ISPRS Journal of Photogrammetry and Remote Sensing 201 (2023) 26–37

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images

Weijia Li [a], Wenqian Zhao [b], Jinhua Yu [a], Juepeng Zheng [c,*], Conghui He [d], Haohuan Fu [e,f,*], Dahua Lin [g]

[a] School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai, China
[b] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[c] School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai, China
[d] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[e] Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing, China
[f] Tsinghua University (Department of Earth System Science)-Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Beijing, China
[g] CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Building extraction
Semantic segmentation
Graph neural networks
High-resolution remote sensing images

ABSTRACT

As a fundamental task for geographical information updating, 3D city modeling, and other critical applications, the automatic extraction of building footprints from high-resolution remote sensing images has been substantially explored and received increasing attention over recent years. Among different types of building extraction methods, the polygonal segmentation methods produce vector building polygons that are in a more realistic format compared with those obtained from pixel-wise semantic labeling and contour-based methods. However, existing polygonal building segmentation methods usually require a perfect segmentation map and a complex post-processing procedure to guarantee the polygonization quality, or produce inaccurate vertex prediction results that suffer from wrong vertex sequence, self-intersections, fixed vertex quantity, etc. In our previous work, we have proposed a method for polygonal building segmentation from remote sensing images that addresses the above limitations of existing methods. In this paper, we propose PolyCity, which further extends and improves our previous work in terms of the application scenario, methodology design, and experimental results. Our proposed PolyCity contains the following three components: (1) a pixel-wise multi-task network for learning the semantic and geometric information via three tasks, i.e., building segmentation, boundary prediction, and edge orientation prediction; (2) a simple but effective vertex selection module (VSM), which effectively bridges the gap between pixel-wise and graph-based models via transforming the segmentation map into valid polygon vertices; (3) a graph-based vertex refinement network (VRN) for automatically adjusting the coordinates of VSM-generated valid polygon vertices, producing the final building polygons with more precise vertices. Results on three large-scale building extraction datasets demonstrate that our proposed PolyCity generates vector building footprints with more accurate vertices, edges, shapes, etc., achieving significant vertex score improvements while maintaining high segmentation and boundary scores compared with the current state-of-the-art. The code of PolyCity will be released at https://github.com/liweijia/polycity.

1. Introduction

Building footprint extraction is a fundamental task in a variety of practical applications, such as urban planning, disaster assessment and environmental management (Sun et al., 2018; Li et al., 2019a). It also provides important information for residential area management, statistics of urban and rural population, maintenance and update of geographic information data, etc. (Demir et al., 2018; Alshehhi et al., 2017). Traditional methods usually require extracting conventional features based on spectral, lines, shadow index, etc. (Ok et al., 2012), which is followed by machine learning classifiers such as Random Forest and SVM (Turker and Koc-San, 2015). In recent years, owing to the rapid progress of deep neural networks (DNN) and the growing

* Corresponding authors.
E-mail addresses: zhengjp8@mail.sysu.edu.cn (J. Zheng), haohuan@tsinghua.edu.cn (H. Fu).

https://doi.org/10.1016/j.isprsjprs.2023.05.010
Received 18 July 2022; Received in revised form 7 May 2023; Accepted 8 May 2023
Available online 24 May 2023

abundance of building annotations, DNN-based building extraction has become the main stream method and widely explored in remote sensing, GIS, computer vision, and other research domains (Li et al., 2019a; Shi et al., 2020; Zhao et al., 2021; Liu et al., 2022; Huang et al., 2021). Nevertheless, due to several challenges such as the great discrepancy in color, shape, area and material of buildings in different regions, as well as the high similarity and unclear boundary between buildings and other object types in remote sensing images (Li et al., 2018), automatic building extraction from remote sensing imagery still suffers from an unsatisfactory accuracy that requires further improvement.

Most DNN-based studies formulate building extraction as a pixel-wise semantic labeling task. Among these studies, various semantic segmentation models have been applied and adapted to building extraction task, based on U-Net (Ronneberger et al., 2015; Li et al., 2021b), FC-DenseNet (Yang et al., 2018; Jégou et al., 2017), HR-Net (Wang et al., 2020; Li et al., 2021a), etc. Similar to the research progress in general object segmentation, several studies designed building extraction methods based on modified instance segmentation models such as Mask R-CNN (He et al., 2017; Mahmud et al., 2020; Wang et al., 2022) and hybrid task cascade (HTC) (Chen et al., 2019; Liu et al., 2021). To further improve the building extraction performance, many studies design effective strategies combined with the deep learning models such as fusing multiple data types (Li et al., 2019a; Sun et al., 2018) or introducing additional tasks (Bischke et al., 2019; Guo et al., 2022).

The above semantic labeling-based methods usually produce building extraction results with curved and irregular boundaries, which are not in the desirable vector format of building polygons. To tackle these limitations, many recent studies propose post-processing or vertex-based methods for polygonal building segmentation. For post-processing-based studies, various contour simplification or polygonization methods have been proposed for transforming the output raster segmentation maps of semantic or instance segmentation models into regularized building polygons in vector format (Li et al., 2020; Zhao et al., 2018). Compared with the pixel-wise and contour-based methods that generate objects with curved boundaries, the above polygonal segmentation methods are more suitable for extracting desirable building footprints that are annotated in a line-based manner. However, these methods usually require multiple steps of complex procedures, such as segmentation map refinement, polygonization, regularization, and other essential steps.

For vertex-based object segmentation studies, several methods are designed for directly predicting a vertex at each time step using a CNN-RNN architecture, such as Polygon-RNN (Castrejon et al., 2017), Polygon-RNN++ (Acuna et al., 2018), PolyMapper (Li et al., 2019b), Zhao et al. (2021), etc. Although the prediction process is similar to the actual annotation procedure of a building polygon, the sequential manner of the recurrent model limits its capability of predicting each vertex for building polygons with complex contours, and the failure case of a former vertex can have a negative impact on latter prediction. In other studies such as PolyTransform (Liang et al., 2019b) and Curve-GCN (Ling et al., 2019), the polygon vertices are first selected from a segment contour using a unified distance or an initial contour given a fixed quantity, which will be further adjusted simultaneously in a regression manner. Such methods effectively improve the segmentation accuracy for general vision datasets such as Cityscapes. Yet, due to the diverse quantity of building vertex (ranging from four to over a hundred), these methods usually generate redundant vertices for buildings with simple contour and insufficient vertices for buildings with complex contour. Another category of building extraction study is based on the active contour model (ACM) (Marcos et al., 2018; Cheng et al., 2019; Xu et al., 2022; Gur et al., 2020), of which most methods are designed for building segmentation from cropped images with a single instance. Similar to Curve-GCN (Ling et al., 2019), these methods require a fixed number of vertices and suffer from vertex redundancy or insufficiency problems.

In our previous work (Li et al., 2021c), we proposed a new approach for polygonal building segmentation, which effectively addresses the limitations of existing post-processing and vertex-based building segmentation methods. We provided preliminary experimental results on SpaceNet and CrowdAI datasets, and comparison with state-of-the-art polygonal segmentation methods (ASIP (Li et al., 2020), PolyMapper (Li et al., 2019b), Frame Field Learning (Girard et al., 2021), etc.). In this work, we further extend and improve our previous work (Li et al., 2021c) in terms of the application scenario, methodology design, and experimental results. Our proposed PolyCity consists of a pixel-wise multi-task network and a graph-based vertex refinement network, as well as a rule-based vertex selection module that bridges the gap between the above two components. The multi-task network is designed for learning both semantic and geometric information of building polygons via building segmentation, boundary prediction, and edge orientation prediction. The vertex selection method is designed for transforming the building segmentation contour into valid polygon vertices based on the three types of network outputs. The vertex refinement network is designed for automatically adjusting the valid polygon vertices to more accurate locations.

We compare our proposed PolyCity with three powerful methods for single object segmentation (i.e., Polygon-RNN (Castrejon et al., 2017), Polygon-RNN++ (Acuna et al., 2018) and Curve-GCN (Ling et al., 2019)), HR-Net followed by traditional polygonization method (baseline) (Wang et al., 2020), and two recently proposed methods dedicated for building extraction (i.e., CVNet (Xu et al., 2022) and Frame Field Learning (Girard et al., 2021)). We also provide extensive experimental results comparison between our method and our previous work (Li et al., 2021c). Results on three popular building datasets demonstrate that our approach improves the vertex prediction accuracy by 3%–4% compared with the current state-of-the-art, producing building polygons with more precise vertices, edges, and shapes. The code of PolyCity will be released at https://github.com/liweijia/polycity.

