MSBR-GNet: A High-Resolution Imagery Generative Optimization Model for Building Rooftop Boundary Guided by Interpretable Statistical Model in Spatial and Spectral Domain

Liu Jianhua, Ning Xiaohe, Wang Mengchen, Wang Xinyu, Liu Yuan, Chen Xiaoyou, and Zeng Shiyi

***Abstract*—Automated extraction of building rooftop information is of great significance in remote sensing of land resources and other related applications. In this article, a building roof boundary generating optimization model called multiscale boundary regulation generative net (MSBR-GNet), guided by interpretability statistical model in the spatial and spectral domains, is proposed to solve the problem of inaccurate boundary segmentation caused by mixed pixel transition region of remote sensing images. Incorporate the boundary loss function guided by statistical models in the spatial and spectral domains into the generator loss calculation of MSBR-GNet, precisely constrain the regularized generation of building rooftop contours by interpretable mechanism. The experiments show that MSBR-GNet can extract more regular building rooftop contours, and the precision values in the INRIA, WHU, and Massachusetts public datasets reached 0.9275, 0.9228, and 0.8779, respectively, which can ensure the accuracy of building extraction while achieving optimal results in the boundary morphology evaluation index.**

***Index Terms*—Building rooftops, contour optimization, edge transition zones, generative networks, instance segmentation, interpretability, Tupu theory.**

I. Introduction

With the rapid development of remote sensing Earth observation technology, especially the significant improvement of image spatial resolution, the detailed information of geometry, structure, spectrum, and texture of features is becoming more and more abundant, which makes the accurate detection and recognition of targets possible. Meanwhile, as a class of extremely important artificial feature targets in high-resolution remote sensing images, in the process of urbanization, the number and height of buildings increase year by year, and the impact on the ecological environment is increasingly evident. Therefore, it is of great significance to study its refined extraction and to research and explore the impact of buildings on the ecological environment, in order to promote the development of high-resolution remote sensing image information mining technology and the future sustainable development of the city [1].

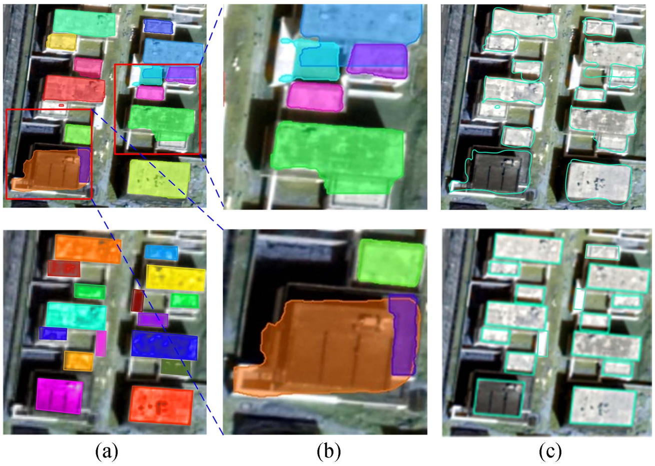


Fig. 1. Comparison of original mask with idealized mask for building recognition. (a) shows the comparison of the original mask and the idealized mask for building recognition. (b) shows the jagged, speckled building outline. (c) shows the comparison of the vectorization results of (a) original mask and (b) idealized mask for building recognition, respectively. (a) Original mask versus idealized mask. (b) Spotty, jagged appearance. (c) Original vectorization versus idealized vectorization.

At present, while deep learning recognition models have made great progress in extracting buildings using geometric, spectral, and textural features, there are still some critical issues that have not been effectively solved. As shown in Fig. 1, first, the edge transition area formed by mixed pixels on both sides of the rooftop boundary leads to inaccurate segmentation results of the deep learning network model at the boundary, especially in the low-contrast boundary region, which also causes secondary

Manuscript received 5 September 2023; revised 27 November 2023 and 3 February 2024; accepted 16 March 2024. Date of publication 28 March 2024; date of current version 17 April 2024. This work was supported in part by the Beijing Key Laboratory of Urban Spatial Information Engineering, in part by the Beijing Institute of Surveying and Mapping Design under Grant 20220112, in part by the Beijing association of higher education under Grant ZD202244, and in part by the Beijing University of Civil Engineering and Architecture Education Scientific Research Projects under Grant Y2111. *(Corresponding author: Liu Jianhua.)*

Liu Jianhua is with the School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, the Beijing Key Laboratory of Urban Spatial Information Engineering, the Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing 100044, China (e-mail: liujianhua@bucea.edu.cn).

Ning Xiaohe, Wang Mengchen, Wang Xinyu, Chen Xiaoyou, and Zeng Shiyi are with the School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing 100044, China.

Liu Yuan is with the Beijing Institute of Surveying and Mapping, Beijing 100038, China.

