Topological Building Extraction With Bidirectional Prediction From Remote Sensing Images

Mingming Zhang, Ye Du, Zhenghui Hu, Wei Wang, Qingjie Liu, *Member, IEEE*, and Yunhong Wang, *Fellow, IEEE*

***Abstract*—Topological building extraction in remote sensing images is vital for city planning, disaster assessment, and other real-world applications. To meet the requirements of real-world applications, existing building extraction approaches predict topological building by vectorization of binary building masks using multiple refinement stages, leading to complex methodology and poor generalization. To tackle this issue, we propose a topological building extraction approach by directly predicting serialized vertices of each building instance. We observe that the order of serialized vertices from one building is inherently bidirectional, which can be clockwise or counterclockwise. By this new observation, the proposed method learns serialized vertices for each building supervised by the bidirectional constraint. Moreover, we design a cross-scale feature fusion module to obtain building representations with rich spatial and context information, facilitating the following serialized vertex prediction. Besides, a merge strategy is adopted to generate the final topological building from serialized vertices of two directions (clockwise and counterclockwise). Experiments are conducted on three building benchmarks to evaluate the effectiveness of our proposed method. Finally, extensive results show that the proposed approach outperforms state-of-the-art methods highlighting its superiority.**

***Index Terms*—Bidirectional constraint, remote sensing images, serialized vertex prediction, topological building extraction.**

I. Introduction

EXTRACTING topological buildings has been essential for many applications, including urban planning, change detection, disaster assessment, and remote sensing cartography, which is studied by more researchers in the remote sensing community. Since buildings of diverse regions (e.g., rural, industrial, and residential regions) vary significantly in shape, material, and size, automatic topological building extraction remains extremely challenging and urgently needs to be solved.

Manuscript received 4 July 2023; revised 5 November 2023, 17 January 2024, and 23 February 2024; accepted 29 April 2024. Date of publication 10 May 2024; date of current version 30 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176017, and in part by the "Leading Goose" R&D Program of Zhejiang under Grant 2022C03107. *(Corresponding author: Qingjie Liu.)*

Mingming Zhang, Ye Du, Qingjie Liu, and Yunhong Wang are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China (e-mail: sara_@buaa.edu.cn; duyee@buaa.edu.cn; qingjie.liu@buaa.edu.cn; yhwang@buaa.edu.cn).

Zhenghui Hu is with the Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China (e-mail: zhenghuihu2021@buaa.edu.cn).

Wei Wang is with the National Disaster Reduction Center of China, Beijing 100124, China (e-mail: 732196152@qq.com).

Digital Object Identifier 10.1109/JSTARS.2024.3399251

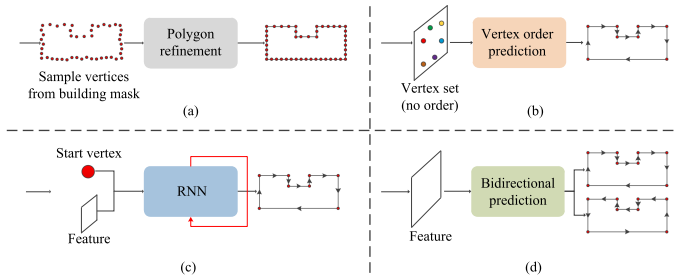

Fig. 1. Topological building extraction with different pipelines. (a) First extract building masks, then perform regularization and polygonization (R&G). (b) Simultaneously predict all vertices of one building, then predict the order of building vertices. (c) Directly predict serialized vertices of one building in one direction. (d) Directly predict serialized vertices of one building under a bidirectional constraint proposed in our article.

Early building extraction studies address this task by segmentation models built with convolutional neural networks (CNNs). These methods first utilize an instance segmentation network to obtain binary building masks by indicating a category for each pixel over remote sensing images. Then, they generate topological building from the predicted masks by regularization and polygonization (R&G) operations, as shown in Fig. 1(a). However, such a paradigm heavily relies on building segmentation masks, from which irregular topological buildings are unexpectedly generated. Therefore, multistage refinement is essential in these methods to pursue high performance.

Building delineation works have recently learned the order of all vertices extracted from the image feature. As illustrated in Fig. 1(b), these methods simultaneously predict order-agnostic vertices using deep learning. Afterward, they predict the order of building vertices typically supervised by additional geometric priors and complex polygonal constraints of buildings in remote sensing images. These approaches have complex model structures and tedious inference processes, resulting in poor generalization capabilities.

To relieve the burden of cumbersome refinement or building vertex's order determination, one straightforward way is to directly predict *serialized vertices* of one building. Fig. 1(c) presents another line of topological building extraction by this observation. Since this idea is simple and intuitive, many works have been studied recently and received encouraging results. However, these works iteratively yield sequential vertices along the building contour in a pre-defined direction (e.g., clockwise).

The modeled dependencies are limitedly *unidirectional*, ignoring the *bidirectional* nature of building vertex prediction. As such, there still exists a risk that outrageous results often occur since unidirectional predictions are susceptible to occlusions and shadows.

We argue that an appropriate building representation and prediction method could guarantee concise and efficient building footprint extraction. To tackle the above issues, our previous work [1] has proposed a building footprint extraction framework to directly generate the vertex sequence of a building. This work extends our previous research [1] in terms of module design and hyper-parameter selection, pipeline in training and inference, and experimental results, which are detailed in Sections III and IV. Fig. 1(d) presents the main pipeline of our proposed topological building extraction framework, which generates serialized vertices of the individual building from remote sensing images. Considering the bidirectional characteristics of building serialized vertices, the proposed method formulates a building instance as serialized vertices and directly learns building serialized vertices under the bidirectional constraint. Finally, a merge strategy is introduced to produce the final result from two directions (i.e., clockwise and counterclockwise). By leveraging *bidirectional* information, the proposed method can generate accurate topological building. In addition, this work embeds the attention constraint to improve prediction accuracy of complex buildings. Moreover, a cross-scale feature fusion module is designed to learn building representations with rich spatial and context information. The cross-scale feature fusion module fuses feature maps of different levels and scales efficiently, facilitating the following serialized vertex prediction.

Similar to Mask RCNN [2], our proposed method adopts a two-stage architecture, embedding a serialized vertex prediction branch into Faster RCNN [3] parallel to building classification and bounding box regression. The proposed method is evaluated on three building extraction benchmarks and compared with state-of-the-art methods. Extensive experiments prove that our method significantly improves building extraction performance, highlighting its effectiveness.

In this article, our method's contributions are summarized as follows.

1) We propose a topological building extraction framework by formulating buildings as serialized vertices with the bidirectional trait. The proposed method leverages *bidirectional* information of serialized vertices to yield accurate topological buildings.
2) We design a cross-scale feature fusion module to fuse multiscale features, which can enhance rich spatial and semantic building representation learning. Moreover, an attention module is embedded into the proposed method to improve prediction accuracy of complex buildings.
3) Our method is evaluated on three building extraction benchmarks with diverse and challenging buildings, including residential, rural, and industrial areas. Compared with instance segmentation methods and polygonal building segmentation methods, our method achieves state-of-the-art on three building benchmarks.

The rest of this article is organized as follows. Section II reviews and summarizes related studies. Section III depicts our proposed approach in detail. Experiment settings and comparisons with different methods are described in Section IV and discussed. Finally, Section V concludes this article.