The new contributions compared with Li et al. (2021c) are summarized as follows:

- We extend (Li et al., 2021c) to a new application scenario, i.e., extracting the individual building instance from the input images that are cropped by the ground truth or predicted bounding boxes.
- We remain the general methodology design following Li et al. (2021c) and modify the main components for the single object segmentation scenario, in terms of network architectures, multi-task definitions, vertex selection rules, etc.
- We provide experimental result comparison with additional methods on more building extraction datasets, using new experimental settings and evaluation metrics that could more directly reflect the polygonization performance, which demonstrates the significant improvement of our method compared with Li et al. (2021c) and other building extraction methods.
- We rewrite the whole paper to provide more details of motivation, methodology, experiments and analysis.

2. Related work

2.1. Pixel-wise building footprint extraction

Building footprint extraction from satellite or aerial images has been broadly studied for decades. Traditional building extraction methods include shadow index (Huang and Zhang, 2011), edge regularity (Chen et al., 2018a), or line fragment (Sun et al., 2014) based methods, etc. Recently, most building footprint extraction studies are based on deep learning methods for pixel-wise semantic labeling. Semantic segmentation models (e.g. U-Net (Ronneberger et al., 2015), FC-DenseNet (Jégou et al., 2017), High-Resolution Network (HR-Net) (Wang et al., 2020)) and instance segmentation models (e.g. Mask-RCNN (He et al., 2017)) have been broadly explored and achieved promising building extraction

results (Yang et al., 2018; Li et al., 2019a, 2021b). For example, Guo et al. (2022) proposed a coarse-to-fine network to progressively refine the building boundaries at different scales, based on the U-Net architecture. In Wu et al. (2022), a topography-aware loss was introduced to boost the network capability for preserving building segmentation boundary, based on the high-resolution network architecture. In addition, various useful strategies (e.g. data fusion (Li et al., 2019a; Sun et al., 2018; Hosseinpour et al., 2022), distance transform (Bischke et al., 2019), boundary regularization (Zhao et al., 2018; Wei et al., 2019), etc.) are combined with the pixel-wise segmentation network to further improve the building segmentation performance.

Although pixel-wise CNN-based methods have achieved increasingly higher prediction accuracies, the building extraction results produced from these methods often have a great discrepancy compared with the desired polygonal buildings for actual applications. The outlines of building footprints predicted from these methods are in a curved shape, while the actual building polygons are manually annotated in a line-based manner, with a relatively small number of edges and vertices. It often requires onerous efforts for processing the prediction results of these methods into the desirable building polygons in vector format.

2.2. Polygonal object segmentation

To solve the above limitations, many studies design polygonal object segmentation methods to produce the building footprints or other types of objects in a desirable vector format. Existing polygonal segmentation methods can be divided into two main categories, i.e., post-processing-based methods (Li et al., 2020; Zhao et al., 2018) and vertex-based methods (Castrejon et al., 2017; Acuna et al., 2018; Ling et al., 2019; Li et al., 2019b; Chen et al., 2020; Zorzi et al., 2022). For post-processing-based methods, raster maps or curved contours are vectorized or simplified via various polygonization strategies, such as Douglas–Peucker (Wu and Marquez, 2003) and polyline decimation (Dyken et al., 2009), and other traditional post-processing methods. In recent years, many studies proposed post-processing-based methods for vectorizing the semantic segmentation or instance segmentation results. In Zhao et al. (2018), a boundary regularization method was proposed for regularizing the building instances produced from the Mask-RCNN model into simplified polygons via multiple steps. Li et al. (2020) proposed ASIP, a polygonal partition refinement method for vectorizing buildings and other objects from the output segmentation map of a U-Net-based model. In Wei et al. (2019), a series of strategies were proposed for transforming the output segmentation maps of an FCN-based model into final regularized building footprints, including segmentation map refinement via post-processing, vectorization and polygonization of the refined segmentation maps, as well as polygon regularization to convert the former polygons into final structured building footprints.

The other category of polygonal segmentation methods designs deep neural networks that directly predict the polygon vertices. Polygon-RNN (Castrejon et al., 2017) is a pioneer vertex-based method for single object segmentation, which designed LSTM-based architecture for predicting a vertex location at each time step. Polygon-RNN++ (Acuna et al., 2018) further extended this work to improve the segmentation accuracy through several strategies including attention mechanism, an evaluator with beam search, reinforcement learning, and upscaling with a graph neural network. PolyMapper (Li et al., 2019b) extended the application scenario of Polygon-RNN, producing multiple building footprints and road typologies in a vector format. Although these methods produce vectorized prediction results, it is hard to correctly predict vertex for polygons with a complex contour and the failure cases of former vertices can result in continuous errors on latter predictions, due to the sequential prediction order of the recurrent model. Curve-GCN (Ling et al., 2019) is another type of vertex-based method for single object segmentation, which simultaneously predicts all vertices using a graph convolutional network and achieves better segmentation accuracy compared with Polygon-RNN++. However, as Curve-GCN represents the object as a graph with a fixed number of vertices, it often generates redundant vertices for simple polygons and insufficient vertices for complex polygons. A recently proposed method, named PolyWorld (Zorzi et al., 2022), regards all building polygons of one image as a whole graph. Different from the CNN and GNN used in our method, PolyWorld uses a graph neural network to predict the connection strength between each pair of vertices detected by CNN, which are further optimized by minimizing a combined segmentation and polygonal angle difference loss. Although producing neat building polygons, the performance drops seriously for buildings with inner holes, small areas, or other complex cases.

The active contour model (ACM) is another type of widely-used method for polygon-based building extraction. Among these studies, most ACM-based methods are proposed for single object segmentation (Marcos et al., 2018; Cheng et al., 2019; Gur et al., 2020; Xu et al., 2022), i.e., extracting a single building instance from a cropped input image. For instance, Marcos et al. (2018) proposed DSAC, a deep structured active contour method that integrates ACMs and CNNs via learning the ACM parameterizations per instance using a CNN for single building segmentation. Cheng et al. (2019) designed DarNet, a deep active ray network for single building segmentation, which avoids the self-intersection and improves the boundary performance via adopting the polar coordinates and a new loss function. On the other hand, several recent studies are designed for multiple building segmentation. For example, Hatamizadeh et al. (2020) proposed trainable deep active contours (TDACs), which intimately unites CNNs and ACMs and devises an implicit ACM formulation for extracting multiple buildings from remote sensing images. Similar to Curve-GCN, ACM-based methods represent the object as a graph with a fixed number of vertices (contour points). Consequently, these methods also suffer from producing redundant vertices for simple polygons and insufficient vertices for complex polygons.

In summary, the post-processing-based methods usually require a complex procedure with multiple steps. The polygonization performance is seriously influenced by the quality of the segmentation map. For vertex-based methods, the recurrent manner of the RNN-based methods limits the vertex prediction capability for complex polygons, while the GCN and ACM-based methods suffer from the fixed vertex quantity. Our approach, on the contrary, combines the pixel-wise CNN model with the vertex-based graph model via a rule-based vertex selection module, which effectively solves the above limitations and produces building polygons with precise vertices even for complex cases.

2.3. Multi-task learning

Multi-task learning strategy has been effectively applied in building and other object segmentation studies. In these studies, additional prediction tasks are introduced and trained jointly with the object segmentation tasks, of which distance prediction (Hui et al., 2018; Bischke et al., 2019; Mahmud et al., 2020) and direction prediction (Guo et al., 2022; Yuan et al., 2020; Girard et al., 2021) are two broadly explored tasks for improving the segmentation and boundary accuracies. For example, Bischke et al. (2019) proposed a multi-task building extraction method that combines distance transform prediction with building segmentation tasks to improve the boundary performance. With a different distance representation, Mahmud et al. (2020) introduced the modified signed distance function prediction that is jointly trained with building instance segmentation, semantic segmentation, and DSM prediction tasks.

Several recent studies introduced direction-related prediction tasks to improve the segmentation boundary (Yuan et al., 2020; Girard et al., 2021; Guo et al., 2022). Yuan et al. (2020) proposed SegFix to refine the segmentation boundary of existing models via additionally learning a direction map, i.e. the direction from the boundary pixel to an interior

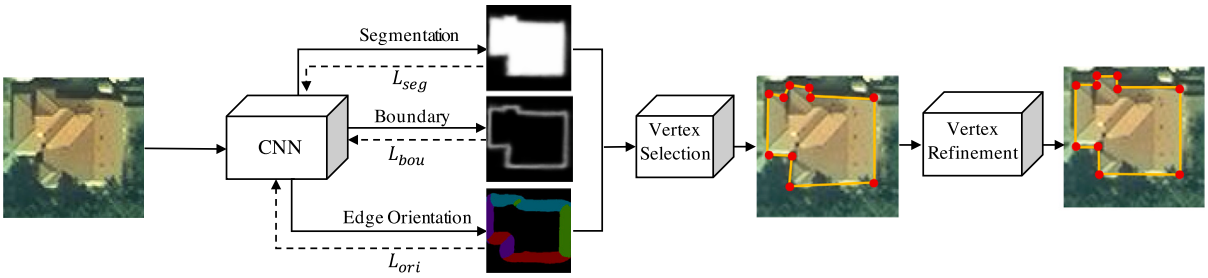

Fig. 1. The overall pipeline of our proposed polygonal building segmentation method, which includes: (1) a multi-task CNN for building segmentation, boundary prediction and edge orientation prediction; (2) a vertex selection module for transforming the network outputs into valid polygon vertices; (3) a vertex refinement network for automatically adjusting the selected valid vertices to more accurate locations.

pixel. Girard et al. (2021) introduced a frame field learning task for polygonal building segmentation, of which the pixel-wise frame field is represented as two directions encoded by complex numbers.