The code of this article is open source at. https://github.com/GHLJH/MSBR-GNet

Team website: https://www.dxkjs.com/

Digital Object Identifier 10.1109/JSTARS.2024.3382636

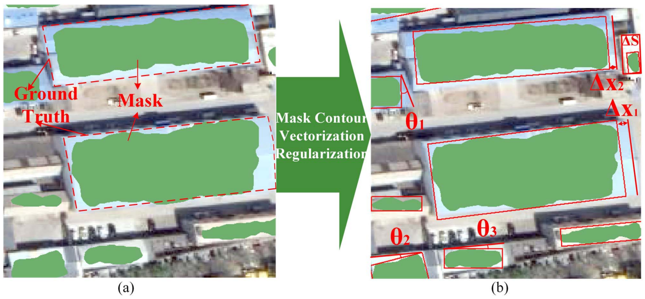


Fig. 2. Mask extraction vectorized contour boundary.

problems, such as the boundary cannot be accurately depicted when the samples are manually constructed; second, the width direction of the rooftop raster boundary is composed of multiple pixels, resulting in polygonal contour lines that appear jagged and speckled during the "raster-vector" conversion process for building rooftop segmentation, which is not in line with the industry standard of GIS building edge vectorization mapping [2], [3], [4]; finally, the method is prone to errors and omissions due to the potential confusion between side and elevation confusion, as well as possible obscuration of tree and shadow with rooftop geometric spectral features.

In response to the above-mentioned problems, postprocessing optimization methods based on building contour extraction results have been proposed, such as those using image morphology processing [5] and optimization methods using image feature information [6], these methods perform overall operations with polygonal contour individuals, and all of them require well-segmented building rooftop contour masks as input. In fact, the building rooftop contour shape obtained by segmentation is generally coarse, making such algorithms difficult to implement and prone to loss of details, which can cause error accumulation problems while regularizing the boundaries [2]. Meanwhile, the traditional deep networks will give rise to blurry boundaries between different semantics among hierarchical features. Therefore, the refinement of boundary features can improve the extraction accuracy [7]. In addition, if the results of the vectorized building roof contour are directly optimized, in other words, the secondary optimization of the model recognition result mask, it will also cause the error accumulation. As shown in Fig. 2, the primary (secondary) directional deviations (θ_1, θ_2, and θ_3), contour (direction and length) deviations (Δx_1 and Δx_2), and area deviations (ADs) (Δs) differ significantly from the real contours of the building rooftop.

In summary, 1) the separation of the building rooftop segmentation and contour regularization processes reduces the accuracy in large-scale processing, so the optimization of the original identification mask generation process for building rooftops directly is the most fundamental and effective method. Adversarial loss in generative adversarial network (GAN) can be used to complement the standard pixel-level loss used in CNN networks, improving the efficiency of model training [8]. In addition, deep learning network models are computationally large and complex, leading to challenges for users to explain the decision mechanisms in the models even if they have a well-defined network structure, which can lead to difficulties in subsequent optimization and generalization of the models. The integration of mechanisms into models has become a hot research concern in the field of computer vision and geomatics.

2) At the same time, deep learning network models need to ensure a certain number and quality of samples if they want to maintain high accuracy of recognition results. Currently, the spatial distribution and types of building samples inadequately capture the coupling factors of diverse regional geographic environments. Moreover, the process of constructing the sample database lacks effective theoretical methodologies to guide it, resulting in overfitting of the model to local area or specific building features. Consequently, this diminishes the generalization ability of deep learning network models when applied on a large scale [9], [10]. Furthermore, the model lacks the ability to quantify the type and quantity of training samples, and it remains unclear which classes of samples (features) contribute to the parameterization process during model training. This can result in repetitive and ineffective labeling of imbalanced training samples, while also lacking guidance from mechanisms in the "black box" construction process of samples. Consequently, there is a steep increase in manual labeling workload without achieving desired outcomes. Therefore, it is necessary to explore how to interpretably and comprehensively to guide the construction of the sample set according to the intrinsic mechanism of sample spectral characteristics [11], and to more efficiently migrate the model with both geographical coupling and global universality [10]. In this study, we propose a generative optimization model multiscale boundary regulation generative net (MSBR-GNet) for building rooftop boundaries guided by interpretable statistical models in the boundary space domain and spectral domain. This module can accurately extract building rooftops from high-resolution remote sensing images and effectively alleviate the problem of model generalization from the above-mentioned contour regularization and interpretability perspectives. MSBR-GNet is based on the MS-CNN model [10], and combines the latest GAN architecture with a library of Tupu samples for training and sample ablation experiments. The main contribution points of this study are as follows.

1) This study proposes a generative optimization network MSBR-GNet for building rooftop boundaries guided by interpretable statistical patterns in the spatial and spectral domains. Based on the statistical patterns of spatial distribution of band grayscale in the edge transition region of high-resolution images, this study uses a GAN as the reference framework and an MS-CNN model as the generator. An innovative boundary loss function is added to the generator mask loss calculation of MSBR-GNet to constrain the regularized generation of building rooftop mask boundaries, and the model training results are made closer to the ground-truth labels by iterative feedback of the discriminator.
2) Based on WHU, INRIA, Massachusetts public building dataset, and national representative city remote sensing image data, we propose a genealogical sample library (dataset) construction method based on the three laws of geography (correlation, heterogeneity, and similarity) and Tupu theory, and accordingly implement BUCEA2.0, a

genealogical Tupu sample library (dataset) for building rooftops.

The rest of this article is organized as follows. Section II describes the progress of existing research on intelligent identification and contour optimization methods for building rooftops. Section III introduces the main methods used in this article. Section IV describes the experiments of this study, including datasets, accuracy evaluation, and experimental parameter settings; experimental results will be presented and discussed in Section V. Finally, Section VI concludes this article.

