CGSANet: A Contour-Guided and Local Structure-Aware Encoder–Decoder Network for Accurate Building Extraction From Very High-Resolution Remote Sensing Imagery

Shanxiong Chen, Wenzhong Shi, Mingting Zhou, Min Zhang, and Zhaoxin Xuan

***Abstract*—Extracting buildings accurately from very high-resolution (VHR) remote sensing imagery is challenging due to diverse building appearances, spectral variability, and complex background in VHR remote sensing images. Recent studies mainly adopt a variant of the fully convolutional network (FCN) with an encoder–decoder architecture to extract buildings, which has shown promising improvement over conventional methods. However, FCN-based encoder–decoder models still fail to fully utilize the implicit characteristics of building shapes. This adversely affects the accurate localization of building boundaries, which is particularly relevant in building mapping. A contour-guided and local structure-aware encoder–decoder network (CGSANet) is proposed to extract buildings with more accurate boundaries. CGSANet is a multitask network composed of a contour-guided (CG) and a multiregion-guided (MRG) module. The CG module is supervised by a building contour that effectively learns building contour-related spatial features to retain the shape pattern of buildings. The MRG module is deeply supervised by four building regions that further capture multiscale and contextual features of buildings. In addition, a hybrid loss function was designed to improve the structure learning ability of CGSANet. These three improvements benefit each other synergistically to produce high-quality building extraction results. Experimental results on the WHU and NZ32km2 building datasets demonstrate that compared with the tested algorithms, CGSANet can produce more accurate building extraction results and achieve the best intersection over union value 91.55% and 90.02%, respectively. Experiments on the INRIA building dataset further demonstrate the ability for generalization of the proposed framework, indicating great practical potential.**

***Index Terms*—Building extraction, fully convolutional network (FCN), hybrid loss function, multitask learning, very high resolution (VHR) remote sensing imagery.**

Manuscript received July 4, 2021; revised November 11, 2021; accepted December 12, 2021. Date of publication December 28, 2021; date of current version February 9, 2022. This work was supported in part by The Hong Kong Polytechnic University under Grants 1-ZVN6, ZVU1, and 4-BCF7, in part by Hong Kong Innovation, and Technology Commission under Grants SST/051/20GP, and in part by the Beijing Key Laboratory of Urban Spatial Information Engineering under Grants 2020101. *(Corresponding author: Shanxiong Chen.)*

Shanxiong Chen is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong (e-mail: shanxiongchen@whu.edu.cn).

Wenzhong Shi and Min Zhang are with the Smart Cities Research Institute and Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong (e-mail: john.wz.shi@polyu.edu.hk; 007zhangmin@whu.edu.cn).

Mingting Zhou is with the Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: mintyzhou@whu.edu.cn).

Zhaoxin Xuan is with the Beijing Institute of Surveying and Mapping, Beijing 100038, China, and also with the Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100038, China (e-mail: 54786060@qq.com).

Digital Object Identifier 10.1109/JSTARS.2021.3139017

I. Introduction

BUILDINGS are one of the main artificial objects on the earth. Extracting buildings automatically and accurately from remote sensing data is of great significance in cadastral mapping, disaster management, urban monitoring, and many other geospatial applications [1], [2]. Remote sensing enables users to collect data with coverage over extensive areas repeatedly and efficiently. Furthermore, with advances in remote sensor technologies, very high-resolution (VHR) remote sensing data can be acquired, making it possible to ameliorate the quality of the detected building boundaries. However, in practical applications, automatic and accurate building extraction from VHR remote sensing data is still challenging [3]. Buildings come in varied shapes, sizes, heights, locations, and materials, leading to large intraclass differences but small interclass variances. Therefore, developing automatic and robust methods for extracting buildings from VHR remote sensing data is a nontrivial and meaningful task in the remote sensing community. Several attempts have been made to extract discriminative features to distinguish buildings from nonbuildings. Thus, the existing building extraction methods can be roughly sorted into manually designed feature-based algorithms and deep-learning (DL)-based algorithms.

Manually designed feature-based building extraction methods mainly rely on hand-crafted features, intuitively utilizing the implicit or inherent characteristics of buildings. These are based on low-/mid-level features, such as geometric features (e.g., key points [4], lines [1], and contours [5]), spatial-spectral features [e.g., morphological building index (MBI)] [6], contextual features (e.g., shadows) [7], shape features [8], and structure features [9] or object-level features [10]. These methods can extract buildings in a specific task. Still, they can hardly capture high-level semantic information, leading to poor performance in complex scenarios, which is now a more normal situation,

especially in VHR remote sensing images with a more complex background. Thus, manually designed feature-based methods usually have limited capabilities for generalization despite many significant achievements since the hand-crafted features are task-specific, over-specified, and incomplete, especially for VHR remote sensing images. Recent DL-based methods have achieved promising feature learning and classification performance and promoted various research studies toward automatic building extraction [11].

The DL-based building extraction methods learn discriminative features from training data automatically without explicit manual feature design. This approach benefits from developments in convolutional neural network (CNN) theory and from the support of many public datasets [12], [13]. The fully convolutional neural networks (FCNs) [14] are the most commonly used CNN architectures in building extraction [11]. FCN extended the original CNN architecture by replacing the fully connected layer with a fully convolutional layer to enable efficient pixels-to-pixels dense prediction [14]. Despite the promising classification performance of the conventional FCN model, it has two inherent limitations. First, the repeated downsampling process and coarse upsampling layers will lose detailed spatial information, resulting in low boundary localization accuracy [15]. Second, the receptive field of FCNs grows linearly with the increase of network depth. The slow growth still fails to capture the global context information, which leads to misclassification of multiscale objects [16]. Much research has been devoted to tackling these two issues in the computer vision community, with solutions such as U-Net [17] and DeepLab series [18]. U-Net employs encoder–decoder architecture cascades low-level features to high-level features through skip connection, which helps restores the spatial information loss caused by downsampling. DeepLab series reduces downsampling through dilated convolution and introduced atrous spatial pyramid pooling (ASPP) module to fuse multiscale context information. These methods have mitigated the two issues to a certain extent and have been typical and widely used FCN architectures. However, there are still problems when applying these classic semantic segmentation methods to extract buildings from VHR remote sensing imagery. For one thing, buildings in VHR remote sensing images have diverse appearances, complex periphery, and larger scale variances than objects in natural images. For another, buildings are typical man-made objects with abundant morphological properties, and the loss of detailed spatial information limits their potential for practical applications. Therefore, the effective extraction of features while retaining the spatial details of the VHR remote sensing data to obtain accurate building boundaries is a research frontier in the remote sensing community.

Many algorithms have been proposed to enhance the extracted building boundary quality. The straightforward one is to add a postprocessing stage, such as probability graph models [19], [20] and empirical rules [11], [21]. The adopted postprocessing step can refine the segmentation results. However, they are usually complicated methods. Some studies improve the extracted boundary quality with semantic edge detection networks [22], [23]. They have achieved high-quality results, but buildings have highly structured shapes and boundaries rather than all the edges of objects. Therefore, these practices increase the complexity of the model while not achieving the optimal results. The highly structured building shape priors can be encoded into the model through building contour learning. There is an imbalanced foreground/background problem, however, in contour learning since the contour only accounts for a small proportion of all the sample pixels. Contour learning, therefore, must preserve and retain the structural properties of buildings to overcome this problem. However, the most used binary cross-entropy (BCE) loss function only focuses on pixel-level similarity, resulting in the loss of structural information and sensitivity to foreground/background imbalance issues. Inspired by these two observations, methods combining a multitask learning framework with a hybrid loss function to learn building regions and outline simultaneously to refine building extraction results were proposed. Wu *et al.* [24] and Zheng *et al.* [15] designed robust loss function and supervised roof region and outline simultaneously, which show a great performance improvement. But they only supervise the building edges in the last layer of the decoder, which still suffers from the loss of detailed spatial information in the encoder. Actually, the encoder layers have richer spatial details about the original input [25]. These methods have improved the extracted building boundary quality but still overlook the building shape priors or the abundant spatial information in the encoder layers.

A framework integrating multitask learning, stepwise weighting deep supervision techniques, and robust loss function design could address these issues. In our proposed method, a contour-guided (CG) module overcomes spatial information loss and preserve building contour-related low-level features. A multiregion-guided (MRG) module was designed for high-level multiscale and contextual feature capture that overcomes frequently occurrences of building scale variance in VHR remote sensing images. The MRG module is similar to the U-Net encoder–decoder architecture comprised of an encoder, ASPP module, and decoder. The encoder is a modified ResNet34 [26] backbone with fewer downsampling (i.e., three times) to further reduce spatial information loss. Parameters from the encoder layers are shared and updated by the CG and MRG modules jointly through multitask learning. The ASPP module is an effective semantic segmentation module for capturing useful image context at multiple scales. The decoder produces multiregion outputs and is deeply supervised by building region ground truth. The CG module is supervised by building contour ground truth. In this way, complementary building edge semantic features and multiscale building region semantic information are captured. In addition, a pixel position-aware and structure-preserving hybrid loss function is introduced to guide the network to learn parameters from the pixel-level similarity, local structural similarity, and global similarity. The introduced loss function supervises the CG and MRG modules simultaneously to further ameliate the structure learning ability of our model. The main contributions of this research work are given as follows.

1) A contour-guided and local structure-aware encoder–decoder network (CGSANet) is proposed for accurate building extraction from VHR remote sensing imagery.

2) An additional CG module at the encoder stage was devised to overcome the loss of low-level spatial information and the overlook of building shape priors.
3) A hybrid loss function was designed to treat pixels differently and guide the model to learn parameters from the pixel-level similarity, local structural similarity, and global similarity.

The rest of this article is organized as follows. Section II introduces a brief review of the existing DL-based building extraction algorithms. Section III presents the proposed framework. Section IV analyses the experimental results. The discussion and conclusion are drawn in Sections V and VI, respectively.

II. Related Work

Remote sensing data analysis has undergone significant advancement since the application of DL algorithms. The DL-based methods can easily resolve typical remote sensing image analysis tasks, such as scene classification [27], object detection [28], change detection, land use land cover classification [29], and semantic segmentation due to their superior performance in hierarchical feature representation. DL-based building extraction is a binary semantic segmentation problem that aims to classify every image pixel into building and nonbuilding pixels. A brief review of the research efforts on DL-based semantic segmentation methods can be seen from the recent review [30]. In this section, we focus on the development of CNN-based building extraction methods in the remote sensing community.

Building extraction methods based on CNN can automatically learn high-level and discriminative features and have been widely adopted in remote sensing communities. Early work [31], [32] trained a simple patch CNN for building labeling, yielding competitive performance when compared to manually designed feature-based methods. However, patch-based CNN will dramatically increase the memory cost when processing larger patch sizes, thereby significantly reducing its processing efficiency [33]. Some approaches extract buildings based on FCNs since the first success of end-to-end FCN [14] for semantic segmentation. Despite the promising performance, the original FCN cannot fully capture the long-range relationships between pixels in an image [16], leading to incomplete extraction of large objects and missing small objects. Furthermore, FCNs need repeated downsampling operations to extract discriminative features, washing out high-frequency spatial details, leading to blurry extraction boundaries [15]. For building extraction from VHR remote sensing imagery, these two challenges are more critical since larger scale variance and a more regular and clearer shape of buildings. Various algorithms were developed to improve the region segmentation and boundary localization accuracy when extracting multiscale buildings from remote sensing data.

