

Waterloo building dataset: a city-scale vector building dataset for mapping building footprints using aerial orthoimagery¹

Hongjie He, Zijian Jiang, Kyle Gao, Sarah Narges Fatholahi, Weikai Tan, Bingxu Hu, Hongzhang Xu, Michael A. Chapman, and Jonathan Li

Abstract: Automated building footprint extraction is an important area of research in remote sensing with numerous civil and environmental applications. In recent years, deep learning methods have far surpassed classical algorithms when trained on appropriate datasets. In this paper, we present the Waterloo building dataset for building footprint extraction from very high spatial resolution aerial orthoimagery. Our dataset covers the Kitchener–Waterloo area in Ontario, Canada, contains 117 000 manually labelled buildings, and extends over an area of 205.8 km². At a spatial resolution of 12 cm, it is the highest-resolution publicly available building footprint extraction dataset in North America. We provide extensive benchmarks for commonly used deep learning architectures trained on our dataset, which can be used as a baseline for future models. We also identified a key challenge in aerial orthoimagery building footprint extraction, which we hope can be addressed in future research.

Key words: building footprint, urban mapping, aerial orthoimagery, building dataset, deep learning.

Résumé : L'extraction automatisée de l'empreinte d'un bâtiment est un domaine de recherche important dans la télédétection avec de nombreuses applications civiles et environnementales. Au cours des dernières années, les méthodes d'apprentissage profond, lorsqu'elles sont concentrées sur les ensembles de données appropriés, ont largement surpassé les algorithmes classiques. Dans la présente communication, nous présentons l'ensemble de données des bâtiments de Waterloo pour l'extraction des empreintes des bâtiments de l'ortho-imagerie à résolution spatiale très élevée. Nos ensembles de données couvrent le secteur de Kitchener–Waterloo en Ontario, au Canada, contiennent 117 000 bâtiments étiquetés manuellement et s'étendent sur un secteur de 205.8 km². À une résolution spatiale de 12 cm, ce sont les ensembles de données de l'extraction des empreintes des bâtiments à la résolution la plus élevée disponible pour le public en Amérique du Nord. Nous offrons des repères approfondis d'architectures d'apprentissage profond fréquemment utilisées et concentrés sur nos ensembles de données qui peuvent être utilisés comme base pour de futurs modèles. Nous identifions également un problème majeur dans l'extraction des empreintes des bâtiments de l'ortho-imagerie

Received 14 July 2021. Accepted 3 December 2021.

H. He, Z. Jiang, K. Gao, S. Narges Fatholahi, W. Tan, B. Hu, and H. Xu. Geospatial Sensing and Data Intelligence Lab, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

M.A. Chapman. Department of Civil Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada.

J. Li.* Geospatial Sensing and Data Intelligence Lab, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada; Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, 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 Eric Guilbert.

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

© 2022 The Author(s). Permission for reuse (free in most cases) can be obtained from [copyright.com](https://www.copyright.com).

aérienne qui, nous l'espérons, peut être réglé dans une recherche future. [Traduit par la Rédaction]

Mots-clés : empreinte d'un bâtiment, cartographie urbaine, ortho-imagerie aérienne, ensemble de données d'un bâtiment, apprentissage profond.

1. Introduction

As a key element in urban areas, buildings are an important indicator for urban change detection (Chen et al. 2021). Building rooftops or footprints (outlines along the exterior walls of buildings; there are few differences between them in orthoimages) are also essential for other urban applications such as urban planning and management, cadastral management, urban geo-database updates, and smart city construction (Rastogi et al. 2020). In addition to these exclusive urban applications, building datasets are essential for population estimation, natural hazards, and damage estimation. By combining building footprints with other building information, such as the number of stories, population and population densities can be estimated efficiently, which is necessary for epidemic or pandemic control (Xie et al. 2015). Furthermore, building maps are of paramount importance for natural hazard management and damage estimation (Thomas et al. 2013; Shu 2014). Accurate building map data are required to estimate earthquake damage and assess risks (Sahar et al. 2010). In addition, to assess the loss from typhoons, floods, and other geological disasters, building maps should be obtained effectively and efficiently post-disaster (Wen et al. 2019). For these applications, remote sensing, especially aerial image-based methods, has achieved significant results.

Automated extraction of building rooftops from aerial orthoimages is a challenging task in remote sensing. Conventional pixel- and object-based image analysis methods are often ineffective because they require expertise in feature engineering or feature collection. The development of deep learning techniques has revolutionized automated building rooftop extraction (Chen et al. 2021). However, deep learning techniques are known to be data intensive. A large number of high-quality pixel-level labelled images are required for the development of new algorithms.

In this study, we constructed the Waterloo building dataset, which covers the Kitchener–Waterloo area in Ontario, Canada. The main contributions of this study are two-fold. First, we released a city-scale vector building dataset, with the building footprints manually labelled on the 12 cm resolution aerial orthoimagery. To the best of our knowledge, this is the first open-access city-scale dataset with such a high spatial resolution in North America. Second, an extensive comparative study was performed to benchmark existing deep learning methods, which can be used to benchmark future methods trained on our dataset.

2. Related work

2.1. Existing datasets

With the development of imaging technologies, an increasing number of high-spatial-resolution satellite and aerial images are being released. Several open-access building datasets based on these images have been developed in recent years.

- The ISPRS Vaihingen and Potsdam datasets (Rottensteiner et al. 2013) are two relatively small building datasets. In the datasets, six classes, including building and background, were provided. The Vaihingen part has 33 image patches with a spatial resolution of 9 cm; spectral bands of red, green, blue, and near-infrared; and a size of approximately 2500×2500 pixels.

The Potsdam part has 38 image patches with a spatial resolution of 5 cm, the same spectral resolution as the Vaihingen part, and a size of approximately 6000×6000 pixels. The corresponding digital surface model (DSM) data are also provided with the image data for each part. These datasets have the highest spatial resolution among the existing datasets, while they cover only a 5 km^2 area.

