Deep Learning Based Building Extraction from High-Resolution Remote Sensing Images

by

Yifan Wu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2022

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 2 of this thesis consists of an article that has been co-authored by myself, Linlin Xu, Yuhao Chen, Alexander Wong, and David Clausi. Linlin Xu and Yuhao Chen helped in guiding the research and planning the experiments. Alexander Wong and David Clausi assisted with the writing of the article.

Chapter 3 of this thesis consists of an article that has been co-authored by myself, Linlin Xu, Lei Wang, Qi Chen, Yuhao Chen, and David Clausi. Linlin Xu helped in guiding the research and planning the experiments. Lei Wang and Qi Chen helped in guiding the research and provided experimental data of previous works in the literature used for comparison. Yuhao Chen and David Clausi assisted with the writing of the article.

Abstract

Building extraction from remote sensing images is a critical task to support various applications such as cartography, disaster response, and urban planning. The automation of this task is an active research area due to the time-consuming nature and high expense associated with the manual approach. However, traditional computer vision methods rely on handcrafted features and human knowledge, leading to the lack of the ability to leverage big remote sensing data. Although recently developed deep learning based methods brought significant advancements in the identification and coarse annotation of buildings, the accuracy and precision of extracted buildings are still insufficient for high-precision applications such as surveying and mapping.

This thesis presents two works aiming at enhanced building extraction from high-resolution remote sensing images by tackling key issues in building footprint extraction and building vectorization. For building footprint extraction, to address the heterogeneous noisy features around building boundaries, this thesis presents a deep learning strategy that incorporates a topography-aware loss (TAL) within a multi-resolution fusion architecture to increase the accuracy of boundaries in building segmentation. For building vectorization, to address the interference caused by noise and obstruction from shadows and trees around buildings and the limited receptive field in deep learning networks, this thesis presents a framework that combines a deep learning based building edge detection strategy and a geometry-guided building polygon reconstruction method for improved building outline vectorization in terms of vertex accuracy. Comparative experimental results on high-resolution remote sensing building datasets demonstrate significant improvements in building boundary accuracy and polygon vertex accuracy respectively over state-of-the-art methods. Hence, both works provide new means to address challenges posed by complex environmental conditions around buildings captured in remote sensing images and enable accurate building segmentation and vectorization towards automatic building extraction for high-precision applications.

Acknowledgements

I would like to thank my supervisors, Prof. David Clausi and Prof. Linlin Xu, for the support and guidance that they provided during my Masters studies. I am grateful to them for giving me the opportunity to study and research at the VIP Lab. I would also like to thank Prof. John Zelek and Prof. Jonathan Li for being members of my thesis committee.

I want to thank my friends and members of the remote sensing group for supporting each other inside and outside the lab during this unusual time.

Dedication

This is dedicated to my parents. Thank you for always being there for me.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problems and Contributions

Building extraction from remote sensing images is an important task with various applications such as cartography, disaster response, and urban planning [41, 39, 7]. The manual approach is extremely time-consuming and expensive due to the large scale of building coverage. Further challenges are posed by the ripped development of urban area and the requirement of up-to-date maps for modern applications, which demand more frequent updates of building information. This leads to active research in the automation of building extraction, which benefits applications in geomatics and environmental science.

Over the decades, many traditional computer vision approaches have been proposed for building extraction. They mostly rely on empirical knowledge of buildings to extract features such as colors, textures, edges, shapes, shadows, and context, combining with one or more knowledge-based methods such as template matching, active contour model, mathematical morphology, graph-based analysis, and dynamic programming [36, 1, 16, 30, 26]. Despite the progress achieved, the performance of those approaches largely depends on the quality of manually designed features, which requires human experience and explicit understanding of building characteristics. However, those approaches do not benefit from the increased resolution of remote sensing images because of the complexity of the environment captured by high-resolution remote sensing images and the difficulty of adopting handcrafted features to different building variations.

Recently, the development of deep learning techniques including convolutional neural networks (CNNs) have shown improvement in computer vision tasks such as image classification, object detection, and semantic segmentation [20, 22, 14]. As a result of this

development and the increased availability of high-resolution remote sensing images, significant advancements have been made in the field of automatic building extraction. Those data-driven approaches can extract discriminative features by learning from labeled building data and demonstrate state-of-the-art performance in detection and segmentation of buildings, e.g., CBR-Net [12] achieves intersection-over-union (IoU) [22] of 91.4%, 74.55%, and 81.1% in the WHU building dataset [18], the Massachusetts building dataset [28], and the Inria aerial image dataset [24] respectively. However, the accuracy of extracted buildings provided by those coarse annotations is still insufficient and limits their applications in mapping and navigation which require high-precision building annotations.

Although boundaries are important features that define the shape and location of buildings, achieving accurate boundary prediction is difficult due to the limitation of typical CNN design and the complex environmental conditions in the immediate neighboring region of building boundaries. The work in Chapter 2 proposes a topography-aware multi-resolution fusion learning strategy specifically designed for enhanced building footprint extraction. A topography-aware loss (TAL) that adapts to building topology and helps the network to learn building boundary features is introduced. It is then incorporated in a multi-resolution fusion architecture that provides high-resolution feature representation to boost segmentation performance. Furthermore, the average thresholded contour accuracy (tCA) is introduced to effectively evaluate the accuracy of building boundaries. The effectiveness of the method is demonstrated and compared with state-of-the-art semantic segmentation models using experimental results on the SpaceNet buildings dataset [37].

Building vector maps are essential in supporting high-precision applications, but its generation through deep learning approaches is prone to errors caused by noise and obstruction around buildings in aerial images. The work in Chapter 3 proposes a multi-task edge detection (MTED) framework for building vectorization countering those challenges. A deep learning based building edge detection approach utilizing rotated bounding boxes is introduced to increase robustness to interference. In order to take advantage of spatial context and regularization, a multi-task learning strategy is designed to integrate building segmentation with building edge detection. Finally, a simple yet effective geometry-guided post-processing method reconstructs building outline vectors by leveraging learned building shape priors and predicted building edges. Significant improvements over state-of-the-art methods in terms of vertex accuracy metrics and overall vector representation are demonstrated by comparative experiments on the very-high-resolution Aerial Imagery for Roof Segmentation (AIRS) dataset [8].

Therefore, this thesis presents two works that focus on improving the quality of automatically extracted building annotations from high-resolution remote sensing images. Both works are motivated by the increased availability of high-resolution remote sensing images

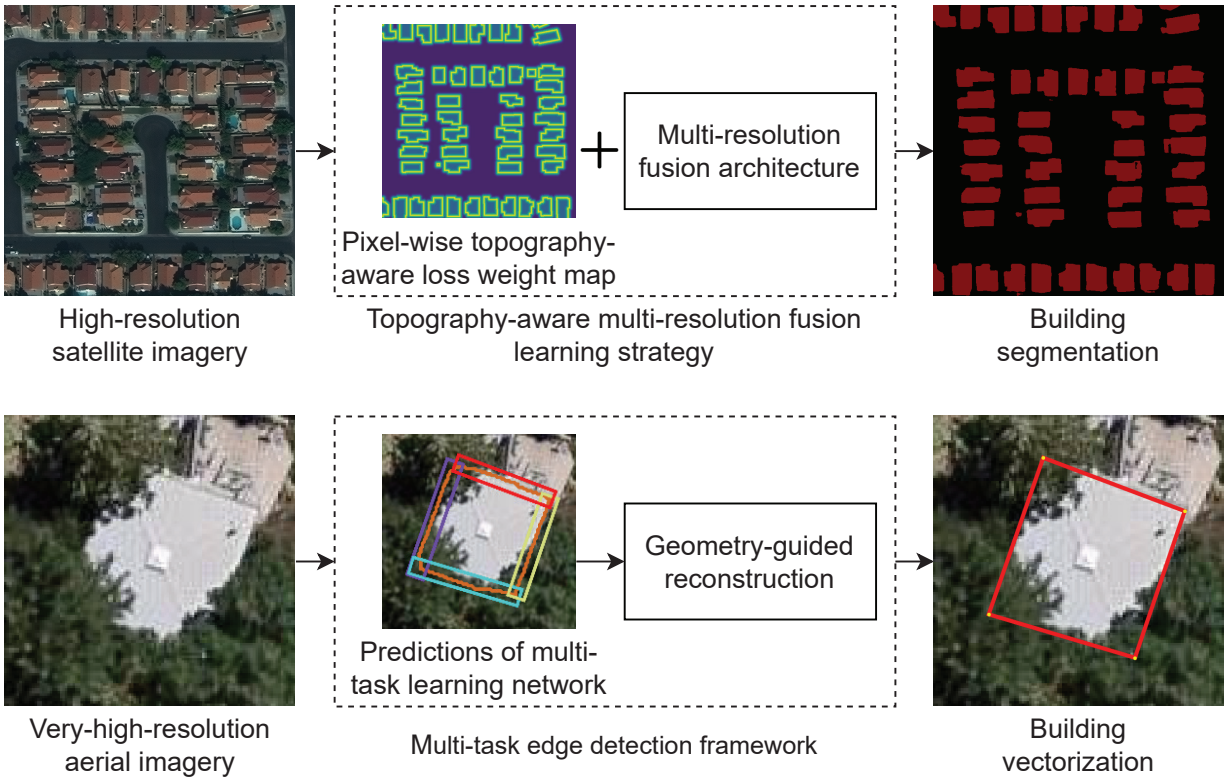

Figure 1.1: Overview of the works in Chapter 2 and Chapter 3. The first work addresses boundary accuracy to improve building segmentation. The second work provides robust edge detection to improve building vectorization.

and the successes of applying deep learning techniques to computer vision tasks. Furthermore, complex environmental conditions around building boundaries and CNNs' limited prediction precision at object boundaries are common challenges faced in those works. The two works are differentiated by their product types used in applications, which are partly determined by the level of precision enabled by the resolution of input images. The work in Chapter 2 improves building annotations extracted by segmentation of satellite images with a spatial resolution of 30 cm. In comparison, the work in Chapter 3 produces building annotations in polygon format for vector map productions from aerial images with a spatial resolution of 7.5 cm. As a result, this leads to the different methodology directions as shown in Figure 1.1, where Chapter 2 proposes a topography-aware loss for buildings and incorporates it in a multi-resolution fusion architecture to improve building segmentation, and Chapter 3 proposes a building edge detector and combines it with a geometry-guided

polygon reconstruction method to improve building vectorization.