Unlike the above direction-related tasks that were introduced for improving the segmentation boundary, the edge orientation proposed by our method is defined as a different representation and designed for generating valid polygon vertices, which is crucial for the further integration of the pixel-wise segmentation model and the graph-based refinement model.

3. Method

As shown in Fig. 1, the overall pipeline of our proposed method contains three main components, i.e., a pixel-wise multi-task network, a rule-based vertex selection module, and a graph-based vertex refinement network. Taking a satellite image cropped by a single object as input, the multi-task network is designed for building area segmentation, building boundary prediction, and edge orientation prediction. Then the vertex selection module effectively leverages the three types of network outputs to transform the segmentation mask into a set of valid polygon vertices. Finally, the vertex refinement network regards the valid polygon vertices as the initial nodes of a graph representation and predicts a displacement for each node, which automatically adjusts the valid polygon vertices to more accurate locations. In the following, we first introduce the design of our proposed multi-task network in 3.1, including the definitions of edge orientation, network architecture and training. Then we introduce the vertex selection module and the vertex refinement network in Sections 3.2 and 3.3. The implementation details are introduced at the end of this section.

3.1. Multi-task learning with edge orientation

We design a pixel-wise multi-task network to learn both semantic and geometric information of building polygons via three different tasks, i.e. building segmentation, boundary prediction, and edge orientation prediction. The representation of our proposed edge orientation property and the network architecture for multi-task learning are introduced as follows.

3.1.1. Representation of edge orientation

Building segmentation and boundary prediction are common tasks in existing building extraction methods. In addition to these tasks, we introduce an extra task to learn the geometric information of building polygons, i.e. edge orientation prediction, which is beneficial for building footprint extraction in many aspects (Li et al., 2021c). We use the normal vector of each edge to calculate the orientation values. The original building footprint annotations are converted into the representation of edge orientation according to the following method.

Let I denote an input image and E the edges of a building footprint annotation. For a pixel i that belongs to edge $E(j)$, its orientation angle α_i is determined by the normal vector of edge $E(j)$, which is denoted by

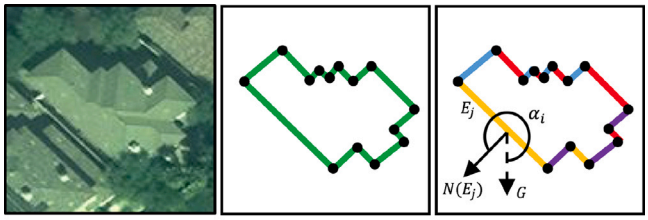

Fig. 2. Representation of edge, vertex and edge orientation. The green lines and black points in (b) denote the edges and vertices. Different colors in (c) denote different edge orientation classes.

$N(E_j)$. Specifically, the orientation angle α_i of pixel i is defined as the angle between the normal vector $N(E_j)$ and the gravity orientation G in a counter-clockwise direction. In our method, the orientation angle α_i is further discretized into the orientation class $y_o(i)$, which is divided into K categories. For each pixel, the orientation class $y_o(i)$ is an integer in the range of $[0, K]$. If pixel i does not belongs to any building edges, then its orientation class $y_o(i)$ equals zero; If pixel i is located at the building corners, then we assign $y_o(i)$ with one of its neighboring edges. In this way, each edge of a building footprint E_j is assigned with an orientation property, and the edge orientation of each pixel can be annotated as one of $K+1$ classes (see Fig. 2).

3.1.2. Network architecture and training

The goal of our multi-task network is to learn both semantic and geometric information of building footprints. Many existing building segmentation studies are based on semantic segmentation models (such as U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), and other encoder–decoder architectures) or instance segmentation models (such as Mask-RCNN (He et al., 2017)). In our previous work (Li et al., 2021c), a modified Res-U-Net architecture is employed to guarantee the consistency of segmentation map for comparison with (Li et al., 2020). Inspired by a recent study (Li et al., 2021a), we adopt the High-Resolution Network (HR-Net) (Wang et al., 2020) as the backbone architecture of our multi-task network in this study, of which the capacity of maintaining high-resolution representations throughout the whole process is important and beneficial for our building segmentation task.

As shown in Fig. 1, our HR-Net-based segmentation network is designed for three different tasks: (1) building area segmentation; (2) building boundary prediction; (3) edge orientation prediction. Each task is formulated as a pixel-wise classification problem and trained with the cross entropy loss (denoted by L) according to formula (1):

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \times log(p(y_{i,k})) \quad (1)$$

where K is the number of classes of the corresponding task; N is the number of pixels of an image; $y_{i,k}$ is a binary indicator that equals 1 if

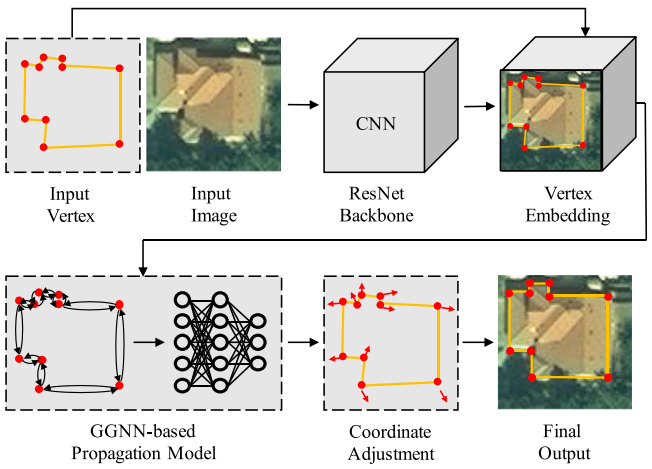

Fig. 3. The overall framework of vertex refinement network, including a ResNet-based backbone for vertex embedding and a GGNN-based propagation model for vertex correction.

k is the ground truth label of pixel i or 0 in other cases; $p(y_{i,k})$ is the predicted probability that pixel i belongs to class k. The total loss L_{total} of the three tasks can be summarized as:

$$L_{total} = \lambda_1 L_{seg} + \lambda_2 L_{bou} + \lambda_3 L_{ori} \tag{2}$$

where L_{seg}, L_{bou} and L_{ori} represent the loss of building area segmentation, boundary prediction, and edge orientation prediction; λ_1, λ_2 and λ_3 are the weight of each task. The multi-task network will be trained jointly to produce different types of prediction maps. The input image for network training is cropped by a bounding box that is corresponding to a single building instance. The hyper-parameter setting of our segmentation network will be introduced in Section 3.4.

3.2. Vertex selection module

We design a simple but effective vertex selection module for transforming the segmentation mask into polygon vertices. Due to the non-nadir views of satellite images, there is misalignment between the building annotations and the actual building outlines, resulting in challenges for directly predicting the accurate vertices or edges of a building polygon (Wang et al., 2022). However, even with the misalignment between annotated and actual building edges, the edge orientation predicted by the multi-task network is still capable of providing effective geometric information and characterizing the topology of a building polygon.

The vertex selection module aims at filtering out the redundant vertices and remain the valid ones, based on the three types of network outputs. First, we extract the pixel coordinates on the contour of the segmentation mask via dense sampling, and select the pixel coordinates of which the boundary prediction probability is larger than a given threshold t_{bon}, constituting a set of initial vertex candidates $C = \{c_0, c_1, c_2, \dots, c_n\}$. We use $y_{ori}(c_i), (i \in \{0, 1, 2, \dots, n\})$ to denote the edge orientation class (predicted by the multi-task network) for each vertex candidate in C, and use $y_{ver}(c_i), (i \in \{0, 1, 2, \dots, n\})$ to indicate whether c_i is selected as a valid vertex ($y_{ver}(c_i) = 1$) or not ($y_{ver}(c_i) = 0$). A candidate c_i will be selected as a valid vertex only if the absolute difference between $y_{ori}(c_i)$ and $y_{ori}(c_{i-1})$ is greater than or equal to a given threshold t_{ori}, constituting the output vertex set V. The selection rule is defined according to the prior knowledge, which can be summarized as follows:

$$y_{ver}(c_i) = \begin{cases} 1, & \text{if } y_{ori}(c_i) - y_{ori}(c_{i-1}) \geq t_{ori} \\ 0, & \text{if } y_{ori}(c_i) - y_{ori}(c_{i-1}) < t_{ori} \end{cases} \tag{3}$$

3.3. Vertex refinement network

The output vertex set V obtained from the vertex selection module contains only valid polygon vertices. Inspired by previous single object segmentation methods (Liang et al., 2019b; Acuna et al., 2018; Ling et al., 2019), we design a vertex refinement network (VRN) to further improve the vertex prediction accuracy. As shown in Fig. 3, VRN contains two main components: (1) a ResNet-based backbone for vertex embedding, which extracts the features of the input image and vertices for further vertex correction; (2) a GGNN-based propagation model for vertex correction, which learns to predict a displacement for each vertex in order to move it to more accurate location. Note that the ResNet-based backbone and the GGNN-based propagation model are trained together in an end-to-end manner. The details of each component are introduced as follows.