- The Massachusetts building dataset (Mnih 2013) classifies images only into building and non-building. It contains 151 aerial images with a spatial resolution of 1 m; spectral bands of red, green, and blue; and 1500×1500 pixels, covering approximately 340 km^2 of the Boston area. A total of 151 aerial orthoimages were further divided into training, validation, and testing sets of 137, 4, and 10 images, respectively. As illustrated by the author, the dataset possesses high accuracy, with less than 5% average omission of building classification.
- The Inria dataset (Maggiori et al. 2017) also includes building and non-building classes. It contains aerial images covering 810 km^2 of 10 cities in the United States and Austria. The training and testing sets captured 360 images with a spatial resolution of 30 cm and spectral bands of red, green, and blue. The dataset aims to explore the generalizability of CNNs; therefore, adjacent images are split into training and testing sets.
- The WHU (Wuhan University) building dataset (Ji et al. 2018) includes an aerial image dataset and a satellite image dataset for building extraction. The aerial image dataset has 8189 tiles with a spatial resolution of 30 cm; spectral bands of red, green, and blue; and a size of 512×512 pixels. The dataset was manually edited and converted from aerial images covering 450 km^2 in Christchurch, New Zealand. The satellite image dataset is composed of two separate datasets. One has 204 images from six cities worldwide with a spatial resolution varying from 30 cm to 2.5 m and a size of 512×512 . The other has 17 388 tiles of six adjacent images, with a spatial resolution of 45 cm and a size of 512×512 pixels, covering 860 km^2 of East Asia. These two satellite image datasets, with different sensors, also have different spectral resolutions.
- The SpaceNet building dataset (Van Etten et al. 2018) was released using two SpaceNet challenges for building detection. Five cities worldwide were considered areas of interest. The WorldView-2 and WorldView-3 images have a size of 650×650 pixels and pixel-wise building labels covering 5555 km^2 on different continents.
- The AIRS (Aerial Imagery for Roof Segmentation) dataset (Q. Chen et al. 2018) was created using the same aerial images as the WHU aerial building dataset, where the original spatial resolution of the images (7.5 cm) was preserved.
- The Sencity Toulouse dataset (Roscher et al. 2020) was created based on WorldView-2 images for building instance segmentation, which covers an 50 km^2 area of Toulouse, France. Images in the dataset are classified into eight classes, including building and background. These images have a spatial resolution of 0.5 m for the panchromatic band and 2 m for the other bands. Each image is split into 16 tiles. Eventually, the panchromatic band has a size of 3504×3452 , and the other bands have a size of 876×863 .

In addition to the datasets described above (as summarized in Table 1), there are other building datasets, such as datasets released for the DeepGlobe Building Extraction Challenge (Demir et al. 2018), Open Cities AI Challenge (<https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/>), and the Crowd-AI Mapping Challenge (<https://www.crowdai.org/challenges/mapping-challenge>). The DeepGlobe building dataset is based on the SpaceNet dataset, in which building footprints, instead of rooftops, were annotated. The dataset for the Open Cities AI Challenge is a building footprint dataset across 10 cities in Africa and is known for its inconsistent annotation accuracy. The dataset for the Crowd-AI Mapping Challenge has more than 40 000 tiles of RGB images with a size of 300×300 pixels; however, the buildings are homogeneous, making it easier

Table 1. Existing building datasets.

Dataset	Location	Spectral bands	Classes	Coverage (km ²)	Spatial resolution (cm)
ISPRS Vaihingen/Postdam	Vaihingen / Potsdam, Germany	NIR, R, G, B, DSM	Six land-cover classes	1.40/3.40	5.00/9.00
Massachusetts	Massachusetts, USA	R, G, B	Building and non-building	340.00	100.00
WHU (aerial)	Christchurch, New Zealand	R, G, B	Building and non-building	457.00	30.00
Inria	10 regions in the United States and Austria	R, G, B	Building and non-building	810.00	30.00
SpaceNet	Four cities around the world	WorldView-3 8 bands	Building, road, and background	5555.00	30.00/50.00
AIRS	Christchurch, New Zealand	R, G, B	Building and non-building	457.00	7.50
ISPRS Semcity Toulouse	Toulouse, France	WorldView-2 8 bands	Eight land-cover classes	50.00	50.00
Waterloo building dataset	Kitchener–Waterloo, Canada	R, G, B	Building and non-building	205.83	12.00

for them to be segmented from the background compared to other datasets (Roscher et al. 2020).

Compared to these existing datasets, our dataset has both a higher spatial resolution and a larger-scale covering, except for the AIRS dataset. However, the AIRS dataset was constructed using images covering the Southern Hemisphere. Both illumination conditions and building styles are limited; models trained on this dataset may struggle with building extraction in the Northern Hemisphere, particularly in Canada and the United States. We provide the highest spatial resolution large-scale building extraction dataset in North America; models trained on our dataset should perform better in Canada and the United States.

2.2. Building detection methods

Building detection from remote sensing imagery was studied prior to the popularization of deep learning; classification strategy-based methods, active contour-based methods, and graph-based methods are the mainstream methods for this task (Ok 2013). Although deep learning methods (Mnih 2013; Shu 2014) were used for building detection before the proposal of fully convolutional networks (FCNs) (Long et al. 2015), deep learning-based image segmentation has become state-of-the-art soon after the invention of FCNs, despite recent work on semi-automated methods (Brooks et al. 2015). Recent classical machine learning methods that combine LiDAR and hyperspectral data have achieved good results (Parsian and Amani 2017). However, when only RGB orthoimages are available, deep learning methods significantly outperform classical methods.

With the development of deep learning methods, powerful semantic segmentation methods have been introduced. SegNet was applied with an active contour model to extract buildings from the ISPRS Potsdam dataset (Sun et al. 2018). Using the same dataset, Xu et al. (2018) and Yang et al. (2018) proposed new networks based on Res-U-Net and an attention mechanism, respectively. Through comparative studies by S. Wang et al. (2020) and Kemker et al. (2018), RefineNet and DeepLab v3 were introduced in this area. In addition to these methods, there are many other deep learning-based semantic segmentation

methods, such as multiple feature reuse network (MFRN) (Li et al. 2018), Deep Encoding Network (DE-Net) (H. Liu et al. 2019), Spatial Residual Inception Convolutional Neural Network (SRI-Net) (P. Liu et al. 2019), ENRU-Net (Kemker et al. 2018), and Capsule Feature Pyramid Network (CapsFPN) (Yu et al. 2020).

Apart from the methods mentioned above, instance segmentation and semi-automatic annotation methods can also be used in this task. Mask R-CNN, the most representative instance segmentation method, and its derivatives are widely used to extract buildings from remote sensing images (Zhao et al. 2018; Wen et al. 2019). In recent years, semi-automatic annotation methods have been introduced for this purpose. In 2019, Li et al. (2019) proposed PolyMapper, which can map buildings from images in an end-to-end manner. An improved PolyMapper was proposed in 2020 by Zhao et al. (2020).