II. Related Works

A. Building Rooftop Identification

In the field of remote sensing interpretation, building recognition refers to the extraction of individual building targets from remote sensing images [11]. Very high-resolution aerial and satellite images, such as IKONOS, QuickBird, GeoEye, WorldView, Pleiades, Ziyuan-3, and Gaofen-2, provide rich spatial geometric detail information for building recognition [12]. Traditional image target recognition of remote sensing methods include the following.

1) Template-based methods: mainly composed of two parts: template making and similarity measurement. It first constructs a standard library through template making, and then extracts feature vectors from the image to be detected and compares and matches with the standard library. The template library needs to be designed manually, heavily relies on *a priori* knowledge, and is computationally intensive [13].
2) Expert knowledge-based method: It can effectively reduce the phenomenon of false detection, but the key of this method is to build expert knowledge, and the manually constructed expert knowledge relies too much on subjective factors, which easily leads to missed detection and makes the final effect insufficient [14], [15], [16].
3) Traditional machine learning methods: First, the features of the region of interest (RoI) are extracted from the image; Subsequently, these features are fed into a trained classifier for redundant candidate region identification and removal, resulting in refined outcomes [17], [18]. This process necessitates meticulous data cleaning and refinement efforts, accompanied by a substantial human–machine interaction workload.

With the rapid development of deep learning-related research, convolutional neural networks (CNNs) have successfully addressed challenges encountered in traditional machine learning approaches and demonstrated remarkable performance in target detection, image segmentation, and image classification tasks. Deep learning-based target detection methods can be divided into two categories, one is region proposal-based target detection methods and the other is regression-based target detection methods. Region proposal-based target detection methods, include R-CNN [19], SPP-Net [20], fast R-CNN [21], faster R-CNN [22], mask R-CNN [23], hybrid task cascade [24], CPNDet [25], CenterNet2 [26], etc. The class of methods referred to as two-stage methods involves the initial generation of a series of sample candidate frames by the algorithm, followed by classification using CNN. This approach exhibits higher network accuracy but relatively slower speed. Regression-based target detection methods, such as YOLOv1 [27], SSD [28], RetinaNet [29], YOLACT [30], TensorMask [31], M2Det [32], and PolarMask [33]. The class of methods referred to as one-stage methods employs CNN for feature extraction and directly regress object class probability and location information, resulting in faster processing but relatively lower accuracy compared to two-stage methods.

In remote sensing images, multiscale detection of targets with “different sizes” and “different aspect ratios” is one of the main technical challenges in target detection. The term “scale” in the context of multiscale detection of image targets generally refers to the perceptual field size, which encompasses various factors, such as target size, range, mapping scale, resolution, geometric and spectral heterogeneity, and particle size [34]. In the last 20 years, multiscale detection has gone through several historical periods: the “feature pyramids and sliding windows (before 2014),” “object proposals-based detection (2010–2015),” “deep regression (2013–2016),” “multireferences detection (post-2015),” “multiresolution detection (post-2016)” [35]. Earlier detection models, such as the VJ detector [36] and the HOG detector [37], were specifically designed to detect objects with a “fixed aspect ratio,” and for more complex objects, only the “hybrid object proposals” were first used for object detection in 2010 to help avoid exhaustive sliding window searches across images [38]. In recent years, with the increase in GPU computing power, it has become more straightforward for people to handle multiscale detection. The multiscale problem is solved by using deep regression, i.e., directly predicting the coordinates of the bounding box based on deep learning features [39]. However, due to the lack of accurate localization, it can prone to some missed detection of extremely small targets. Multireference and multiresolution detection have become two fundamental frameworks in state-of-the-art target detection systems. Meanwhile, attention mechanism [40] is introduced to aggregate multiscale features to facilitate achieving multiscale symbiosis detection of buildings. *n* of remote sensing images, it is crucial to further investigate the optimization approach for constructing rooftop contours in deep learning network models. This investigation aims to improve the ultimate accuracy of building recognition outcomes while accounting for the impact of multiscale parameters on recognition models. The optimization method of building rooftop contour for deep learning network models should be further investigated to enhance the ultimate accuracy of building recognition outcomes.

B. Building Contour Optimization

The instance segmentation results of general high-resolution images often exhibit jagged and speckled morphological features, which do not conform to the geometric characteristics of building contours or the standard specifications of the GIS mapping industry. Currently, numerous optimization methods have been proposed by researchers for delineating recognition mask contours in high-resolution remote sensing images. These methods can be broadly categorized into three main groups, and this article specifically focuses on enhancing the third category of techniques.

1) Optimization Methods Using Image Morphology Processing: This method mainly uses the operations of image morphology processing, such as erosion and expansion to morphologically correct the contours.

2) Optimization Methods Using Image Feature Information: This method mainly uses the direction of the straight lines in the building outline, the turning points, and the bounding rectangle to rasterize the building outline and reconstruct it.

3) Optimization Methods Using Deep Learning: The exceptional capability of deep feature extraction has positioned deep learning as a prominent approach for extracting buildings from high-resolution remote sensing images. These methods are broadly divided into two categories, one is the optimization method based on boundary loss function, the other is the optimization method based on point generation.