1.2 Thesis Outline

In this thesis, two main works aiming at enhanced building extraction from high-resolution remote sensing images are presented as shown by the overview in Figure 1.1. Both works focus on resolving key issues that prevent the application of automatic building extraction as described in Chapter 1. Chapter 2 presents the first work addressing the precision issues of CNN-based semantic segmentation approaches at the delineation of building boundaries. Chapter 3 presents the second work improving accuracy and precision of building vectorization from very-high-resolution aerial images. Finally, Chapter 4 summarizes the works and briefly outlines the direction of future works.

Chapter 2

TAL: Topography-Aware Multi-Resolution Fusion Learning for Enhanced Building Footprint Extraction

The following article has been published in the journal IEEE Geoscience and Remote Sensing Letters, with the full reference being: Y. Wu, L. Xu, Y. Chen, A. Wong and D. A. Clausi, "TAL: Topography-Aware Multi-Resolution Fusion Learning for Enhanced Building Footprint Extraction," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 6506305, doi: 10.1109/LGRS.2022.3149709. The text was reformatted to fit the thesis format.

This article documents a topography-aware multi-resolution fusion learning strategy aiming at enhanced building footprint extraction from high-resolution satellite images and compares it with state-of-the-art methods with experimental results.

TAL: Topography-Aware Multi-Resolution Fusion Learning for Enhanced Building Footprint Extraction

Yifan Wu, Linlin Xu, Yuhao Chen, Alexander Wong, and David A. Clausi

Abstract

Automatic building footprint extraction from remote sensing imagery is a challenging task with important applications in geomatics and environmental science. Significant advances have been made in this field as a result of the emergence of deep convolutional neural networks (CNNs) designed for semantic segmentation. Although CNNs have demonstrated state-of-the-art performance in coarse annotation and identification of buildings, the boundary accuracy of extracted building footprints is still insufficient for high-precision applications such as mapping and navigation [8, 39, 7]. We propose the topography-aware multi-resolution fusion learning strategy tailored to the problem of enhanced building footprint extraction. More specifically, we introduce a topography-aware loss (TAL) for enhancing a deep CNN's ability to learn heterogeneous building features for better boundary preservation during segmentation. We then incorporate the proposed TAL loss within a multi-resolution fusion architecture to boost high-resolution segmentation performance. Finally, we introduce a novel metric named average thresholded contour accuracy (tCA) which specifically measures the accuracy of segmentation boundaries. The experimental results on the SpaceNet buildings dataset [37] show significant improvements in boundary integrity of extracted building footprints when compared with previously proposed methods. Hence, this method enables accurate boundary annotation toward automatic production of building footprint maps for high-precision applications.

2.1 Introduction

Building footprint extraction from satellite images and aerial images is an important task in remote sensing. The automation of this task is an active research area with applications such as cartography, disaster response, and urban planning. The development of convolutional neural networks (CNNs) such as fully convolutional networks (FCNs) [22] and

U-Net [33] benefits it by providing a semantic segmentation approach. In this approach, a label is predicted for every pixel of an image. While pixel-based predictions give promising coverage, they have precision issues at the delineation of object boundaries [19, 5, 39]. However, accurate boundary annotation is a pre-requirement for applications in automatic mapping and land surveying. In addition, the widely adopted intersection-over-union (IoU) [22] metric only evaluates the area coverage and lacks the ability to assess boundary accuracy.

Building boundaries are considered the most important features of a building footprint because they define the shape and location of a building [39, 7]. However, accurate predictions with sharp corners and straight walls are difficult to achieve due to limitations of typical CNN architectures [27, 10]. Typical CNNs use a series of encoding stages which reduce the spatial resolution of feature maps [22, 2]. This low-resolution representation leads to reduced localization accuracy. To make predictions at the input resolution, decoding stages are used to recover the lost information by upsampling the low-resolution representation in multiple stages, usually with the aid of additional information from the encoding stages [33, 21, 6]. U-Net improves upon FCNs by introducing skip connections and multiple upsampling stages. FPN [21] combines low-resolution features and high-resolution features via lateral connections. DeepLab [5] uses a fully connected Conditional Random Field (CRF) [19] to capture fine edge details. However, the lack of high-resolution representations still hinders the accurate prediction of pixel labels on object boundaries [6].

Furthermore, the complex environmental conditions in remote sensing imagery also raise challenges in boundary-accurate building footprint extraction [25]. The immediate neighboring region of building boundaries contains heterogeneous features due to environmental conditions including visible sidewalls, partial coverage from trees, shadows, and small irregular structures. Figure 2.1 shows some examples. In contrast, the inner region of buildings is often homogeneous with simple texture. This imbalance in the feature variety at different topology levels makes the building boundaries difficult to predict. The softmax loss function used in CNNs penalizes incorrect prediction equally across the whole image. While this is desired for an even distribution of features in nature images, buildings in remote sensing imagery require a different design.

Early works on building footprint extraction focus on adopting generic segmentation CNNs. Maggiori *et al.* [25] uses the FCN architecture to target building extraction. Ji *et al.* [18] and Iglovikov *et al.* [17] modify the popular U-Net to improve the performance. Wei *et al.* [39] add a multiscale aggregation in an FCN and use a polygon regularization algorithm to refine building boundaries. Shao *et al.* [35] combine an encoder-decoder structure and a residual refinement module which enhance the accuracy of boundaries. Other recent works use additional building information for footprint extraction. Yuan [41] proposes the signed distance function as an output representation for building footprint

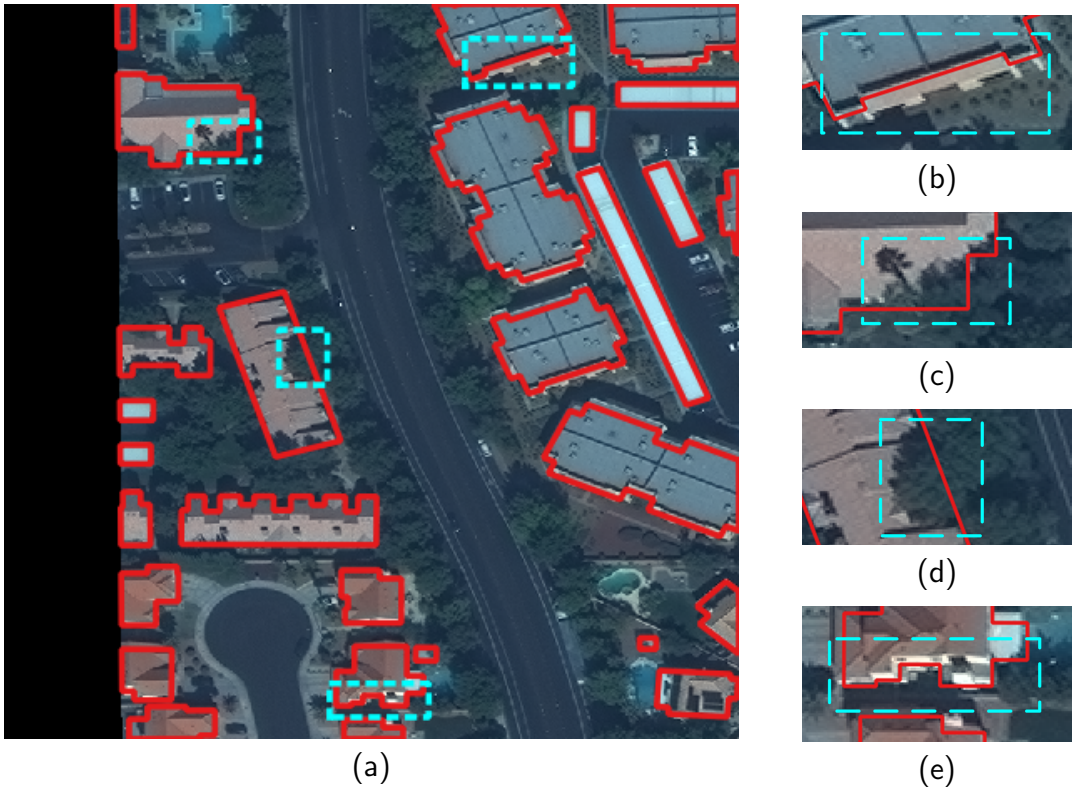

Figure 2.1: (a) High-resolution RGB spectral satellite image with building boundary annotation overlay in red. Cyan boxes highlight environmental conditions including (b) visible sidewalls, (c)(d) partial coverage from trees, shadows, and (e) small irregular structures.

labels. Bittner *et al.* [3] fuse features extracted from the three-band, panchromatic, and normalized digital surface model images in hybrid FCN. Guo *et al.* [13] consider the scene prior knowledge in a multitask parallel attention network to improve robustness. The works mentioned above add complexity to the existing CNNs by requiring additional inputs and modifying network architectures, making them difficult to adapt to various state-of-the-art segmentation networks.