II. Related Work

Topological building extraction is generally seen as an instance segmentation problem. Traditional methods [4], [5], [6], [7], [8], [9] use hand-crafted features (e.g., textures, geometry, and shadows), thus suffering from poor generalization capabilities. Recently, most studies have addressed this task with a deep learning framework [10], [11], [12], [13], [14] thanks to the robust feature learning capability. Maltezos et al. [14] extracted buildings using a CNN-based deep learning framework from orthoimages and exploited height information as an additional feature to provide potential for building detection. However, they extract buildings by pixel-wise segmentation, resulting in blob-like masks and an inability to distinguish instances. This section mainly reviews literature closely related to our research, which can be categorized into two classes based on the output format (i.e., binary segmentation masks and polygonal building vectors).

A. Building Instance Segmentation

Building segmentation has been a long-standing research topic in the remote sensing community. Early works perform pixel-wise classification to address this problem by grouping pixels to discriminate different instances. Early approaches [15], [16], [17], [18] employ multisource data (e.g., DEM, DOM, and LiDAR) to extract robust building representations, improving the accuracy of building extraction. Awrangjeb et al. [15] presented a building detection framework to extract buildings from LiDAR and multispectral images. Li et al. [16] adopted a variant of U-Net [19] to extract buildings using multiple data sources by combining multispectral images with public GIS datasets. However, these methods generally require fusion either at the feature level or data level, making feature engineering complicated and degrading the model's performance to some extent.

As deep learning has succeeded in computer vision, building segmentation has been typically addressed by instance segmentation. For instance, motivated by Mask RCNN [2], many building instance segmentation approaches have been developed by researchers rapidly. Li et al. [20] presented a cascaded deep neural network architecture, which incorporates region proposal prediction of multiple stages and the Hough transformation to learn better semantic features for building and is jointly trained by multiple losses end-to-endly. Zhao et al. [21] utilized Mask R-CNN [2] to generate building instances and perform boundary regularization to produce topological buildings. Zorzi and Fraundorfer [22] combined adversarial and regularized losses to supervise a fully convolutional neural network (FCN [23]) for boundary refinement and regularization. GSMC [24] proposed a two-stage instance segmentation network and adds a

centroid-aware head to regress the building's geometric center. The network introduces a gated spatial memory module to enhance essential information and add information lacking.

Many real-world applications typically require building layers in vector format rather than building masks in raster format. Therefore, these methods that generate binary building masks in raster format can only serve as an intermediate step. To satisfy requirements in real-world applications, complex post-processing procedures are designed to vectorize building outlines by fitting, regularization, and optimization from binary building masks. However, this pipeline heavily relies on binary segmentation masks, which may produce irregular topological buildings, and usually has a poor generalization. In addition, building segmentation and vectorization are not end-to-end, leading to poor accuracy of building footprint extraction.

B. Topological Building Extraction

Topological building extraction represents buildings as polygons and extracts buildings in vector format. PolygonCNN [25] first extracts building contours through a fully convolutional network and then uses a modified PointNet [26] to adjust the sampling vertices to refine building polygons. PolyTransform [27] generated building masks by a segmentation network and exploited a deforming network to transform vertices sampled from building masks to better fit the building polygon boundaries. FrameField [28] predicted a frame field output to ground truth contours and then combined the frame field output and raster segmentation to achieve building polygons. Zorzi et al. [29] adopted a generative adversarial network (GAN [30]) to regularize building boundaries. Chen et al. [31] refined building polygons using a Relative Gradient Angle (RGA) Transform to project building contours and quantize angles in the RGA domain space. Wei et al. [32] used a contour initialization module to generate an initial polygon, then adopted a contour evolution module to refine polygon vertices. These methods first extract building outlines or masks and then refine the polygon shapes, which are heavily influenced by binary building masks and produce irregular topological buildings.

By representing the building as the building vertex sequence, current methodologies directly predict serialized vertices from the corresponding building feature map. Li et al. [33] used a fully convolutional network to obtain the heatmap of keypoints and then group keypoints into polygon boundaries under the polygonal geometric constraint. APGA [34] determined the order of building vertices by using position and orientation information of building boundaries. Li et al. [35] is a multitask segmentation model integrating different building information to get serialized vertices and then uses a polygon refinement network to predict offset for refining the vertex's position. PolyWorld [36] used a graph neural network (GNN) to organize all the building vertices and then formulated the vertex connection prediction as the optimal transport problem. Although these methods have achieved high accuracy, they typically decompose the topological building extraction task into subtasks and need complex polygon priors, leading to computational intensive and poor generalization.

To tackle the issue above, PolyRNN [37] and PolyRNN++ [38] produce polygonal annotations in a unified CNN-RNN architecture. These two interactive annotation tools allow interactive annotation correction in a human-in-the-loop manner. Curve-GCN [39] simultaneously yields all building vertices by a graph convolutional network, greatly alleviating the sequential nature of Polygon-RNN. These three works require ground truth bounding boxes to train, which are not end-to-end frameworks. PolyMapper [40] uses a ConvLSTM [41] to predict building vertices iteratively and is trained end-to-endly. Zhao et al. [42], based on PolyMapper, combined the building boundary refinement module with channel-wise and spatial-wise attention to improve the model's effectiveness. RNN-based models can only capture information in one direction and are difficult to obtain long dependencies, which may be difficult in complex building extraction.

III. Method

This section describes the proposed method in detail. The proposed method directly predicts building vertices sequentially by representing a building with serialized vertices and leverages *bidirectional* information of serialized vertices to generate accurate building footprints.

A. Overview

Fig. 2 shows the overall pipeline of our method, including three modules: 1) a feature extraction module; 2) a cross-scale feature fusion module (CSFF); and 3) a bidirectional building polygon prediction module with attention mechanism (A-Bi-BP). Multiscale building features are extracted by a feature extraction network instantiated with a CNN-FPN architecture (e.g., the ResNet50 [43] and the feature pyramid network [44]). Subsequently, the enhanced multiscale features are fed into the cross-scale feature fusion module along with building proposals produced by a region proposal network to extract the building representations with high spatial and rich context information. Afterward, the bidirectional building polygon prediction module takes in building representations and outputs two building serialized vertices by a fusion strategy. Moreover, the bidirectional building polygon prediction module can be flexibly embedded into any detection network and trained end-to-endly.