Among them, the methods based on the boundary loss function introduce a boundary function on the basis of CNN to guide the generation of building rooftops boundaries. In the initial stage of building boundary extraction based on deep learning methods, many scholars have used CNN to extract building rooftops. However, the CNN fails to adequately address edge detection, leading to potential issues, such as missed and false detections due to occlusion from vegetation or shadows. Consequently, it proves ineffective in accurately extracting buildings with well-defined boundaries. To address this issue, Gregoris and Stavros [41] developed a new active contour model for building extraction using descriptors from grayscale values and saturated images. In addition, they introduced a new energy term to enhance the accuracy of contour segmentation results. Zhao et al. [42] proposed a new deep neural network that can jointly detect building instances and normalize noisy building boundary shapes from a single satellite image. As the number of parameters increases with the model representation capability, many scholars have enhanced the efficiency and accuracy of model task processing by incorporating an attention mechanism to focus on key information and solve the information overload problem. Jin et al. [43] proposed a new network BARNet embedded with special boundary-aware loss to solve the problem of incomplete segmentation of building boundaries. The network incorporates gated attention refinement fusion units, denser spatial pyramid pool modules, and boundary-aware loss. Li et al. [44] used boundary-aware attractor fields to represent building footprints in remotely sensed images, which helped to enhance building boundaries while suppressing the effects of background clutter. In addition to computation, multiscale features of objects are also a challenge for CNN. Zhu et al. [45] proposed a MAP-Net for efficient and accurate extraction of multiscale building footprint boundaries by parallel localization-preserving convolutional networks, and introduced a channel-level attention module to adaptively compress multiscale features extracted from multipath networks. Liu et al. [10] proposed a multiscale building rooftop recognition model, MS-CNN, to enhance multiscale symbiosis recognition of building rooftops by introducing multiscale parameters in the residual network. The second category is the optimization method of building rooftops contour extraction based on point generation, which considers image segmentation as a point-rendering problem. Kirillov et al. [46] proposed a point-based rendering (PointRend) neural network, which optimizes image segmentation of object edges by iteratively refining the segmentation prediction from the points of the target contour region, and it uses an order of magnitude less floating-point operations than direct dense computation. Wei et al. [47] proposed CLCNN, which initially employs an edge detector to extract the coarse contour boundaries of the RoI and subsequently refines the polygon vertices using a concentric ring convolutional network with bidirectional pairwise loss. This point-based generation method is suitable for building rooftops edges that are difficult to segment, or for the scenes that require high accuracy in edge segmentation.

In recent years, multitask learning in the field of deep learning has provided researchers with novel insights for developing contour optimization methods that exploit the similarity between different tasks to simultaneously solve several different tasks, thereby enhancing the quality of building contour extraction while ensuring the accuracy of building recognition. Shao et al. [48] proposed a building residual refinement network consisting of a prediction module and a residual refinement module for accurate and complete building extraction in remote sensing images. Zhou et al. [49] proposed a new segmentation network framework, BOMSC-Net, which introduces a directional feature optimization module to further sense the orientation information of buildings to refine the results of building boundary segmentation. However, most of these methods are postprocessing optimization methods based on deep learning recognition results, and their optimization results are heavily dependent on the input building segmentation mapping and have a strong dependence on the labeled data, therefore the potential of the method is limited. In addition, the separation of building rooftop segmentation and contour regularization makes the whole method inefficient in large-scale processing.

The GAN is a framework that employs alternating training between generation and adversarial processes. During training, the generation method learns from samples and labels to obtain a model with a distribution similar to the original data. Subsequently, the discriminative network assesses whether generated instances are true or false through continuous confrontation and competition, ultimately capturing the underlying patterns within the data. However, when confronted with complex distributions, controlling the generated outcomes during GAN model training becomes challenging. To address this issue, current GAN models are striving to enhance stability and proposing novel training techniques to elevate model quality.

Currently in the area of adversarial learning strategies for building rooftop extraction, the segmentation model can be viewed as a generative network that can learn building rooftop segmentation results in an adversarial manner by using CNN discriminators. Ding et al. [50] proposed an adversarial shape learning network (ASLNet) to learn shape regularized building extraction results. The ASLNet employs a shape discriminator to eliminate redundant information and prioritize the construction of shape details, followed by a shape regularizer that expands