Large-scale building extraction datasets are required for the application of deep learning techniques in building footprint extraction. This work is intended to further support the research and benchmark of new methods for automated building extraction, especially those that are deep learning based. Extensive comparative studies can also provide a reference for algorithm selection in practice and act as a benchmark for future algorithms trained on the Waterloo building dataset.

3. Waterloo building dataset

3.1. Study site

As shown in Fig. 1, the Kitchener–Waterloo area is in southeastern Ontario, Canada. The dataset covers 205.83 km² and includes both urban and rural areas with buildings of different shapes, heights, and colors.

3.2. Images

We first collected aerial images covering the study site in 2014 from the Geospatial Centre at the University of Waterloo and obtained permission from the Regional Municipality of Waterloo. As documented, these images were collected using a Vexcel UltraCam-D camera with a standard deviation of 3.5 mm for all image points. The North American Datum of 1983 (NAD 83) was used as the geographic system, and the Universal Transverse Mercator (UTM) Zone 17°N was used as the projection system.

Duplicate images across the boundary between Kitchener and Waterloo were removed from the entire set of images. Images without buildings at the boundary were also removed from the dataset. As a result, 242 images were acquired for building mask labelling from the total of 307. These images had a spatial resolution of 12 cm; spectral bands of red, green, and blue; and a size of 8350 × 8350.

3.3. Dataset generation

The entire dataset generation process is illustrated in Fig. 2. For building annotation, we used a triple-check scheme to ensure annotation accuracy. The three steps were self-checking after labeling, followed by partner and supervisor checking. In each step, the accuracy was controlled to within three pixels, which means that the distance between the boundaries of the polygons to the building boundaries was within three pixels. After several iterations of checking and revising, the manually edited building polygons were converted to masks. Approximately 117 000 independent buildings were extracted and labeled from these images.

Considering the computational resources and aiming to match most of the existing datasets, such as the WHU building dataset and Inria building dataset, we further tiled these images and masks into paired patches with a size of 512 × 512 pixels. In total, 69 938 patches were generated after tiling 242 images. They were randomly split into training, validation,

Figure 1. Map of the Kitchener–Waterloo area (Tool: ArcGIS; Data source: administrative areas shapefiles from Hijmans et al. (2004) (<http://www.diva-gis.org/gdata>) and basemap from ArcGIS Online). [Colour online.]



Figure 2. Flowchart of dataset generation.

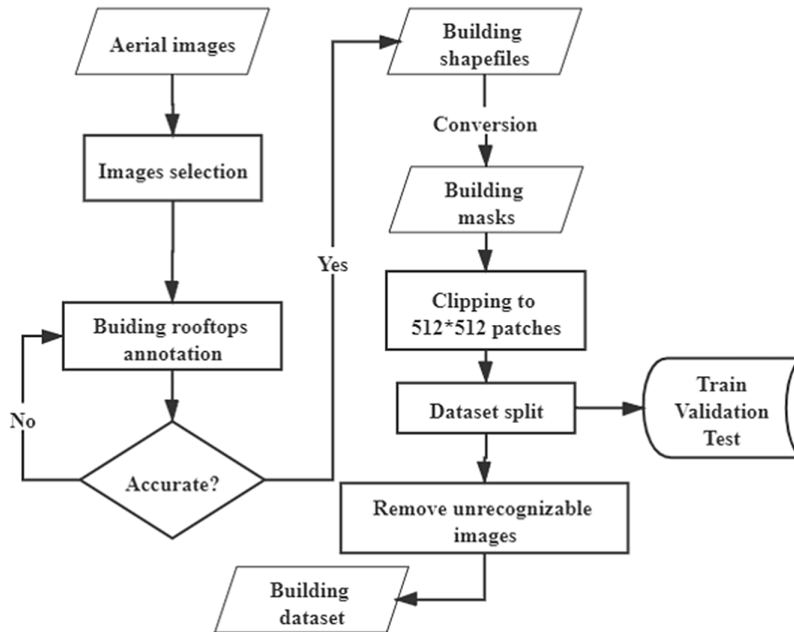


Figure 3. Examples of geometric distortion in aerial images (Tool: Microsoft Word; Data source: Waterloo building dataset (He et al. 2021)). [Colour online.]



and testing sets at a ratio of 7:1:3. The number of independent buildings (or parts of buildings at the boundaries) in each set was 66 289, 12 464, and 40 490. We removed the patches affected by geometric distortions, where the boundaries of the rooftops could not be recognized accurately by a human, as shown in Fig. 3. Finally, we obtained 42 147, 6887, and 18 945 patches for training, validation, and testing, respectively.

3.4. Dataset description

With permission from the municipality of Waterloo, we released the original large images with our shapefiles and masks (as shown in Fig. 4) and close-ups of tiled patches and matched masks (as shown in Fig. 5). Both the raw and georeferenced imagery are freely

Figure 4. Example of an annotated image. Top: original image, bottom left: mask, and bottom right: shapefile (Tool: Microsoft Word and ArcGIS; Data source: Waterloo building dataset (He et al. 2021)). [Colour online.]