B. Feature Extraction Module

Feature extraction module consists of a deep CNN as the CNN backbone, a feature pyramid network enhancing feature maps of different scales, and a region of building proposal network. In this article, we adopt a ResNet50 [43] as the CNN backbone to extract multiscale feature maps $\{C_2, C_3, C_4, C_5\}$ from an input $I \in R^{3\times H\times W}$, where $C_i \in R^{c_i\times H/r_i\times W/r_i}$ ($c_i \in [256, 512, 1024, 2048]$ and $r_i \in [4, 8, 16, 32]$). To improve multiscale building segmentation performance, especially for small and dense buildings, a feature pyramid network [44] fuses multiscale feature maps C_i of different resolutions and obtains the enhanced pyramid feature maps $\{P_2, P_3, P_4, P_5, P_6\}$. Finally, a

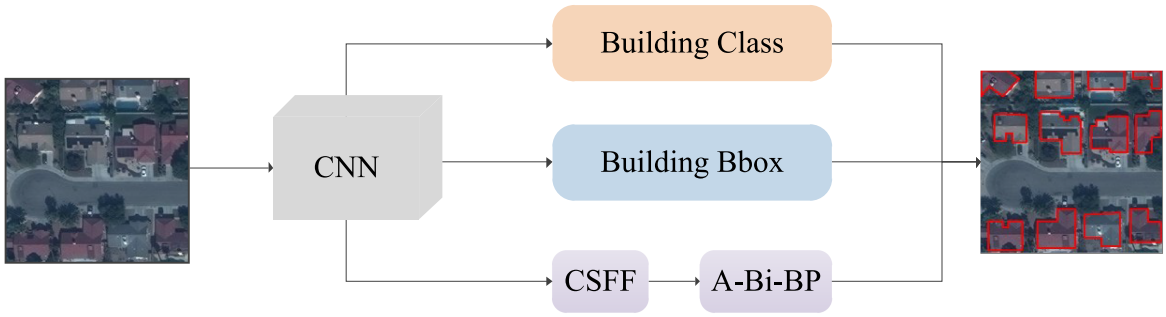

Fig. 2. Overview of our method, an end-to-end framework for topological building extraction from remote sensing images.

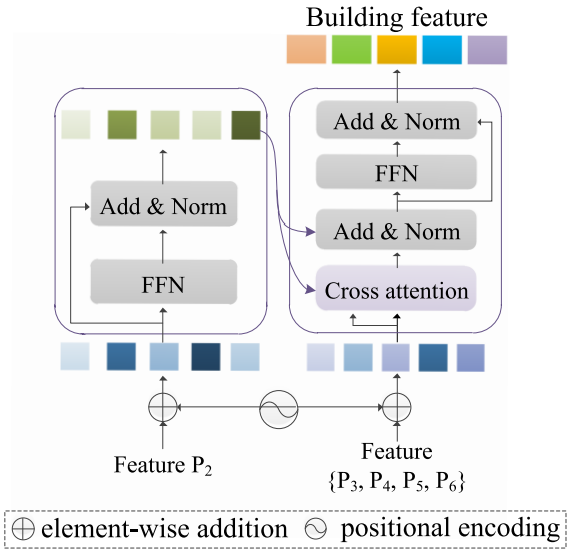

Fig. 3. Pipeline of the cross-scale feature fusion module, enhancing high spatial and rich context building representation learning.

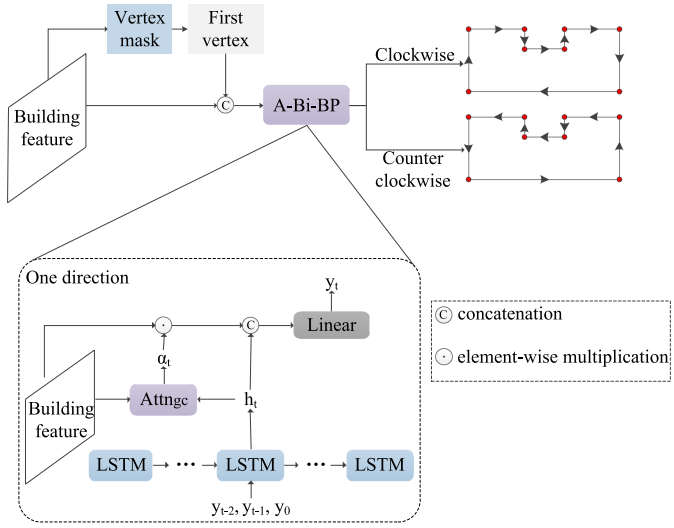

Fig. 4. Pipeline of the bidirectional building polygon prediction with an attention mechanism module (A-Bi-BP). This module directly predicts building serialized vertices in a two-directional manner.

region of building proposal network generates candidate building regions from each feature map P_i of multiscale feature maps.

C. Cross-Scale Feature Fusion Module

Serialized vertex prediction requires building features with high spatial and rich context information. Most existing strategies resize feature maps by up- or downsampling and fuse them by simple pixel-wise summation, which will lose some details. As shown in Fig. 3, the proposed cross-scale feature fusion module utilizes a transformer-based architecture to produce building representations with high spatial resolution and rich context information. Specifically, the cross-scale feature fusion module adopts a cross-attention mechanism to fuse multiscale features coarse-to-finely. Moreover, the positional encoding is embedded with projected features, which can better localize building instances and boundaries.

CSFF encoder: Given enhanced multiscale features $\{P_2, P_3, P_4, P_5, P_6\}$, the CSFF encoder first takes in the feature map $P_2 \in R^{c_2 \times H/4 \times W/4}$ and its positional encoding. Then, it flattens the spatial dimension to the size of H/4×W/4 since the CSFF encoder requires a sequence as input. Subsequently, a position-wise feed-forward network (FFN) is used to get the building query B_q. Moreover, the short-cut connection along with a layer normalization operation is added after FFN. The calculation procedure of the encoder is defined as follows:

$$B_q = \text{FFN}(P_2 \oplus \text{PE}(P_2)) \tag{1}$$

where PE is the positional encoding operation, $\oplus$ means the element-wise summation, and FFN represents the position-wise FFN.

CSFF decoder: CSFF decoder consists of four identical decoder blocks that each block includes a cross-attention layer and an FFN layer. CSFF decoder blocks take in feature maps $\{P_3, P_4, P_5, P_6\}$ from FPN, respectively. Each decoder block first performs a positional encoding operation with one feature P_i $(i \in [3, 4, 5, 6])$. Then, a cross-attention module aggregates multiscales semantic information between building query B_q from the CSFF encoder and feature P_i $(i \in [3, 4, 5, 6])$ from the FPN. In this way, building query B_q can capture rich spatial and semantic information from features of different levels, improving the performance of building serialized vertex prediction. Besides, a short-cut connection and a layer normalization operation are also employed. Finally, we can get the building feature