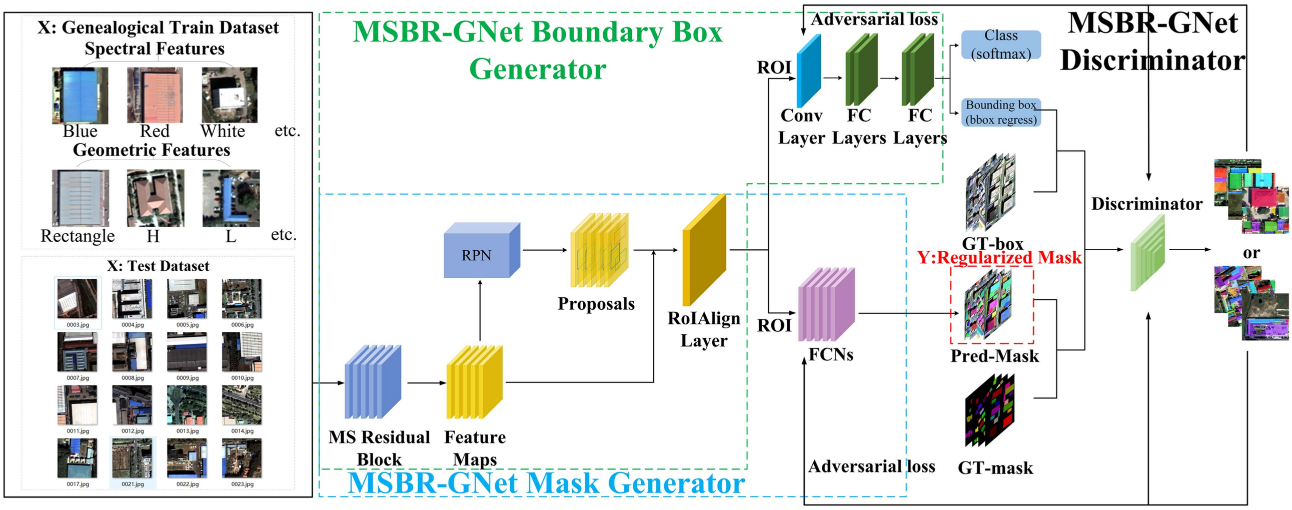


Fig. 3. MSBR-GNet structure diagram.

the receptive field and explicitly models local shape patterns. Li et al. [51] proposed an end-to-end network RegGAN for building footprint generation that utilizes a multiscale discriminator to distinguish between false and true inputs, and a generator to learn from the discriminator responses to generate more realistic building rooftop boundaries. Most of these methods are based on relevant paradigms in the traditional computer vision field with modular and migratory improvements to deep learning models, and have not yet resulted in a dedicated deep learning network model that fits the characteristics of remote sensing data. The remote sensing image features extracted by deep learning models need to consider the geology spatial semantic relation and the physical knowledge of quantitative remote sensing to improve the interpretability and reliability of the models [52]. In this article, we use the above-mentioned features of GAN to guide the construction of boundary loss functions through boundary models based on spatial and spectral domains. This approach aims to enhance the accuracy of building rooftop mask extraction at the boundary while ensuring interpretability of recognition results. Ultimately, our proposed method generates building rooftop masks that are oriented to engineering practical applications and meet the quality standards of GIS data production.

III. Methods

A. Overview

Aiming at the building rooftop recognition and its boundary regularization objective, based on the boundary model (i.e., the spatial statistical distribution pattern that the pixel grayscale in the edge transition region have), this article proposes an MSBR-GNet model that integrates instance segmentation and boundary regularization in an end-to-end network. As shown in Fig. 3, MSBR-GNet is a GAN composed of two modules. The generator aims to learn regularized building rooftop masks and the discriminator distinguishes ideal building masks; these two modules compete with each other in training, and both eventually reach Nash equilibrium to produce the optimal output. The final output of the generator is a building rooftop mask closer to the ground-truth version with clear boundaries and corners.

B. Bounding Box Section

Bounding box generator: As shown in Fig. 4, the bounding box header of MS-CNN [10] is utilized as the bounding box generator of MSBR-GNet. Instead of taking random noise as input like the conventional GAN, the bounding box generator uses RoI features and outputs class and bounding box predictions. During the training process, only the predicted bounding boxes and ground-truth bounding box labels are exclusively selected and subsequently forwarded to the bounding box discriminator.

Bounding box discriminator: As shown in Fig. 4, the purpose of the bounding box discriminator is to receive the bounding box predictions from the generator and evaluate their merit. In fact, it is difficult for any discriminator to evaluate the bounding box predictions of an object by observing at four coordinates. Therefore, our solution is to send the feature maps of the bounding box predictions and their ground-truth values to the discriminator as a pair of false and true sample labels. For this purpose, we use an accurate RoI pool on the feature map output of the FPN backbone to extract the RoI for bounding box prediction. The discriminator network consists of five convolutional layers, and then each layer is immediately followed by a BN layer and a LeakyReLU layer. It accepts a multidimensional image of size 512 × 512 and outputs a score between [0, 1] to indicate the superiority of the prediction result (higher score is better).

C. Mask Section

Mask generator: As shown in Fig. 5, like the prediction head, MSBR-GNet uses the mask head of MS-CNN as the mask generator. The network contains four convolutional layers followed by a transposed convolutional layer and a 1 × 1 convolutional layer (see Fig. 5). It takes MS-CNN RoI features as input and outputs a binary mask prediction of size 512 × 512 for each object class. During the training process, logs are extracted from the channels of the prediction classes and sent to the mask discriminator.