In this letter, we propose a topography-aware loss (TAL), a topography-aware multi-resolution fusion learning strategy for enhanced building footprint extraction. TAL is a simple yet effective loss weight function suitable for any network without additional input requirements. More specifically, TAL adapts to building topology and helps the network to learn boundary features without adding complexity from extra data or a change of

output representation. We incorporate TAL into an architecture that provides a high-resolution feature representation by fusing multi-resolution parallel convolution streams for information exchange at different scales. As a result, the semantic representation becomes more precise without the loss of localized information. To effectively evaluate the building boundary accuracy, we introduce the average thresholded contour accuracy (tCA) as an evaluation metric in addition to IoU. The experimental results on the SpaceNet buildings dataset [37] demonstrate the effectiveness of our method compared with the existing semantic segmentation models.

2.2 Methodology

2.2.1 Topography-Aware Loss (TAL) Function

Buildings in remote sensing images often have a feature imbalance problem associated with the nature of building distribution and characteristics. To solve this problem, we propose a TAL function that adapts to building topography characteristics in remote sensing imagery. More specifically, each pixel in the segmentation mask is weighted by two components. The first component addresses the learning imbalance problem associated with background and building regions which can lead to predictive bias. The number of pixels associated with background regions is often many times the number of pixels with building regions and thus can cause much of the network's capacity to be spent learning the background instead of the buildings. We introduce a region weight for each class, which is inversely proportional to the pixel occurrence of the class in the training set. Reflecting on the binary segmentation natural of building footprint extraction, this region weight is further balanced with a factor to cooperate with the design of the TAL function.

The second component addresses the imbalanced feature characteristics of buildings. The neighboring region of building boundaries often contains difficult to learn features with high variation, while the inner region of buildings has relatively simple features. We introduce a boundary weight, which is assigned relative to the distance from the pixel to the nearest building boundary. It gives extra weight to pixels close to the boundaries and forces the network to learn features in those areas. Figure 2.2(b) shows an example of pixel-wise topography weight visualization for a training image.

Based on the proposed TAL, we compute the loss weight w for each pixel x of each image in the training set

$$w(x) = c(x) \cdot w_r(x) + w_b \cdot \exp\left(-\frac{d(x)^2}{2\alpha^2}\right) \tag{2.1}$$

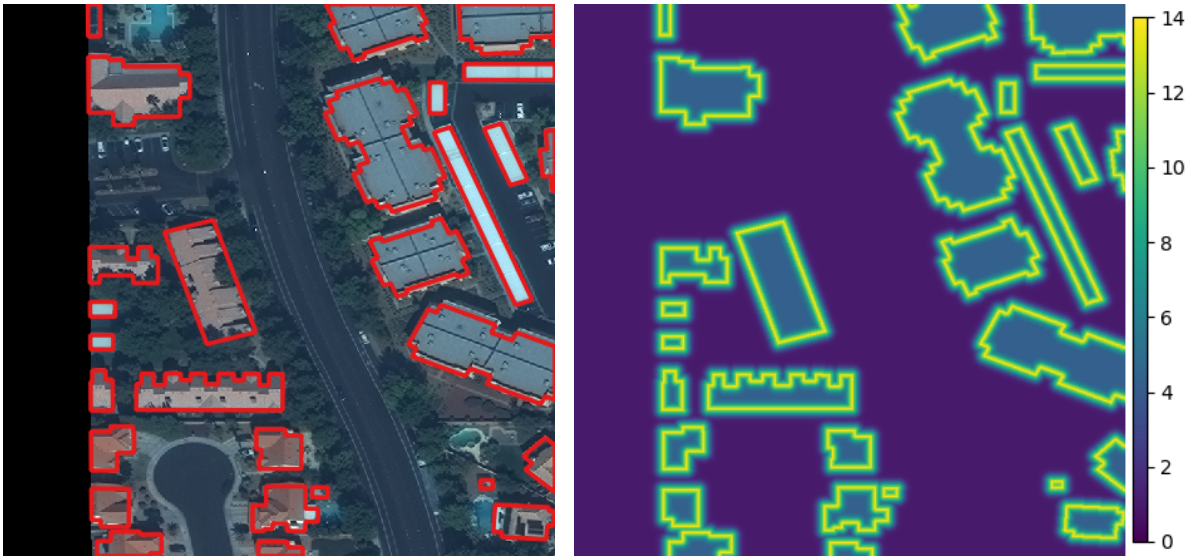

Figure 2.2: (a) Original image with ground truth (GT) building boundary annotation overlay. (b) Pixel-wise TAL weight map derived from GT building boundaries with $w_b = 10$ and $\alpha = 5$. As can be seen from (b), by assigning very high loss weights to pixels close to GT building boundaries, the TAL weight map helps the CNN model focus on learning the building boundary characteristics for more accurate boundary delineation. Moreover, by assigning higher loss weights to building pixels than background pixels that greatly outnumber the building pixels, TAL helps alleviate the imbalanced classes issue.

where c is the class balancing factor, w_r is the region weight, w_b is the boundary weight coefficient, d denotes the distance to the nearest building boundary, and α^2 denotes the variance of the boundary weight distribution. We design the region weight component and boundary weight component to work cooperatively in a single equation by introducing the class balancing factor c to avoid extreme difference in loss weight w between pixels after combining the two components.

2.2.2 Multi-Resolution Fusion Learning

For multi-resolution fusion learning, we leverage a multi-resolution fusion architecture based on HRNet [38] with the incorporation of the proposed TAL loss as shown in Fig-

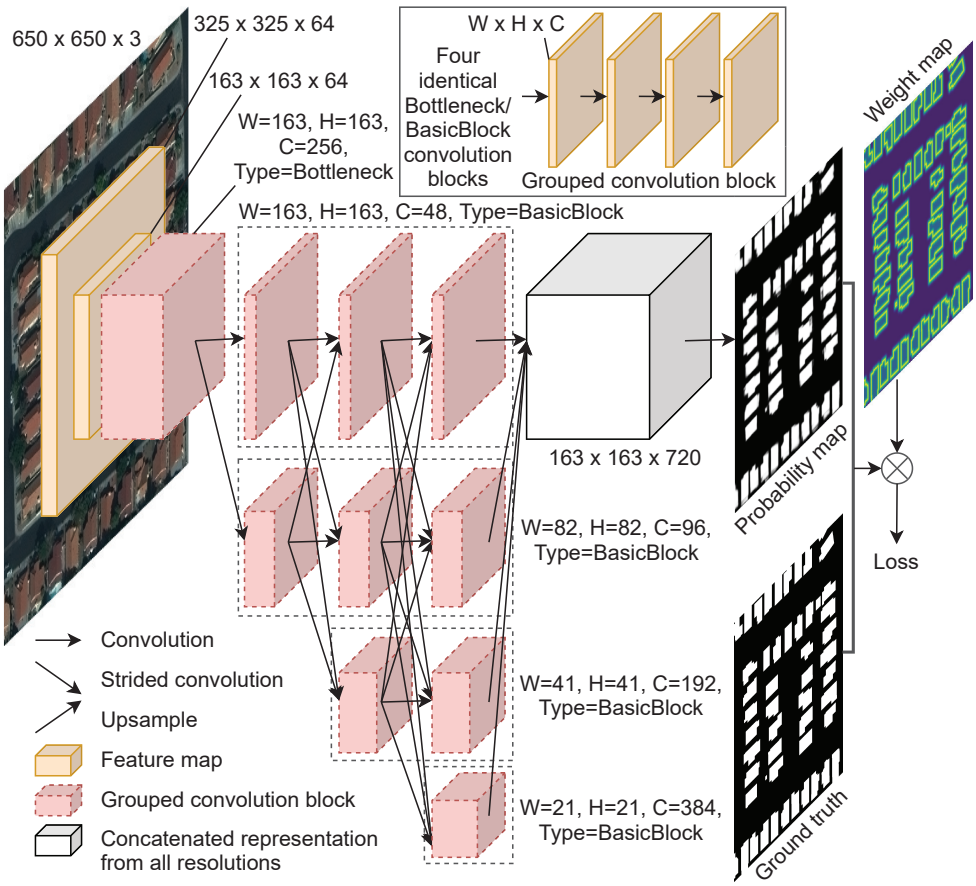

Figure 2.3: Multi-resolution fusion architecture with the incorporation of the proposed TAL loss.

ure 2.3. The multi-resolution convolution streams include four parallel streams with high-to-low resolution. The network starts with a high-resolution stream and adds lower resolution stream branches while keeping the higher resolution streams through the entire network. In this way, the network maintains feature representation at different resolution levels. Fusion modules are used to aggregate feature representation from different streams using striding or upsampling, which enables the exchange of information between multi-resolution streams. It is used between multi-resolution streams repeatedly to form a multistage fusion design. Thus, the multi-resolution fusion architecture provides a representation with information combined from different spatial resolutions.