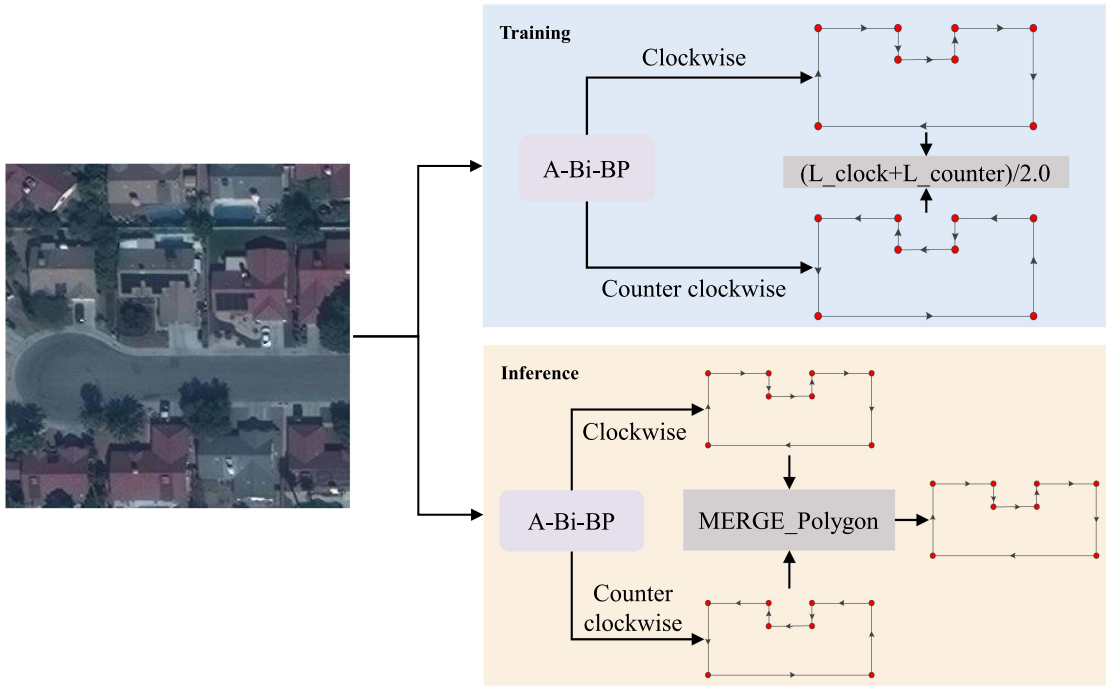

Fig. 5. Training and inference of the A-Bi-BP module. Top: Training stage. Bottom: Inference stage (Polygon selection strategy: First sum the score of building vertices in each direction, then select the one with the highest score as the final building polygon.).

B, which is defined as

$$B = \text{FFN}\left(\text{softmax}\left(\frac{B_q P_i^k}{\sqrt{d}}\right) \cdot P_i^v\right) \tag{2}$$

where FFN is the position-wise feed-forward network. B_q is the building query from the CSFF encoder. P_i^k and P_i^v ($i \in [3, 4, 5, 6]$) are projected from features of different levels from FPN as the key and value vector.

D. Bidirectional Building Polygon Prediction Module

As shown in Fig. 4, the bidirectional building polygon prediction module first generates the first building vertex y_0 from the building feature B. Then, the module produces serialized vertices of two directions. Next, step t is described to show the building polygon prediction with the attention mechanism (A-BP), which is one direction of the A-Bi-BP module as shown in the bottom side of Fig. 4.

First, the building polygon prediction module concatenates building feature B from the previous CSFF module, y_{t-2}, y_{t-1} and the first building vertex y_0 to get the input, where y_{t-2} and y_{t-1} represent the output of step t-2 and step t-1, respectively. Then, building polygon prediction module outputs the hidden state h_t from the previous input. Furthermore, an attention module with Gaussian constraints [45] calculates the attention weight α_t to integrate building feature B and hidden state h_t, which can better focus on the local region of the previously predicted vertices y_{t-2} and y_{t-1}. Afterward, the region-related coefficient is calculated from attention weight α_t and building feature B using an element-wise product. Finally, building vertex y_t at step t or the end signal is generated from hidden state h_t and region-related coefficient. The procedure of building polygon prediction module at step t is defined as

$$\begin{aligned}(h_t, c_t) &= \text{LSTM}(B, y_{t-2}, y_{t-1}, y_0),\\ \alpha_t &= \text{attn}_{\text{gc}}(B, h_t),\\ y_t &= \text{softmax}(W[h_t; \alpha_t \odot B])\end{aligned} \tag{3}$$

where y_0 is the first building vertex, y_{t-2} and y_{t-1} represent output of step t-2 and step t-1 respectively, and B is the building feature map. attn_{gc} means the attention module with Gaussian constraints. $\alpha_t \odot B$ calculates the dot product between α_t and B. $[;]$ is a concatenation operation. W is the trainable parameter.

As depicted in Fig. 5, the pipeline of bidirectional building polygon prediction module between training and inference is different. In the training stage, the bidirectional building polygon prediction module is trained by the mean loss between bidirectional directions and the corresponding ground truth. In the inference stage, we introduce the MERGE_Polygon operation to select the final building polygon. In MERGE_Polygon, it first gets the score of building serialized vertices by computing the sum of all vertex confidences in each direction. Then, it selects the one with the highest score to get the final building polygon from two directions.

E. Training Objective

Our proposed approach includes three branches for binary classification, bounding box regression, and building serialized vertex prediction. Consequently, the loss consists of three parts: 1) a binary cross entropy loss for building classification loss L_{cls}; 2) a L1 loss for building bounding box regression loss

TABLE I
DETAILS OF THREE BUILDING DATASETS

Dataset	GSD (cm)	Image size	Sensor	Training		Testing	
				Images	Buildings	Images	Buildings
SpaceNet (LV)	30	650×650	WorldView-3	3080	87534	385	11269
5M-Building	80	512×512	GaoFen-2	2961	62487	799	15550
CNData	30	512×512	WorldView-3	3360	81104	420	10163

Fig. 6. Some samples in three building datasets. Row (a): SpaceNet (Las Vegas) containing urban and suburban areas. Row (b): 5M-Building containing buildings with considerable variation, intraclass diversity, and interclass similarity. Row (c): CNData containing residential, rural, and industrial areas resulting in different data distribution, especially in rural and urban villages.

L_{reg}; and 3) a building serialized vertex prediction loss L_{ver} by calculating the mean loss of two directions, in which the loss of each direction uses the cross entropy loss between the corresponding prediction with ground truth. The final loss is calculated by:

$$L_{\text{ver}} = (\text{L}_{\text{ce}}(pred, gt) + \text{L}_{\text{ce}}(pred_c, gt_c))/2.0,$$
$$L = L_{\text{cls}} + L_{\text{reg}} + L_{\text{ver}} \tag{4}$$

where $pred$ and gt ($pred_c$ and gt_c) represent building serialized vertices and ground truth in clockwise (counterclockwise).

IV. EXPERIMENTS

A. Building Datasets

Our proposed method is evaluated on three building datasets, i.e., the 5M-Building dataset, the SpaceNet (Las Vegas) dataset, and the CNData dataset. Table I provides details of the following three building datasets in terms of Ground Sampling Distance (GSD), image size, sensor, image, and building number for train/test subsets.

1) SpaceNet (Las Vegas) dataset [46] is a building benchmark dataset consisting of 3851 images across Las Vegas collected from WorldView-3 satellite. The size of images is 650×650 with a GSD of 0.3 m/pixel. SpaceNet dataset contains 151 367 building polygon footprints in GeoJSON format. In our experiment, this dataset is randomly divided into train/test/validation subsets with the ratio of 8:1:1. With high spatial resolution, this dataset mainly consists of urban and suburban areas as shown in Fig. 6(a), and is well annotated in vector format.
2) 5M-Building [47] contains 109 panchromatic (PAN) images and multispectral (MS) images of resolution 0.8 and 3.2 m, respectively. It covers the Shandong province of China and is acquired by the GaoFen-2 satellite, which includes residential buildings, factory buildings, and other buildings. Image size of 5M-Building ranges from 2000×2000 to 5000×5000. In our experiment, we fuse PAN and MS images using the Brovery fusion

TABLE II
RESULTS ON SPACENET (LAS VEGAS) TEST DATASETS

Method	AP ↑	AP_{50} ↑	AP_{75} ↑	AR ↑	AR_{50} ↑	AR_{75} ↑	$F1_{75}$ ↑	C-Area ↑	MTA ↓
PANet [51]	46.9	85.1	45.9	54.6	87.8	60.9	52.35	–	–
PolyMapper [40]	51.6	**87.4**	59.6	58.3	**89.5**	68.9	63.91	63.4	32.5
FrameField [28]	**53.6**	84.5	**63.1**	58.5	87.9	68.8	65.83	22.6	53.6
Baseline [2]	47.0	85.9	46.4	54.6	88.0	60.5	52.52	–	–
Ours	$53.2_{+6.2}$	$87.2_{+1.3}$	$62.1_{+15.7}$	**59.3**$_{+4.7}$	**89.5**$_{+1.5}$	**70.1**$_{+9.6}$	**65.86**$_{+13.34}$	**65.7**$_{+2.3}$	**32.1**

The best results are marked in bold.