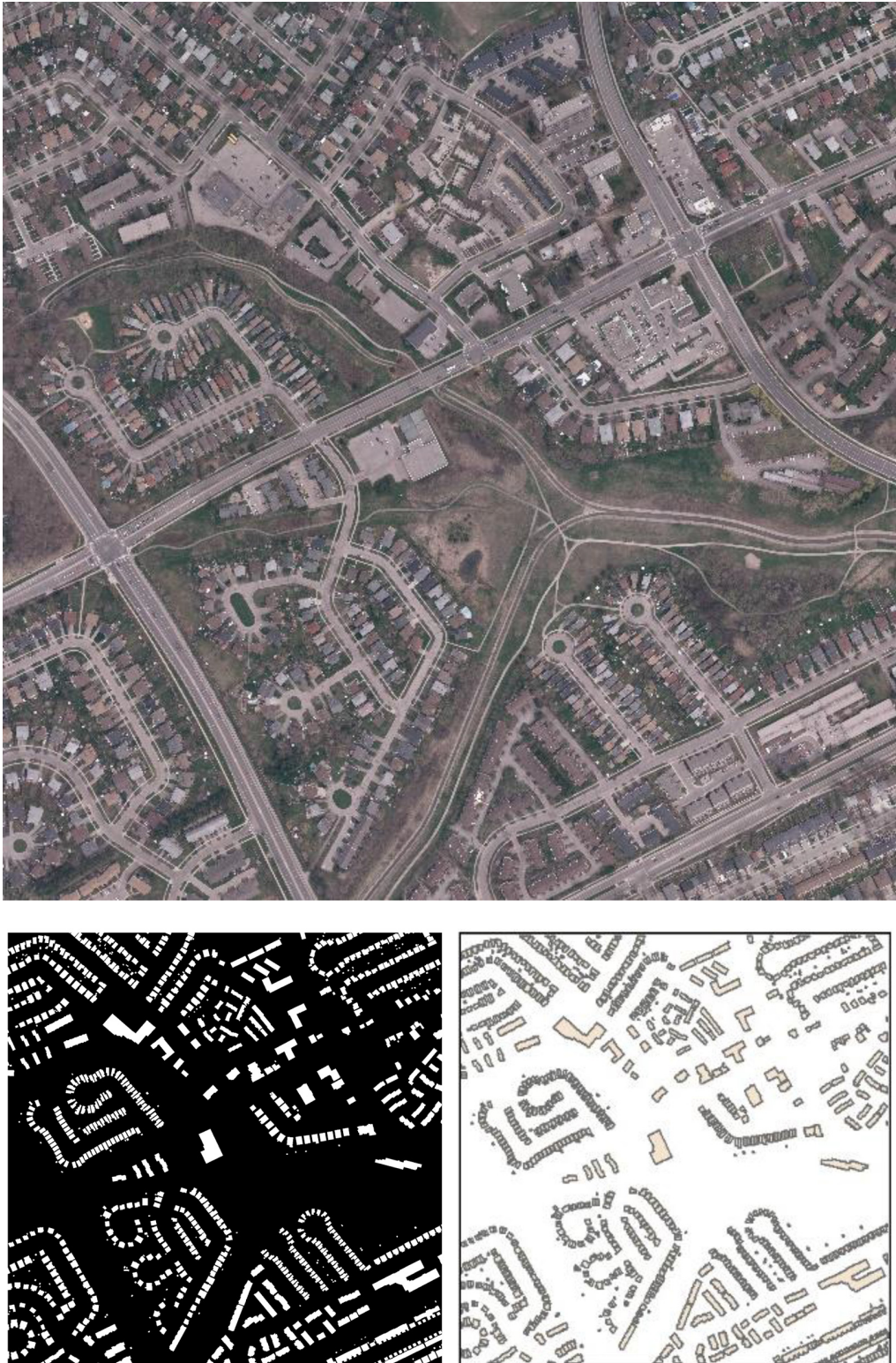
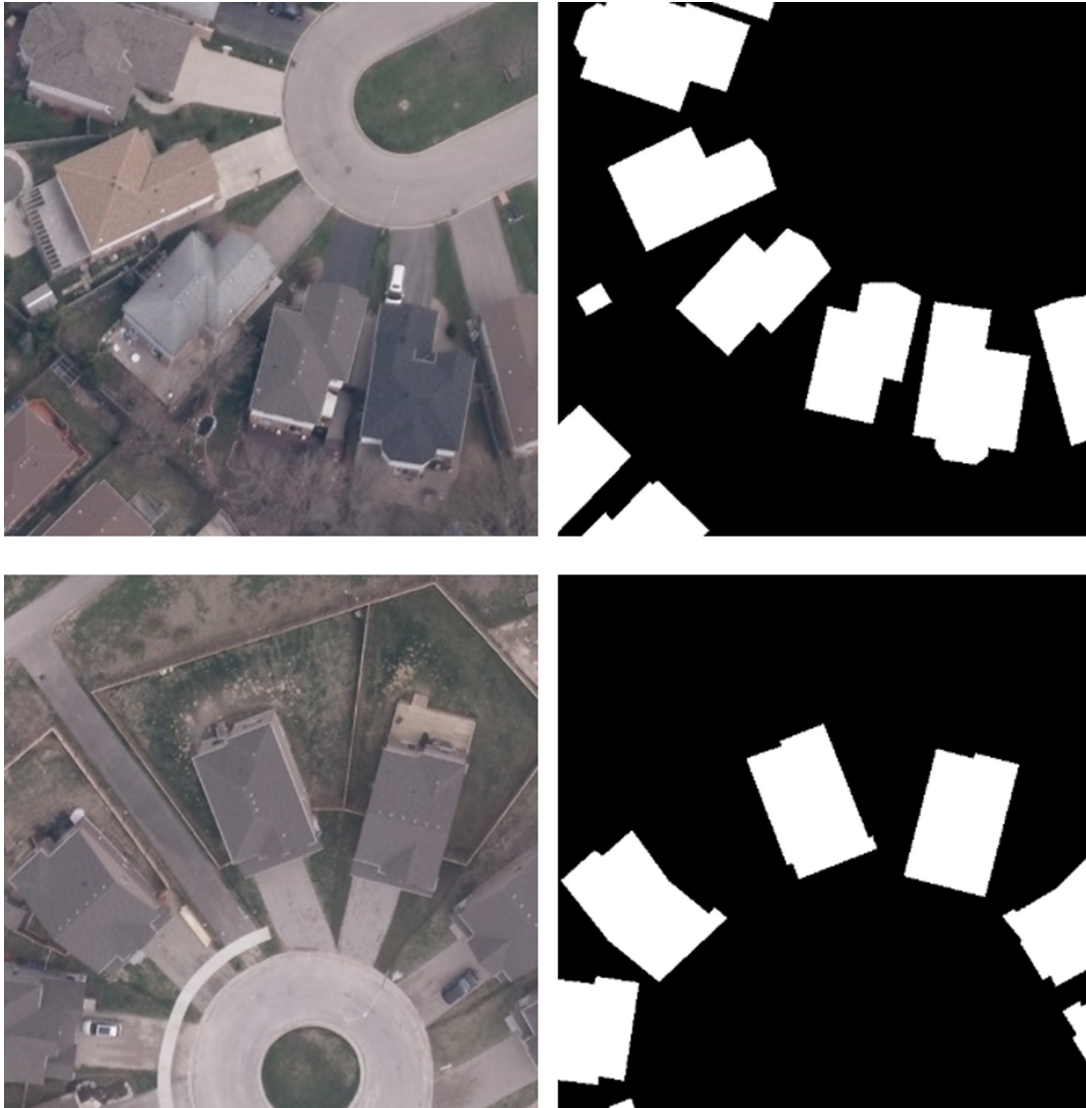