Fusing multisource data, such as combining LiDAR data or nDSM with spectral images, is another approach to extract buildings [3], [34]. Huang *et al.* [3] proposed a gated residual refinement network with LiDAR data and aerial imagery fusion to extract buildings. The fusion of height and spectral information is able to achieve promising results on complex data. Another stream in data-fusion-based methods employ GIS data to assist building extraction [35], [36]. Huang *et al.* [35] created a dataset based on ground truth from the OpenStreetMap project and trained a two-stream deep deconvolution network with RGB and NRG fusion to extract buildings holistically. These data-fusion-based approaches have yielded promising results, but are limited by the difficulty of obtaining large-scale, high-quality co-registered multisource data. Another stream in FCN-based methods employ only a single data source and improve the building extraction results by improving the feature representation in the FCN model.

To extract multiscale buildings more completely, many approaches extract buildings by feature-enhanced FCN model, such as fusing multiscale features [2], [37], [38], adding multi-constraints on additional predictions [33], and introducing attention mechanisms [39], [40]. To fuse multiscale features, Yuan [2] designed an FCN that integrated activation from multiple layers. To enhance the multiscale feature representation ability, Wu *et al.* [33] proposed a multiconstraint FCN (MC-FCN) with extra constraints applied on the intermediate layers. Yang *et al.* [38] proposed a dense-attention encoder–decoder network comprising lightweight DenseNets and spatial attention fusion modules to integrate different level features rationally. By exploiting multiscale hierarchical and contextual features, these feature enhancement-based methods boost the performance of multiscale building region segmentation. However, these methods are still limited due to the complexity of VHR remote sensing images and variation in the morphological properties of buildings. The extracted building boundary accuracy needs further improvement.

Specific strategies and FCN architectures were developed for building extraction with more accurate boundaries. Apart from the feature enhancement-based methods, recent methods mainly include postprocessing methods, semantic edge-assisted methods, and loss function design methods. Postprocessing methods refine the building edges by adding a postprocessing stage [11], [19]–[21]. Many methods employ probability graph models [e.g., Markov random field (MRF) model and conditional random fields (CRFs) model] to optimize preliminary results [19], [20]. Vakalopoulou *et al.* [19] employed a CNN integrating additional spectral information with pretrained features from ImageNet to calculate deep features for automatic building extraction and utilized the MRF model to optimize the extraction results. Shrestha and Vanneschi [20] adopted the CRFs model to improve the extracted building boundaries. Some approaches deploy empirical rules to refine the initial results [11], [21]. Zhao *et al.* [21] introduced a boundary regularization process to optimize segmentation maps from Mask R-CNN to generate regularized building polygons. Wei *et al.* [11] developed a framework for building segmentation and structured footprint extraction; an empirical polygon regularization postprocessing algorithm was designed to refine preliminary FCN-based building segmentation results into structured building footprints. The additional postprocessing step does improve the initial segmentation results, but may improve the computational cost of the methods and cannot be done end-to-end, which limits the application.

The semantic edge-assisted methods perform edge refinement through a combined classic edge detection model and semantic segmentation/object detection networks [22], [23]. Marmanis *et al.* [22] developed deep CNN models combining the SegNet model and the HED edge detection model to explicitly extract the class boundaries to boost the semantic segmentation performance of VHR aerial images. Xia *et al.* [23] proposed a refined building footprint extraction approach based on the Faster R-CNN and the CASENet edge detection model. In addition, they proposed a boundary repair algorithm that further refines incomplete building edges with distinct advantages in terms of the quality of extracted building boundaries against Mask R-CNN. These approaches have achieved promising results, but the buildings are artificial objects with highly structured semantic edges rather than all image edges; therefore, these methods can be improved by explicitly encoded the building shape priors into the model.

Building contours implicitly represent building shape and structure features. The contours are learnable since the building region label can easily generate building contour labels [41]. Researchers adopted a multitask learning framework and designed a hybrid loss function to learn building structure features [15], [24], [42], [43]. The multitask learning paradigm aims to leverage valuable information in multiple related tasks to maximize the performance of one or all of the tasks [44]. Building region segmentation and building outline extraction are two highly dependent tasks. Wu *et al.* [24] proposed a boundary-regulated network (BR-Net), which supervised building regions and outlines simultaneously. They designed a shared U-Net backend to extract local and global features and employed boundary information to regulate the parameter updating, resulting in conspicuous performance improvements. Zheng *et al.* [15] proposed an edge-aware neural network (EaNet) to optimize semantic segmentation boundaries in VHR urban scene images and achieved promising performance in the VHR ground/aerial images. These methods have greatly improved the quality of extracted building boundaries; but they only supervise the building edges in the last layer of the decoder, which is still limited by the loss of detailed spatial information in the encoder layers. Liao *et al.* [41] proposed a boundary-preserved network for building extraction by simultaneously learning building structure and contour. A structural prior constraint module combined with the dice loss function was designed to learn building contour information from the gradient image. This method showed superior performance on building edges, especially for adjacent buildings. However, an additional gradient image is needed for the input data.

The semantic edge-assisted methods focus on spatial information preservation but utilizing all the image edges, which is not optimal as buildings are artificial objects with highly implicit semantic contours. Feature enhancement-based methods often add modules at the end of the encoder. Postprocessing and loss function design-based methods are often performed in the last layer of the decoder. They ignore the abundant spatial information in the encoder layers, and actually, the encoder layers retain finer spatial details about the original input [25]. Inspired by these observations, a CG and local structure-aware encoder–decoder network (CGSANet) is proposed in this article. A CG module appended in the encoder enhances the ability of the model to preserve building contour-related low-level spatial features. A local structure-aware hybrid loss function optimizes the model parameters from the pixel-level similarity, local structural similarity, and global similarity. The architecture of CGSANet, CG module, and hybrid loss function are introduced in Section III.

III. Methodology

CGSANet solves the inaccurate building boundaries problem and corrects incomplete extraction results caused by varied building scales in VHR remote sensing imagery. This is achieved through the network architecture and optimized with a novel hybrid loss function. In this section, we present the framework of the proposed CGSANet and its key components in detail. The whole architecture of CGSANet is introduced in Section III-A. The encoder CG module and loss function are depicted separately in Sections III-B and III-C.

A. Architecture of CGSANet

The overall framework of CGSANet is shown in Fig. 1. The basic architecture of CGSANet is similar to the U-Net [17] encoder–decoder structure, with a skip connection. Our framework integrates the deep supervision technique, multitask learning, and hybrid loss function into a system to obtain accurate building extraction results. Taking a VHR remote sensing imagery as the input, CGSANet outputs four building region predictions and a building edge map, i.e., D0, D1, D2, D3, and D4 in Fig. 1. The former four predictions are deeply supervised by the building region ground truth, and the D4 output is supervised by building edge ground truth to guide the network focus on building semantic edge-related spatial information, thus improving the boundary quality of extracted buildings.

It can see from Fig. 1 that CGSANet is comprised of two modules, a CG module and an MRG module. The MRG module shown in the large aqua green box in Fig. 1 is comprised of an encoder module, an ASPP module, and a decoder module. A modified ResNet34 backbone was selected as the encoder of the MRG module to extract multilevel feature and achieve consistent training [25]. Other networks such as ResNet18 can be applied as the encoder as well. To reduce the loss of spatial information, the first two downsampling layers of ResNet34 were replaced with a plain $3 * 3$ convolution block. The orange box in Fig. 1 indicates a plain convolution block, including a $3 * 3$ convolution (Conv) operator, a batch normalization (BN) layer, and a ReLU operator in a sequential order. The gray box in Fig. 1 is the basic ResNet block. A basic ResNet block comprises two layers of $3 * 3$ Conv/BN/ReLU, as shown with an enlarged grey box at the left bottom of Fig. 1. The four stages of ResNet34 include [3, 4, 6, 3] basic ResNet blocks, shown by the corresponding number of gray boxes in Fig. 1. The red box in the MRG module is the ASPP module. In the encoder stage, the output of the first convolution block is fed into the four standard ResNet34 stages, during which the feature map is downsampled three times. Multiscale features are obtained by repetitive convolution operation on the local receptive field in different scales.

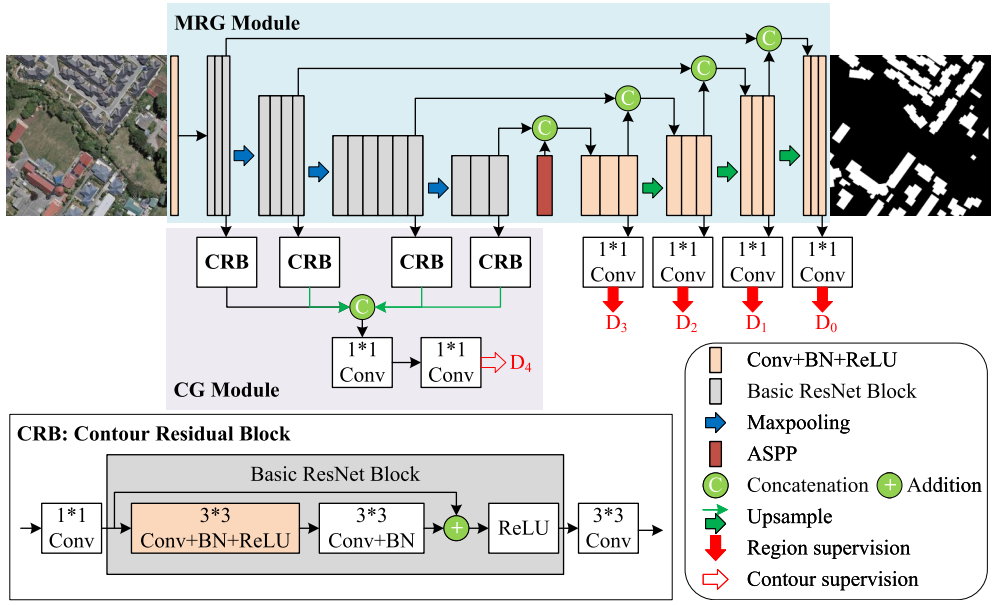

Fig. 1. Architecture of our proposed CGSANet

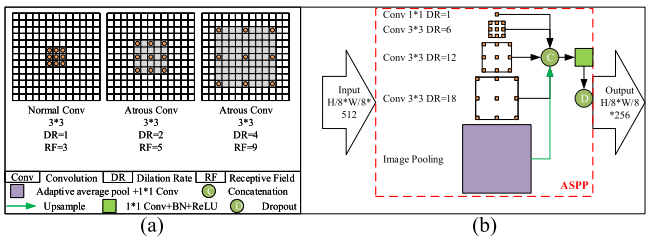

Fig. 2. Illustration of (a) normal convolution and atrous convolution with kernel size three and different dilation rate and (b) the atrous spatial pyramid pooling (ASPP) module within the DeepLabv3+ [18] architecture.