TABLE III
RESULTS ON 5M-BUILDING TEST DATASETS

Method	AP ↑	AP_{50} ↑	AP_{75} ↑	AR ↑	AR_{50} ↑	AR_{75} ↑	$F1_{75}$ ↑	C-Area ↑	MTA ↓
PANet [51]	31.3	59.6	28.9	45.8	74.7	48.5	36.22	–	–
PolyMapper [40]	32.0	62.9	30.1	47.5	82.1	51.1	37.88	64.8	26.8
FrameField [28]	18.4	36.4	16.4	31.0	54.7	31.0	21.45	11.1	51.1
Baseline [2]	31.5	60.5	29.0	45.9	75.4	48.3	36.24	–	–
Ours	**32.7**$_{+1.2}$	**63.8**$_{+3.3}$	**31.1**$_{+2.1}$	**48.7**$_{+2.8}$	**83.9**$_{+8.5}$	**52.9**$_{+4.6}$	**39.17**$_{+2.93}$	**65.4**$_{+0.6}$	**26.4**

The best results are marked in bold.

TABLE IV
RESULTS ON CNDATA TEST DATASETS

Method	AP ↑	AP_{50} ↑	AP_{75} ↑	AR ↑	AR_{50} ↑	AR_{75} ↑	$F1_{75}$ ↑	C-Area ↑	MTA ↓
PANet [51]	35.1	68.8	34.0	47.5	81.3	50.3	40.57	–	–
PolyMapper [40]	36.4	70.6	35.7	50.6	86.1	54.6	43.17	63.3	31.8
FrameField [28]	21.7	40.7	21.2	32.9	54.9	34.4	26.23	41.6	36.1
Baseline [2]	35.1	68.4	33.7	47.7	81.6	50.2	40.33	–	–
Ours	**37.9**$_{+2.8}$	**71.5**$_{+3.1}$	**37.7**$_{+4.0}$	**52.7**$_{+5.0}$	**88.3**$_{+6.7}$	**57.3**$_{+7.1}$	**45.48**$_{+5.15}$	**64.7**$_{+1.4}$	**31.4**

The best results are marked in bold.

TABLE V
MODEL COMPUTATIONAL COMPLEXITY

Method	#Params (M) ↓	FLOPs (G) ↓
Mask RCNN [2]	43.8	266.9
PANet [51]	47.3	292.8
PolyMapper [40]	53.8	869.7
FrameField [28]	76.7	833.2
Ours	75.1	923.2

M and G denote million and gillion, respectively.

method [48] to generate large-scale aerial images of high spatial and spectral resolution. Afterward, the fused images are cropped into 512 × 512 and split for training and testing with the ratio of 7:3. Some samples are shown in Fig. 6(b), from which we can see that buildings in the dataset have considerable variation, intraclass diversity, and interclass similarity.

3) CNData contains 4200 images of 512 × 512 (at 30 cm resolution), a very challenging dataset. Images of this dataset cover most provinces in China, including industrial, residential, and rural regions, resulting in large building diversity in shape, material, and size. Especially, urban and rural buildings are typically dense and small. CNData has 101 430 buildings annotated with polygonal labels, which are split by 8:1:1 for train/test/validation subsets. Unlike 5M-Building, CNData contains different areas, resulting in different data distributions as shown in Fig. 6(c).

B. Implementation Details

The proposed model is implemented with PyTorch and trained end-to-endly with the SGD [49] optimizer. ResNet50 [43] is adopted as our backbone, and its learning rate is set to 1e-5, and the other part of the model is set to 1e-4. The weight decay is set to 1e-4. Finally, the proposed model is trained for 24 epochs, with the learning rate decreasing by ten at the 16th and 22nd epochs.

In our experiments, we use two evaluation metrics in the raster and vector levels to evaluate our method. In the raster level, metrics proposed by MSCOCO [50] under different Intersection over Union (IoU) thresholds of segmentation masks are used to evaluate results. In our experiments, AP with the average precision over ten IoU thresholds from 0.50:0.05:0.95, AP_{50} with 0.5 IoU threshold, and AP_{75} with 0.75 IoU threshold are calculated to evaluate the proposed model's precision. Like AP, AP_{50}, and AP_{75}, AR, AR_{50}, and AR_{75} are calculated to evaluate the robustness of our model. Finally, the $F1_{75}$ is calculated from AP_{75} and AR_{75}, which can comprehensively compare different methods and reflect a higher positioning standard.

Finally, we use two metrics to measure the polygon generation of the extracted buildings. The MTA [28] calculates the tangent angles from the lines between a predicted polygonal building and the ground truth, which is lower when the extracted building is similar to the ground truth. Following Truong-Hong and Laefer [52], we use the C-Area to evaluate the polygonal complexity in terms of the vertex number and polygonal area.

Fig. 7. Qualitative results on three building test datasets. We underline and scale up the large and complex buildings for convenient comparison. Column (a): Corresponding ground truth. Column (b)–(f): Building footprint extraction results of our method, Mask RCNN, FFL, PolyMapper, and PANet, respectively. Row 1st and 2nd: SpaceNet (Las Vegas). Row 3rd: CNData. Row 4th: 5M-Building.

The performance of the polygonal building extraction is better when the C-Area indicator is higher.

C. Comparison With State-of-the-Art

To evaluate our approach's performance, we compare it with other state-of-the-art approaches on three challenging building datasets. Considering that building extraction generally is an instance segmentation problem, we compare it with two representative instance segmentation methods (the baseline model Mask R-CNN [2], and PANet [51]). Besides, polygonal building extraction methods are also compared to further verify the effectiveness of our method, which includes PolyMapper [40] and the SOTA Framefield [28].

Quantitative Results: Tables II, III, and IV show extensive experimental results on three challenging building datasets for different methods. We can see from experimental results that our method in this article performs better, demonstrating its superiority in topological building extraction from challenging scenes.

For building instance segmentation, we can see that our method outperforms the baseline method (Mask RCNN) by large margins on all metrics. Specifically, $F1_{75}$ on three datasets are significantly improved by 13.34%, 2.93%, and 5.15%, comprehensively proving that our method can extract buildings accurately. Moreover, AP_{75} on three datasets are improved by 15.7%, 2.1%, and 4%, indicating that our method can extract buildings more precisely. PANet performs similarly to Mask RCNN, as they all output binary building masks in raster format by using the instance segmentation network rather than building geometric information in vector format, making it challenging to learn building boundaries.