3.3.1. Backbone and vertex embedding

The backbone architecture of VRN is a variant of ResNet-50 (He et al., 2016) following Li et al. (2021c), Acuna et al. (2018), Ling et al. (2019), which employs a skip-connection structure to upsample and concatenate the feature maps obtained from four skip layers, constituting the final feature map for vertex embedding. For the final feature map with a larger size, the vertex coordinates can be represented more precisely due to the higher resolution of the feature map. On the other hand, for the final feature map with a smaller size, a grid often has a larger receptive field, which benefits the displacement prediction for each vertex. Taking both aspects into consideration, the size of the final feature map of the backbone is set as half of the original size of the input image.

Besides taking the cropped satellite image as input, which is the same as the multi-task network, VRN also requires a set of vertex coordinates as extra inputs. As mentioned previously, the segmentation mask obtained from the multi-task network is converted into a set of valid polygon vertices by VSM. These vertices are then mapped to the final feature map of the ResNet backbone according to their coordinates, which are represented by the red points on the volume of Fig. 3. In the vertex embedding process, each vertex is assigned with a feature vector, which is extracted from the volume in the channel direction.

3.3.2. Propagation model based on GGNN

In our proposed VRN, the building polygon is represented in a graph format. Each vertex obtained from VSM constitutes the node and each neighboring vertex pair constitutes the edge. Inspired by previous work (Acuna et al., 2018; Li et al., 2021c), the propagation model for vertex correction is based on the gated graph neural network (GGNN) (Li et al., 2015). Different from the pixel-wise segmentation network, the GGNN-based propagation model is capable of utilizing extra information such as the feature of each node (vertex) and the relation between each node of the graph. In addition, we add two fully-connected layers to GGNN, which outputs a value indicating the predicted displacement of each node.

During the training phase, the target of the GGNN is to learn the displacement between each valid polygon vertex (the output of VSM) and its nearest ground truth vertex. The whole process is formulated as a classification problem and trained with the cross entropy loss. Since the vertices selected by VSM are already close to the building corners, we use a fixed range of $[-k, k]$ to adjust the vertices in x and y directions. Within this range, each case of displacement coordinates (δ_x, δ_y) is encoded as a class value, constituting $(2k + 1)^2$ categories in total. The whole training process is much easier compared with regression without a range limitation. The class value predicted by VRN is further converted into the displacement coordinates, and added to the corresponding vertex coordinates of VSM to produce the final prediction results. In this way, the GGNN-based VRN automatically moves the polygon vertices to more accurate locations in the inference phase.

3.4. Implementation details

For our multi-task network, the numbers of channels in the four stages of HR-Net are set as 24, 48, 96, and 192, respectively. For the training dataset, each input image cropped by the bounding box is resized to 224 × 224 pixels following Acuna et al. (2018), Ling et al. (2019) to guarantee a fair comparison. The weights of three tasks $(\lambda_1, \lambda_2, \lambda_3)$ are simply set as 1. The value of K for edge orientation prediction is set as 36, indicating that the bin width of the edge orientation angle is 10. For the vertex selection module, we set the value of t_{ori} as 3 since the angle between two edges of a building polygon could rarely be greater than 150 or smaller than 30 in the actual scenario according to prior knowledge, and set the boundary probability threshold t_{bon} as 0.5. In this way, redundantly selecting the invalid polygon vertices due to edge orientation prediction noises and mistakenly removing the valid polygon vertices can be avoided in a balanced manner. For the ResNet backbone of the vertex refinement network, the number of channels of the final feature map for vertex embedding is set as 256. The dimension sizes of the two fully-connected layers of the GGNN-based propagation model are also set as 256. In addition, we set the vertex moving range as $[-7, +7]$ pixels, and set the output dimension of two fully-connected layers as 225, accordingly.

4. Experimental settings

4.1. Datasets

Various building datasets have been used in previous building segmentation studies, of which most datasets only provide ground truth masks without the coordinates of polygon vertices, such as ISPRS semantic labeling datasets (Paisitkriangkrai et al., 2016), ARIS (Chen et al., 2018b), etc. To ensure the network training and accurate evaluation of the polygon vertex, we evaluate our proposed method using SpaceNet building footprint dataset (SpaceNet) (Van Etten et al., 2018) and Microsoft US building footprint dataset (MSUS) (Microsoft, 2022). Both datasets provide specific geographic information (e.g. the longitude and latitude of each vertex) in GeoJson format, so that it could be flexibly used as the annotation for remote sensing images of different resolutions and sizes. In addition, we conduct experiments on Inria Aerial Image Labeling Dataset (Inria-building) (Maggiori et al., 2017), which has been used in several recently proposed methods for polygon-based building extraction (Xu et al., 2022; Girard et al., 2021). Following these methods, we convert the initial mask annotations of Inria-building dataset into the pixel coordinates of polygon vertices for the training and evaluation of different polygon-based methods. The details of each dataset are introduced as follows.

The SpaceNet building dataset contains satellite images and building footprints of several cities located in different continents. We select all annotated building instances of Las Vegas for evaluating the polygon vertex, of which most buildings are annotated in a unified and consistent standard (compared with other cities) (Wu et al., 2022). The dataset of Las Vegas contains 3,851 images (in 650 × 650 pixels, with a spatial resolution of 0.3 m) and around 108,000 manually annotated building footprints. The MSUS building dataset contains over a hundred million computer-generated building footprints in all 50 US states, with better or similar metrics compared to OpenStreetMap building metrics against the labels. As the dataset only provides the building polygons without original images, we download the corresponding Google Earth high-resolution images. The dataset of Salt Lake City (SLC) contains 500 images (in 2048 × 2048 pixels, with a spatial resolution of about 0.6 m) and around 150,000 building footprints. The Inria building dataset (Maggiori et al., 2017) contains 180 large-scale satellite images (in 5000 × 5000 pixels, with a spatial resolution of about 0.3 m) and over 20,0000 annotated buildings that are located in five cities. Compared with the above two datasets, Inria-building dataset is more challenging in many aspects. Many buildings are with holes, i.e., one building may contain multiple polygon contours. The distribution pattern is more diverse, including sparse and dense building areas. Moreover, there are more buildings with serious parallax effect, i.e., the footprint annotation has a deviation from the roof. These challenges enables analyzing the performance of our method in different complex cases.

For Vegas and SLC datasets, the whole datasets are randomly divided into training/validation/test images using a ratio of 8:1:1 following Li et al. (2021c). For the Inria-building dataset, we divide the whole dataset into train, validation and test sets with the ratio of 6:2:2 following Xu et al. (2022). Each large-scale image is cropped by the ground truth (GT) bounding box corresponding to each building instance during both training and inference phase following the existing single building extraction methods (Liu et al., 2022; Chen et al., 2020; Marcos et al., 2018; Cheng et al., 2019; Xu et al., 2022; Gur et al., 2020). Each cropped image is further uniformly resized into 224 × 224 pixels following Acuna et al. (2018), Ling et al. (2019), constituting the input images of our proposed building segmentation method. When we obtain the output building polygons for each dataset or method, the predicted vertex coordinates relative to the cropped image are further transformed into those relative to the large-scale image, constituting the final large-scale building extraction results.

4.2. Comparison methods

In this study, to precisely evaluate the polygonal segmentation performance of each building instance, we compare our method with another three polygonal segmentation methods, i.e., Polygon-RNN (Castrejon et al., 2017), Polygon-RNN++ (Acuna et al., 2018) and Curve-GCN (Ling et al., 2019), which are the state-of-the-art for polygonal single object segmentation and have been broadly used as the comparison methods in recent single building extraction studies (Liu et al., 2022; Huang et al., 2021; Chen et al., 2020). We also provide the experimental results of the multi-task HR-Net model (Wang et al., 2020). The raster segmentation results are converted into polygon vertices using the Douglas–Peucker algorithm (Wu and Marquez, 2003) following Li et al. (2021c) for vertex performance evaluation, which will be introduced in Section 5.1. To ensure a fair comparison, we carefully evaluate different settings and strategies of the comparison methods, and optimize these methods in order to fit the characteristics of the building segmentation task. For Polygon-RNN, we use the residual encoder with skip-connection following Acuna et al. (2018), which obtains better performance compared with the original VGG encoder. For Polygon-RNN++, results demonstrate that using an evaluator with beam search improves the prediction performance, while reinforcement learning and upscaling with a graph neural network deteriorate the prediction results. Thus we did not apply these two strategies to Polygon-RNN++ in our building segmentation task. For Curve-GCN, we adopt the Polygon-GCN instead of Spline-GCN as the buildings are line-based objects. Considering the actual distribution of vertex numbers, we set the number of the vertex as 20. Additionally, we use the point matching loss (using $K = 1280$ for sampling points following Ling et al. (2019)) as it obtains better results compared with the differentiable accuracy loss for our task. The detailed qualitative and quantitative comparisons with the above methods will be introduced in Section 5.1.