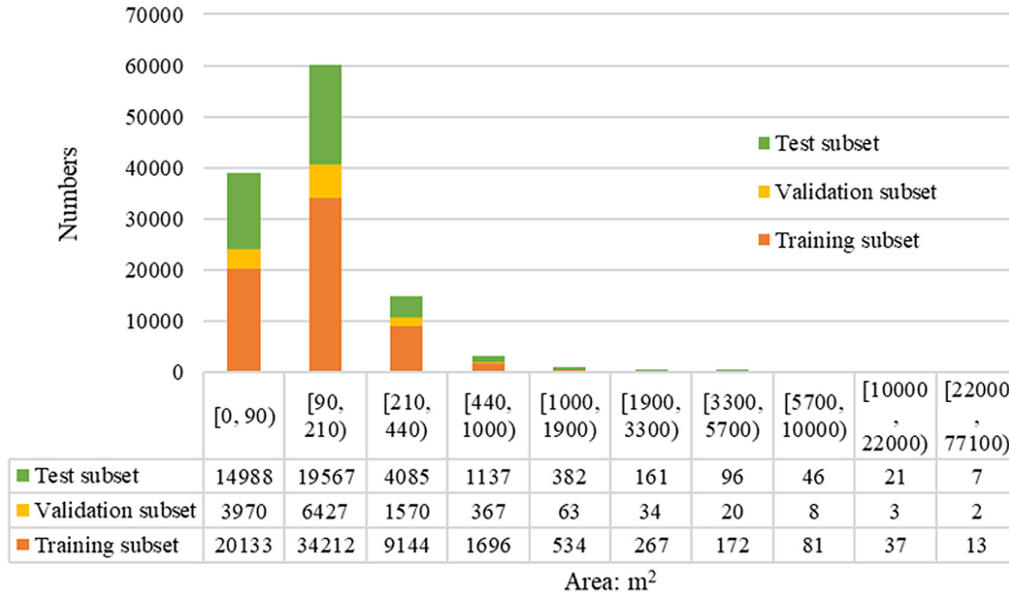
Figure 5. Close-ups of tiled patches (left) and matched masks (right) in our dataset (Tool: Microsoft Word; Data source: Waterloo building dataset (He et al. 2021)). [Colour online.]



accessible to the public. The removed patches with geometric distortion and fishnet files used for clipping images to patches were also released to recover the full images.

Figure 6 shows the distribution of the building footprint area for the training, validation, and test subsets. Most building footprint areas fall within the range of 0–1000 m², and each subset shares the same distribution. The distribution of the building footprint area is in line with common sense: most buildings are small and medium-sized, and only a few buildings for public infrastructure, such as schools, hospitals, shopping malls, and factories, have footprint areas larger than a few hundred square meters.

Figure 6. Distribution of building footprint area for training subset, validation subset, and test subset. [Colour online.]



4. Methods and metrics

An extensive comparative study was performed to benchmark the existing methods for the development of new methods. Both semantic segmentation and instance segmentation methods were tested. We leave PolyMapper and its derivation to future work.

4.1. Methods

Semantic segmentation methods

In this study, five semantic segmentation networks, namely, Fully Convolutional Networks-8s (FCN-8s) (Long et al. 2015), U-Net (Ronneberger et al. 2015), DeepLab v3+ (L.C. Chen et al. 2018), Fast Segmentation Convolutional Neural Network (Fast SCNN) (Poudel et al. 2019), and High-Resolution Network (HRNet) v2 (J. Wang et al. 2020) were selected as benchmarks. For method selection, we considered three criteria: first, the methods are commonly used according to the number of citations reported by Google Scholar; second, the models are easy to reimplement; and third, the methods outperform others in terms of accuracy or speed. For example, although CapsFPN exhibits a high performance in semantic segmentation tasks, it relies on a sophisticatedly designed architecture with more parameters and better hardware environments. Methods such as RefineNet, DeepLab v3, and SRI-Net have rarely been used in recent publications, especially in the remote sensing field. In contrast, our selected methods, such as FCN-8s and U-Net, are the most representative simple semantic segmentation methods that are widely used in different fields and for benchmarking existing building datasets. DeepLab v3+ and HRNet v2 were selected because of their high number of citations and high performance in published semantic segmentation studies. Fast SCNN was selected as a simple and fast semantic segmentation method that can be easily implemented on most computers. Herein, we briefly review their architectures. Readers are referred to the corresponding literature for detailed information.

FCN-8s is the first fully convolutional network for semantic segmentation, in which the final fully connected layer is replaced by convolutional layers. Deconvolutional layers were

introduced to up-sample the feature maps. Skip architecture is applied to preserve low-level and high-resolution features for final pixel-wise classification via feature-wise addition. In our study, we adopted VGG16 as the backbone of FCN-8s.

U-Net is an architecture that was developed at almost the same time as FCN-8s. For the first time, it employs a U-shaped architecture for semantic segmentation. The symmetric architecture preserves low-level and high-resolution features at the down-sampling part. The persevered features are further concatenated with up-sampled high-level features to achieve high accuracy in the segmentation results.

DeepLab v3+ is based on DeepLabv3. A decoder part is added based on DeepLab v3 for feature map up-sampling. Dilated convolutional layers were applied following the former version of the DeepLab algorithm. In DeepLab v3+, depth-wise separable convolution is applied in the atrous spatial pyramid pooling (ASPP) and decoder parts. In our study, the Xception network was used as the backbone.

A fast SCNN has been developed for real-time semantic segmentation. The entire architecture is divided into four parts: down-sampling, global feature extractor, feature fusion, and classification. Depthwise separable convolution and improvements in architecture allow Fast SCNN to reach the above real-time inference speed.

HRNet v2 can be used for different tasks with different heads. The four stages with multi-resolution convolution and repeated multi-scale feature fusion are the same as the main body in HRNet v1. The proposed head makes HRNet v2 different from the former version and adapts pose estimation to semantic segmentation and other computer vision tasks.

Instance segmentation methods

Among all instance segmentation methods in the computer vision field, we selected Mask R-CNN because it and its derivations are widely used in building footprint extraction. Specifically, we used a model trained on our own dataset and the model released by ESRI to benchmark our dataset.