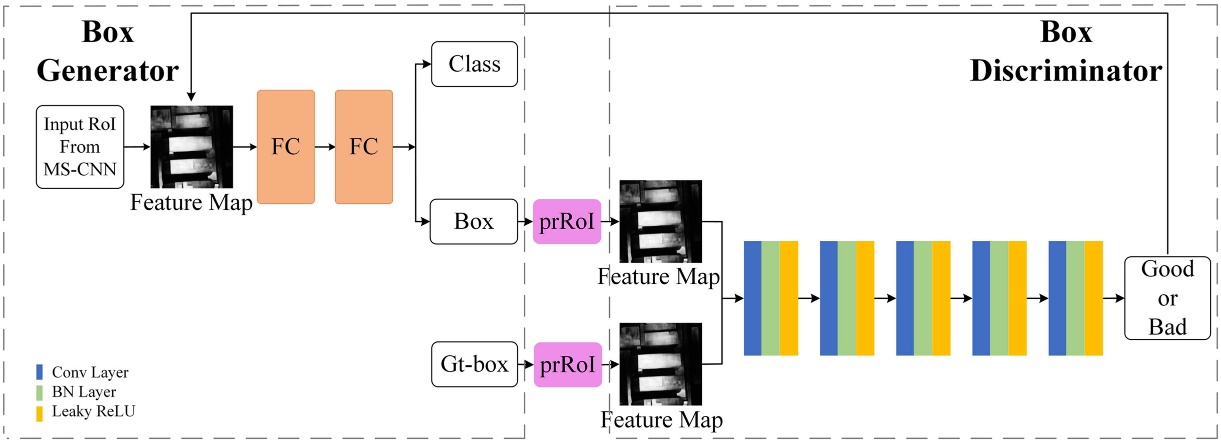


Fig. 4. Box prediction architecture.

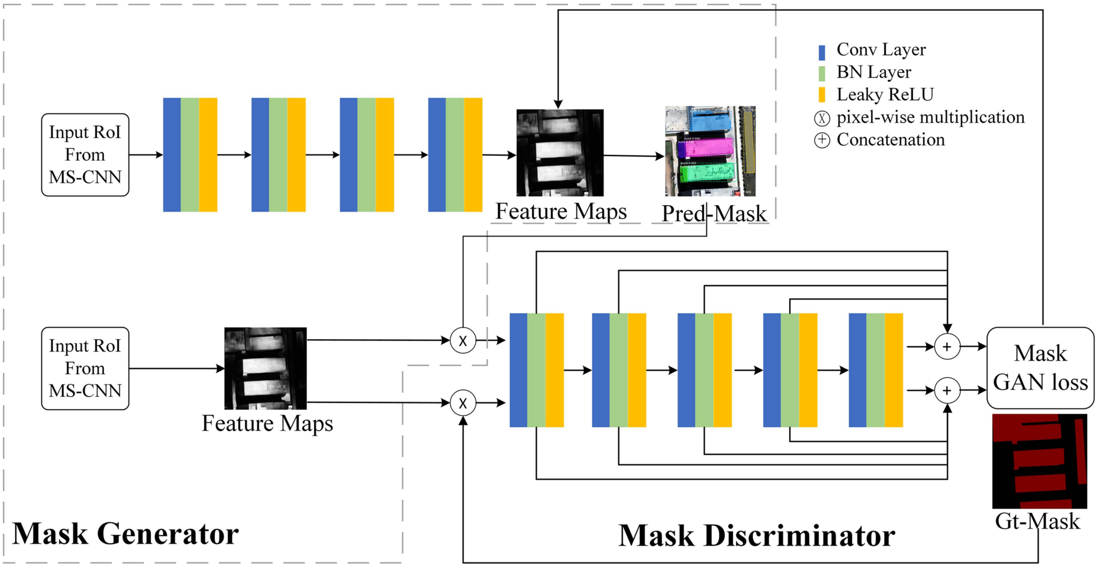


Fig. 5. Mask head architecture.

Mask discriminator: As shown in Fig. 5, the mask and Gt-mask, i.e., the binary mask predictions and their ground-truth values, are input to the mask discriminator, and both are multiplied with RoI features of size 512 × 512. The mask discriminator network consists of five convolutional layers, designed in a manner similar to an encoder. However, due to the inherent resolution degradation problem caused by the convolution operation, we employ a residual link-like approach to connect the features of each convolutional layer with an output. This strategy effectively preserves the geometric detail information of the rooftop and enables its utilization for calculating the adversarial loss.

D. Loss Function

The loss function of the model is a combination of generator loss and discriminator loss. The generator loss function is formulated as follows:

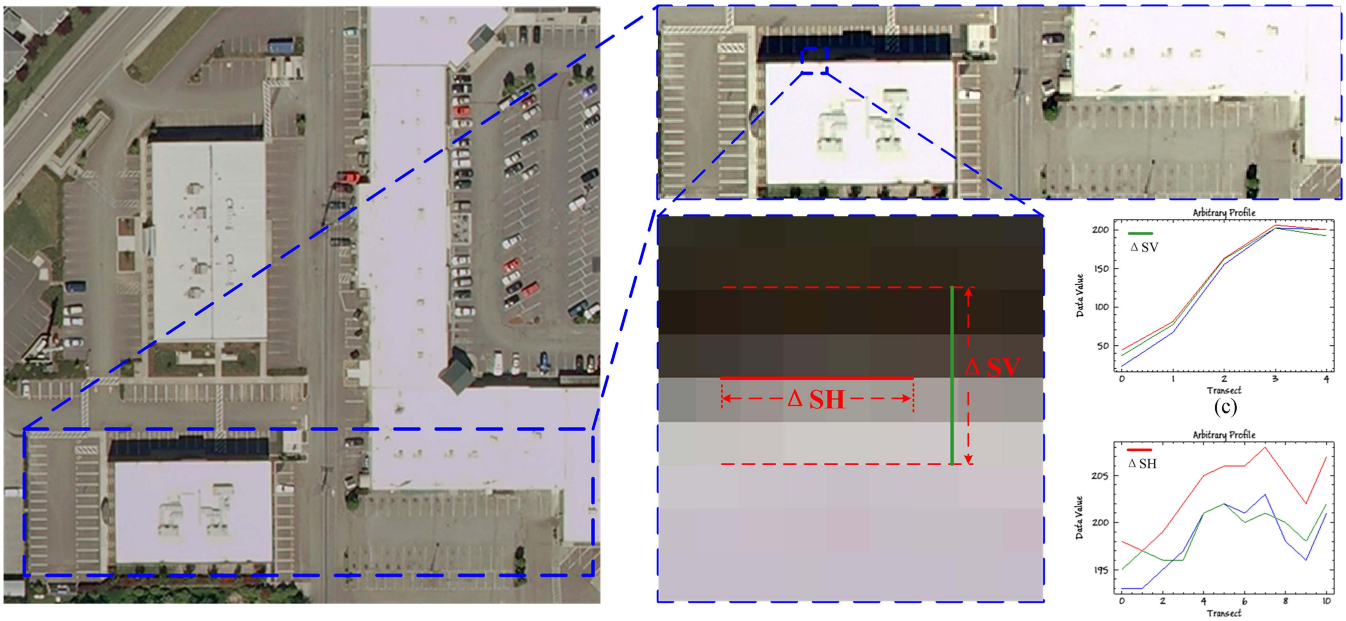
$$L_{\text{Generator}} = L_{\text{cls}} + L_{\text{bbox}} + L_{\text{mask}} + L_{\text{adv}}^{G_b} + L_{\text{adv}}^{G_m} \quad (1)$$