In addition to the above methods, we compare our proposed method with some recently proposed methods that are dedicated for polygon-based building extraction, i.e., CVNet (Xu et al., 2022), Frame Field Learning (Girard et al., 2021), and our previous work (Li et al., 2021c). As CVNet (Xu et al., 2022) uses a fixed number of polygon vertices like Curve-GCN (Ling et al., 2019), we reduce the number of the vertex from 60 (by default) to 20 to improve the vertex prediction scores while maintaining a relatively high IoU and boundary score. For FFL (Girard et al., 2021), we select the Inria polygonized dataset, Frame Field polygonization, Active Contour Model from the provided datasets and models for comparison with our proposed method. For our previous

Table 1
The overall results of Vegas dataset obtained from Polygon-RNN (Castrejon et al., 2017), Polygon-RNN++ (Acuna et al., 2018), Polygon-GCN (Ling et al., 2019), HR-Net + DP (Wang et al., 2020). The results are evaluated in terms of IoU, Boundary F-score (B-F-score), and Vertex F-score (V-F-score).

Method	IoU	B-F-score	V-F-score
Polygon-RNN (Castrejon et al., 2017)	0.8676	0.8346	0.6882
Polygon-RNN++ (Acuna et al., 2018)	0.8758	0.8441	0.6875
Polygon-GCN (Ling et al., 2019)	**0.8871**	0.8559	0.2759
HR-Net + DP (Wang et al., 2020)	0.8804	0.8486	0.6240
Ours	0.8862	**0.8635**	**0.7172**

Table 2
The overall results of SLC dataset obtained from Polygon-RNN (Castrejon et al., 2017), Polygon-RNN++ (Acuna et al., 2018), Polygon-GCN (Ling et al., 2019), HR-Net + DP (Wang et al., 2020). The results are evaluated in terms of IoU, Boundary F-score (B-F-score), and Vertex F-score (V-F-score).

Method	IoU	B-F-score	V-F-score
Polygon-RNN (Castrejon et al., 2017)	0.8623	0.8976	0.8001
Polygon-RNN++ (Acuna et al., 2018)	0.8671	0.9006	0.7967
Polygon-GCN (Ling et al., 2019)	**0.8805**	0.9089	0.2862
HR-Net + DP (Wang et al., 2020)	0.8743	0.8932	0.7294
Ours	0.8771	**0.9114**	**0.8015**

work (Li et al., 2021c), we use the default experimental settings on SpaceNet-Vegas dataset. Besides the original results of our previous work using GT bounding boxes (denoted by ours (gt-bbox)), we also provide the results using predicted bounding boxes (denoted by ours (pr-bbox)). Note that the predicted bounding boxes are generated from the segmentation mask of the multi-task network of Li et al. (2021c). The detailed qualitative and quantitative comparisons with the above methods will be introduced in Section 5.2.

4.3. Evaluation metrics

We use three evaluation metrics to measure the prediction performance, i.e., IoU, boundary F-score (denoted by B-F-score), and vertex F-score (denoted by V-F-score). For each metric, we calculate the mean value of each building instance. IoU (Intersection over Union) has been widely used for evaluating segmentation results in previous object segmentation studies (Acuna et al., 2018; Ling et al., 2019). Boundary F-score is also a common metric for evaluating object segmentation boundaries. Similar to Ling et al. (2019) and Cheng et al. (2019), we calculate the Boundary F-score according to Perazzi et al. (2016), which calculates the precision and recall between the boundary of the predicted polygon and GT polygon. Moreover, we adopt a vertex evaluation metric following Homayounfar et al. (2018), Liang et al. (2019a), Chen et al. (2020), Li et al. (2021c) to calculate the precision, recall, and F-score between the predicted vertex set and GT vertex set. The boundary and vertex F-scores are measured using the threshold of 2 and 3 pixels, which can precisely reflect the performance of vertex prediction and polygonal object segmentation.

5. Experimental results and analysis

5.1. Comparison with state-of-the-art methods for polygonal single object segmentation

Tables 1 and 2 list the overall results of Vegas and SLC obtained from different methods. To further analyze the vertex prediction performance in detail, Table 3 provides additional vertex evaluation metrics of the Vegas dataset, including the precision, recall, and F1-score with the threshold of 2 and 3 pixels (denoted by V-P @2pix, V-R @2pix, V-F @2pix, V-P @3pix, V-R @3pix, and V-F @3pix). In general, for both Vegas and SLC datasets, our method achieves the best performance among all methods in terms of the boundary F-score and all vertex metrics. For the vertex evaluation metrics, our method achieves a V-F-score improvement of 3%–4% compared with the second highest score. Regarding the IoU evaluation metric, although Polygon-GCN obtains a slightly higher IoU of 0.09% and 0.34%, its vertex F-scores are 44.13% and 51.53% lower than ours on Vegas and SLC dataset, respectively. This fatal shortage makes Polygon-GCN barely a usable method for polygonal building segmentation in practical scenarios. In addition, the building polygons of the Vegas dataset are carefully annotated by human annotators, which have a more complex structure (with finer vertices and edges) compared with the computer-generated SLC dataset. Compared with the state-of-the-art methods, the performance improvement of our method is more obvious on the Vegas dataset than the SLC dataset, which also verifies the superiority of our proposed method for complicated cases.

To qualitatively evaluate the performance of different methods, we further provide the building segmentation results of our method and comparison with Polygon-GCN and Polygon-RNN++, which achieve relatively higher performance among all comparison methods. Fig. 4 shows some examples of the Vegas and SLC dataset, respectively. Polygon-RNN++ obtains promising results for simple buildings with a small number of vertices, but the results deteriorate seriously for buildings with large size or complex contour. Specifically, the failure cases of the former prediction (such as missing vertices) can result in consecutive errors on the latter prediction, due to the sequential prediction order of the recurrent model. For Polygon-GCN, the predicted vertices scatter on the boundary of buildings instead of clustering around the corners. As the number of predicted vertices is fixed, it usually results in redundant vertices for buildings with simple contours (e.g. the fourth row of Fig. 4) and insufficient or inaccurate vertices for buildings with complex contours (e.g. the second row of Fig. 4). For both Vegas and SLC datasets, our method shows obvious advantages over Polygon-RNN++ and Polygon-GCN, especially for buildings with large size or complex contour.

5.2. Comparison with dedicated methods for polygon-based building extraction

In this section, we provide the detailed qualitative and quantitative comparisons with three dedicated methods for polygon-based building extraction, i.e. CVNet (Xu et al., 2022), FFL (Girard et al., 2021) and our previous work (Li et al., 2021c). Fig. 5 and Table 4 show the qualitative and quantitative comparisons between our method and CVNet (Xu et al., 2022) and FFL (Girard et al., 2021) methods. We provide the results comparison on inria-building dataset, which is a challenging dataset that has been used in CVNet (Xu et al., 2022) and FFL (Girard et al., 2021). Both qualitative and quantitative results demonstrate the advantages of our proposed PolyCity compared with CVNet (Xu et al., 2022) and FFL (Girard et al., 2021) methods. Although FFL (Girard et al., 2021) produces building polygons with accurate boundaries in most cases, there are many redundant polygon vertices in the prediction results. Meanwhile, the adjacent buildings are always regarded as an entirety and cannot be extracted separately. For the buildings with holes, FFL (Girard et al., 2021) fails to extract every inner contour in most cases. Similar to Polygon-GCN, CVNet (Xu et al., 2022) uses a fixed number of polygon vertices for all buildings. For buildings with simple contours, the prediction results of CVNet (Xu et al., 2022) usually have redundant polygon vertices. Moreover, the performance drops seriously for buildings with complex contours, producing building polygons with insufficient vertices and inaccurate shapes. The above limitations result in the low vertex prediction scores especially for the vertex precision. Compared with CVNet (Xu et al., 2022) and FFL (Girard et al., 2021), our method is more capable of extracting adjacent buildings individually and extracting every inner contour for buildings with holes, producing building polygons with more accurate shape and vertices. The quantitative results also show that our method achieves the best performance in terms of IoU and all vertex metrics