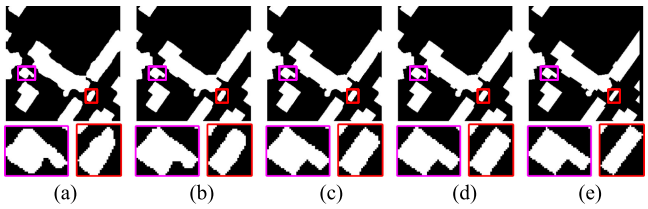

Fig. 3. Four region outputs of decoder layer and the corresponding building region ground truth. Only the bottom-left corner of the sample image is shown for better viewing. (a) D3 8*up output. (b) D2 4*up output. (c) D1 2*up output. (d) D0 output. (e) Building region.

Based on the modified ResNet34 backbone encoder, we introduce ASPP module for capturing multiscale context information to upgrade the model's capability to perceive varied scale targets. The ASPP module is proven to be an effective component in DeepLabv3+ networks [18]. In this work, we employ an ASPP module as a connector between the encoder and decoder module to capture the multiple scales contextual and image-level features. Fig. 2 illustrates the normal convolution, atrous convolution, and ASPP module.

Fig. 2(a) shows that the input imagery is independently convolved with different dilation settings from left to right. The big square box of 13^*13 represents the input image, and each small square box indicates a pixel. The orange circle represents the 3^*3 convolution kernel, and the square with gray color represents the receptive field after convolution. It can see clearly in Fig. 2(a) that the receptive field of the normal convolution with the dilation rate of one is three, whereas the atrous convolution with the dilation rate of two and four have a receptive field of five and nine, respectively. Atrous convolution inserting "holes" into the convolution kernel can effectively provide a larger receptive field without extra downsampling. Normal convolution is a special case of atrous convolution with a dilation rate of one.

Given the feature maps sized $H/8^*W/8*512$ outputted from the last encoder layer, the ASPP module produces a feature map sized $H/8^*W/8^*256$. The H and W represent the height and width of the input imagery. As shown in Fig. 2(b), the ASPP module has four parallel dilated convolutions of different dilation rates (1, 6, 12, 18) that maintain the same feature map. In addition, ASPP introduced global average pooling to capture image-level features, as shown in the bottom of Fig. 2(b). The output feature maps are fused through concatenation, 1^*1 plain convolution block, and dropout layer at the end.

The decoder part of CGSANet is comprised of stacked plain convolution blocks. The plain convolution block is composed of convolutional layers, BN layers, and ReLU activation layers. Each decoder stage is comprised of three consecutive plain convolution blocks. The spatial size and number of filters for each convolution layer are adjusted with the corresponding encoder stage. Bilinear interpolation is applied for upsampling. The classic U-Net like encoder–decoder structure suffers from the problem that the update priority of the middle layer features is lower than the last layer [33]. The deep supervision technique

is known to easily optimize and refine the model feature. Therefore, the deep supervision technique is introduced to solve the above updating priority problem. Furthermore, considering the outputs of different decoder layers have different edge accuracy, we designed a stepwise weighting deep supervision training strategy.

Fig. 3 shows the four outputs we used for deep supervision and the ground truth building region. Note that only the bottom-left part of the sample image is shown for better viewing. The deeper the output layer, the larger the upsampling factor. Thus, the output from the deeper layer will have relatively low boundary accuracy. As shown in Fig. 3, D3 is the closest to the encoder layer and requires eight times of upsampling to restore to the original input image size. It can be seen in the enlarged magenta and red rectangle that the object boundaries in D3 are too smooth. The building edges in D2 and D1 have similar performance, and D0 achieved the best boundary accuracy. The observation from the quantitative evaluation of the WHU test dataset is consistent with the qualitative analysis. The intuitive reason for the stepwise weighting strategy is that as the accuracy increases, the corresponding output image should be highlighted [45]. Therefore, these outputs should be assigned with different weights. D0 is the highest weight, D3 the lowest, and D1 and D2 are equal weights. In our experiments, we achieved the best performance when setting the weights of D0, D1, D2, and D3 to 1, 0.5, 0.5, and 0.3.

An encoder–decoder framework equipped with an ASPP module and deep supervision technique can extract hierarchical contexts and fuse the multilevel information efficiently for multiscale building extraction. However, it still suffers from the overlook of abundant building shape priors and the loss of detailed spatial information, leading to inaccurate and irregular building boundaries. We propose a CG module appended at the encoder part to mitigate this problem. The CG module will be introduced in Section III-B.

B. Building Semantic Edge Information Capture With Encoder CG Module

Building contours indicate the shapes of buildings. The goal in building extraction is to find the contours that distinguish buildings and nonbuilding areas. From this perspective, the accuracy of building extraction depends on the degree of match between the contour of an extracted building and the ground truth. Therefore, the quality of the building contour matters. However, current methods still overlook much of the abundant morphological properties of buildings and spatial information in the encoder layers. Inspired by this observation, residual learning [26], and multitask learning, a CG module appended to the encoder and supervised with building contour ground truth is proposed to effectively extract building semantic edge information. In this section, we detail the proposed CG module.

The CG module is comprised of four contour residual blocks (CRB) and two $1 * 1$ convolution layers. The CRB is the kernel of the CG module. The CRB module extracts spatial features from the encoder layers. Specifically, we expanded the encoder by linking the last layer of each encoder stage with one of the four CRB modules. Each CRB module consists of one $1 * 1$ convolution layer, one basic ResNet block, and one $3 * 3$ convolution layer. The input feature channels of the four CRB modules are 64, 128, 256, and 512, respectively, and the output feature channels of the three parts of each CRB module are 21, 21, and 1. The output of the four CRB modules was upsampled to the original input image size, concatenated, and followed by two processing in $1 * 1$ convolution layers. The final outputs are supervised by building contour. The building contour labels are generated from the building region labels via a Laplacian operator [15], require no additional manual labeling. The intermediate layer output of the CG module on a WHU aerial building test dataset example image is shown in Fig. 4.

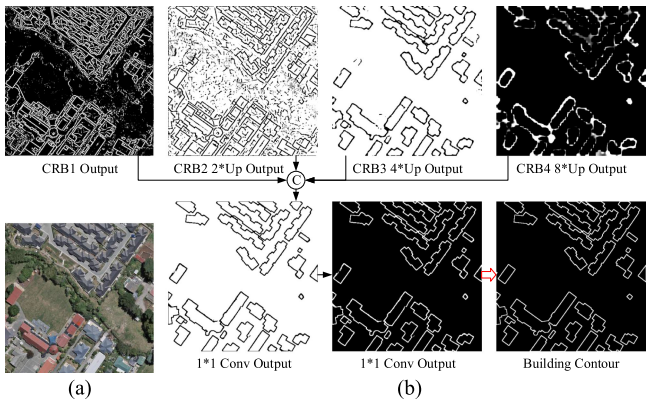

Fig. 4. Illustration of feature maps captured by the CG module on a WHU aerial building dataset image, in which Conv is short for convolution, the red hollow arrow represents contour supervision, and number*Up means number times upsampling. (a) Image. (b) Intermediate output of the CG module and the building contour supervision.

As shown in Fig. 4, CRBs at different stages can obtain different degrees of edge information. The CRB1 and CRB2 capture abundant edge information but have much redundant low-level information. The output of CRB3 is close to the final output, but there are problems given the insufficient distinction between adjacent buildings and incomplete extraction of complex buildings. CRB4 contains little low-level information and contains more regional information. The combination of four CRBs can produce a regular output close to the outline of buildings. It removes redundant low-level information, retains the regular contours, and can effectively distinguish adjacent buildings as well as extract complex buildings. The features acquired by the CG module can ameliorate building extraction results through a multitask learning framework.

The multitask learning paradigm aims to learn multiple related tasks jointly while maximizing performance on one or all tasks [44]. We fuse the building contour and building region tasks to achieve accurate building extraction results. Furthermore, we added the building contour task to the encoder and the building region task to each decoder stage. The parameters of the encoder are shared and updated by the CG module and decoder jointly through multitask learning. This can capture complementary building edge semantic features and multiscale building region semantic information. The whole architecture is a synchronous end-to-end network supervised by the same hybrid loss function that can easily train and fuse the two highly

related tasks for accurate building extraction. The hybrid loss function is introduced in Section III-C.

C. Loss Function

Loss function is essential for model optimization. The pixel-wise BCE loss function is the most applied solution for building extraction. However, The BCE loss function is entirely local and treats every pixel separately, leading to ineffective structure learning ability. Moreover, BCE treats all pixels equally and, thus, is sensitive to imbalanced foreground/background problems. These issues adversely affect building extraction results more than other semantic segmentation tasks as buildings have abundant morphological properties.

Various structure-aware loss functions have been proposed to strengthen the structure capturing ability [15], [46], [47], but these still consider pixels separately, without taking account of the local surroundings of the pixel. Wei *et al.* [45] proposed pixel position-aware hybrid loss function, which treats pixels differently to extend the BCE and intersection over union (IoU) loss function. Inspired by this work, we propose a novel hybrid loss function defined as the summation of weighted BCE (wBCE), structural similarity index metric (SSIM), and weighted IoU (wIoU) to optimize the model parameters from the position-aware pixel-level similarity (wBCE), local structural similarity (SSIM), and position-aware global similarity (wIoU).

The wBCE loss extends BCE by treating pixels differently, which calculates a weight $w_{(r,c)}$ based on the difference between the pixel and its neighborhoods to determine whether the pixel is a hard pixel or a plain pixel, thereby assigning different weights. In this way, $L_{w\text{BCE}}$ pays less attention to simple pixels and vice versa. In addition, $L_{w\text{BCE}}$ integrates local structure information by considering neighborhood pixels [45]. $L_{w\text{BCE}}$ is defined in (1), shown at the bottam of this page, where *r* and *c* represent the row and column of the image; *H* and *W* represent height and width of the image; and λ is a hyperparameter to revise the ratio of hard pixels. The value $w_{(r,c)}$ indicates the assigned weight for each pixel; it is calculated based on the difference between the pixel and its neighborhoods, as shown in Equation (2). $I(\cdot)$ is the indicator function. The notation $l \in \{0, 1\}$ indicates nonbuilding and building. $g_{(r,c)}$ and $p_{(r,c)}$ are ground truth and prediction of the pixel at location (*r*, *c*). Prob$(p_{(r,c)} == l|\psi)$ denotes the predicted probability.

$$w_{(r,c)} = \left| \frac{\Sigma_{i,j \in N_{r,c}} g_{i,j}}{\Sigma_{i,j \in N_{r,c}} 1} - g_{r,c} \right| \tag{2}$$

where $|\cdot|$ indicate absolute operation and $N_{r,c}$ represents the neighborhood of the pixel (*r*, *c*). $g_{(r,c)}$ is the ground truth of the pixel at location (*r*, *c*). For any pixel, $w_{r,c} \in [0, 1]$. A large $w_{(r,c)}$ value indicates a pixel at (*r, c*) is very inconsistent with its neighborhoods. It is a discriminative pixel (e.g., boundary pixels) and should be paid more attention and vice versa. Paying more attention to these challenging pixels can further enhance model generalization.