For polygonal building extraction, our method outperforms PolyMapper by about 1.95%, 1.29%, and 2.31% on three building datasets in terms of the comprehensive metric $F1_{75}$, reflecting the effect of a higher positioning standard. The significant improvement on CNData shows that our approach is more adaptive to complex shapes since CNData contains different building types of residential, rural, urban villages, and industrial areas. Since 5M-Building and CNData contain buildings in considerable variation, intraclass diversity, and interclass similarity (e.g., continuous urban villages and factory buildings and low and dense urban buildings) that are challenging to extract building boundaries accurately, FrameField performs worse than the baseline method as well as our method. For the SpaceNet (Las Vegas) dataset, where the building boundaries are clear and buildings are basically separate urban buildings with a high spatial resolution, FrameField still performs worse than ours, although it exceeds the baseline method. Furthermore, AR, AR_{50}, and AR_{75} on three building datasets are improved by our

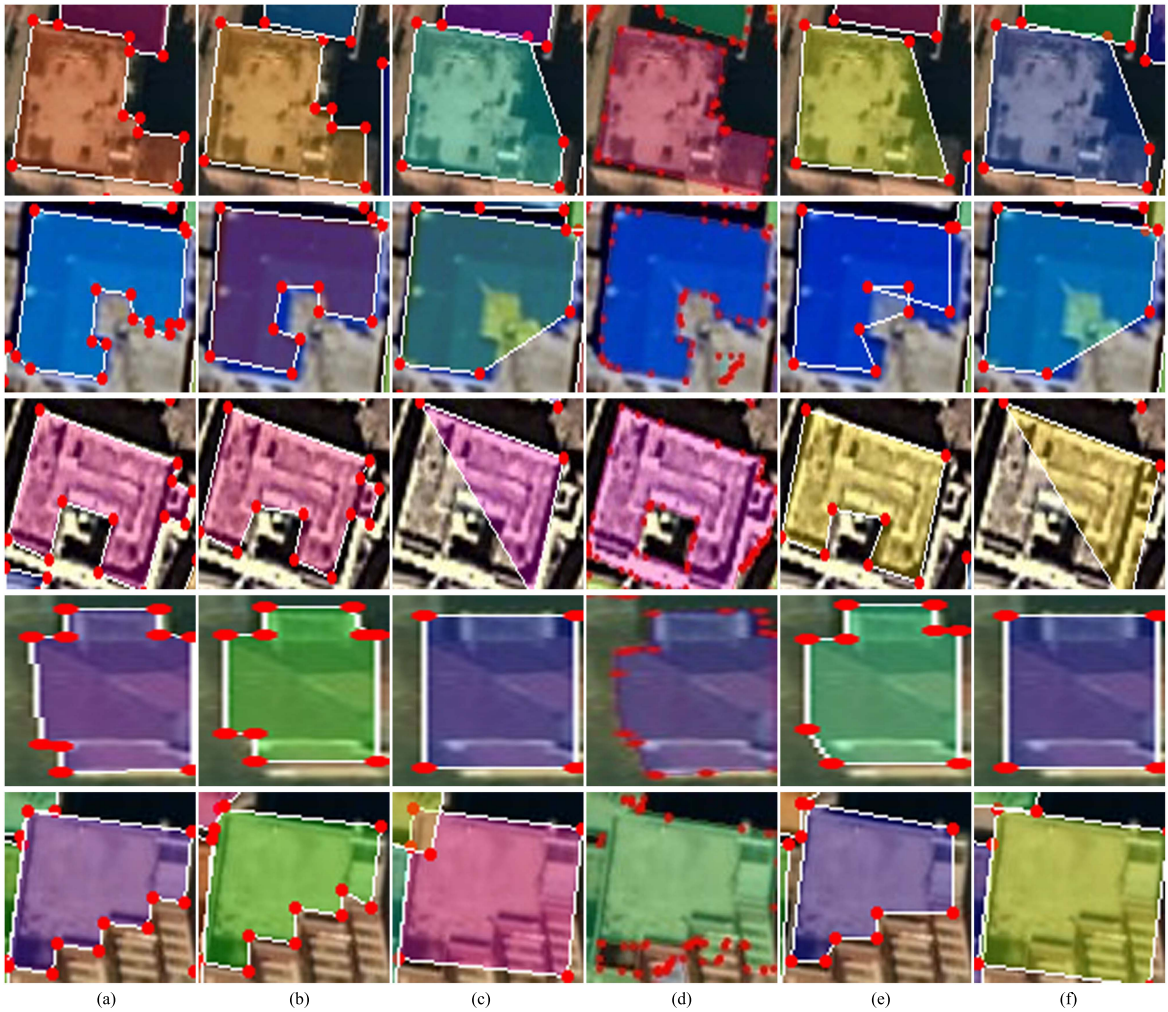

Fig. 8. Qualitative results on CNData. Column (a): GT. Column (b)–(f): Building footprint extraction of our method, Mask RCNN, FFL, PolyMapper, and PANet.

method, especially on SpaceNet (+9.6% in the indicator AR_{75}). For the vector metrics, our method improves the performance of generating polygons with the higher C-Area and the lower MTA, demonstrating that the proposed approach balances the shape similarity and complexity well. The greater improvement in building footprint delineation suggests that our approach performs better in extracting topological buildings, even for building instances with complex shapes.

Finally, we report the number of model parameters (#Params (M)) and floating point operations (FLOPs (G)) of our model and the compared methods to further compare the model complexity. We test an image with a resolution of 1333×800 on 1 GPU for the model complexity evaluation. As reported in Table V, our model has higher parameters than instance segmentation methods since a bidirectional building polygon prediction module is proposed to directly generate serialized vertices. However, our method outputs vectorized building outlines rather than binary building masks. Compared with topological building extraction methods, our model has higher parameters and FLOPs than PolyMapper [40] since it designs a feature fusion module to enhance building features and a bidirectional building polygon prediction module to generate polygonal buildings. In addition, our model has comparable parameters to FrameField [28].

Qualitative Results: Fig. 7 displays the qualitative results of some examples obtained by our method and the comparison methods, which qualitatively illustrate that our approach generates high-quality topological building extraction. From Fig. 7(c), we can see that the baseline method (Mask RCNN) can only give the binary building mask, missing the building boundary information required by the applications in geographic information systems. Although building instance segmentation results are post-processed via traditional contour simplification methods (e.g., Douglas–Peucker [53]), the polygonization process may lose some structural details, thus generating simple building polygons. FrameField [FFL in Fig. 7(d)] can generate building vertices since it can output the frame field aligning to building contours for the polygonization of building masks. However, FrameField predicts many redundant vertices for the individual building and detects some false buildings under the complex background. Therefore, it cannot satisfy the practical applications. PolyMapper can produce building polygons using a ConvLSTM [41] to iteratively predict building vertices, as shown in Fig. 7(e). However, it only captures information in one

Fig. 9. Qualitative results on SpaceNet (Las Vegas). Column (a): GT. Column (b)–(f): Building footprint extraction of our method, Mask RCNN, FFL, PolyMapper, and PANet.

direction and is susceptible to shadow or occlusion, resulting in the error order of building vertices for complex building shapes. Fig. 7(f) reports the prediction results of PANet. PANet is similar to Mask RCNN and requires vectorization of predicted binary building masks.

Our proposed method achieves better performance as shown in Fig. 7(b), because the proposed method considers the bidirectional trait of serialized vertex prediction and yields serialized vertices under the bidirectional constraint. Therefore, the proposed method leverages *bidirectional* information of building polygons, and thus, it can generate accurate building footprints. Besides, a cross-scale feature fusion module generates building representations with high spatial resolution and rich context information, enhancing the ability to predict long sequences of complex buildings.