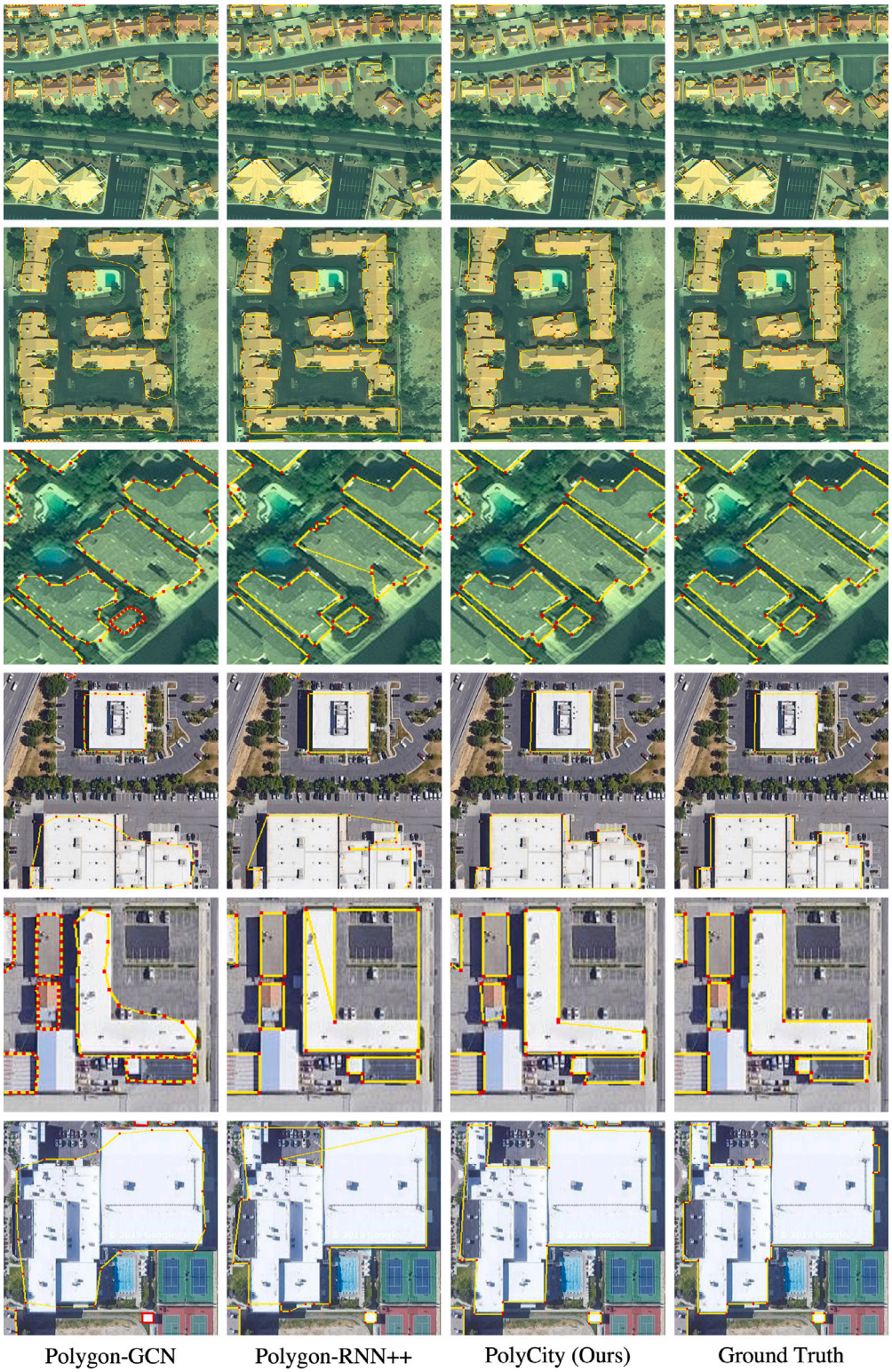

Fig. 4. Results comparison of different methods on SpaceNet-Vegas and MSUS-SLC datasets. The red points and yellow lines denote the vertices and edges of the building polygons predicted by different methods.

(with a slight drop of B-F-score compared with FFL (Girard et al., 2021)), and significantly improves the vertex prediction scores by over 14%.

Fig. 6 and Table 5 show the qualitative and quantitative comparisons between our PolyCity and our previous work (Li et al., 2021c). We conduct experiments on the SpaceNet-Vegas dataset, which was also used in our previous work (Li et al., 2021c). We also provide some large-scale building extraction results of our PolyCity in Fig. 7. The first row of Table 5 shows the results obtained from our previous work (Li et al., 2021c). In the second row, i.e., ours (pr-bbox), we provide the results of our method using predicted bounding boxes, which are generated from the segmentation mask of the multi-task network of Li et al. (2021c). In the third row, i.e., ours (gt-bbox), we provide the original results of our method using GT bounding boxes. We

Table 3
The detailed vertex prediction performance of the Vegas dataset obtained from different methods, in terms of Vertex Precision (V-P), Vertex Recall (V-R), and Vertex F-score (V-F).

Method	V-P @2pix	V-R @2pix	V-F @2pix	V-P @3pix	V-R @3pix	V-F @3pix
Polygon-RNN (Castrejon et al., 2017)	0.4662	0.4259	0.4451	0.7183	0.6606	0.6882
Polygon-RNN++ (Acuna et al., 2018)	0.4511	0.4310	0.4408	0.7079	0.6683	0.6875
Polygon-GCN (Ling et al., 2019)	0.0931	0.2976	0.1418	0.1879	0.5194	0.2759
HR-Net + DP (Wang et al., 2020)	0.4274	0.4182	0.4227	0.6312	0.6169	0.6240
Ours	**0.5133**	**0.4641**	**0.4885**	**0.7606**	**0.6785**	**0.7172**

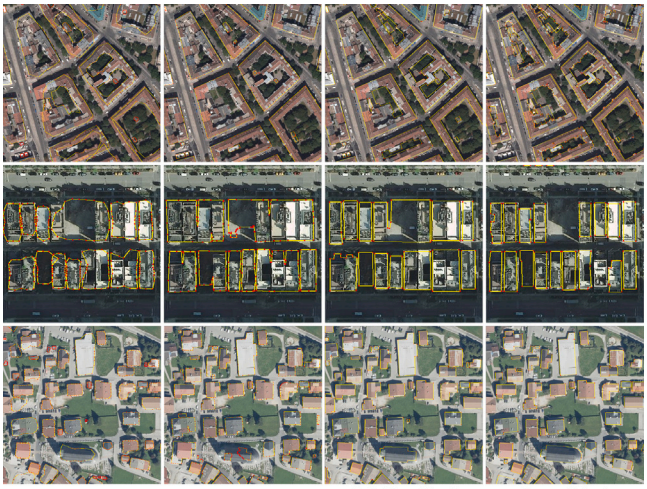

Fig. 5. Qualitative comparison with CVNet (Xu et al., 2022) and FFL (Girard et al., 2021). The red points and yellow lines denote the vertices and edges of the building polygons predicted by different methods.

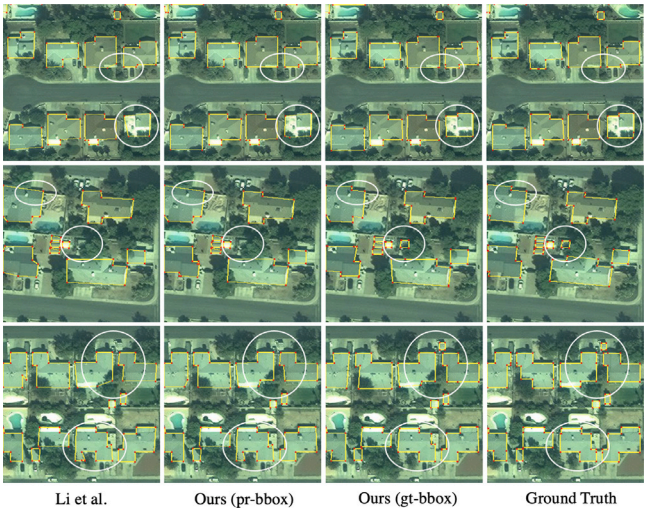

Fig. 6. Qualitative comparison with our previous work (Li et al., 2021c). The red points and yellow lines denote the vertices and edges of the building polygons predicted by different methods.

can find that ours (gt-bbox) achieves the best performance in terms of all evaluation metrics, which is followed by ours (pr-bbox). The vertex prediction score is significantly improved by 6% compared with our previous work. From the qualitative comparison of Fig. 6, we can also find that our proposed outperforms our previous work in many aspects. PolyCity improves the polygonization performance on small buildings owing to the provided GT bounding box, which is also benefitted from the fact that the cropped image of the small building is resized to a larger image as the network input. In addition, the polygonal building extraction results of PolyCity (whether using GT or predicted bounding box) has finer boundary with more accurate short edges, which is owing to the usage of HR-Net (instead of Res-U-Net) on cropped and resized input images (instead of the complete large input images).