Buildings are artificial objects and highly structured. Every building has unique morphological properties. The SSIM loss was added to the loss function to preserve building shape properties. Following [47], a local SSIM loss function (L_{SSIM}) was introduced to assess the structural similarity of the extracted buildings. Let $p = \{p_i, \text{i} = 1, ..., N^2\}$ and $g = \{g_i, \text{i} = 1, ..., N^2\}$ denote the pixel values of two matching square patches cropped from the prediction map and the ground truth mask, and N denotes the size of the sliding window; the L_{SSIM} of p and g is defined as follows:

$$L_{\text{SSIM}} = 1 - \frac{(2^* \mu_p^* \mu_g + C_1)^* (2^* \sigma_{pg} + C_2)}{(\mu_p^2 + \mu_g^2 + C_1)^* (\sigma_p^2 + \sigma_g^2 + C_2)} \tag{3}$$

where σ_p, σ_g and μ_p, μ_g are the standard deviations and mean of p and g, and σ_{pg} is their covariance. $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are two constants utilized to avoid dividing by zero. The average of L_{SSIM} of all the cropped square patches represents the total SSIM loss of the whole predicted map.

The IoU loss function was inspired by performance measure criterion and is widely applied in semantic segmentation [46]–[48]. The IoU measure accounts for the class imbalance issue usually present in the binary semantic segmentation [48], and IoU loss function optimizes the global structure at image level. Combined with attention to pixel position, the wIoU loss function allocates different weights to challenging/simple pixels to differentiate their importance [45], as shown in the following equation.

The formulated hybrid loss function directs attention according to pixel position, but also learns the difference between the prediction and the reference map at the pixel level (wBCE), local level (SSIM), and global level (wIoU). The designed loss is robust with imbalanced background/foreground problem. Therefore, for every side output, we utilize the same loss function, and just change the ground truth, i.e. for edge side output, we utilize the building edge ground truth. Our total loss function L_{total} is defined as the weighted summation over all side outputs

$$\begin{aligned} L_i &= L_{\text{wBCE}} + L_{\text{SSIM}} + L_{\text{wIoU}} \\ L_{\text{total}} &= w_i^* L_i^R + L_1^E, i = 1, 2, 3, 4 \end{aligned} \tag{5}$$

where $w_i = [0.3, 0.5, 0.5, 1]$ is the weight for multiregion outputs, as stated in Section III-A. L_i^R represents region output loss and L_1^E indicates there is only one edge output loss.

IV. Experiments and Analysis

Experiments and analysis are presented in this section. Section IV-A1 describes the experimental datasets, implementation details, and evaluation metrics. Section IV-B presents the experimental results and visual and quantitative analysis of the tested algorithms for the three datasets. An ablation analysis

$$L_{w\text{BCE}} = -\frac{\sum_{r=1}^{H} \sum_{c=1}^{W} (1 + \lambda^* w_{(r,c)}) \sum_{l=0}^{1} I(g_{(r,c)} == l) \log \text{Prob}(p_{(r,c)} == l|\psi)}{\sum_{r=1}^{H} \sum_{c=1}^{W} \lambda^* w_{(r,c)}} \tag{1}$$

of the proposed architecture and loss function are presented in Section IV-C.

A. Dataset, Implementation Details, and Evaluation Metrics

1) Dataset: To assess the performance of our approach, we conduct experiments on the WHU aerial building dataset (WHU) [13], NZ32km2 dataset [24], and INRIA aerial image labeling dataset (INRIA) [12]. They are three open benchmark datasets provided at websites[1,2,3], which are briefly described in the following.

The WHU dataset [13] is divided into three parts: a 4736 tiles training set (130 500 buildings), a 1036 tiles validation set (14 500 buildings), and a 2416 tiles test set (42 000 buildings). A total of 8188 images and 187 000 buildings are included. Each image tile has 512 $*$ 512 pixels. The ground resolution was downsampled to 0.3 m.

The NZ32km2 dataset for Christchurch, New Zealand, was provided by Xia *et al.* [24]. It contains aerial images at the 0.075 m resolution. The corresponding building roofs cover 32 km^2. The training dataset contains four big images with 32366 $*$ 25218 pixels. The test dataset contains four big images with 33445 $*$ 26841 pixels. The big images were split into 512 $*$ 512 pixels tiles and slices with a building coverage rate lower than 15% were removed [24]. In our experiments, the training dataset was further divided into training and validation datasets at a ratio of 4:1. There are 6594 tiles for training, 1611 for validation, and 7569 for testing.

The INRIA dataset [12] includes 360 images covering ten cities, each city with 36 images. Each image tile has a spatial resolution of 0.3 m at a size of 5000 $*$ 5000 pixels. The dataset, including a coverage area of $\sim$810 km^2, covers highly different and representative terrain, landforms, and buildings type. Only Austin, Chicago, Kitsap County, Vienna, and West Tyrol have public labels, with 180 images. Our experiments were conducted on these five cities with disclosed ground truth. The label quality of INRIA dataset is lower than the NZ32km2 and WHU datasets. The ultimate goal was to assess the generalization power of the techniques. Therefore, we tested our approach on this dataset to assess the generalization ability of the proposed framework. Following [49] and [50], the first five big images of each town are chosen as the test dataset, and the rest were utilized for training. The big images are split into 512 $*$ 512 pixels tiles. A total of 12 555 tiles were generated for training and 2025 tiles for testing.

2) Implementation Details: We implemented our method as well as U-Net [17], DeepLabv3+ [18], MC-FCN [33], and BR-Net [24] on the PyTorch [51] library for detailed qualitative and quantitative analysis. For a fair comparison, considering the number of parameters, we implemented another version of CGSANet termed CGSANet-ResNet18. The CGSANet-ResNet18 taken the modified ResNet18 as the encoder. Each decoder stage was composed of two consecutive plain convolution blocks. The CGSANet took the modified ResNet34 as the encoder. Our Pytorch source code will be available at GitHub.[4] For U-Net implementation, the encoder was comprised of five stages, each stage has two repeated plain convolution blocks (with convolution, BN, and ReLU in sequence). The channel for the output feature map was [64, 128, 256, 512, 1024]. Maxpooling was applied with downsampling for four times. The decoder was the symmetrical to the encoder, and transposed convolution was utilized for upsampling. The DeepLabv3+ implementation was taken from an open source repository.[5] The MC-FCN and BR-Net implementations were taken from the Paszke's [52] public released code. To verify the effectiveness of our method further, we performed a quantitative analysis with the other published state-of-the-art algorithms using the WHU and INRIA datasets. These methods were MA-FCN [11], MAP-Net [53], and EaNet [15] for WHU dataset. The GAN-SCA [49], Building-A-Nets [50], and AMUNet [54] were the comparative methods for the INRIA dataset. For a fair comparison, U-Net, DeepLabv3+, MC-FCN, and BR-Net and our CGSANet and CGSANet-ResNet18 were evaluated. Experiments on the corresponding datasets were done using the same settings described in the following.

We used a NVIDIA GeForce RTX 3090 graphics card with 24 GB GPU memory for training and testing. For optimization, we utilized the Adam optimizer with default values [55] to train our model. The hyperparameter λ for calculating wBCE and wIoU was set to five, and the sliding window size N for local SSIM was 11 pixels. The image values were rescaled to [0, 1] before inputting into the network, and no data augmentation techniques were applied in our experiments. The weights that performed best on the validation dataset in terms of IoU value were saved for testing. Due to the different sizes and complexity of the three datasets, we applied different weight initialization strategies and training epochs. For the WHU dataset, we initialized the weights with pretrained ResNet34 on the ImageNet dataset. All the tested algorithms were trained for up to 60 epochs. For experiments on the INRIA and the NZ32km2 datasets, the pretrained weights on the WHU dataset were adopted for initialization and trained 100 epochs and 30 epochs. The batch size for all the experiments was set to four.

[1][Online]. Available: http://gpcv.whu.edu.cn/data/building_dataset.html

[2][Online]. Available: https://drive.google.com/file/d/1PNkGLRT8J9h4Cx9iyS0Bh9vamQS_KOTz/view

[3][Online]. Available: https://project.inria.fr/aerialimagelabeling/

[4][Online]. Available: https://github.com/MrChen18/CGSANet

[5][Online]. Available: https://github.com/whuchenlin/Khaos/tree/ff5b4ef4810331ad681be2eba5f66cae67f4de18/deeplabv3plus

$$L_{\text{wIoU}} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W}(g^*_{(r,c)}p_{(r,c)})(1+\lambda^* w_{(r,c)})}{\sum_{r=1}^{H}\sum_{c=1}^{W}(g_{(r,c)}+p_{(r,c)}-g^*_{(r,c)}p_{(r,c)})(1+\lambda^* w_{(r,c)})} \tag{4}$$

3) Evaluation Metrics: Five commonly used accuracy evaluation indicators were chosen to evaluate the region segmentation performance of our proposed method. They were overall accuracy (OA), precision, recall, $F1$-score, and IoU. The OA, precision, recall, $F1$-score, and IoU equations are shown in (6).

$$\begin{aligned} \text{OA} &= (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \\ \text{Precision} &= \text{TP}/(\text{TP}+\text{FP}) \\ \text{Recall} &= \text{TP}/(\text{TP}+\text{FN}) \\ F1 &= 2^{*}(\text{Recall}^{*}\text{Precision})/(\text{Recall}+\text{Precision}) \\ \text{IoU} &= \text{TP}/(\text{TP}+\text{FP}+\text{FN}) \end{aligned} \tag{6}$$

where TP (i.e., true positive) represents the number of building pixels correctly classified; TN (i.e., true negative) represents the number of nonbuilding pixels correctly classified; FP (false positive) represents the number of nonbuilding pixels classified as building pixels; and FN (i.e., false negative) represents the number of building pixels classified as nonbuilding pixels.

The regional accuracy evaluates the extraction performance of all the building pixels. The boundary accuracy evaluation can better show the extraction performance of pixels near the edge of the building. For boundary accuracy assessment, the Boundary $F1$-score (BF-Score) was calculated. The BF-Score evaluates the degree of match between the boundary of a predicted object and the ground truth boundary. It is defined as the $F1$-score of boundary pixels with an error tolerance buffer [56]. A MATLAB built-in implementation function *bfscore* was applied for BF-Score calculation.

B. Comparative Analysis

Comparative experiments and analysis on the three datasets are presented in this section. Visual and quantitative analysis of the tested algorithms on the WHU dataset, the NZ32km2 dataset, and the INRIA dataset are presented separately in Section IV-B1 through Section IV-B3.

1) Experiments on WHU Aerial Building Dataset: To compare the tested methods intuitively, five representative image tiles of WHU dataset were selected for qualitative evaluation, as shown in Fig. 5. The rows of Fig. 5 are images and results for five samples in the WHU test dataset. Each column from left to right represents the image, the visual accuracy evaluation results of the five tested algorithms, and the ground truth. The pixels colored in yellow, red, black, and green are TP, FP, TN, and FN.

As shown in Fig. 5, fewer pixels are colored in green and red in column (f) than in the other columns. This indicates that our method has fewer omission errors and less commission errors compared with the other tested algorithms. There is perceptual variance in building scales and appearance in the WHU dataset. The proposed CGSANet can extract complex buildings more completely than the comparative algorithms, such as the region **A** marked by the red square in Fig. 5. The four comparative methods all have omission errors for the complex buildings. Our CGSANet showed more robust performance under complex scenarios, as shown in the regions **B** and **C** marked by the red square in the second and third rows. Buildings in the marked regions **B** and **C** are built with different materials, shapes, and sizes, and seriously affect the performance of the tested methods. Our method, however, showed stronger performance, with no omission errors and only a few commission errors.