2.3 Experiments

2.3.1 Dataset and Training Setup

We evaluate and compare our method using the SpaceNet buildings dataset [37] because of its high-resolution RGB spectral satellite images and fine-grained building footprint annotations. We choose the area of interest in Las Vegas for its large number of building samples and good consistency across building labels. It covers an area of 216 km^2 and contains 151,367 building labels. The RGB spectral images are collected from the WorldView-3 satellite at about 1.3-m spatial resolution and pansharpened to 0.3-m spatial resolution. They are provided as 3,851 tiles with the size of 650 $\times$ 650 pixels and the corresponding building polygons in GeoJSON format. We randomly choose 60% of tiles for training, 20% of tiles for validating and 20% of tiles for testing. The polygon format building labels are preprocessed to produce ground-truth binary masks and pixel-wise weight maps for training.

The weights of our network are initialized with a model pretrained on ImageNet [34] to take advantage of the generic features learned and speed up convergence. For training, we use an initial learning rate of 0.007, SGD optimizer with a momentum of 0.9, and a weight decay of 0.0005. Random cropping, scaling, and flipping are disabled for all the models because they are not beneficial for remote sensing images with fixed spatial resolution and viewing angle. The models are trained for 30 epochs with a batch size of 2 on an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of memory. The proposed TAL method has a similar training time and the same inference time compared with HRNet. We use the inverse of pixel occurrence in the training set for region weight w_r of background and building pixels in (2.1). The class balancing factor c normalizes w_r so that the region weight component of any background pixels is 1. This normalization technique is necessary to prevent degradation of performance caused by extreme loss weight values. Our experiments find $w_b = 10$ and $\alpha = 5$ in (2.1) yield strong result.

2.3.2 Average Thresholded Contour Accuracy (tCA)

In semantic segmentation, predictions are stored in the form of binary masks. Each pixel in the mask belongs to either the foreground class or the background class. Sets of contours for the foreground class can be derived from each binary mask, and we consider each contour the boundary of a building. We use a bipartite graph matching with morphology approximation [31] between ground-truth contour and predicted contour to classify pixels

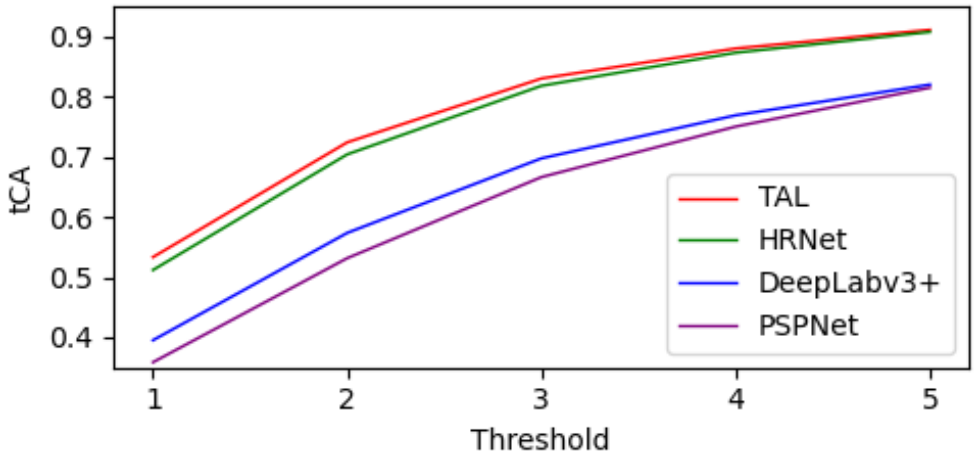

Figure 2.4: Contour accuracy at thresholds from 1 to 5 pixels. The proposed TAL method is in red; HRNet [38] in green; DeepLabv3+ [6] in blue; PSPNet [43] in purple.

into true positives (TP), false positives (FP), and false negatives (FN). A threshold distance in pixels is applied to this calculation, where matches within this threshold buffer are considered true positives. The average thresholded contour accuracy (tCA) based on F-score is defined as

$$\text{tCA}_{\text{average}} = \frac{1}{n}\sum_{t=1}^{n}\frac{2P_tR_t}{P_t + R_t} \tag{2.2}$$

where $P_t = \text{TP}_t/(\text{TP}_t + \text{FP}_t)$ and $R_t = \text{TP}_t/(\text{TP}_t + \text{FN}_t)$ are the contour precision and recall at t pixel threshold, respectively [31]. We use multiple threshold values to generate multiple tCA values as an unbiased comprehensive metric that tends to be more independent of threshold distances. More specifically, we use the averages of these tCA values, i.e, $\text{tCA}_{\text{average}}$ as quantitative metric (as in Table 2.1), and also use the tCA curve as visual evaluation, as shown in Figure 2.4.

We measure the segmentation performance with both IoU and tCA for building footprint prediction. IoU evaluates segmentation in terms of labeled area, but it lacks the ability to localize the variation in performance. In contrast, the proposed tCA is computed from contour precision and recall between predicted boundaries and ground truth. We take measurements at thresholds from 1 to 5 pixels and calculate the average tCA for unbiased comprehensive boundary accuracy evaluation.

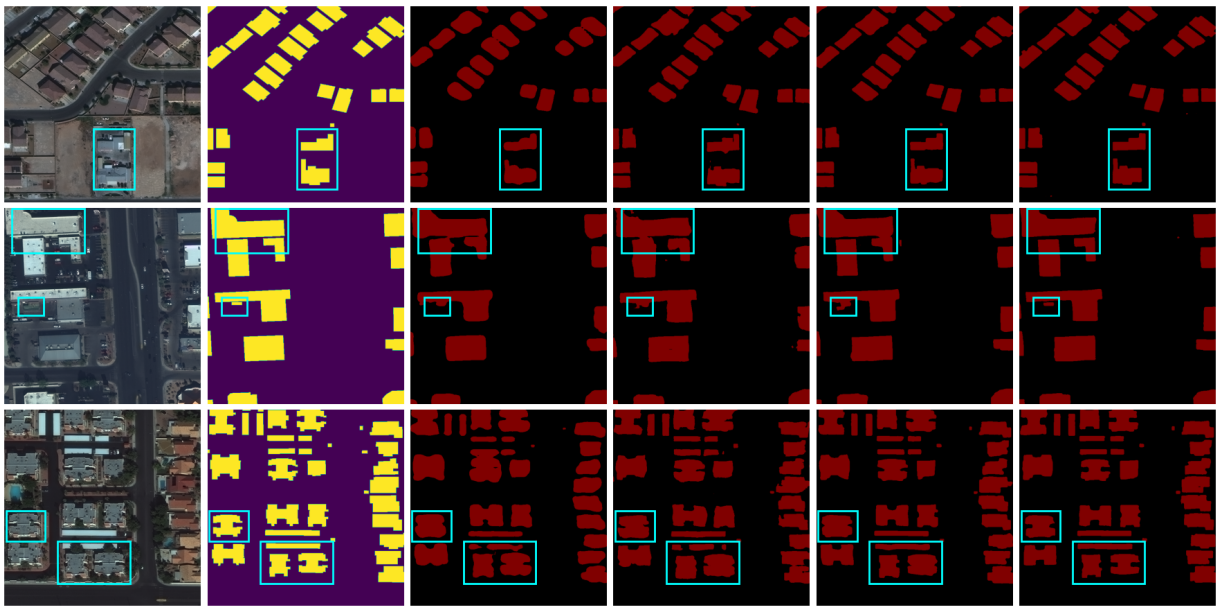

Figure 2.5: Three test scenes. From left to right are RGB images, GT binary masks, results by PSPNet [43], DeepLabv3+ [6], HRNet [38], and the proposed TAL method, respectively.

2.3.3 Results and Analysis

We compare the proposed TAL method with the state-of-the-art semantic segmentation networks, namely DeepLabv3+ and HRNet. Figure 2.5 shows segmentation results in three scenes. The first scene shows that the proposed TAL method is able to better preserve small building structures and predicate sharp building corners. In the second scene, shadows and visible sidewalls between adjacent buildings lead to difficulties for PSPNet, DeepLabv3+, and HRNet to correctly define building boundaries. Segmentation from our method shows clean separation between those adjacent buildings while maintaining straight boundaries and sharp corners. The third scene includes buildings with partial tree coverage and complex small structures with shadows and visible sidewalls. Our method shows an overall improvement in reproducing "H"-shaped buildings with correct structures. Furthermore, if we pay attention to the long rectangle building highlighted by the box in the middle, both PSPNet and DeepLabv3+ fail at inferring the building segmentation under tree coverage, and HRNet makes a prediction with curved building boundaries. In comparison, our method produces a building segmentation with straight edges without being affected by the partial tree coverage.

Table 2.1: Segmentation results on 770 images in the test set. $\text{IoU}_{\text{building}}$ is taken for building predictions over all image pixels. $\text{tCA}_{\text{average}}$, $\text{P}_{\text{average}}$, and $\text{R}_{\text{average}}$ are averaged for thresholds from 1 to 5 pixels.