Table 4
Quantitative comparison with CVNet (Xu et al., 2022) and FFL (Girard et al., 2021) on Inria-building dataset. The results are evaluated in terms of IoU, Boundary F-score (B-F-score), Vertex F-score (V-F-score), Vertex Precision (V-P @3pix), and Vertex Recall (V-R @3pix).

Method	IoU	B-F-score	V-F-score	V-P @3pix	V-R @3pix
CVNet (Xu et al., 2022)	0.7887	0.4032	0.1515	0.1079	0.2543
FFL (Girard et al., 2021)	0.7814	**0.6171**	0.3895	0.3219	0.4929
Ours	**0.8040**	0.6115	**0.5315**	**0.5728**	**0.4958**

Table 5
Quantitative comparison with our previous work (Li et al., 2021c) on SpaceNet-Vegas dataset. The results are evaluated in terms of IoU, Boundary F-score (B-F-score), Vertex F-score (V-F-score), Vertex Precision (V-P @3pix), and Vertex Recall (V-R @3pix).

Method	IoU	B-F-score	V-F-score	V-P @3pix	V-R @3pix
Li et al. (2021c)	0.8655	0.8501	0.6537	0.6950	0.6170
Ours (pr-bbox)	0.8672	0.8524	0.6819	0.7316	0.6386
Ours (gt-bbox)	**0.8862**	**0.8635**	**0.7172**	**0.7607**	**0.6785**

5.3. Evaluation of each module

Table 6 lists the performance of each module evaluated on Vegas and SLC datasets, in terms of IoU, B-F-score, and V-F-score. The first row shows the prediction performance of the multi-task learning stage. To enable a fair comparison in terms of all metrics, the mask contour obtained from the pixel-wise multi-task network is converted into polygon vertices through the Douglas–Peucker algorithm (Wu and Marquez, 2003) following the same manner as Li et al. (2021c). The second row shows the results obtained from the vertex selection module (VSM), while the third row shows the final results obtained from the vertex refinement network (VRN).

Experimental results show that the vertex selection module improves the prediction performance of baseline results in terms of all evaluation metrics. Through effectively leveraging the multi-task network outputs to filter out the redundant polygon vertices and remain the valid polygon vertices, the vertex selection module achieves a better vertex prediction performance while maintaining a high IoU and Boundary F-score of the polygon, indicating the effectiveness of VSM and its superiority to Douglas–Peucker algorithm. Moreover, VRN further promotes the prediction performance and significantly improves the vertex F-scores by 7.81% and 4.59%, which is achieved via adjusting the valid polygon vertices generated from VSM to more accurate places. From the qualitative comparison shown in Fig. 8, we can also find the visual improvement of the predicted building polygons after using the vertex refinement network.

5.4. Evaluation of robustness to bbox noise

In our previous experiments, the input image of each building instance is cropped by bbox without noise, which is obtained via enlarging the ground truth polygon by a fixed value of 15% following Acuna

Fig. 7. Building extraction results of PolyCity in large-scale areas. PolyCity generates vector building footprints with accurate vertices, edges, and shapes even for complex scenes.

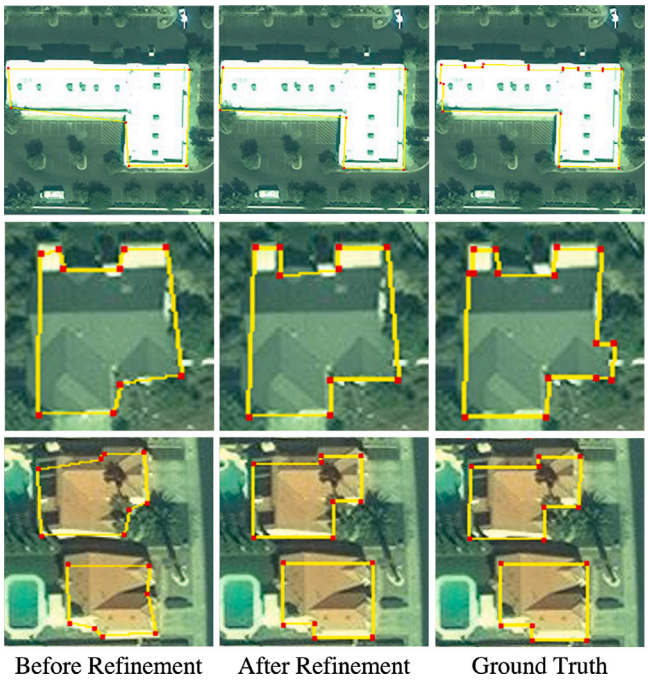

Fig. 8. Qualitative comparisons of the predicted building polygons before and after the vertex refinement network.

Table 6
The performance of each module on Vegas and SLC datasets, in terms of IoU, B-F-score, and V-F-score.

Dataset	Method	IoU	B-F-score	V-F-score
Vegas	Baseline	0.8804	0.8486	0.6240
	+ VSM	0.8836	0.8493	0.6391
	+ VRN	**0.8862**	**0.8635**	**0.7172**
SLC	Baseline	0.8743	0.8932	0.7294
	+ VSM	0.8751	0.8967	0.7556
	+ VRN	**0.8771**	**0.9114**	**0.8015**

Table 7
Evaluation of robustness to bbox noise, in terms of IoU, B-F-score, and V-F-score.

Noise level	Methods	IoU	B-F-score	V-F-score
None	Polygon-RNN++	0.8758	0.8441	0.6875
	Polygon-GCN	**0.8871**	0.8559	0.2759
	Ours	0.8862	**0.8635**	**0.8015**
Moderate	Polygon-RNN++	0.8655	0.8197	0.6301
	Polygon-GCN	**0.8816**	0.8451	0.2869
	Ours	0.8799	**0.8609**	**0.6978**
Large	Polygon-RNN++	0.8395	0.7608	0.5357
	Polygon-GCN	0.8596	0.7845	0.2709
	Ours	**0.8729**	**0.8453**	**0.6709**

et al. (2018), Ling et al. (2019). In this section, we further evaluate and compare the robustness of different methods to bbox noise.

Table 7 shows the results of the Vegas dataset obtained from different methods, when provided with GT bounding boxes in different scales. Specifically, the side length of a GT bounding box is enlarged by: (1) a fixed value of 15% (No noises); (2) a random value in the range of 15% to 30% (Moderate noise); (3) a random value in the range of 10% to 50% (Large noise). For all cases, our method achieves better boundary F-scores and Vertex F-scores compared with Polygon-RNN++ and Polygon-GCN. Although Polygon-GCN shows a slightly higher IoU (smaller than 0.2%) in the first two cases, its vertex F-scores are significantly lower than our method. Results demonstrate the superiority of our method regarding the robustness to bounding box noise compared with state-of-the-art methods.

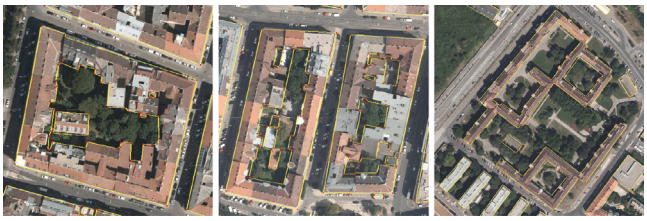

Fig. 9. Extraction results of our method for buildings with holes.

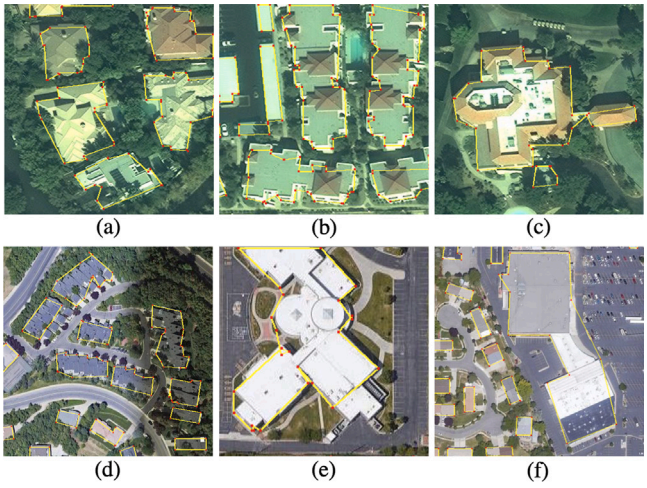

Fig. 10. Typical failure cases of our method on different datasets.