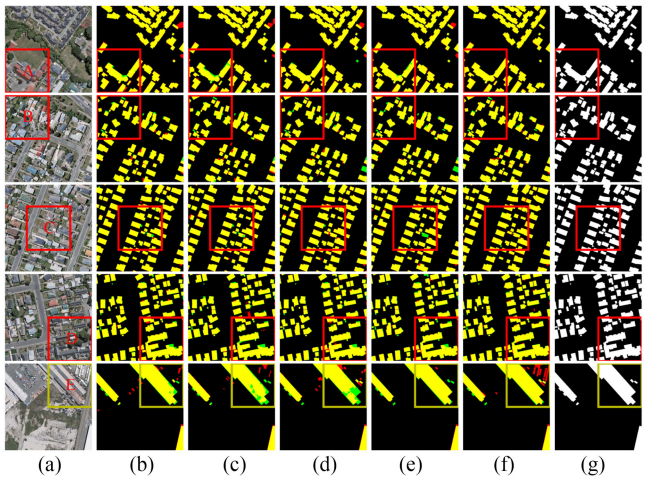

Fig. 5. Building extraction accuracy evaluation results on WHU aerial building test dataset. The pixels colored in yellow, red, black, and green are TP, FP, TN, and FN. (a) Image. (b) U-Net. (c) DeepLabv3+. (d) MC-FCN. (e) BR-Net. (f) Proposed. (g) Label.

Buildings in the fourth-row marked area **D** have complex surrounding environments, e.g., spectral ambiguities and occlusion. On the left side of the region **D**, the spectral difference between the adjacent object and the building is minimal. U-Net, DeepLabv3+, MC-FCN, and BR-Net have commission errors in this area. On the right side, the complex building in the marked area is obscured by the adjacent tree partly, leading to the omission errors of the tested methods. Our CGSANet yielded the highest quality result. The extracted result is close to the ground truth.

There is a failed example as the region **E** marked by the yellow square in Fig. 5. Our method incorrectly detects containers as buildings because they are similar to buildings in spectrum, shape, and even surrounding shadows, and there is no sign to determine whether they are buildings in terms of this independent image tile. Even with manual interpretation, only looking at this image would be hard to correctly determine whether the object is a building. Automatic algorithms cannot easily extract buildings in such a situation at high quality.

To assess the building extraction results quantitatively, we utilized six assessment indexes, including recall, precision, IoU, $F1$-score, OA, and BF-Score, to evaluate the accuracy of 2416 images in the WHU test dataset, as given in Table I. The buffer for calculating the BF-Score is set to three pixels. The first column of Table I tabulates the nine approaches. Three other recent state-of-the-art approaches, including MA-FCN [11], MAP-Net [53], and EaNet [15], were quantitatively compared to our CGSANet additionally. The accuracy evaluation results of the three methods are cited from the relevant paper. The methods are arranged in the ascending order of IoU value. The strongest values per column are labeled in bold, and the secondary values are underlined. "-" indicates that no relevant value is disclosed.

It can see clearly that our CGSANet achieved the highest IoU value of 91.55% compared to other tested algorithms, indicating

TABLE I
WHU AERIAL BUILDING TEST DATASET EVALUATION RESULTS OF THE PROPOSED CGSANET AND SEVEN COMPARATIVE METHODS (%)

Method	Recall	Precision	IoU	F1-score	OA	BF-Score
DeepLabv3+	94.13	94.37	89.13	94.25	98.72	88.83
MC-FCN	94.94	94.52	89.99	94.73	98.82	89.67
U-Net	95.37	94.60	90.45	94.98	98.88	89.51
BR-Net	94.62	95.54	90.62	95.08	98.91	90.06
MA-FCN [11]	95.10	95.20	90.70	-	-	-
MAP-Net [52]	94.81	**95.62**	90.86	95.21	-	-
EaNet [15]	**96.09**	94.63	91.11	95.35	-	-
CGSANet-ResNet18	95.38	95.39	91.18	95.39	98.97	91.35
CGSANet	96.07	95.11	**91.55**	**95.59**	**99.01**	**91.84**

Strongest values per column are labeled in bold, and the secondary values are underlined.

that the proposed method balances the recall and precision optimally. The CGSANet-ResNet18 ranked second in terms of IoU and the BF-Score. EaNet achieved the highest recall value, but its precision was lower than our method, by approximately 0.5%. The recall value of our approach was slightly lower than the EaNet 0.02%; thus, the IoU value of our CGSANet was higher than EaNet. MAPNet achieved the highest precision value, but its recall value was much lower than our CGSANet, while CGSANet outperformed its IoU value by approximately 0.7%. These two methods are recent powerful models. They are representative methods of multiscale feature enhancement model and robust loss function design building extraction model. Given that these methods are already powerful, the improvement is substantial.

Among the four comparison methods we implemented, BR-Net achieved the highest $F1$ value, reaching 95.08%. Our CGSANet-ResNet18 outperformed their $F1$ index by 0.31%. Furthermore, the BF-Score of our CGSANet-ResNet18 outperformed BR-Net by 1.29%. Much higher than region improvement, these results demonstrate that our method improves the quality of the extracted building region, especially the edge quality. To further verify the quality of the building boundary extracted by our method, we conducted a qualitative and quantitative evaluation of the boundary quality.

The qualitative result for boundary accuracy assessment is shown in Fig. 6. We selected the marked area of the same sample images as the region evaluation for boundary qualitative evaluation. The buffer size was set to one pixel—a rigorous setting that will not tolerate any localization error between the predicted boundary and the ground truth.

From Fig. 6, it is clear that the misclassification could cause severe visual degradation on boundaries. Our method still outperforms the other tested methods visually with fewer missed detections on the contours of each building and false detections on all buildings. Most of the extraction results of the comparative methods have omission and commission errors simultaneously, which seriously affects the usability of the extraction results. Actually, a one-pixel buffer size is too strict and would consequently over penalize algorithms since even the ground truth data may contain boundary localization errors. So, it was rational to set a tolerance buffer. A comparative analysis of boundary performance with varied buffer size is shown in Fig. 7.

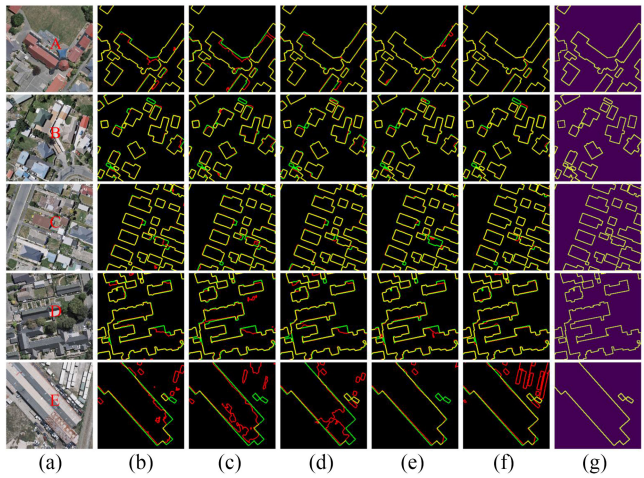

Fig. 6. Building extraction boundary accuracy evaluation results on WHU aerial building test dataset. The pixels colored in yellow, red, black, and green are TP, FP, TN, and FN. (a) Image. (b) U-Net. (c) DeepLabv3+. (d) MC-FCN. (e) BR-Net. (f) Proposed. (g) Label.

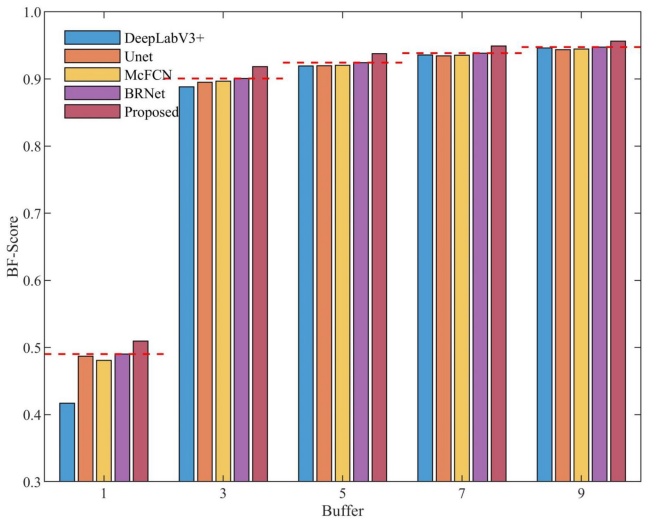

Fig. 7. Comparative analysis of boundary performance with varied buffer size

Fig. 7 shows a quantitative comparison of BF-Score in different methods under different buffer size settings. It shows that no matter how large the buffer is, the BF-Score of our approach was higher than that of the compared algorithms indicating that our method can extract boundaries closer to the ground truth. When the buffer size is set to three pixels, the BF-Score of the comparative methods are less than 90%, while our method is close to 92%. This fully demonstrates that our method can extract more accurate buildings than the counterpart methods.

A detailed qualitative and quantitative analyses demonstrate that our method achieved competitive performance on the high-quality labeled WHU dataset. The intuitive qualitative evaluation illustrates that CGSANet has fewer missed detections and can produce more accurate building boundaries than the counterpart methods. The objective quantitative analysis demonstrates that our method outperforms the compared algorithms in both region and boundary accuracy. The WHU dataset downsampled

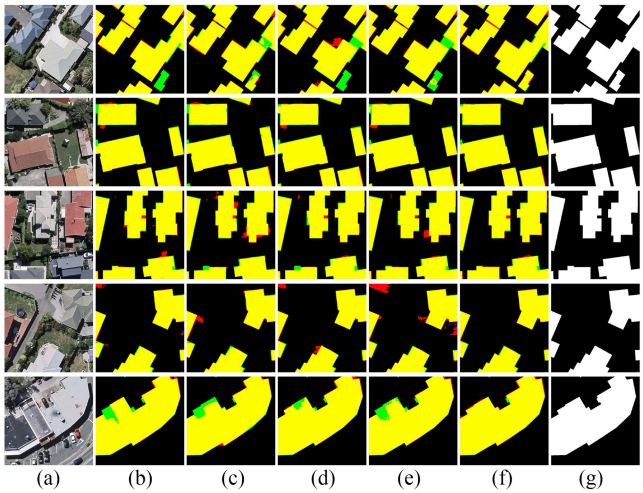

Fig. 8. Building extraction accuracy evaluation results on NZ32km2 test dataset. The pixels colored in yellow, red, black, and green are TP, FP, TN, and FN. (a) Image. (b) U-Net. (c) DeepLabv3+. (d) MC-FCN. (e) BR-Net. (f) Proposed. (g) Label.

the image resolution to 0.3 m, to further demonstrate the effectiveness of our proposed method on ultrahigh-resolution images and more complex scenes, experiments on the NZ32km2 dataset and the INRIA dataset were conducted.

2) Experiments on NZ32km2 Dataset: To validate the effectiveness of CGSANet on ultrahigh-resolution images, experiments were conducted on the NZ32km2 dataset. The imagery spatial resolution of the NZ32km2 dataset is 0.075 m, four times the resolution of the images in the WHU dataset. We implemented four comparative methods for experiments on the NZ32km2 dataset. The weights that performed best on the WHU dataset were utilized as the pretraining parameters. Five representative images were selected for visual accuracy evaluation, as shown in Fig. 8.