Method	$\text{IoU}_{building}$	$\text{tCA}_{average}$	$\text{P}_{average}$	$\text{R}_{average}$
PSPNet [43]	78.47%	62.42%	66.43%	58.89%
DeepLabv3+ [6]	78.75%	65.09%	64.86%	65.33%
HRNet [38]	85.21%	76.27%	77.57%	75.03%
TAL region weight only	85.49%	77.27%	**78.26%**	76.30%
TAL boundary weight only	85.61%	77.52%	78.10%	76.96%
TAL	**85.62%**	**77.55%**	78.09%	**77.03%**

The results summarized in Table 2.1 show that the proposed method improves on both building IoU and average tCA. Ablation experiments with TAL boundary weight only and TAL region weight only show the improvements brought by each component in (2.1). Although the improvements in building IoU and average tCA are numerically small when comparing with HRNet, both segmentation results in Figure 2.5 and contour accuracy curves in Figure 2.4 demonstrate the effectiveness of TAL loss. Figure 2.4 shows that the TAL performs particularly well in the low threshold value range where the allowed error margin is small and the tCA of the proposed TAL method and HRNet only coverage as the threshold increases. The proposed TAL function assigns higher loss weights to pixels that are close to GT building boundaries as shown in Figure 2.2(b). This helps the CNN model to focus on learning the building boundary characteristics for more accurate boundary delineation. The most significant improvement of segmentation result characterized by contour accuracy is found in the area with the highest loss weight assigned.

Although the SpaceNet buildings dataset provides high spatial resolution images and building footprint annotations with good overall accuracy, there are some inherent limitations because it is designed for an object detection problem. In such a problem, achieving a true positive building detection only requires a building footprint IoU greater than 50%. Most small errors and inconsistencies in annotations would not affect the evaluation result. However, here we evaluate building segmentation based on pixel-wise classification results in terms of IoU and tCA. Figure 2.6 shows some problems of building footprint annotations in the dataset that would cause inaccurate evaluation. Other similar datasets suffer from similar problems and often more extensively [28, 24, 29].

Figure 2.6: (a) (b) Building footprint annotations include visible sidewalls and exclude part of the roof for tall buildings because orthorectification is not applied to satellite images. (c) (d) Inconsistent building footprint annotations for buildings with similar structures. (d) Oversimplification of small building structures.

2.4 Conclusion

TAL, a topography-aware multi-resolution fusion learning method for enhanced building footprint extraction has been designed and implemented, and then compared with other state-of-the-art methods. A TAL function was proposed that adapts to the underlying building topography to better learn and preserve building boundaries characterized in remote sensing imagery. Moreover, the proposed loss function was incorporated into a multi-resolution fusion architecture that better captures information at different spatial resolutions to enable a more precise high-resolution representation when compared with encoder-decoder architectures. Finally, we defined an average tCA which is tailored for measuring of boundary accuracy to supplement IoU metric. The proposed TAL method was demonstrated on the SpaceNet buildings dataset to exhibit enhanced building boundary accuracy and overall footprint extraction improvement when compared with state-of-the-art methods. Future work includes investigations of additional priors that can be incorporated to further improve trade-offs between boundary preservation and footprint coverage. We will also investigate data argumentation techniques, e.g., synthetic image-label pair generation via generative adversarial networks, which could reduce the impact of building footprint annotation errors in our task.

Chapter 3

Multi-Task Edge Detection for Building Vectorization from Aerial Images

The following article has been submitted to the journal IEEE Geoscience and Remote Sensing Letters. The text may be modified later for the journal submission.

This article presents a multi-task edge detection framework aiming at building outline vectorization from very-high-resolution aerial images and compares it with state-of-the-art methods with experimental results.

Multi-Task Edge Detection for Building Vectorization from Aerial Images

Yifan Wu, Linlin Xu, Lei Wang, Qi Chen, Yuhao Chen, and David A. Clausi

Abstract

The extraction of building outline vectors is an essential task in supporting various applications. Although the recent development of deep learning based techniques has made advancements in the automation of this task, the accuracy and precision are insufficient due to errors caused by abundant noise and obstruction from shadows and trees around buildings in aerial images. To better address this issue, this paper presents a new approach called the multi-task edge detection (MTED) for building vectorization with the following characteristics. First, instead of detecting building corner points that are very sensitive to noise effects, a deep learning based rotated bounding box detector is introduced for building edge detection to increase robustness to interference. Second, a multi-task learning strategy is designed to integrate building segmentation inside the METD framework to closely guide the edge detection using spatial context. Third, a simple yet effective geometry-guided post-processing method is designed to reconstruct vectorized building outlines based on the detected edges and learned building shape prior knowledge. The comparative experiments conducted on a benchmark of very-high-resolution optical aerial images indicates that the proposed approach can significantly outperform the state-of-the-art methods in terms of vertex-based building outline accuracy metrics.

3.1 Introduction

High-precision building vector maps are essential to support various applications such as cartography, urban planning, and disaster response [41, 39, 7]. Many machine learning and computer vision approaches have been designed for building detection to support automatic vector map generation. Traditional approaches rely on the extraction of building signature information using textures, edges, shapes, shadows, and context information [36, 16, 26]. They utilize feature engineering and cannot adaptively accommodate the data characteristics for discriminative feature mining. Recently, deep learning techniques have been used to better extract discriminative features in a data-driven manner, leading to the

state-of-the-art performance in building outline vectorization [44, 27, 10, 7]. Nevertheless, although these deep learning based approaches have significant improvements compared with traditional feature engineering approaches, they still struggle with building signature ambiguity and noise effect in aerial images, leading to some key research gaps.

First, although detecting building edges, instead of detecting building corner points, can better accommodate noise and obstruction effects, effective edge detection approaches have not been tailor-designed to enhance building vectorization. Most building outline vectorization approaches highlight the role of building corner points in constructing the building polygon, and try to accurately estimate and locate building corner points, which are then connected to form the building polygon [27, 10, 7]. However, building corner points are subtle targets that are incredibly vulnerable to interference. Moreover, corner points are small targets that do not enable the large-receptive-field-detector to better leverage large-scale spatial context [42, 45]. In contrast, building edges are more salient and less susceptible to noise and complete coverage from shadows and nearby trees. Furthermore, edges are elongated and larger targets allow the object detector to fully leverage large-scale spatial context, as shown in Figure 3.1, to improve detection robustness and accuracy. Nevertheless, given the benefit of edge detection, it is difficult to design a deep learning based edge detector due to the randomly oriented and aligned nature of building edges. Recently, a rotated bounding box representation has been successfully utilized for arbitrary-oriented scene text detection [23], and other rotated object detection [40], which have the potential to be adapted for building edge detection. Therefore, it is vital to explore rotated bounding box representations for effective building edge detection that can better resist the noise and obstruction for enhanced building outline vectorization.

Second, without building segmentation regularization and guidance, edge detection tends to be either misled by isolated non-building edges, e.g., roadside, or weak building edges with tree coverage, leading to big commission errors and omission errors. Although a multi-task learning network that combines building edge detection and building segmentation can enable better usage of segmentation information to improve edge detection, this approach has not been sufficiently investigated for enhanced building outline vectorization. Early CNN-based object detection approaches, e.g., Faster R-CNN [32], tend to not leverage segmentation masks for improving performance. Although there exists segmentation masks incorporated object detection approaches such as Mask R-CNN [14], they are not tailor-designed for building edge detection, e.g., they do not use a rotated bounding box proposal, and as such, do not allow the accurate detection of rotated objects. Some rotated bounding box based approaches have been designed, but they tend not to use guidance from the segmentation task [23, 40]. Therefore, it is critical to design a multi-task learning strategy that integrates building segmentation with the rotated bounding box based

Figure 3.1: (a) Building polygon with corner points highlighted in green. Colored dots in (b)-(d) illustrate feature sampling points grouped for different points or edges. (b) Point detector has the most limited field-of-view. (c) Line based edge detector has improved field-of-view. (d) In comparison, rotated bounding box based edge detector has the largest receptive field to leverage spatial context.

building edge detection to allow mutual performance improvement of edge detection and segmentation.

Third, given detected building edges and learned building shape priors, a simple yet effective geometry-guided building polygon reconstruction method is required to achieve building outline vectorization. Although a geometry-guided post-processing approach based on accurate edges could lead to an effective and fast building polygon reconstruction, it has not been fully explored. Most recent deep learning based approaches tend to ignore this explicit knowledge integration process [44, 7, 9], and thereby cannot effectively leverage the learned knowledge regarding the prior geometry shape for building outline vectorization. In fact, given accurate building edges with relatively low commission and omission errors, a geometry-guided approach can effectively locate and estimate building corners by leveraging the interaction among adjacent edges and shape prior knowledge learned from building segmentation. Therefore, it is important to investigate how to effectively use building shape information to achieve simple yet effective reconstruction that transforms the detected building edges into building polygons.