4.2. Evaluation metrics

The metrics derived from the confusion matrix are not suitable for evaluating building rooftop/footprint extraction methods using high spatial and very high spatial resolution images (Shu 2014). In previous studies, methods for building footprint or rooftop extraction were evaluated using intersection over union (IoU), mean IoU (mIoU), precision, recall, F_1 -score, accuracy, and frame per second (FPS). Specifically, IoU represents the percentage of overlap between the ground truth and the prediction output. mIoU represents the average of positive and negative objects IoU. Precision indicates the extent to which the predicted positive objects are correct compared with all predicted positive objects. Recall shows how many positive objects are predicted accurately compared to all positive objects from the ground truth. The F_1 -score or F_1 measure is the harmonic mean of the precision and recall. The accuracy or average accuracy calculates the percentage of correctly classified pixels or other objects in the images. These metrics are defined as follows:

$$(1) \text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

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

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

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

$$(5) \text{ F1} = \frac{2TP}{2TP + FP + FN}$$

$$(6) \text{ Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$(7) \text{ FPS} = \frac{TI}{t}$$

where TP indicates true positive, denoting a correct prediction of the positive class (presence of building); FP refers to false positive, which occurs when the model predicts the positive class as the negative class; FN stands for false negative, in which the model classifies the positive class as the negative class; TN represents true negative, in which the model predicts the negative class correctly in the output; TI is the number of images that have been tested; and t is the total time cost for testing.

4.3. Implementations

Except for HRNet v2, we utilized Adam as the optimizer with a stable learning rate of $1e-4$, a batch size of 5, and training for 100 epochs. Binary cross-entropy and accuracy were used as the loss function and training metrics, respectively. For HRNet v2, we implemented the Jaccard loss function following the original configuration. The training metrics included Jaccard loss, binary cross entropy, joint loss including soft Jaccard loss, mean square error, and accuracy. Considering the computational resources, we set the batch size to four for HRNet v2. All algorithms were trained and tested on a single GeForce RTX 2080ti GPU and CUDA 10.2.

ESRI recently released a pre-trained Mask R-CNN model². The model file can be downloaded, and the building footprints can be directly extracted using ArcGIS Pro. In this study, we tested the model on a laptop with a single GeForce GTX 1650 GPU and CUDA 11.0. All parameters, except batch size (set as 1), were set to default. For comparison, we trained a Mask R-CNN³ on our own dataset from scratch, following the initial parameter setting. To successfully train the model, we changed the learning rate from 0.01 at epoch 65 to 0.00001 to avoid an Not a Number (NaN) loss error. In addition, to make full use of the GPU resources, we set the batch size to 5 and 1 in the training and testing phases, respectively. The model was trained and tested on a single Nvidia TITAN XP with CUDA 11.4.

5. Results and discussion

Both qualitative and quantitative evaluations of building rooftop extraction results are provided and discussed in this section. A discussion of the open challenge in building an extraction dataset is also included.



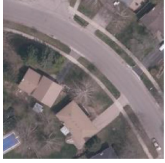
























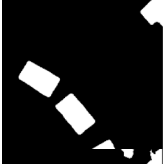




5.1. Qualitative evaluation

We randomly selected two patches from the testing set as examples and compared the extraction results with the ground-truth images. We also colorized the ground truth, predicted pixels, and wrongly classified pixels in one image to visualize the extraction results (Ji et al. 2018), in which blue, green, and red represent the predicted results, ground truth, and wrongly classified results, respectively. Examples and extraction results are presented in Table 2. For each example, the images in the first column are the original images

²<https://www.arcgis.com/home/item.html?id=a6857359a1cd44839781a4f113cd5934>.

³https://github.com/matterport/Mask_RCNN.

Table 2. Two examples of extraction results. [Colour online.]

		Example 1		Example 2	
Model					
Semantic segmentation	FCN-8s				
	U-Net				
	DeepLab v3+				
	Fast SCNN				
	HRNet v2				
Instance segmentation	Mask R-CNN (ESRI)				
	Mask R-CNN (ours)				

and the colorized extraction results. The second column presents the ground truth and the predicted masks.

As shown in [Table 2](#), among the semantic segmentation methods, more recent methods show a higher performance, as expected. The boundaries of buildings are more accurate

with newer algorithms, except for Fast SCNN. The number of incorrectly classified pixels in the background also decreased significantly. This result is reasonable because Fast SCNN focuses on the speed of the extraction process at the cost of accuracy. With shared low-level features and feature fusion using only high-level features, the architecture of Fast SCNN is similar to that of FCN-8s. DeepLab v3+ is known for its atrous spatial pyramid pooling and depth-wise separable convolution. The former enlarges the field-of-view of the network, whereas the latter increases its efficiency. The combination of these two techniques renders the algorithm accurate and efficient for semantic segmentation. Multi-resolution convolution and repeated multi-scale features fusion of HRNet v2 preserve high-resolution low-level features and provide highly accurate extraction results. With the aid of soft Jaccard loss, the imbalanced data distribution problem in the dataset is relieved, further benefiting building footprint extraction.

Table 2 provides a visual comparison of the results obtained using the different methods. In this table, the green and red pixels are false negatives and false positives, respectively. In these two examples, the red pixels decreased sharply in the newer algorithms, except for the Fast SCNN. DeepLab v3+ and HRNet v2 in particular achieved high accuracy in these examples. In the first example, both methods omitted parts of the buildings and mistakenly detected a building on the top-right side of the patch. In the second example, they also missed the building on the top-right of the patch. In the second example, Fast SCNN exhibited a higher performance than FCN-8s. Visually, it achieves the same performance as DeepLab v3+ and HRNet v2. However, in the first example, its performance degrades to the level of FCN-8s.

For instance, the segmentation results show that the two Mask R-CNN also exhibit high performance as DeepLab v3+ and HRNet v2. Multi-task learning of Mask R-CNN and the nature of instance segmentation give it high performance and remove most of the background noise. Both Mask R-CNN models successfully detected the top right and left buildings and avoided the wrong classification of the top-right object in the first example compared to DeepLab v3+ and HRNet v2. In addition, the ESRI-trained model outputs less background noise compared to all the methods in the second example, whereas it shows poor accuracy on the large building of the first example. Qualitative visual analysis cannot properly compare the different models. Therefore, in the next section, we provide a quantitative analysis using the metrics described in Section 4.2.