Fig. 8 shows the building extraction accuracy evaluation results of five $512 * 512$ size images of the five tested methods, including the proposed method. Because the spatial resolution is ultrahigh (0.075 m), a $512 * 512$ size imagery can only contain a few buildings, and the outline of the buildings is clear. Fig. 8 contains five rows and seven columns, and each row contains images, accuracy evaluation results of the tested method, and the ground truth. The images in the first four rows are taken from four large test images separately. The sample in the fifth row is to show the extraction effect of large and complex buildings.

Compared with other methods, CGSANet can extract more accurate buildings. As shown in the first and last rows of Fig. 8, our method has fewer missed detections than comparative methods on large complex buildings. U-Net, MC-FCN, and BR-Net failed to detect the building in the bottom right corner of the first row in Fig. 8. DeepLabv3+ cannot detect the building completely. Our CGSANet has almost perfectly detected the building outline. For the building in the upper right corner of the first row in Fig. 8, DeepLabv3+, MC-FCN, and BR-Net tended to detect this building as two buildings, and our method only has a small amount of missed detection at the building boundary. For the sample image of the second row, all methods have achieved high-quality detection results. But our method outperformed the other tested methods because there are fewer green pixels and red pixels in column (f) than in other columns.

TABLE II
NZ32KM2 TEST DATASET EVALUATION RESULTS OF THE PROPOSED CGSANET AND FOUR COMPARATIVE METHODS (%)

Method	Recall	Precision	IoU	F1-score	OA	BF-Score
MC-FCN	93.25	94.60	88.54	93.92	96.41	73.15
BR-NET	93.97	94.08	88.73	94.03	96.45	72.70
DeepLabv3+	93.72	94.55	88.92	94.13	96.53	71.44
U-NET	94.01	94.45	89.09	94.23	96.58	73.25
CGSANet-ResNet18	94.01	**95.02**	89.60	94.51	96.75	76.04
CGSANet	**95.36**	94.14	**90.02**	**94.75**	**96.86**	**76.47**

Strongest values per column are labeled in bold, and the secondary values are underlined.

The accuracy evaluation results in the third and fourth rows demonstrate that CGSANet can achieve the highest quality results when extracting buildings with complex surrounding environments or with abundant outline details. In contrast, the other comparative methods are prone to missing or false detections. The sample image in the fifth row contains only one building. This building is huge and has a complicated shape. All four comparative methods failed to detect the building contour accurately, but our method has the least missed detections. The qualitative evaluation of these methods shows that our method can extract buildings more accurately, especially in complex scenes and complex building outlines. The quantitative evaluation results are given in Table II.

The accuracy assessment result of 7569 test images in the NZ32km2 building dataset is given in Table II. The first column of Table II represents six methods, and the second to seventh columns represent the six accuracy evaluation indicators. The methods are arranged in an ascending order of IoU value. From Table II, our CGSANet achieved the highest IoU value of 90.02%, which is 1–2% higher than other methods, indicating that our approach achieved a balance between precision and recall. Although our CGSANet was lower in precision than U-Net, it achieved a higher IoU, by nearly 1%. The performance of CGSANet-ResNet18 was also better than the comparative methods. Apart from our proposed methods, DeepLabv3+ achieved second-ranked precision value and third-ranked IoU value. MC-FCN achieved the highest precision but the lowest recall, thus it only achieved the lowest IoU. BR-NET was not outstanding in the precision index and the recall index, but it has achieved a better balance in the two indexes and achieved a higher IoU than MC-FCN. Although the $F1$-score of Deeplabv3+ is higher than that of MC-FCN and BR-Net, its BF-Score is lower than these two methods.

MC-FCN represents building extraction methods with deep supervision techniques, and BR-Net represents simultaneously supervises building edges and regions at the last decoder layer. The performance improvement of CGSANet relative to theirs further demonstrates the usefulness of the proposed modules. Furthermore, our method performed stable on the WHU dataset and the NZ32km2 dataset with images of different resolutions. The four comparative algorithms have obvious differences in the performance of processing images of different resolutions. For

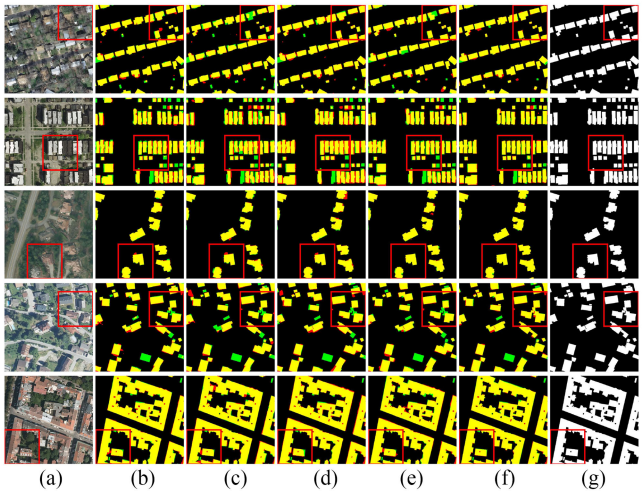

Fig. 9. Building extraction accuracy evaluation results on INRIA dataset. The pixels colored in yellow, red, black, and green are TP, FP, TN, and FN. (a) Image. (b) U-Net. (c) DeepLabv3+. (d) MC-FCN. (e) BR-Net. (f) Proposed. (g) Label.

example, on the WHU dataset, BR-NET achieved the highest IoU value among the compared methods, but on the NZ32km2 dataset, its IoU performance only ranked the third and is worse than DeepLabv3+ and U-NET. Our method, however, outperformed the other tested methods on both datasets. This further demonstrates the robustness of our proposed method against the resolution change of the VHR imagery. To further demonstrate the effectiveness of our proposed method under more complex scenes, experiments on the INRIA dataset were conducted.

3) Experiments on INRIA Dataset: Visual and quantitative analysis of the tested algorithms on the INRIA dataset are presented in this section. We introduced the qualitative evaluation of five representative images and then quantitatively analyzed the whole official validation dataset. Five representative images of five cities were selected for qualitative analysis. The visual accuracy evaluation results are shown in Fig. 9.

Fig. 9 shows the building extraction accuracy evaluation results of five $512 * 512$ size images of the five tested methods. Each column from left to right represents the image, the five tested methods, and the ground truth. The images in each row are taken from a different city. The scenes and building styles of different cities are different. The five-row sample images were from Austin, Chicago, Kitsap County, Austrian Tyrol, and Vienna. As shown in the first row of Fig. 9, buildings in Austin were evenly distributed and surrounded by trees that often occluded buildings. Therefore, the comparative methods tended to miss detect parts of buildings, as shown by the area marked by the red square. However, our method can generate more regular building detection results because the loss function considers structural information and utilizes the building edge for supervision.

As shown from the second row of Fig. 9, buildings in Chicago had large-scale variation, and the distribution was very dense; therefore, missed detection of small buildings and false detection of multiple adjacent buildings as one building, as shown by the area marked by the red square. Our method considers the building edge information and assigns a higher weight to the more difficult pixels, so we only had a few false detections among the dense buildings. Most of the buildings were detected independently. However, our method missed five buildings. All the tested methods missed the building in the bottom middle. A visual comparison revealed that it was actually an error in the ground truth. There is no building in the image. One of the missed detections of the other two small buildings was because they were completely obscured, and the other was because of the spectral reflection close to the road. Other comparative methods had more missed detections than our proposed CGSANet.

TABLE III
INRIA VALIDATION DATASET BUILDING EXTRACTION EVALUATION RESULTS OF THE PROPOSED CGSANET AND SEVEN COMPARATIVE METHODS (%)

Method	Recall	Precision	IoU	F1-score	OA	BF-Score
DeepLabv3+	84.24	87.42	75.14	85.80	96.17	63.01
MC-FCN	86.96	86.41	76.50	86.69	96.33	65.09
BR-NET	85.56	88.96	77.34	87.22	96.56	67.08
U-NET	86.59	87.97	77.42	87.28	96.53	66.38
GAN-SCA [48]	-	-	77.75	-	96.61	-
Building-A-Nets [49]	-	-	78.73	-	96.71	-
CGSANet-ResNet18	88.07	89.32	79.68	88.69	96.91	71.95
AMUNet [53]	-	-	79.76	-	96.73	-
CGSANet	**88.68**	**90.22**	**80.90**	**89.44**	**97.12**	**74.65**

Strongest values per column are labeled in bold, and the secondary values are underlined.

The third row of Fig. 9 is the image of Kitsap County and its accuracy evaluation results. Kitsap County had more vegetation and an uneven building distribution. All the tested methods achieved high-quality extraction results, with only a few false detect pixels. The fourth row of Fig. 9 is the image of Austrian Tyrol and its accuracy evaluation results. Austrian Tyrol is a low-density urban settlement. The shapes of the buildings in Tyrol were more complicated than those in other images. As shown in the area marked by the red square, all the comparative methods miss detected the link pixels in the complex building, which resulted in falsely detecting it as two or three buildings. Our method completely detected the building at the cost of a small number of falsely detected pixels. There are many missed detections in the Austrian Tyrol sample image. We found that the missed detections of larger buildings were ground truth errors, and there are actually no buildings on the corresponding location in the image.

The last row of Fig. 9 is the image of Vienna and its accuracy evaluation results. Buildings in Vienna are large, complex in shape, and densely distributed. As shown in the area marked by the red square, U-Net, DeepLabv3+, and BR-Net made false detections in the bottom left corner, MC-FCN made missed detections in the bottom right corner, but our method can detect buildings more correctly, with fewer missed detections and false detections.

The qualitative evaluation of these methods demonstrates that our algorithm can extract buildings more accurately even when dealing with different scenes and building types. The quantitative evaluation results of the tested approaches on the INRIA validation dataset are given in Table III. In addition to the qualitative comparative algorithms in Fig. 9, three other

recent state-of-the-art approaches, including GAN-SCA [49], Building-A-Nets [50], and AMUNet [54], were further compared to our CGSANet. The accuracy evaluation results of the three methods were directly quoted from the relevant paper. The methods are arranged in ascending order of IoU values.

From Table III, the recall, precision, IoU, and OA for our CGSANet without overlapping prediction were 88.68%, 90.22%, 80.90%, and 97.12%, respectively. The $F1$-score of our CGSANet-ResNet18 in the INRIA validation dataset was 88.69%, indicating that the method has specific adaptability. Even without the overlapping strategy, CGSANet achieved the highest recall, precision, and IoU scores among the tested algorithms, indicating that our CGSANet can optimally balance the completeness and correctness of the extracted results. The detected buildings were the most accurate as our method achieved the highest BF-Score as well.

The INRIA dataset covers the affluent and representative areas, from densely populated areas to low-density towns in high mountains. This dataset is challenging. In general, our method achieved more convincing results than many other algorithms, which further demonstrates the robustness and generalization ability of the proposed method. The extraction performance on the INRIA dataset was lower than in the previous two datasets. The INRIA dataset has various scene changes; the image quality of the INRIA dataset is worse than the previous two datasets, and the label has some errors. Our method relies on ground truth building edge supervision to improve the extracted features. However, CGSANet still achieved more convincing results than comparative methods with less missed detection and fewer false detections.

