

Lu, Y., Chen, Y., Zhao, D., and Chen, J. 2019. "Graph-FCN for image semantic segmentation." In *Proceedings of ISNN*, pp. 97–105.

Ma, J., Xu, Z., Zheng, E., and Fan, Q. 2020. "Accurate road segmentation in remote sensing images using dense residual learning and improved focal loss." *Journal of Physics: Conference Series*, Vol. 1544: pp. 012101. doi:10.1088/1742-6596/1544/1/012101.

Maggiori, E., Tarabalka, Y., and Charpiat, G. 2015. "Optimizing partition trees for multi-object segmentation with shape prior." *Proceedings of BMVC*, Vol. 64: pp. 1–12.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. 2017. "Convolutional neural networks for large-scale remote-sensing image classification." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55(No. 2): pp. 645–657. doi:10.1109/TGRS.2016.2612821.

Maggiori, E., Tarabalka, Y., Charpiat, G., and., and Alliez, P. 2017. "Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark." *Proceedings of IGARSS*, pp. 3226–3229.

Mnih, V. 2013. *Machine learning for aerial image labeling*. PhD Thesis. University of Toronto. http://www.cs.toronto.edu/~vmnih/docs/Mnih_Volodymyr_PhD_Thesis.pdf

Nie, D., Wang, L., Xiang, L., Zhou, S., Adeli, E., and Shen, D. 2019. "Difficulty-aware attention network with confidence learning for medical image segmentation." *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33: pp. 1085–1092. doi:10.1609/aaai.v33i01.33011085.

Pan, X., Yang, F., Gao, L., Chen, Z., Zhang, B., Fan, H., and Ren, J. 2019. "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms." *Remote Sensing*, Vol. 11(No. 8): pp. 917. doi:10.3390/rs11080917.

Papandreou, G., Kokkinos, I., and Savalle, P.A. 2015. "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection." *Proceedings of CVPR*, pp. 390–399.

Ronneberger, O., Fischer, P., and Brox, T. 2015. "U-Net: Convolutional networks for biomedical image segmentation." In Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. doi:10.1007/978-3-662-54345-0_3.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. 2015. "ImageNet large scale visual recognition challenge." *International Journal of Computer Vision*, Vol. 115(No. 3): pp. 211–252. doi:10.1007/s11263-015-0816-y.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. 2014. "OverFeat: Integrated recognition, localization and detection using convolutional networks." In 2nd International Conference on Learning Representations, ICLR 2014.

Shao, Z., Zhang, Y., and Zhou, W. 2016. "Long-term monitoring of the urban impervious surface mapping using time series Landsat imagery: A 23-year case study of the city of Wuhan in China." In *Proceedings of 4th International Workshop on Earth Observation and Remote Sensing Applications EORSA*, pp. 212–216.

Shelhamer, E., Long, J., and Darrell, T. 2017. "Fully convolutional networks for semantic segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39(No. 4): pp. 640–651. doi:10.1109/TPAMI.2016.2572683.

Shu, Y. 2014. *Deep convolutional neural networks for object extraction from high spatial resolution remotely sensed imagery*. Ph.D. Thesis. Canada: University of Waterloo,.

Simonyan, K., and Zisserman, A. 2015. "Very deep convolutional networks for large-scale image recognition." In Proceedings of 2015 International Conference on Learning Representations (ICLR), pp. 1–14. arXiv: 1409.1556.

Sirmacek, B., and Unsalan, C. 2008. "Building detection from aerial images using invariant color features and shadow information." In Proceedings of the 23rd International Symposium on Computer and Information Sciences, pp. 1–5. doi:10.1109/iscis.2008.4717854.

Statistics Canada. 2017. Waterloo, CY [Census subdivision], Ontario and Kitchener – Cambridge –Waterloo [Census metropolitan area], Ontario, table. Census Profile. 2016 Census. Data published on Statistics Canada website by Canadian government.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., and Rabinovich, A. 2015. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 1–9. doi:10.1109/cvpr.2015.7298594.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., and Cai, G. 2020. "Toronto-3D: a large-scale mobile LiDAR dataset for semantic segmentation of urban roadways." *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 797–806. doi:10.1109/cvprw50498.2020.00109.

Tiwari, P.S., and Pande, H. 2008. "Use of laser range and height texture cues for building identification." *Journal of the Indian Society of Remote Sensing*, Vol. 36(No. 3): pp. 227–234. doi:10.1007/s12524-008-0023-1.

Volpi, M., and Tuia, D. 2017. "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55(No. 2): pp. 881–893. doi:10.1109/TGRS.2016.2616585.

Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., and Bhaduri, B. 2018. "Building extraction at scale using convolutional neural network: mapping of the United States." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 11(No. 8): pp. 2600–2614. doi:10.1109/JSTARS.2018.2835377.

Ye, C., Wang, W., Zhang, S., and Wang, K. 2019. "Multi-depth fusion network for whole-heart at image segmentation." *IEEE Access.*, Vol. 7: pp. 23421–23429. doi:10.1109/ACCESS.2019.2899635.

Yun, P., Tai, L., Wang, Y., Liu, C., and Liu, M. 2019. "Focal loss in 3D object detection." *IEEE Robotics and Automation Letters*, Vol. 4(No. 2): pp. 1263–1270. doi:10.1109/LRA.2019.2894858.

Zhang, Y. 1999. "Optimization of building detection in satellite images by combining multispectral classification and texture filtering." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 54(No. 1): pp. 50–60. doi:10.1016/S0924-2716(98)00027-6.

Zhao, Y., Lin, F., Liu, S., Hu, Z., Li, H., and Bai, Y. 2019. "Constrained-focal-loss based deep learning for segmentation of spores." *IEEE Access.*, Vol. 7: pp. 165029–165038. doi:10.1109/ACCESS.2019.2953085.

Zhong, C., Xu, Q., Yang, F., and Hu, L. 2015. "Building change detection for high-resolution remotely sensed images based on a semantic dependency." *Proceedings of IGARSS*, pp. 3345–3348. doi:10.1109/igarss.2015.7326535.

Zhong, S. H., Huang, J. J., and Xie, W. X. 2008. "A new method of building detection from a single aerial photograph." In Proceedings of the 9th International Conference on Signal Processing, pp. 1219–1222. doi:10.1109/icosp.2008.4697350.