5.5. Complex case analysis

In this section, we analyze the performance of our method for extracting building polygons in complex scenarios. Fig. 9 shows the results of our method for buildings containing multiple polygons (with holes). In our method, the inner and outer contours corresponding to the same building are processed individually for both vertex selection and vertex refinement modules. Consequently, the complex multi-polygon cases can be effectively handled using our proposed method. Although the proposed method achieves promising building extraction results and outperforms state-of-the-art methods in many aspects, there are still some failure cases that should be further improved. Fig. 10 provides some typical failure cases obtained from our proposed method. First, for buildings with curved walls, our method can only predict the general shape of the contour and the performance is worse than those without curved walls (see Fig. 10-(e)(f)). Second, our method has difficulty in precisely extracting the footprint polygons for off-nadir buildings, in which a portion of polygon vertices are invisible due to the parallax effect (see Fig. 10-(b)). Moreover, for some extremely challenging cases, e.g., buildings that are seriously sheltered by trees (see Fig. 10-(a)), buildings with a large number of short edges (see Fig. 10-(d)),infrequent shapes and appearance (see Fig. 10-(c)(e)(f)), etc., the prediction results of our method are still not accurate enough and require further improvements in our future work.

6. Conclusions

In this work, we have presented an effective approach for extracting building polygons from high-resolution remote sensing images, which solves the limitations of existing polygonal object segmentation methods and produces vector buildings that are in a desirable format for actual applications. The complete building segmentation pipeline of our proposed approach contains the following three components. First, a multi-task deep neural network is designed for pixel-wise building segmentation, boundary prediction and edge orientation prediction. Second, a vertex selection module is proposed for transforming the segmentation mask into valid polygon vertices using the three types of network outputs and prior knowledge-based selection rules. Finally, a graph-based vertex refinement network is designed for further adjusting the valid polygon vertex coordinates to more accurate locations, producing the final vector buildings with more precise vertices, edges and shapes.

We conduct substantial experiments on two large-scale building extraction datasets, i.e., the Las Vegas dataset of SpaceNet challenge and the SLC dataset of Microsoft US building footprints. Results show that our approach achieves promising prediction accuracy in terms of IoU and boundary scores, and improves the vertex scores by 3%–4% compared with current state-of-the-art methods. We also provide a detailed analysis of the effect of each module, the robustness to bounding box noise, and the typical failure cases of the proposed method. In our future work, we would like to design new strategies and methods to further improve the polygonal building segmentation performance, such as improving the network architecture and multi-task learning strategies, using the instance segmentation models, etc. We will also improve the proposed method and apply it to larger-scale study areas and cross-city application scenes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant No. 42201358 and T2125006), and the Jiangsu Innovation Capacity Building Program (Grant No. BM2022028).

References

Acuna, D., Ling, H., Kar, A., Fidler, S., 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 859–868.

Alshehhi, R., Marpu, P.R., Woon, W.L., Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 130, 139–149.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

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

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5230–5238.

Chen, R., Li, X., Li, J., 2018a. Object-based features for house detection from RGB high-resolution images. Remote Sens. 10 (3), 451.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al., 2019. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983.

Chen, Q., Wang, L., Waslander, S.L., Liu, X., 2020. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. ISPRS J. Photogramm. Remote Sens. 170, 114–126.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., Waslander, S.L., 2018b. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. arXiv preprint arXiv:1807.09532.

Cheng, D., Liao, R., Fidler, S., Urtasun, R., 2019. Darnet: Deep active ray network for building segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7431–7439.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, IEEE, pp. 172–17209.

Dyken, C., Dhlen, M., Sevaldrud, T., 2009. Simultaneous curve simplification. J. Geogr. Syst. 11 (3), 273–289.
Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5891–5900.
Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 183, 240–252.
Gur, S., Shaharabany, T., Wolf, L., 2020. End to end trainable active contours via differentiable rendering. In: International Conference on Learning Representations.
Hatamizadeh, A., Sengupta, D., Terzopoulos, D., 2020. End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery. In: European Conference on Computer Vision. Springer, pp. 730–746.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Homayounfar, N., Ma, W.-C., Kowshika Lakshmikanth, S., Urtasun, R., 2018. Hierarchical recurrent attention networks for structured online maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3417–3426.
Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115.
Huang, W., Tang, H., Xu, P., 2021. OEC-RNN: Object-oriented delineation of rooftops with edges and corners using the recurrent neural network from the aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–12.
Huang, X., Zhang, L., 2011. 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.
Hui, J., Du, M., Ye, X., Qin, Q., Sui, J., 2018. Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network. IEEE Geosci. Remote Sens. Lett. 16 (5), 786–790.
Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 11–19.
Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019a. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. Remote Sens. 11 (4), 403.
Li, M., Lafarge, F., Marlet, R., 2020. Approximating shapes in images with low-complexity polygons. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8633–8641.
Li, L., Liang, J., Weng, M., Zhu, H., 2018. A multiple-feature reuse network to extract buildings from remote sensing imagery. Remote Sens. 10 (9), 1350.
Li, W., Meng, L., Wang, J., He, C., Xia, G.-S., Lin, D., 2021a. 3D building reconstruction from monocular remote sensing images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12548–12557.
Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021b. Building footprint generation through convolutional neural networks with attraction field representation. IEEE Trans. Geosci. Remote Sens. 60, 1–17.
Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R., 2015. Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.
Li, Z., Wegner, J.D., Lucchi, A., 2019b. Topological map extraction from overhead images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1715–1724.
Li, W., Zhao, W., Zhong, H., He, C., Lin, D., 2021c. Joint semantic–geometric learning for polygonal building segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 3. pp. 1958–1965.
Liang, J., Homayounfar, N., Ma, W.-C., Wang, S., Urtasun, R., 2019a. Convolutional recurrent network for road boundary extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9512–9521.
Liang, J., Homayounfar, N., Ma, W.-C., Xiong, Y., Hu, R., Urtasun, R., 2019b. PolyTransform: Deep polygon transformer for instance segmentation. arXiv preprint arXiv:1912.02801.
Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S., 2019. Fast interactive object annotation with curve-gcn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5257–5266.
Liu, X., Chen, Y., Wei, M., Wang, C., Gonçalves, W.N., Marcato, J., Li, J., 2021. Building instance extraction method based on improved hybrid task cascade. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Liu, Z., Tang, H., Huang, W., 2022. Building outline delineation from VHR remote sensing images using the convolutional recurrent neural network embedded with line segment information. IEEE Trans. Geosci. Remote Sens. 60, 1–13.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: 2017 IEEE International Geoscience and Remote Sensing Symposium. IGARSS, IEEE, pp. 3226–3229.
Mahmud, J., Price, T., Bapat, A., Frahm, J.M., 2020. Boundary-aware 3D building reconstruction from a single overhead image. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.
Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R., 2018. Learning deep structured active contours end-to-end. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8877–8885.
Microsoft, 2022. Microsoft US building footprints. URL https://github.com/microsoft/USBuildingFootprints.
Ok, A.O., Senaras, C., Yuksel, B., 2012. Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery. IEEE Trans. Geosci. Remote Sens. 51 (3), 1701–1717.
Paisitkriangkrai, S., Sherrah, J., Janney, P., Van Den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (7), 2868–2881.
Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 724–732.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. ISPRS J. Photogramm. Remote Sens. 159, 184–197.
Sun, X., Christoudias, C.M., Fua, P., 2014. Free-shape polygonal object localization. In: European Conference on Computer Vision. Springer, pp. 317–332.
Sun, Y., Zhang, X., Zhao, X., 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.
Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, hough transformation and perceptual grouping. Int. J. Appl. Earth Obs. Geoinf. 34, 58–69.
Van Etten, A., Lindenbaum, D., Bacastow, T.M., 2018. Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.
Wang, J., Meng, L., Li, W., Yang, W., Yu, L., Xia, G.-S., 2022. Learning to extract building footprints from off-nadir aerial images. IEEE Trans. Pattern Anal. Mach. Intell..
Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.
Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. IEEE Trans. Geosci. Remote Sens..
Wu, S.T., Marquez, M.R.G., 2003. A non-self-intersection douglas-peucker algorithm. In: Computer Graphics and Image Processing, 2003. SIBGRAPI 2003. XVI Brazilian Symposium on.
Wu, Y., Xu, L., Chen, Y., Wong, A., Clausi, D.A., 2022. TAL: Topography-aware multi-resolution fusion learning for enhanced building footprint extraction. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Xu, Z., Xu, C., Cui, Z., Zheng, X., Yang, J., 2022. CVNet: Contour vibration network for building extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1383–1391.
Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., Xu, Y., 2018. Building extraction in very high resolution imagery by dense-attention networks. Remote Sens. 10 (11), 1768.
Yuan, Y., Xie, J., Chen, X., Wang, J., 2020. Segfix: Model-agnostic boundary refinement for segmentation. In: European Conference on Computer Vision. Springer, pp. 489–506.
Zhao, K., Kang, J., Jung, J., Sohn, G., 2018. Building extraction from satellite images using mask R-CNN with building boundary regularization.. In: CVPR Workshops. pp. 247–251.
Zhao, W., Persello, C., Stein, A., 2021. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. ISPRS J. Photogramm. Remote Sens. 175, 119–131.
Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1848–1857.