The detailed experimental analysis of the three datasets demonstrates that our method is highly competitive. To analyze the effectiveness of the specific modules of our proposed framework, ablation experiments with the proposed architecture and loss function were conducted on the WHU dataset. The ablation analyses are introduced in Section IV-C.

C. Ablation Analyses

In this section, the key modules of CGSANet, including the encoder CG module and the stepwise weighting deep supervision strategy (WDS) of the multiscale region supervision, were analyzed. In addition, the robustness of the hybrid loss function was further discussed by comparing with BCE, wBCE+wIoU, BCE+SSIM+IoU, and wBCE+SSIM+IoU loss functions. Experiments were conducted on the WHU dataset and NZ32km2 dataset. The experimental setting on the corresponding dataset is consistent with Section IV-B.

We evaluated the two modules of CGSANet by gradually subtracting components. The model that removes the CG module and the WDS module was taken as the baseline. The baseline is comprised of the modified ResNet34, the ASPP module, and the symmetrical decoder with only one building region output. All the models were optimized by the proposed hybrid loss function. Table IV tabulates the quantitative evaluation results of the ablation experiment using the proposed architecture. W/O is short for without.

TABLE IV
ARCHITECTURE ABLATION ANALYSIS BUILDING EXTRACTION EVALUATION RESULTS ON THE WHU DATASET AND NZ32KM2 DATASET (%)

Method	WHU				NZ32km2			
	Rec	Prec	IoU	BF	Rec	Prec	IoU	BF
Baseline	95.78	94.81	91.00	91.15	<u>94.58</u>	94.69	89.82	75.99
W/O CG	95.18	**95.56**	91.15	91.61	94.03	**95.33**	<u>89.89</u>	76.04
W/O WDS	<u>95.81</u>	<u>95.14</u>	<u>91.34</u>	<u>91.71</u>	94.39	<u>94.92</u>	89.85	<u>76.34</u>
CGSANet	**96.07**	95.11	**91.55**	**91.84**	**95.36**	94.14	**90.02**	**76.47**

W/O, Rec, Prec, and BF are short for without, recall, precision, and BF-Score. Strongest values per column are labeled in bold, and the secondary values are underlined.

As given in Table IV, CGSANet achieved the highest value in three of the four evaluation indicators in both datasets, which demonstrates the rationality of the architecture design. Specifically, the baseline achieved an IoU of 91.00% on the WHU dataset, an IoU of 89.82% on the NZ32km2 dataset, indicating that the basic architecture with the modified encoder combined with the ASPP module delivers high building extraction performance. The proposed method achieved an IoU of 91.55%, an increase of 0.55% over the baseline on the WHU dataset. This shows that the proposed two modules further improve the effectiveness of the model.

From the perspective of a single module, the CG module has a greater impact on the model performance. The approach without the CG module achieved the highest precision value but the lowest recall value on both datasets. As a result, without the CG module will bring 0.40% IoU loss on the WHU dataset, and 0.43% BF-Score loss on the NZ32km2 dataset. The WDS module can increase the IoU of the model by 0.21% on the WHU dataset, and 0.17% on the NZ32KM2 dataset. Therefore, each module contributes, and the combination of two modules can achieve the highest IoU value. In terms of boundary quality, the last column reveals that compared with the baseline model, the proposed framework improved the results by nearly 0.7% on the WHU dataset, and 0.5% on the NZ32km2 dataset, simultaneously indicating the high edge and regional accuracy of the extracted buildings.

To visually compare the differences between the ablation experiments, we show heat maps utilizing the Grad-CAM [57] technique to better understand the features captured by different models. The results for a sample image in the WHU dataset are shown in Fig. 10. The first row of Fig. 10 shows the heatmap of Baseline, Baseline+CG, Baseline+WDS, and the ground truth. The second row shows the four regional outputs of CGSANet. As shown in Fig. 10, the Baseline model can capture the characteristics of multiscale buildings due to the introduced ASPP module but have false detections on objects similar to buildings. The introduction of the CG module can eliminate the false detections from the Baseline model, but it may lead to the missed detection of the building area with inconspicuous edges. The introduction of the WDS module can make the backpropagation of the gradient more stable. It can see from the second row that all four outputs of CGSANet can more accurately identify multiscale buildings, indicating that

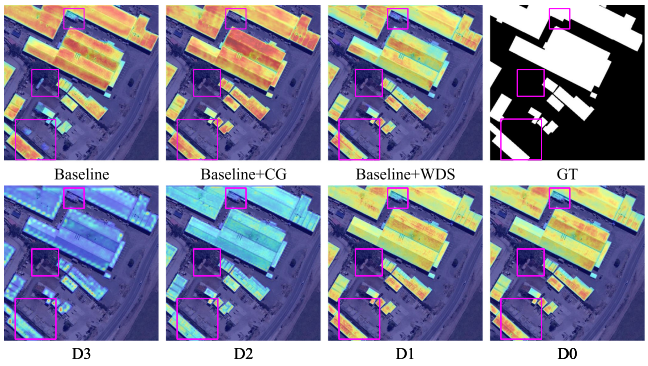

Fig. 10. Heatmaps of different models and four regional branches of CGSANet.

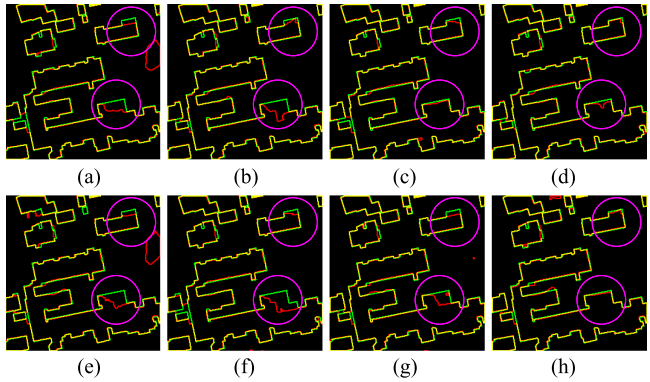

Fig. 11. Ablation experimental edge accuracy evaluation results of the fourth-row marked area **D** in Fig. 5. (a) BCE. (b) wBCE+wIoU. (c) BCE+SSIM+IoU. (d) wBCE+SSIM+IoU. (e) Baseline. (f) W/O CG. (g) W/O WDS. (h) Proposed.

TABLE V
LOSS FUNCTION ABLATION ANALYSIS BUILDING EXTRACTION EVALUATION RESULTS ON THE WHU DATASET AND NZ32KM2 DATASET (%)

Loss function	WHU				NZ32km2			
	Rec	Prec	IoU	BF	Rec	Prec	IoU	BF
BCE	95.44	95.53	91.36	91.46	93.77	**95.55**	89.85	75.50
wBCE+wIoU	95.66	95.25	91.31	91.48	94.39	94.97	89.90	76.05
BCE+SSIM+IoU	95.52	95.39	91.31	91.50	94.42	94.95	89.90	76.03
wBCE+SSIM+IoU	95.32	**95.79**	91.49	91.81	94.36	94.96	89.86	76.26
wBCE+SSIM+wIoU	**96.07**	95.11	**91.55**	**91.84**	**95.36**	94.14	**90.02**	**76.47**

Strongest values per column are labeled in bold, and the secondary values are underlined.

the introduction of a stepwise WDS can improve the learning efficiency of the network.

To verify the hybrid loss function, CGSANet was trained with different loss functions on the WHU dataset and NZ32km2 dataset. The loss functions for comparison include BCE, wBCE+wIoU, BCE+SSIM+IoU, wBCE+SSIM+IoU, and the proposed wBCE+SSIM+wIoU. The quantitative results on the WHU dataset and NZ32km2 dataset are given in Table V.

As given in Table V, the wBCE+SSIM+wIoU achieved the highest values in Recall, IoU, and BF-Score index on both datasets, indicating that the proposed hybrid loss function optimally balances the precision and recall and generates the highest quality building boundaries. For experiments on the WHU dataset, the BCE and wBCE+wIoU achieved the second-highest recall and precision, respectively, but the BF-Score index based on the BCE alone is the lowest. For experiments on the NZ32km2 dataset, the BCE achieved the highest Precision but the lowest IoU and BF-Score. On both datasets, the IoU score of wBCE+wIoU and BCE+SSIM+IoU was equivalent, indicating that pixel position attention strategy is essential for improving the loss function. At the same time, SSIM is a powerful supplement as wBCE+SSIM+IoU outperformed other loss functions on BF-Score index.

In terms of edge accuracy, when the buffer is set to three pixels, the wBCE+SSIM+wIoU had the highest BF-Score value. The edge accuracy of wBCE+SSIM+IoU was close to the proposed loss function. However, the difference between the BF-Score of other loss functions and the wBCE+SSIM+wIoU was larger than the difference of the $F1$-score based on the region, especially for the BCE loss function. This further proves the superiority of the proposed loss function in improving boundary quality. To visually compare the differences between the ablation experiments, we selected the area marked by the red square in Fig. 5 **D** for clearer edge qualitative evaluation. The results are shown in Fig. 11. All the images are dilated for better viewing.

Fig. 11 shows the ablation experimental edge accuracy evaluation results of the fourth-row marked area **D** in Fig. 5. The first row shows ablation experimental results of different loss functions based on the proposed CGSANet. The second row shows ablation experimental results of the proposed architecture based on the proposed hybrid loss function. From the area marked by the purple circle in Fig. 11, it can be seen clearly that without the SSIM loss function [i.e., Fig. 11(a) and (b)] or the CG module [i.e., Fig. 11(f)], the edge quality of the extraction result is much lower. Without WDS module Fig. 11(g) will lead to incomplete extraction of complex buildings. Fig. 11(c) and (d) shows the results of loss function BCE+SSIM+IoU and wBCE+SSIM+IoU, which considered pixel-, local- and global-level information simultaneously. Hence, the edges produced by these two methods are high quality but still not as accurate as the edges extracted by our method. As shown in the marked area by the purple circle on the top of Fig. 11(c), the BCE+SSIM+IoU method missed some building pixels and was less effective than our method on this building. The marked area by the purple circle in the bottom of Fig. 11(d) shows the wBCE+SSIM+IoU experiment produced irregular building extraction results, the edge quality was lower than our method. Therefore, each module of the proposed method contributed to the result, and their combination achieves optimal performance.

D. Computational Complexity Analysis

To quantitatively evaluate the computational complexity of different models, the number of parameters and inference speeds of the tested models were counted, as given in Tables VI and VII. Params are the abbreviation of Parameters, and M is short for Million, and frame per second (FPS) counts the number of processed image patches per second. We conducted experiments on the test dataset of the WHU building dataset. The input image size is $512 * 512$ pixels.