This paper therefore presents the MTED framework consisting of a deep learning based building edge detection strategy and a geometry-guided building polygon reconstruction method for improved building outline vectorization as illustrated in Figure 3.2, with the following key contributions:

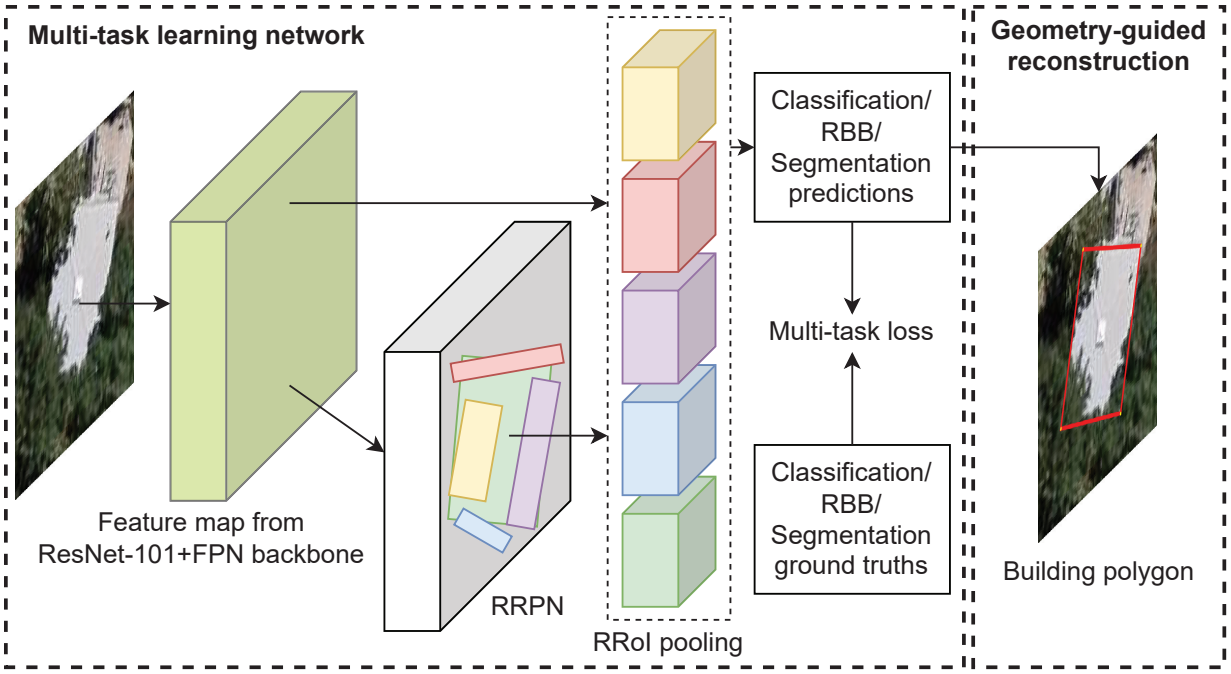

Figure 3.2: The proposed MTED framework consists of a deep learning based building edge detection strategy described in Section 3.2.1, 3.2.2 and a geometry-guided building polygon reconstruction method described in Section 3.2.3.

1. A rotated bounding box based building edge detection approach is introduced to overcome the drawbacks of detecting building corner points by better accommodating noise and obstruction effects in aerial images.

2. A multi-task learning strategy is designed to integrate building segmentation with building edge detection to improve edge detection and segmentation performance through mutual guidance and regularization.

3. A simple yet effective geometry-guided building polygon reconstruction method is designed to effectively leverage building shape information for transforming detected building edges into building polygons.

This paper is organized as follows. Section 3.2 introduces the detailed implementation of the proposed framework. Section 3.3 presents the results of comparative experiments with several state-of-the-art building outline vectorization approaches. Section 3.4 concludes this study.

3.2 Methodology

As illustrated in Figure 3.2, the proposed MTED framework consists of three key components: (1) rotated bounding box detector for building edge detection, (2) building segmentation in multi-task learning network, and (3) a geometry-guided building polygon reconstruction, which are described in detail below.

3.2.1 Rotated Bounding Box for Building Edge Detection

To address the randomly orientated nature and also to better accommodate the noise effect, a rotated bounding box (RBB) approach is designed to represent building edges for enhanced edge detection. Specifically, each RBB is represented by a tuple (x, y, l, w, θ), where (x, y) denotes the center of the RBB, l, w and θ are respectively the length, width, and orientation of the RBB. Therefore, a building edge with endpoints (x_1, y_1) and (x_2, y_2), is represented as

$$\begin{aligned} x &= \frac{x_1 + x_2}{2} \\ y &= \frac{y_1 + y_2}{2} \\ l &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\ w &= A \\ \theta &= \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \end{aligned} \tag{3.1}$$

where A is a hyperparameter applied to all bounding boxes. A sufficiently large A value enables RBBs bigger field-of-view to better accommodate noise and obstruction as shown in Figure 3.1(d).

To predict RBBs, two loss functions for respectively bounding box classification and regression are designed. Given K true classes beside background class, the classification loss for true class $u \in \{1, 2, \ldots, K\}$ is defined as

$$L_{class}(p, u) = -\log p_u \tag{3.2}$$

where $p = (p_0, p_1, \ldots, p_K)$ is a discrete probability distribution computed by a softmax over the $K + 1$ outputs of a fully connected layer. The regression loss is defined between

a ground truth tuple $v = (v_x, v_y, v_l, v_w, v_\theta)$ and a predicated tuple $t = (t_x, t_y, t_l, t_w, t_\theta)$

$$L_{box}(t, v) = \sum_{i \in \{x,y,l,w,\theta\}} \text{smooth}_{L_1}(t_i - v_i)$$
$$\text{where } \text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3.3}$$

As illustrated in Figure Figure 3.2, we use ResNet-101 [15] with Feature Pyramid Network (FPN) [21], which combines semantic features from different resolutions, as the backbone to build a feature map from the input image. The Rotation Region Proposal Network (RRPN) [23] is adopted to generate candidate bounding boxes of building edges. RRPN is capable of sampling anchors with three sets of parameters, namely scales, aspect ratios, and angles. In comparison to the Region Proposal Network (RPN) [32] used in Mask R-CNN, RRPN enables the proposal of anchors at different orientations instead of axis-aligned anchors only. To pair with RRPN, the Rotation Region-of-Interest (RRoI) [23] pooling layer extracts features inside any valid region of interest (RoI) from the feature map produced by the backbone and converts them into a feature map with a fixed spatial extent determined by hyperparameters $H \times W$. The features from the ResNet-101+FPN backbone are pooled for each RBB to generate a RBB tuple representation, which is optimized through backpropagation based on the above loss functions.

3.2.2 Building Segmentation in Multi-Task Learning Network

In addition to the RBB classification task and regression task defined in the previous section, a building semantic segmentation task is introduced to enhance building edge detection through mutual guidance and regularization. The segmentation mask loss is an average binary cross-entropy loss that is only defined for the true class associated with each RoI as

$$L_{mask}(b^*, b) = -\frac{1}{m^2} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} [b_{ij} \log b^*_{ij} + (1 - b_{ij}) \log(1 - b^*_{ij})] \tag{3.4}$$

where m is the width and height of the mask, b^*_{ij} and b_{ij} denote the values of the label at location (i, j) for the predicted mask and the ground truth mask respectively.

In the multi-task learning network, we define the multi-task loss $L = L_{class} + L_{box} + L_{mask}$ to train the three tasks simultaneously.

3.2.3 Geometry-Guided Reconstruction

Given detected edges, we post-process them to achieve building polygon reconstruction based on prior knowledge concerning building shapes and edge interaction patterns. Specifically, we first determine the relative order of building edges based on learned building shape information as shown by the pseudo-code in Algorithm 1. The predictions are preprocessed before being used as inputs. The segmentation mask is traced to produce a contour of the building consisting of a sequence of points $\{C_Y\}$. Any bounding boxes that do not intersect with the contour are ignored as anomalies. By reversing the calculations in (3.1), we get a sequence of building edge line segments $\{L_X\}$ originally represented by the rotated bounding boxes. In the next step, we find the index of the contour point that is closest to the midpoint M of each line segment. Lastly, input line segments are sorted based on those indexes to output the sequence of ordered line segments $\{O_Z\}$.

Algorithm 1 Relative order of building edges

Input: Line segments $\{L_X\}$, Contour points $\{C_Y\}$
Output: Ordered line segments $\{O_Z\}$

$X \leftarrow \text{length}(L_X)$
$Y \leftarrow \text{length}(C_Y)$
for $x \leftarrow 1, X$ **do** ▷ Find the index of the contour point that is closest to the midpoint of each line segment
    $M \leftarrow \text{midpoint}(L_x)$
    for $y \leftarrow 1, Y$ **do**
        $distance_y \leftarrow \text{distanceBetween}(M, C_y)$
    end for
    $index_x \leftarrow \text{indexOfMin}(distance_1, distance_2, \ldots, distance_Y)$
end for
for $y \leftarrow 1, Y$ **do** ▷ Sort line segments based on indexes found
    for $x \leftarrow 1, X$ **do**
        if $index_x = y$ **then**
            Appended L_x *to* O_Z
        end if
    end for
end for

Based on this relative order, we use simple yet effective geometry guidance to reconstruct the building polygon as shown by the pseudo-code in Algorithm 2. For every pair of adjacent building edge line segments O_{first} and O_{second} in $\{O_Z\}$, we compute the point of

intersection I of the lines of which the two line segments are respectively a part of. Then we find the two points C_1 and C_2 that are closest to each other on the two line segments O_{first} and O_{second}. If the distance between the point I and one of two points C_1 and C_2 is less than or equal to 10 pixels, the point I is appended to the sequence of vertex points $\{V_A\}$. Otherwise, the midpoint M between C_1 and C_2 is found and the two points P_1 and P_2 that are the closest points to M on O_{first} and O_{second} respectively are appended to the sequence of $\{V_A\}$. The final sequence of vertex points $\{V_A\}$ composes the reconstructed building outline polygon.