To further demonstrate the effectiveness of our method, Figs. 8 and 9 show additional qualitative results on SpaceNet (Las Vegas) and CNData datasets. For instance segmentation followed by the polygonization postprocess, Mask RCNN and PANet can only produce simple building polygons, which generally contain four vertices for each building and can not meet the accuracy requirements. For polygonal building segmentation, instead of error vertex order produced by PolyMapper or redundant vertices predicted by FFL, our method can generate concise and accurate building vertex sequences. Since these two datasets are typical and representative, these comparative

TABLE VI
ABLATION STUDY

Dataset	CSFF	Bi-BP	Attn_{gc}	$F1_{75}$
SpaceNet (Las Vegas)				52.52
	✓			64.79
		✓		65.32
			✓	65.44
	✓	✓		65.02
	✓		✓	65.43
		✓	✓	65.54
	✓	✓	✓	**65.86**
5M-Building				36.24
	✓			38.50
		✓		38.58
			✓	39.03
	✓	✓		38.63
	✓		✓	38.72
		✓	✓	39.06
	✓	✓	✓	**39.17**
CNData				40.33
	✓			45.06
		✓		43.35
			✓	44.80
	✓	✓		44.77
	✓		✓	45.37
		✓	✓	45.25
	✓	✓	✓	**45.48**

"✓" represents that the corresponding module is added to the baseline. The final row of every dataset represents our method. The best results are marked in bold for each dataset.

TABLE VII
RESULTS ON CNDATA SHOW THE SELECTION OF THE HYPERPARAMETER *ffn_channel*

Method	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}	$F1_{75}$
base [2]	35.1	68.4	33.7	47.7	81.6	50.2	40.33
base+fc_512	37.3	70.9	37.3	51.6	86.8	**56.2**	44.84
base+fc_1024	**37.8**	**71.7**	**37.6**	**52.0**	**87.6**	**56.2**	**45.06**
base+fc_2048	37.2	70.7	37.2	51.1	86.4	55.4	44.51

The base method is mask RCNN [2]. The base+fc_512, base+fc_1024, and base+fc_2048 mean that the hidden channel of the FFN is set to 512, 1024, and 2048, respectively. The best result is marked in bold.

TABLE VIII
RESULTS ON CNDATA REPORT THE ROLE OF THE CSFF ENCODER

Method	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}	$F1_{75}$
base [2]	35.1	68.4	33.7	47.7	81.6	50.2	40.33
base+noffn	37.5	71.1	37.4	51.7	87.4	**56.2**	44.91
base+fc_1024	**37.8**	**71.7**	**37.6**	**52.0**	**87.6**	**56.2**	**45.06**

The base method is mask RCNN [2]. The base+noffn means that the csff does not have the encoder module. On the contrary, the base+fc_1024 means that the CSFF has the encoder module. The best result is marked in bold.

results demonstrate that the proposed method can generate better building footprints than other methods.

D. Ablation Study

This section analyzes the influence of CSFF, Bi-BP, and Attn_{gc} in Bi-BP, which is added to the baseline method [2] in ablation studies, respectively. Besides, the baseline method with all modules added presents the proposed model. The experimental results evaluated on different building datasets are reported in Table VI and are as follows.

1) *CSFF:* As shown in Table VI, the performance without CSFF decreases by 12.27%, 2.26%, and 4.73% in the indicator $F1_{75}$ on three building datasets. Results of CSFF ablation on different datasets indicate that CSFF is an essential module to aggregate multiscale features efficiently, which is vital for building serialized vertex prediction.
   The proposed CSFF generates building representations with high spatial and rich semantic information by utilizing the transformer-based architecture, which can avoid lacking details in the existing feature aggregation methods. In the CSFF encoder and decoder stages, the hyperparameter *ffn_channel* can affect the building feature learning. Therefore, *ffn_channel* is set as different values (i.e., 512, 1024, and 2048) to explore its impact in this ablation. Table VII shows ablation results about the different selections of the hyperparameter *ffn_channel*. We can see that our model achieves the best result when *ffn_channel*=1024.
   Since the feature map fed to the CSFF encoder has a high spatial resolution, the CSFF encoder does not perform the self-attention operation. Therefore, we verify whether it needs the position-wise FFN or not. We set the hyperparameter *ffn_channel* to 1024 and conduct experiments with and without FFN operation. From the statistical results shown in Table VIII, we can see that the proposed approach is insensitive to the operation. The CSFF encoder, with or without the FFN operation, outperforms the baseline method due to the high-resolution feature map.
2) *Bi-BP:* From Table VI, Bi-BP significantly improves performance on different building datasets by leveraging *bidirectional* information of building polygons. Significantly, Bi-BP improves the baseline by 12.8% in terms of $F1_{75}$ on SpaceNet (Las Vegas), showing its effectiveness.
3) *Attention mechanism:* Results on attention ablation are shown in Table VI, showing that Bi-BP with an attention constraint surprisingly performs better. After adding all modules to the baseline method, the proposed method improves performance on three building datasets from 52.52%, 36.24%, and 40.33% to 65.86% (+13.34%), 39.17% (+2.93%), and 45.48% (+5.15%) in terms of $F1_{75}$, respectively.

In addition, we conduct more ablation experiments to verify the influence of any two modules in CSFF, Bi-BP, and Attn_{gc}, as shown in Table VI. It can be seen that the baseline method with any two modules consistently improves the performance on three building datasets. From ablation experiments, robust experimental results demonstrate the effectiveness of our method for topological building extraction by designing CSFF, Bi-BP, and Attn_{gc} modules.

V. Discussion

In this section, we further analyze our method in terms of model structure and performance then discuss limitations about our method.

A. Analysis

Our method is a two-stage architecture with three specific branches, including classification, box regression, and serialized vertex prediction. The feature extraction module can adopt CNNs or Transformers to extract multiscale feature maps. The bidirectional building polygon prediction module directly generates serialized vertices and can be integrated into other two-stage detection architectures. Unlike previous topological building extraction methods (e.g., PolyMapper [40]), the generated serialized vertices by our approach are more concise since it leverages the bidirectional characteristics of building serialized vertices. Based on our previous BiSVP [1], we have systematically reviewed related literature, summarized existing problems, and added each module's structural design and formula description in detail. In addition to the instance evaluation criteria used in BiSVP [1], we employ additional metrics (MTA and C-Area) to further measure the polygon similarity and complexity in terms of angle, vertex number, and area in the experimental setting. Moreover, we have shown more qualitative results from different methods for convenience reading. In addition, we have conducted more ablation experiments for hyper-parameter selection and performance evaluation.

The experimental results demonstrate the superior performance of our method comprehensively. Firstly, our method can accurately and comprehensively extract buildings. As shown in Tables II, III, and IV, our method achieves the highest $F1_{75}$ and the lowest MTA scores, illustrating that the serialized vertices with bidirectionality can better represent building outlines. Second, our method has similar parameters and computation complexity to other methods, especially 923.2 G/75.1 M vs 833.2 G/76.7 M, as reported in Table V, demonstrating its effectiveness. In addition, as shown in Figs. 7, 8, and 9, our method can better extract polygonal buildings in vector format from different scenes, which is more suitable for downstream tasks than rasterized results.