TABLE VI
COMPUTATIONAL COMPLEXITY ANALYSIS OF DIFFERENT COMPARATIVE MODELS ON THE WHU TEST DATASET (%)

Method	Params (M)	FPS	IoU	BF-Score
DeepLabv3+	59.34	38.02	89.13	88.83
MC-FCN	24.23	37.70	89.99	89.67
U-NET	39.40	25.40	90.45	89.51
BR-NET	24.24	37.90	90.62	90.06
CGSANet-ResNet18	27.46	18.59	91.18	91.35

TABLE VII
COMPUTATIONAL COMPLEXITY ANALYSIS OF DIFFERENT ABLATION MODELS ON THE WHU TEST DATASET (%)

Method	Params (M)	FPS	IoU	BF-Score
Baseline	42.96	12.41	91.00	91.15
WO CG	42.97	12.39	91.15	91.61
WO WDS	43.02	12.22	91.34	91.71
CGSANet	43.03	11.85	91.55	91.84

As given in Table VI, MC-FCN and BR-NET have the least number of parameters, and their FPS value is relatively high, indicating that their inference speed is fast. DeepLabV3+ has the largest number of parameters but the highest FPS value. However, the IoU and BF-Score index of DeepLabV3+ are the lowest. The proposed method achieved the highest IoU and BF-Score values. When we select modified ResNet18 as the backbone encoder, the proposed method has 27.46 M parameters, and the FPS is 18.59, which is equivalent to one-half of deeplabv3+, and the performance was the best among all the tested methods. It can see from the bottom row of Table VII that when taking the modified ResNet34 as the backbone, the parameters of the proposed method increased by 15.57 M, and the FPS decreased to 11.85, but a better performance can be achieved. Furthermore, the designed CG module and WDS module only increased the number of parameters slightly but significantly improved the model performance. The limitation of the proposed method was that the computation cost is relatively high. The training and inference time of our method was longer than other comparative methods. However, our method yielded higher performance in automatic building extraction, especially on boundary localization. Furthermore, we can select an appropriate backbone encoder according to practical limitations; when taking a lighter backbone, such as ResNet18, the performance of our method still can outperform other tested algorithms.

V. DISCUSSION

The boundary unreliability of the FCN-based methods for building extraction has been a long-standing problem. Repeated downsampling and coarse upsampling cause detail degradation, and the BCE loss function cannot learn structural information. Detail degradation and weak structural-preservation ability affect the reliability of the resulting building boundaries, limiting the practicality of these methods. Early approaches reduced the number of downsampling, designed an upsampling module that retains position information, and introduced structure preservation loss function to improve boundary reliability. Recent methods further improve the reliability of the extracted building edges by introducing a multitask learning framework and deep supervision technique. However, they only supervise the building contour in the last decoder layer, ignoring the rich spatial information of the encoder layers. Moreover, the newly designed loss function still treats the pixel independently without considering the neighborhood information of the pixel. Our method integrates the advantages of the other methods and simultaneously overcomes shortcomings. Specifically, we improved the preservation of spatial information by adding a building contour supervision module to the encoder layers. Residual learning is introduced to improve the stability of the learning process. Our loss function design takes pixel position, local structure, and global information into consideration. These combinations further improved the reliability of the extracted edges. Many experiments were conducted on three public datasets to illustrate the robustness of the proposed approach.

Experiments on the three challenging public datasets have shown that the proposed algorithm has competitive performance, despite the severe foreground/background imbalance of the WHU dataset, the ultrahigh resolution of the NZ32km2 dataset, and the complex scenarios INRIA dataset. The WHU dataset has various building scales and appearances, and the annotations are of high quality. There is a severe building/nonbuilding imbalance in the WHU dataset, as buildings only account for about 18.7% of the number of pixels in the training dataset [15]. We conducted many experiments on the WHU dataset, including a comparative visual analysis, boundary accuracy evaluation, and ablation experiments. Compared with the recent state-of-the-art approaches, our method achieved the highest IoU value. The boundary accuracy evaluation further demonstrates that the proposed method has improved the reliability of the extracted building boundaries to the comparative algorithms. The ablation experiment on the WHU dataset indicates the rationality of each module design and the superiority of the proposed hybrid loss function. They benefit each other synergistically to yield improved building extraction performance. Experiments on the NZ32km2 dataset demonstrate that our algorithm can still obtain robust results when processing images with a spatial resolution of less than 0.1 m. In contrast, the performance of other tested methods was affected by the change of image resolution. This verifies the effectiveness of our approach against the resolution change of the input imagery. The performance on the INRIA dataset shows that our method can also achieve competitive results in complex scenarios. In conclusion, a detailed quantitative and qualitative analysis demonstrates that our algorithm can extract multiscale and complex-shaped buildings in a complex surrounding environment more effectively than the other tested approaches.

The limitation of the proposed method was that the number of model parameters is relatively large. We argue that the increase in the number of model parameters stems from the decrease in downsampling times. Since our model only downsamples the input image three times, the computational cost and the number of parameters in the model increased. However, our method has the capability to preserve spatial details, and the improvement on the boundary localization is substantial. In our proposed framework, the backbone can be adjusted flexibly according to practical requirements to balance accuracy against computation costs. With the improvement of GPU performance, such processing is acceptable.

VI. Conclusion

A novel CGSANet for accurate building extraction from VHR remote sensing imagery is proposed in this article. The proposed method employs multitask learning, deep supervision techniques, and a new hybrid loss function to extract building boundaries at high quality. An efficient encoder CG module was designed to preserve and refine detail spatial feature maps. A framework with encoder–decoder structure, combined with ASPP module, was designed for multiscale contextual feature capture. A robust hybrid loss function was introduced to guide the model to learn parameters from the pixel-level similarity, local structural similarity, and global similarity.

The performance of CGSANet was tested on three challenging datasets, the WHU dataset, NZ32km2 dataset, and INRIA dataset. Experimental results showed that even when extracting buildings with complicated shapes and buildings in images with complex backgrounds, CGSANet can generate regular and crisp building boundaries. CGSANet building extraction outperformed other tested algorithms, with the IoU of above 91.5% in the WHU dataset, an IoU of 90.4% in the NZ32km2 dataset. Experiments on the INRIA dataset further indicate that the proposed method has specific adaptability. An ablation analysis on the architecture modules and loss function demonstrates the rationality of the designed framework. The limitation of the proposed method was that the model parameters were relatively large because of the preservation of the low-level spatial information. When considering computational efficiency, switching to a lighter backbone is an option. Our future investigations on building extraction will improve building boundary regularity with a computationally efficient model.

Acknowledgment

The authors would like to thank the reviewers and editors for their valuable comments and efforts that helped improve this paper. They would also like to thank Stephen C. McClure for linguistic assistance.

References

[1] S. Noronha and R. Nevatia, "Detection and modeling of buildings from multiple aerial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 5, pp. 501–518, May 2001.

[2] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[3] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, 2019.

[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] S. Chen, W. Shi, M. Zhou, M. Zhang, and P. Chen, "Automatic building extraction via adaptive iterative segmentation with LiDAR data and high spatial resolution imagery fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 5, pp. 2081–2095, May 2020, doi: 10.1109/JSTARS.2020.2992298.

[6] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," *Photogrammetric Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.

[7] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogrammetry Remote Sens.*, vol. 86, pp. 21–40, 2013.

[8] A. J. Fazan and A. P. D. Poz, "Rectilinear building roof contour extraction based on snakes and dynamic programming," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 25, pp. 1–10, 2013.

[9] K. Karantzalos and N. Paragios, "Recognition-driven two-dimensional competing priors toward automatic and accurate building detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 133–144, Jan. 2009.

[10] A. Moussa and N. El-Sheimy, "A new object based method for automated extraction of urban objects from airborne sensors data," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 39, pp. 309–314, 2012.

[11] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[12] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

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

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

[15] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 15–28, 2020.

[16] X. Li *et al.*, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 435–452.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[19] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1873–1876.

[20] S. Shrestha and L. Vanneschi, "Improved fully convolutional network with conditional random fields for building extraction," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1135.

[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 Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 247–251.

[22] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[23] L. Xia, X. Zhang, J. Zhang, W. Wu, and X. Gao, "Refined extraction of buildings with the semantic edge-assisted approach from very high-resolution remotely sensed imagery," *Int. J. Remote Sens.*, vol. 41, no. 21, pp. 8352–8365, 2020.

[24] G. Wu *et al.*, "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sens.*, vol. 10, no. 8, p. 1195, 2018.

[25] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

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

[27] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[28] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[29] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2021.

[30] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[31] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2013.

[32] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery," *Proc. SPIE*, E. Y. Lam and K. S. Niel, Eds., 2015, vol. 9405, pp. 153–164.
[33] G. Wu *et al.*, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 407.
[34] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.
[35] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1835–1838.
[36] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 403.
[37] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 830.
[38] 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.
[39] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1768.
[40] Z. Ye, Y. Fu, M. Gan, J. Deng, A. Comber, and K. Wang, "Building extraction from very high resolution aerial imagery using joint attention deep neural network," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2970.
[41] C. Liao *et al.*, "Joint learning of contour and structure for boundary-preserved building extraction," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1049.
[42] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1480–1484.
[43] Z. Jiang, Z. Chen, K. Ji, and J. Yang, "Semantic segmentation network combined with edge detection for building extraction in remote sensing images," N. Sang, J. K. Udupa, Y. Wang, and Z. Liu, Eds., *Proc. SPIE*, 2020, pp. 60–65.
[44] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, 2021.
[45] J. Wei, S. Wang, and Q. Huang, "F^3 Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12321–12328.
[46] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BasNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.
[47] M. Zhou, H. Sui, S. Chen, J. Wang, and X. Chen, "BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 168, pp. 288–306, 2020.
[48] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds., 2016, pp. 234–244.
[49] X. Pan *et al.*, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 917.
[50] X. Li, X. Yao, and Y. Fang, "Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.
[51] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.
[52] G. Wu, Z. Guo, X. Shao, and R. Shibasaki, "GEOSEG: A computer vision package for automatic building segmentation and outline extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 158–161.
[53] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2020.
[54] M. Guo, H. Liu, Y. Xu, and Y. Huang, "Building extraction based on U-Net with an attention block and multiple losses," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1400.
[55] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015.
[56] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 32.1–32.11.
[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Shanxiong Chen received the B.S. degree in remote sensing science and technology from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2015. He is currently working toward the joint Ph.D. degree with the School of Remote Sensing and Information Engineering, Wuhan University, and Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong.

His research interests include object extraction, change detection, and deep learning for remote sensing.

Wenzhong Shi received the Ph.D. degree in GISci and remote sensing from the University of Osnabrück in Vechta, Osnabrück, Germany, in 1994.

He is Otto Poon Charitable Foundation Professor in Urban Informatics, Chair Professor in GISci and remote sensing, Director of Smart Cities Research Institute, The Hong Kong Polytechnic University Hong Kong. He has authored or coauthored more than 250 academic papers that are indexed by SCI and 15 books. His current research interests include urban informatics and smart cities, GISci and remote sensing, intelligent analytics and quality control for spatial data, artificial-intelligence-based object extraction and change detection from satellite imagery, and mobile mapping and 3-D modeling based on LiDAR and remote sensing imagery.

Mingting Zhou received the B.S. degree in photogrammetry and remote sensing in 2015 from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

Her research interests include high spatial resolution remote sensing image road detection, image processing, and deep learning.

Min Zhang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2020.

He is currently a Research Assistant Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University Hong Kong. His research interests include spatial data quality, deep learning, artificial intelligence, change detection, and object recognition in remote sensing.

Zhaoxin Xuan graduated in remote sensing science and technology from Wuhan University, Wuhan, China. He is/was a Professor-level Senior Engineer. He is currently with the Beijing Institute of Surveying and Mapping, Beijing, China.

His research research interests include new basic surveying and mapping, remote sensing data processing, and engineering surveying.