Algorithm 2 Building polygon reconstruction

Input: Ordered line segments $\{O_Z\}$
Output: Polygon vertex points $\{V_A\}$

$Z \leftarrow \text{length}(O_Z)$
for $first \leftarrow 1, Z$ **do** ▷ Apply algorithm to every pair of adjacent building edge line segments
    $second \leftarrow (first + 1) \mod Z$
    $I \leftarrow \text{pointOfIntersectionOfLines}(O_{first}, O_{second})$
    $C_1, C_2 \leftarrow \text{closestPointsOnLineSegments}(O_{first}, O_{second})$
    $distance_1 \leftarrow \text{distanceBetween}(I, C_1)$
    $distance_2 \leftarrow \text{distanceBetween}(I, C_2)$
    if $\min(distance_1, distance_2) \leq 10$ **then**
        Appended I *to* V_A
    else
        $M \leftarrow \text{midpoint}(C_1, C_2)$
        $P_1 \leftarrow \text{closestPointOnLineSegment}(O_{first}, M)$
        $P_2 \leftarrow \text{closestPointOnLineSegment}(O_{second}, M)$
        Appended P_1 *to* V_A
        Appended P_2 *to* V_A
    end if
end for

3.3 Experiments

3.3.1 Dataset

We evaluate and compare our method using a benchmark selection of building instances from the Aerial Imagery for Roof Segmentation (AIRS) dataset [8] which provides very-high-resolution optical aerial images at 7.5 cm resolution and strictly aligned building ground truths. This selection contains 10,164 building instance samples that have been previously benchmarked in studies [7]. We follow the same experimental data setup to compare performance with previous studies on building vectorization. Each sample has a background padding region equal to 40% of the building bounding box size capturing surrounding context, plus an additional 30 pixel padding to accommodate a random 0 to 30 pixels offset from the center of the building to avoid the bias of center location. We use the same training set and testing set separation as in the previous studies, where 80% and 20% of the samples are used respectively. In this study, we further divide the training set and use 75% and 25% of the samples for training and validating respectively. The ground truth building polygons are pre-possessed to use rotated bounding box representations for building edges and mask representations for buildings.

3.3.2 Training Setup

The minimum and maximum input image sizes of the multi-task learning network are set to 256 pixels and 512 pixels respectively to accommodate buildings with different dimensions. Input images with both sides smaller than 256 pixels or larger than 512 pixels are scaled into the input range without changing the aspect ratio to minimize image degradation. For sampling anchors, we use scales of 16, 32, 64, 128 and 256, aspect ratios of 1:1, 1:2, 1:4, 1:8 and 1:16, and angles of $\frac{\pi}{2}$, $\frac{\pi}{3}$ and $\frac{\pi}{6}$ to cover building edges with various length and orientation combinations. We find that a constant bounding box width A of 10 pixels works with those sampling parameter sets giving strong results. We set base learning rate of 0.0025, SGD optimizer with momentum of 0.9, and weight decay of 0.0001. Data augmentations are disabled, and other hyperparameters use the same values from the Mask R-CNN implementation. The ResNet-101+FPN backbone of our multi-task learning network is initialized with weights from a model pretrained on ImageNet [34]. Our model is trained for 3 epochs with 400 warn-up iterations using a batch size of 4 on an NVIDIA GeForce RTX 2080 Ti GPU with 11GB of memory.

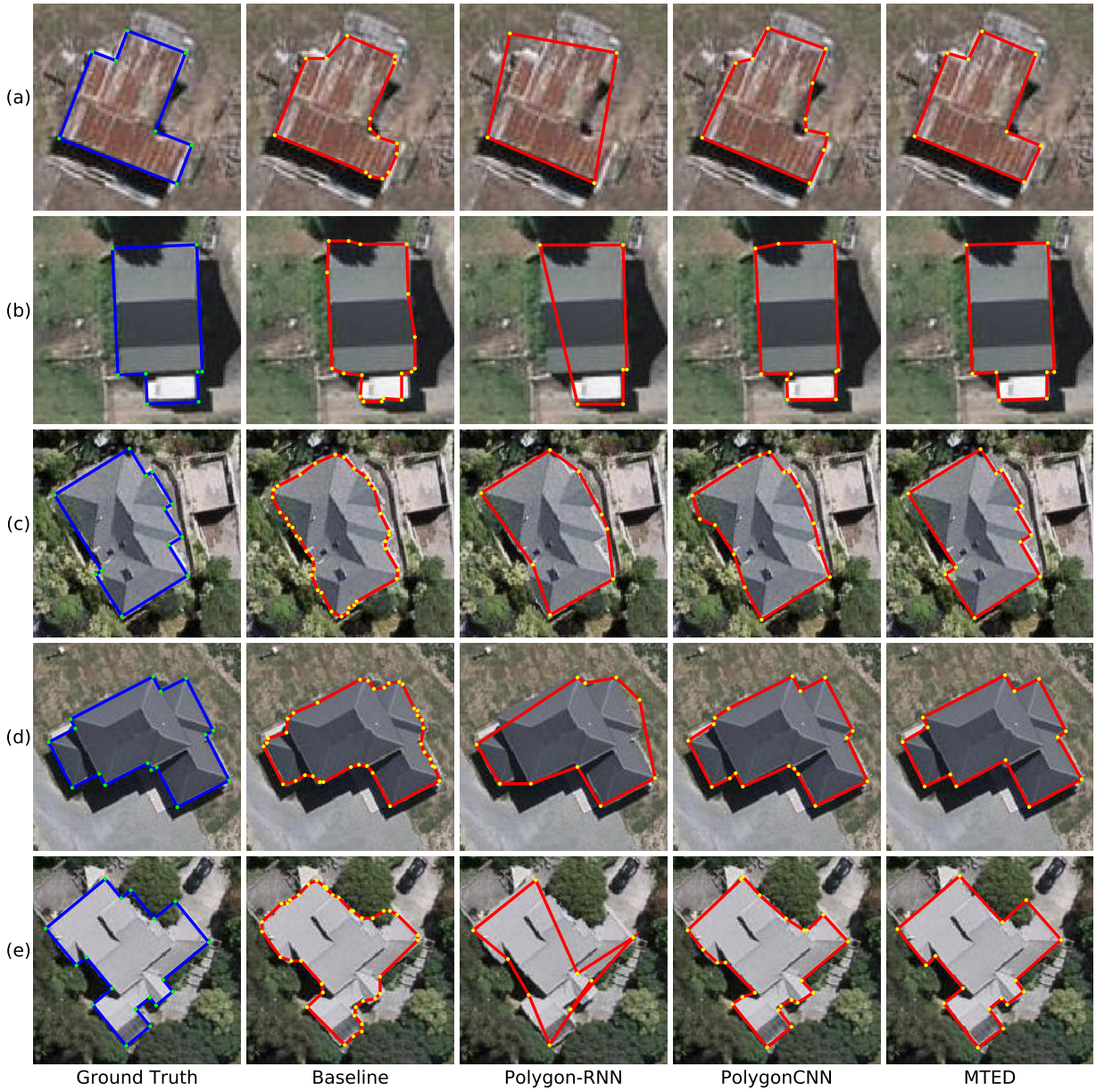

Figure 3.3: Examples of building vectorization results produced by the baseline method, Polygon-RNN [4], PolygonCNN [7], and the proposed MTED framework. Ground truth polygons are shown in blue with corner points highlighted in green. Generated result polygons are shown in red with corner points highlighted in yellow.

3.3.3 Results and Analysis

Figure 3.3 shows examples of building vectorization results produced by the proposed MTED framework compared with the baseline method, Polygon-RNN [4], and state-of-the-art PolygonCNN [7]. The baseline method uses the same building segmentation produced by the MTED method to trace building contours, followed by applying the Douglas–Peucker (DP) algorithm [11] with a threshold of 1 pixel to generate building polygons. Polygons generated by the baseline method reproduce the building outlines with good accuracy and have high coverage of the buildings, which shows that the building segmentation produced by the multi-task learning network can provide accurate building shape prior knowledge for polygon reconstruction purposes. In contrast, Polygon-RNN generates over-simplified polygons with poor coverage and missing building corner points, resulting in distorted building outlines. Although building polygons generated by PolygonCNN have high coverage and simplified vector representations compared with the baseline method that uses the DP algorithm for polygon simplification, it fails to correctly locate building corner points under the effect of noise and obstruction. For example, Figure 3.3(b)(d) shows redundant points that are generated around shadows, Figure 3.3(c)(e) shows building outlines with irregular curves at the location of tree coverage and weak building edges. In comparison, the MTED method generates building polygons using clean vector representations with sharp structures and well-located corner points.

Table 3.1 summaries the experimental results on the testing set between the baseline method, DARNet [10], Polygon-RNN [4], PolygonCNN [7], and the proposed MTED framework. We compare the methods in terms of intersection-over-union (IoU) [22] and vertex accuracy proposed in previous studies [7]. The vertex-based metrics of F1-score, precision, and recall are named VertexF, VertexP, and VertexR respectively. The MTED method outperforms the state-of-the-art PolygonCNN in all vertex-based metrics by a large margin, translating to fewer missing building corner points and fewer incorrectly predicted points. Although IoU of the MTED method lands below PolygonCNN, the standing of the MTED method is not hindered due to the fact that IoU is a metric that evaluates segmentation performance in terms of pixel counts. The limitation of IoU in evaluating vectorization performance can be shown visually by comparing the baseline method, PolygonCNN, and MTED method in Figure 3.3. Our MTED method shows the best overall building vectorization results despite the small loss of IoU. The MTED with separate trainings method is an ablation experiment, where building edge detection task and building segmentation task are trained separately, which shows the performance improvement brought by training the two tasks simultaneously in the MTED framework.