where L_{cls} denotes the loss of the category in the generator MS-CNN; L_{bbox} denotes the loss of the bounding box in the generator MS-CNN; L_{mask} denotes the loss of the mask in the generator MS-CNN. The last two terms $L_{\text{adv}}^{G_b}$ and $L_{\text{adv}}^{G_m}$ are complementary losses, that help the model to further optimize the results of baseline generation. $L_{\text{adv}}^{G_b}$ denotes the generator loss of the bounding box; $L_{\text{adv}}^{G_m}$ denotes the generator loss of the mask. To ensure that the loop set in the GAN achieves closure, a loss of discriminator for bounding boxes $L_{\text{adv}}^{D_b}$ is needed to compete with $L_{\text{adv}}^{G_b}$, and a loss of discriminator for masks $L_{\text{adv}}^{D_m}$ is needed to compete with $L_{\text{adv}}^{G_m}$

$$L_{\text{Discriminator}} = L_{\text{adv}}^{D_b} + L_{\text{adv}}^{D_m} \quad (2)$$

where $L_{\text{adv}}^{G_b}$, $L_{\text{adv}}^{D_b}$, and $L_{\text{adv}}^{G_m}$ are defined as follows:

$$L_{\text{adv}}^{G_b} = \frac{1}{N}\sum_{i=1}^{N} -\log(D_b(G_b(\text{RoI}_i))) \quad (3)$$



(a) (b) (c) (d)

Fig. 6. Example graph of the statistical pattern of grayscale distribution in the edge transition area. (a) Original image. (b) Example of pixel value change in edge transition area. (c) Change in pixel value in the direction of the vertical building rooftop edge ΔSV. (d) Change in pixel value along the building rooftop edge direction ΔSH.

$$L_{\text{adv}}^{D_b} = \frac{1}{N}\sum_{i=1}^{N} -(\log(D_b(bb_i^{gt})) + \log(1 - D_b(G_b(\text{RoI}_i)))) \tag{4}$$

$$L_{\text{adv}}^{G_m} = \frac{1}{N}\sum_{i=1}^{N} \parallel D_m(\text{mask}_i^{gt}) - D_m(G_m(\text{RoI}_i)) \parallel \tag{5}$$