B. Limitations and Future Work

Although our proposed method achieves promising performances in topological building extraction, buildings with holes may be detected with the wrong vertex sequence in large-scale and complicated scenes. In future work, we will further consider the sequence characteristics of building vertices and seek more appropriate sequence models to solve the problem.

VI. Conclusion

This article has presented an end-to-end topological building extraction method, which can directly generate serialized vertices of each building instance from remote sensing images. The proposed method formulates topological building extraction as predicting building serialized vertices with two directions by the novel observation that the order of building serialized vertices is inherently bidirectional (i.e., clockwise or counterclockwise). Therefore, the proposed method predicts serialized vertices for each building supervised by the bidirectional constraint. Besides, an attention mechanism with Gaussian constraint is integrated with building serialized vertex prediction, enhancing prediction ability for complex buildings. Moreover, a cross-scale feature fusion module is introduced to generate building representations with rich spatial and context information by aggregating multiscale feature maps, essential for building serialized vertex prediction. Finally, a merge strategy is proposed to merge building polygons clockwise and counterclockwise, leveraging bidirectional information to generate accurate buildings. Extensive experiments highlight the proposed method's superiority in topological building extraction.

References

[1] M. Zhang, Y. Du, Z. Hu, Q. Liu, and Y. Wang, "BiSVP: Building footprint extraction via bidirectional serialized vertex prediction," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process.*, 2023, pp. 1–5.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[4] B. Sirmacek and C. Unsalan, "Urban-area and building detection using SIFT keypoints and graph theory," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1156–1167, Apr. 2009.

[5] B. Sirmacek and C. Unsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 211–221, Jan. 2011.

[6] S. Cui, Q. Yan, and P. Reinartz, "Complex building description and extraction based on hough transformation and cycle detection," *Remote Sens. Lett.*, vol. 3, no. 2, pp. 151–159, 2012.

[7] A. O. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[8] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.

[9] M. Turker and D. Koc-San, "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 Observ. Geoinf.*, vol. 34, pp. 58–69, 2015.

[10] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from lidar data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[11] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 371.

[12] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3308–3322, 2019.

[13] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[14] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *J. Appl. Remote Sens.*, vol. 11, no. 4, pp. 042620–042620, 2017.

[15] M. Awrangjeb, M. Ravanbakhsh, and C. S. Fraser, "Automatic detection of residential buildings using lidar data and multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 5, pp. 457–467, 2010.

[16] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 238–241.

[17] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[18] D. Brunner, "Advanced methods for building information extraction from very high resolution SAR data to support emergency response," Ph.D. dissertation, Univ. Trento, Trento, Trento (Italy), 2009.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2017, in *Porc. Int. Conf. Med. Image Comput.Comput.-Assisted Intervention*, Springer, 2015, pp. 234–241.

[20] Q. Li, Y. Wang, Q. Liu, and W. Wang, "Hough transform guided deep feature extraction for dense building detection in remote sensing images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1872–1876.

[21] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 247–251.

[22] S. Zorzi and F. Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5140–5143.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[24] L. Xu, Y. Li, J. Xu, and L. Guo, "Gated spatial memory and centroid-aware network for building instance extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4402214.

[25] Q. Chen, L. Wang, S. L. Waslander, and X. Liu, "An end-to-end shape modeling framework for vectorized building outline generation from aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 114–126, 2020.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[27] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "Polytransform: Deep polygon transformer for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9131–9140.

[28] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building extraction by frame field learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5891–5900.

[29] S. Zorzi, K. Bittner, and F. Fraundorfer, "Machine-learned regularization and polygonization of building segmentation masks," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3098–3105.

[30] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2172–2180.

[31] Y. Chen, Y. Wu, L. Xu, and A. Wong, "Quantization in relative gradient angle domain for building polygon estimation," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 8360–8367.

[32] S. Wei, T. Zhang, S. Ji, M. Luo, and J. Gong, "BuildMapper: A fully learnable framework for vectorized building contour extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 87–104, 2023.

[33] Q. Li et al., "Instance segmentation of buildings using keypoints," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1452–1455.

[34] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen, "Adaptive polygon generation algorithm for automatic building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4702114.

[35] W. Li, W. Zhao, H. Zhong, C. He, and D. Lin, "Joint semantic-geometric learning for polygonal building segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1958–1965.

[36] S. Zorzi, S. Bazrafkan, S. Habenschuss, and F. Fraundorfer, "Polyworld: Polygonal building extraction with graph neural networks in satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1848–1857.

[37] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5230–5238.

[38] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 859–868.

[39] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-GCN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5257–5266.

[40] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1715–1724.

[41] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2015, pp. 802–810.

[42] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 119–131, 2021.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[45] Z. Qiao, X. Qin, Y. Zhou, F. Yang, and W. Wang, "Gaussian constrained attention network for scene text recognition," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3328–3335.

[46] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2019, *arXiv:1807.01232*.

[47] Z. Lu, T. Xu, K. Liu, Z. Liu, F. Zhou, and Q. Liu, "5M-Building: A large-scale high-resolution building dataset with CNN based detection analysis," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell.*, 2019, pp. 1385–1389.

[48] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II channel ratio and "chromaticity" transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

[49] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.

[50] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[51] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[52] L. Truong-Hong and D. F. Laefer, "Quantitative evaluation strategies for urban 3D model generation from remote sensing data," *Comput. Graph.*, vol. 49, pp. 82–91, 2015.

[53] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica, Int. J. Geographic Inf. Geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

Mingming Zhang received the B.S. degree in information and computer science from Liaoning University, Shenyang, China, in 2012, and the M.S. degree in software engineering, in 2020, from Beihang University, Beijing, China, where she is currently working toward the Ph.D. degree in computer science with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering.

Her research interests include remote sensing image analysis and computer vision.

Ye Du received the B.S. degree in computer science and engineering from the Ocean University of China, Qingdao, China, in 2020, and the M.S. degree in computer science and engineering from Beihang University, Beijing, China, in 2023.

His research interests include conputer vision, pattern recognition, and image processing.

Zhenghui Hu received the B.S. degree in computer science from the Zhejiang University of Technology, Hangzhou, China, in 2011, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2020.

She is currently a Senior Research Associate with Hangzhou Innovation Institute, Beihang University. Her research interests include computer vision, hybrid intelligence, and crowdsourcing-based software engineering.

Wei Wang received the Ph.D. degree in cartography and geographic information science from the Chinese Academy of Sciences, Beijing, China, in 2007.

She has been with the National Disaster Reduction Center of China since 2007, mainly engaged in disaster remote sensing application research and application, including satellite and airborne remote sensing platform and payload, remote sensing data processing, disaster related information extraction, UAV disaster prevention and mitigation application research, and business system construction and promotion.

Qingjie Liu (Member, IEEE) received the B.S. degree in computer science from Hunan University, Changsha, China, in 2007, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. He is also a Distinguished Research Fellow with the Hangzhou Institute of Innovation, Beihang University, Hangzhou. His current research interests include image fusion, object detection, image segmentation, and change detection.

Yunhong Wang (Fellow, IEEE) received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively.

From 1998 to 2004, she was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.

Dr. Wang is a Fellow of the IAPR and CCF.