5.2. Quantitative evaluation

As mentioned in Section 4.2, we leveraged seven metrics to evaluate the performance of all the algorithms employed in this study. Table 3 presents a quantitative comparison of the benchmark models. Accuracy, or overall pixel accuracy, is the ratio of correctly classified pixels to total pixels. As expected, DeepLab v3+ exhibits the highest accuracy. Here, IoU refers to the intersection of the predicted building masks and ground truth building masks with their union. mIoU is the average of the IoUs for building masks (foreground) and background. Because of the imbalance between pixels labeled as “building” and pixels labeled as “background”, mIoU was higher than IoU for all algorithms. IoU and mIoU share the same trend as accuracy. Precision and recall, or correctness and completeness (Shu 2014), represent the ratio of correctly classified building masks to all predicted building masks, and the ratio of correctly classified building masks to all ground-truth building masks, respectively. Similar to IoU and mIoU, precision exhibited the same trend as accuracy. It is worth noting that the recall fluctuates among the algorithms. U-Net had the highest recall, which means that it generated the highest completeness. The precision-recall curve is widely used to demonstrate the correctness and completeness of algorithms. F_1 -score is an alternator for the curve, which is a harmonic mean of precision and recall (Shu 2014).

Table 3. Quantitative performance evaluation (%).

Model	Accuracy	IoU	mIoU	Precision	Recall	F ₁ -score	FPS
FCN-8s	77.13	24.99	50.12	26.10	85.50	39.99	19.61
U-Net	86.72	37.25	61.42	39.16	88.43	54.28	14.93
DeepLab v3+	97.32	72.72	84.92	88.55	80.27	84.21	17.60
Fast SCNN	77.31	23.02	49.34	24.81	76.10	37.42	24.01
HRNet v2	97.78	76.63	87.12	92.48	81.72	86.77	18.19
Mask R-CNN (ESRI)	96.57	64.56	80.45	89.15	70.06	78.46	—
Mask R-CNN (ours)	95.27	59.39	77.15	71.73	77.53	74.52	—

From the F₁-score, we can confirm that HRNet v2 is the best algorithm and Fast SCNN is the worst algorithm in this study. The FPS evaluates the speed of the algorithms by counting the number of patches predicted per second. Among all semantic segmentation methods, Fast SCNN has the highest value, and U-Net has the lowest value. In other words, Fast SCNN possesses the highest efficiency, and U-Net has the lowest efficiency. Because we tested instance segmentation models using different machines, the FPS for the two models is not comparable here. However, given their more complicated architectures and parameters, these two models are expected to be slow.

Based on the accuracy, IoU, mIoU, and F₁-score, HRNet v2 was the most accurate among all methods. For efficiency, Fast SCNN is the best one, as it is designed for this purpose. The ESRI-trained Mask R-CNN showed a higher performance than our own trained one, but a lower recall value. This can be explained by the difference between the data volume and data distribution. From this successful benchmark, we conclude that our dataset is suitable for the training and evaluation of deep learning-based semantic segmentation and instance segmentation models.

5.3. Some Challenges

The out-of-distribution performance of the deep learning model is often lower than that indicated by its in-distribution test score; that is, a model trained and tested on a specific dataset would perform worse than expected on a different dataset and, of course, during practical use. Addressing this problem is key to the practical application of deep learning models. Specific to building footprint extraction from VHSR aerial orthoimages, deep learning models that generalize well across datasets must account for differences in resolution, image and label quality, image differences induced by different sensors, and other difficult-to-detect differences between datasets. Each of these differences poses a unique challenge for the out-of-distribution generalizability of deep learning models. To address these problems, researchers can train on a mixed training set containing image patches from a large variety of source datasets in the hope that models learn to ignore the peculiarities of any individual dataset and learn building extraction features that are common across all datasets. The Inria dataset (Maggiori et al. 2017) uses this approach and shows that the out-of-distribution performance decreases. Touzani and Granderson (2021) also worked in this direction and proposed a new framework for building footprint extraction by automatically generating building datasets with more variability from openly available data in the United States. Further research in this direction is necessary. Another solution is to try a large variety of data augmentations during training to mimic the changes in resolution, image and label quality, and any other differences across datasets. In addition to the strategies used in the training phase, Nguyen et al. (2020) proposed a super-resolution-based snake model for post-processing to overcome the generalization problem. In their study, both LiDAR data and optical images were used for building footprint extraction. Experiments on the ISPRS Vaihingen benchmark datasets and the City of

Quebec, Canada, solidify the success of the model to overcome this problem. It is also possible that each of these cross-dataset differences should be addressed individually. We hope that future research will shed light on this problem.

We pose the following open questions to the remote sensing research community and hope to address them in our future work. For common deep learning architectures applied to remote sensing tasks, how well does out-of-distribution perform with the inclusion of additional datasets in the training set? Does a set of image augmentations produce a near-perfect model generalizability across datasets?

6. Conclusions

In this study, we introduced the Waterloo building dataset. A city-scale vector building dataset for mapping building footprints using aerial orthoimagery. Our dataset covers the Kitchener–Waterloo area and extends over 205.83 km². Both the original 8350 × 8350 and tiled 512 × 512 images are available. To ensure label quality, the labels were manually generated by experts under multiple consistency checks. We conducted a comparative study on popular semantic segmentation methods trained on our dataset and demonstrated their applicability to the training and evaluation of deep learning algorithms. We pose the question of out-of-distribution generalizability to the research community and hope that our dataset can benefit future research towards high-performance and generalizable deep learning building footprint extraction models.

Acknowledgements

We would like to thank Yuwei Cai, Qitong Yu, Kun Zhao, Junbo Wang, Liyuan Qing, Yan Liu, Hasti Andon Petrosians, Zhehan Zhang, and Siyu Li of the Geospatial Sensing and Data Intelligence Lab, University of Waterloo, for their contributions to building labeling work. The first author also thanks the China Scholarship Council for its support through a doctoral scholarship (No. 201906180088).