Table 3.1: Summary of experimental results on 2,033 images in the testing set. IoU is computed based on the segmentation masks delineated by the building polygons. VertexF, VertexP and VertexR are computed as the average values at buffer sizes from 1 to 5 pixels.

Method	IoU	VertexF	VertexP	VertexR
Baseline	88.4%	20.3%	14.5%	33.5%
DARNet [10]	77.1%	5.9%	4.9%	7.3%
Polygon-RNN [4]	67.7%	37.1%	47.8%	30.4%
PolygonCNN [7]	**88.6%**	41.7%	40.8%	42.6%
MTED with separate trainings	83.1%	48.8%	54.0%	44.4%
MTED	85.8%	**51.7%**	**57.8%**	**46.8%**

3.4 Conclusion

In this paper, we present the multi-task edge detection framework, combining deep learning based building edge detection strategy and geometry-guided polygon reconstruction method for improved building outline vectorization. A rotated bounding box based building edge detection approach is introduced to increase robustness to noise and obstruction. Moreover, a multi-task learning strategy is designed to integrate building segmentation with building edge detection in the MTED framework to take advantage of spatial context and regularization. Finally, a simple yet effective geometry-guided building polygon reconstruction method is designed to effectively leverage learned building shape prior knowledge and transform predicted building edges into building polygons. The experimental results on the testing set of the very-high-resolution AIRS dataset demonstrate that our MTED framework improves the state-of-the-art performance significantly in terms of vertex accuracy metrics VertexF, VertexP, and VertexR, as well as produces the best overall vectorization results with clean vector representations, sharp structures, and well-located corner points.

Chapter 4

Conclusion

4.1 Summary

In conclusion, automatic building extraction from remote sensing images is an essential task in supporting various applications. Existing traditional knowledge-based approaches and generic deep learning based approaches provide insufficient accuracy for high-precision applications. This thesis proposes two deep learning based methods aiming at enhanced building extraction from high-resolution remote sensing images. In Chapter 2, the topography-aware multi-resolution fusion learning method for building footprint extraction is designed and implemented to overcome the limitation of CNN-based building segmentation methods on the accurate delineation of boundaries. Experimental results on the SpaceNet buildings dataset show improvement in boundary integrity and overall accuracy of extracted footprints compared with state-of-the-art methods. In Chapter 3, the multi-task edge detection framework for building outline vectorization is designed and implemented with increased robustness to noise and obstruction, leading to enhanced accuracy and precision of extracted building polygons. Comparative experimental results on the AIRS dataset demonstrates improvement over state-of-the-art methods in vertex accuracy and overall quality of vector representation generated.

4.2 Future work

There are many directions I want to explore for future work in this field. I want to point out a few limitations of the works presented in this thesis. First, high pixel-based cover-

age of building annotation is hard to achieve with a high level of shape regularization at the same time. The work in Chapter 3 demonstrates signification improvement in vertex-based metrics but is subjected to lower IoU. At the same time, the work in Chapter 2 demonstrates improvement in both coverage-based metric IoU and boundary-based metric tCA though in less remarkable amounts. The possibility of incorporating additional priors such as graph-based relationships into deep learning networks to further improve trade-offs between shape preservation and building coverage can be investigated. Second, the building vectorization framework proposed in Chapter 3 is only capable of extracting building polygons for single building instances. Autonomic extraction of building polygons where multiple buildings appear in a single image is not well studied in the literature. Modifications that have to be made to the proposed framework to accommodate such changes in input data can be investigated. Third, both works in this thesis utilize deep learning models based on CNNs, which are widely adopted in computer vision applications. In the meantime, recent development in deep learning architectures has brought transformers, which are deep learning models initially designed for natural language processing, into the field of computer vision. Those new architectures adopt the self-attention mechanisms of transformers and have shown competitive results in image processing tasks. Adopting transformers based deep learning models in the proposed frameworks has the potential to improve overall performance.

4.3 Final remarks

This thesis has presented two works that enhanced the quality of building annotations extracted from high-resolution remote sensing images. The methods in the works utilized effective CNN-based deep learning techniques to improve the accuracy and precision of building segmentation and building vectorization. The works covered two common categories of optical remote sensing imagery, i.e., satellite imagery and aerial imagery, with different levels of spatial resolutions and thus have a high potential for adaptation to various input data. Evaluation results showed significant improvement brought by the two methods in building boundary integrity and building vector representation accuracy respectively. The improvement provided by the works in the automation of building extraction could benefit many high-precision applications including mapping, navigation, and surveying.

References

[1] Salman Ahmadi, MJ Valadan Zoej, Hamid Ebadi, Hamid Abrishami Moghaddam, and Ali Mohammadzadeh. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *Int. J. Appl. Earth Obs. Geoinf.*, 12(3):150–157, June 2010.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, Dec. 2017.

[3] Ksenia Bittner, Fathalrahman Adam, Shiyong Cui, Marco Körner, and Peter Reinartz. Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 11(8):2615–2629, Aug. 2018.

[4] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5230–5238, July 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, Apr. 2018.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proc. Eur. Conf. Comput. Vis.*, pages 801–818, Sept. 2018.

[7] Qi Chen, Lei Wang, Steven L Waslander, and Xiuguo Liu. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS J. Photogramm. Remote Sens.*, 170:114–126, Dec. 2020.

[8] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L Waslander. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS J. Photogramm. Remote Sens.*, 147:42–55, Jan. 2019.

[9] Yuhao Chen, Yifan Wu, Linlin Xu, and Alexander Wong. Quantization in relative gradient angle domain for building polygon estimation. *Proc. Int. Conf. Pattern Recognit.*, pages 8360–8367, Jan. 2021.

[10] Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. Darnet: Deep active ray network for building segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7431–7439, June 2019.

[11] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122, Dec. 1973.

[12] Haonan Guo, Bo Du, Liangpei Zhang, and Xin Su. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.*, 183:240–252, Jan. 2022.

[13] Haonan Guo, Qian Shi, Bo Du, Liangpei Zhang, Dongzhi Wang, and Huaxiang Ding. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 59(5):4287–4306, May 2021.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2961–2969, Oct. 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, June 2016.

[16] Xin Huang and Liangpei Zhang. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 5(1):161–172, Feb. 2012.

[17] Vladimir Iglovikov, Selim Seferbekov, Alexander Buslaev, and Alexey Shvets. Ternausnetv2: Fully convolutional network for instance segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 233–237, June 2018.

[18] Shunping Ji, Shiqing Wei, and Meng Lu. 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, Jan. 2019.

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.*, 24, 2011.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 25, 2012.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2117–2125, July 2017.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3431–3440, June 2015.

[23] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.*, 20(11):3111–3122, Nov. 2018.

[24] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pages 3226–3229, July 2017.

[25] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.*, 55(2):645–657, Feb. 2017.

[26] Andrea Manno-Kovács and Ali Ozgun Ok. Building detection from monocular vhr images by integrated urban area knowledge. *IEEE Geosci. Remote Sens. Lett.*, 12(10):2140–2144, Oct. 2015.

[27] Diego Marcos, Devis Tuia, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, and Raquel Urtasun. Learning deep structured active contours end-to-end. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8877–8885, June 2018.

[28] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, Department of Computer Science, University of Toronto, 2013.

[29] Sharada Prasanna Mohanty, Jakub Czakon, Kamil A. Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad, Sascha Fleer, Jan Philip Göpfert, Akshat Tandon, Guillaume Mollard, Nikhil Rayaprolu, Marcel Salathe, and Malte Schilling. Deep learning for understanding satellite imagery: An experimental survey. *Front. Artif. Intell.*, 3, Nov. 2020.

[30] Ali Ozgun Ok. Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.*, 86:21–40, Dec. 2013.

[31] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 724–732, June 2016.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.*, 28, 2015.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 234–241, Oct. 2015.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, Dec. 2015.

[35] Zhenfeng Shao, Penghao Tang, Zhongyuan Wang, Nayyer Saleem, Sarath Yam, and Chatpong Sommai. Brrnet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.*, 12(6):1050, Mar. 2020.

[36] Beril Sirmacek and Cem Unsalan. Urban-area and building detection using sift keypoints and graph theory. *IEEE Trans. Geosci. Remote Sens.*, 47(4):1156–1167, Apr. 2009.

[37] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, July 2018.

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, Oct. 2021.

[39] Shiqing Wei, Shunping Ji, and Meng Lu. Toward automatic building footprint delineation from aerial images using cnn and regularization. *IEEE Trans. Geosci. Remote Sens.*, 58(3):2178–2189, Mar. 2020.

[40] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. *Proc. IEEE Int. Conf. Comput. Vis.*, pages 8232–8241, Oct. 2019.

[41] Jiangye Yuan. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2793–2798, Nov. 2018.

[42] Ziheng Zhang, Zhengxin Li, Ning Bi, Jia Zheng, Jinlei Wang, Kun Huang, Weixin Luo, Yanyu Xu, and Shenghua Gao. Ppgnet: Learning point-pair graph for line segment detection. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7105–7114, June 2019.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2881–2890, July 2017.

[44] Kang Zhao, Jungwon Kang, Jaewook Jung, and Gunho Sohn. Building extraction from satellite images using mask r-cnn with building boundary regularization. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 247–251, June 2018.

[45] Yichao Zhou, Haozhi Qi, and Yi Ma. End-to-end wireframe parsing. *Proc. IEEE Int. Conf. Comput. Vis.*, pages 962–971, Oct. 2019.