References

- Brooks, R., Nelson, T., Amolins, K., and Hall, G.B. 2015. Semi-automated building footprint extraction from orthophotos. *Geomatica*, **69**(2): 231–244. doi:[10.5623/cig2015-206](https://doi.org/10.5623/cig2015-206).
- Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision*. pp. 801–818.
- Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., and Waslander, S.L. 2018. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *arXiv 2018*. arXiv preprint arXiv:1807.09532.
- Chen, Z., Li, D., Fan, W., Guan, H., Wang, C., and Li, J. 2021. Self-attention in reconstruction bias U-Net for building extraction. *Remote Sens.* **13**: 2524. doi:[10.3390/rs13132524](https://doi.org/10.3390/rs13132524).
- Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., et al. 2018. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proc. CVPR Workshops*, pp. 172–181. doi:[10.1109/CVPRW.2018.00031](https://doi.org/10.1109/CVPRW.2018.00031).
- He, H., Jiang, Z., Gao, K., Narges Fatholahi, S., Cai, Y., Tan, W., et al. 2021. Waterloo building dataset. *Harvard Dataverse*, V1. doi:[10.7910/DVN/EXRA2V](https://doi.org/10.7910/DVN/EXRA2V).
- Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrantes, I., and Rojas, E. 2004. DIVA-GIS. A geographic information system for the analysis of species distribution data. Available from <http://www.diva-gis.org/>.
- 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](https://doi.org/10.1109/TGRS.2018.2858817).
- Kemker, R., Salvaggio, C., and Kanan, C. 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **145**: 60–77. doi:[10.1016/j.isprs.2018.04.014](https://doi.org/10.1016/j.isprs.2018.04.014).
- Li, L., Liang, J., Weng, M., and Zhu, H. 2018. A multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote Sens.* **10**(9): 1350. doi:[10.3390/rs10091350](https://doi.org/10.3390/rs10091350).
- Li, Z., Wegner, J.D., and Lucchi, A., 2019. Topological map extraction from overhead images. In *Proc. IEEE/CVF International Conference on Computer Vision*. pp. 1715–1724.
- Liu, H., Luo, J., Huang, B., Hu, X., Sun, Y., Yang, Y., et al. 2019. DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery. *Remote Sens.* **11**(20): 2380. doi:[10.3390/rs11202380](https://doi.org/10.3390/rs11202380).
- Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., and Zhang, Y. 2019. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.* **11**(7): 830. doi:[10.3390/rs11070830](https://doi.org/10.3390/rs11070830).

- Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *In Proc. Computer Vision and Pattern Recognition*. pp. 3431–3440. arXiv: 1411.4038.
- Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. 2017. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. *In Proc. IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*. pp. 3226–3229. doi:[10.1109/IGARSS.2017.8127684](https://doi.org/10.1109/IGARSS.2017.8127684).
- Mnih, V. 2013. Machine learning for aerial image labeling. PhD Thesis, University of Toronto.
- Nguyen, T.H., Daniel, S., Guériot, D., Sintès, C., and Le Caillec, J.M., 2020. Super-resolution-based snake model—An unsupervised method for large-scale building extraction using airborne LiDAR data and optical image. *Remote Sens.* **12**(11): 1702. doi:[10.3390/rs12111702](https://doi.org/10.3390/rs12111702).
- 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](https://doi.org/10.1016/j.isprsjprs.2013.09.004).
- Parsian, S., and Amani, M. 2017. Building extraction from fused LiDAR and hyperspectral data using Random Forest Algorithm. *Geomatica*, **71**(4): 185–193. doi:[10.5623/cig2017-401](https://doi.org/10.5623/cig2017-401).
- Poudel, R.P., Liwicki, S., and Cipolla, R. 2019. Fast-SCNN: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502.
- Rastogi, K., Bodani, P., and Sharma, S.A. 2020. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto Int.* doi:[10.1080/10106049.2020.1778100](https://doi.org/10.1080/10106049.2020.1778100).
- Ronneberger, O., Fischer, P., and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. *In Proc. MICCAI*. pp. 234–241. arXiv: 1505.04597.
- Roscher, R., Volpi, M., Mallet, C., Drees, L., and Wegner, J.D. 2020. SemCity Toulouse: A benchmark for building instance segmentation in satellite images. *ISPRS Ann.* **5**: 109–116. doi:[10.5194/isprs-annals-V-5-2020-109-2020](https://doi.org/10.5194/isprs-annals-V-5-2020-109-2020).
- Rottensteiner, F., Sohn, G., Gerke, M., and Wegner, J.D. 2013. ISPRS test project on urban classification and 3D building reconstruction. https://www2.isprs.org/media/komfssn5/complexscenes_revision_v4.pdf
- Sahar, L., Muthukumar, S., and French, S.P. 2010. Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories. *IEEE Trans. Geosci. Remote Sens.* **48**(9): 3511–3520. doi:[10.1109/TGRS.2010.2047260](https://doi.org/10.1109/TGRS.2010.2047260).
- Shu, Y. 2014. Deep convolutional neural networks for object extraction from high spatial resolution remotely sensed imagery. PhD Thesis, University of Waterloo.
- 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](https://doi.org/10.3390/rs10091459).
- Thomas, J., Kareem, A., and Bowyer, K.W. 2013. Automated post-storm damage classification of low-rise building roofing systems using high-resolution aerial imagery. *IEEE Trans. Geosci. Remote Sens.* **52**(7): 3851–3861. doi:[10.1109/TGRS.2013.2277092](https://doi.org/10.1109/TGRS.2013.2277092).
- Touzani, S., and Granderson, J., 2021. Open data and deep semantic segmentation for automated extraction of building footprints. *Remote Sens.* **13**(13): 2578. doi:[10.3390/rs13132578](https://doi.org/10.3390/rs13132578).
- Van Etten, A., Lindenbaum, D., and Bacastow, T.M. 2018. SpaceNet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.
- Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10): 3349–3364. doi:[10.1109/TPAMI.2020.2983686](https://doi.org/10.1109/TPAMI.2020.2983686).
- Wang, S., Hou, X., and Zhao, X. 2020. Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access*, **8**: 7313–7322. doi:[10.1109/ACCESS.2020.2964043](https://doi.org/10.1109/ACCESS.2020.2964043).
- 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](https://doi.org/10.3390/s19020333).
- Xie, Y., Weng, A., and Weng, Q. 2015. Population estimation of urban residential communities using remotely sensed morphologic data. *IEEE Geosci. Remote Sens. Lett.* **12**(5): 1111–1115. doi:[10.1109/LGRS.2014.2385597](https://doi.org/10.1109/LGRS.2014.2385597).
- Xu, Y., Wu, L., Xie, Z., and Chen, Z. 2018. Building extraction in very high-resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **10**(1): 144. doi:[10.3390/rs10010144](https://doi.org/10.3390/rs10010144).
- 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](https://doi.org/10.3390/rs10111768).
- Yu, Y., Ren, Y., Guan, H., Li, D., Yu, C., Jin, S., and Wang, L. 2020. Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery. *IEEE Geosci. Remote Sens. Lett.* **18**(5): 895–899. doi:[10.1109/LGRS.2020.2986380](https://doi.org/10.1109/LGRS.2020.2986380).
- Zhao, K., Kang, J., Jung, J., and Sohn, G. 2018. Building extraction from satellite images using mask R-CNN with building boundary regularization. *In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 247–251.
- Zhao, W., Ivanov, I., Persello, C., and Stein, A. 2020. Building outline delineation: From very High resolution remote sensing imagery to polygons with an improved end-to-end learning framework. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **43**: 731–735.