$$L_{\text{adv}}^{D_m} = L_{\text{adv}}^{G_m} \tag{6}$$

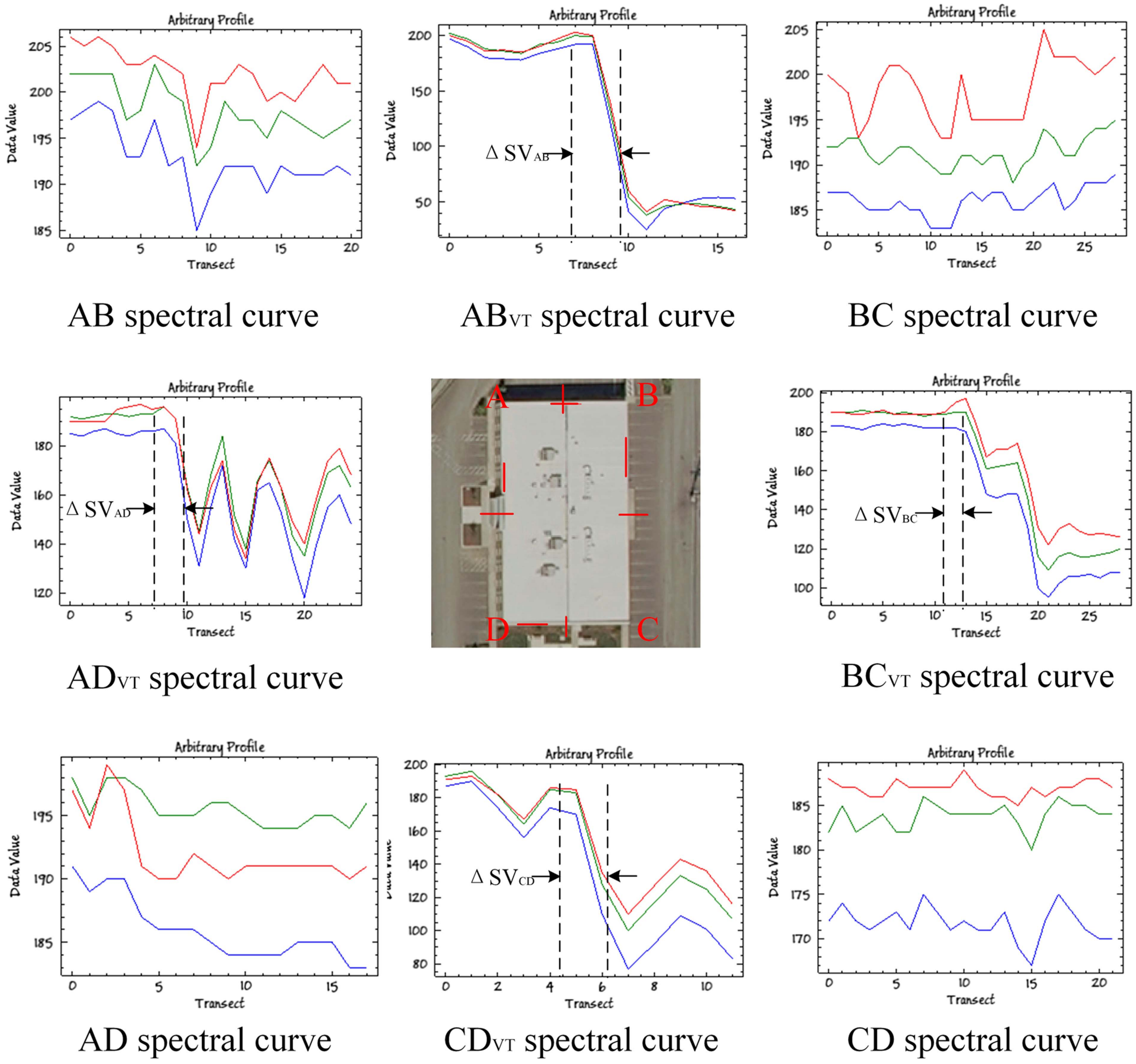
where N is the batch size; $G_b(\text{RoI}_i)$ denotes the predicted value of the RoI of the ith bounding box, denotes the ground-truth value corresponding to the ith bounding box; $D_b(G_b(\text{RoI}_i))$ denotes the probability that the image is true. In training, $L_{\text{adv}}^{G_b}$ encourages G_b to generate a bounding box that can be spoofed to D_b, and $L_{\text{adv}}^{D_b}$ enhances the ability of D_b to distinguish between real and false bounding boxes.

As shown in Fig. 6, in the process of building rooftop mask prediction, the edge transition area $(\Delta SV, \Delta SH)$ formed by the mixed pixels on both sides of the rooftop boundary leads to the inaccurate segmentation results of the model at the boundary. Therefore, in the generator of the mask loss L_{mask}, we add a boundary loss function L_{BR} to constrain the generation of the building rooftop mask boundary in order to improve the accuracy of the building rooftop mask boundary extraction. From the perspective of the loss function, the correspondence between the measured ground-truth boundary vertices and the predicted contour vertices is very complex for pixel-level segmentation. To solve this problem, this article proposes a boundary model based on the spatial and spectral domains for the construction of the boundary loss function L_{BR}.

As shown in Figs. 7–9, the boundary statistical patterns in the spatial and spectral domains describe the fluctuation range of pixels in the value domain and the distribution characteristics in the spatial domain within the edge transition zone of the image, and the boundary model can be established accordingly. According to the spectral profiles of the internal pixels in the edge transition zone along the boundary direction of the building rooftop and perpendicular to the boundary direction in the remote sensing image, we can statistically find that in the edge transition zone $(\Delta SV, \Delta SH)$, the grayscale (R, G, B channels) distribution of the building rooftop along the boundary direction will have a smaller range of fluctuation, and the fluctuation of the grayscale (R, G, B channels) distribution in the perpendicular boundary direction has "step phenomenon."

As shown in Tables I and II, the fluctuation range of pixel grayscale values along the building rooftop edge direction (ΔSV) and perpendicular to the building rooftop edge direction (ΔSV) were statistically analyzed within the edge transition zone $(\Delta SV, \Delta SH)$.

Based on the statistical analysis results, the ΔSH values along the building rooftop edge are generally between 5 and 30, with the mean value of 13.4 and the minimum value of 6, and the fluctuation is small, and the variance is 34.9. The ΔSV values perpendicular to the building rooftop edge are generally above 30, with the mean value of 74.6 and the maximum value of 163, and the fluctuation is large, and the variance is 1835.2. Based on the above-mentioned edge transition area The loss function is constructed based on the above-mentioned statistical feature pattern of the grayscale distribution, first, the loss function L_{BH} is constructed for the pixel value feature distribution pattern along the building rooftop edge direction, and then the loss function L_{BV} is constructed for the pixel value feature distribution



AB spectral curve

ABVT spectral curve

BC spectral curve

ADVT spectral curve

BCVT spectral curve

AD spectral curve

CDVT spectral curve

CD spectral curve

Fig. 7. Profile with pixel values (with white rooftop).

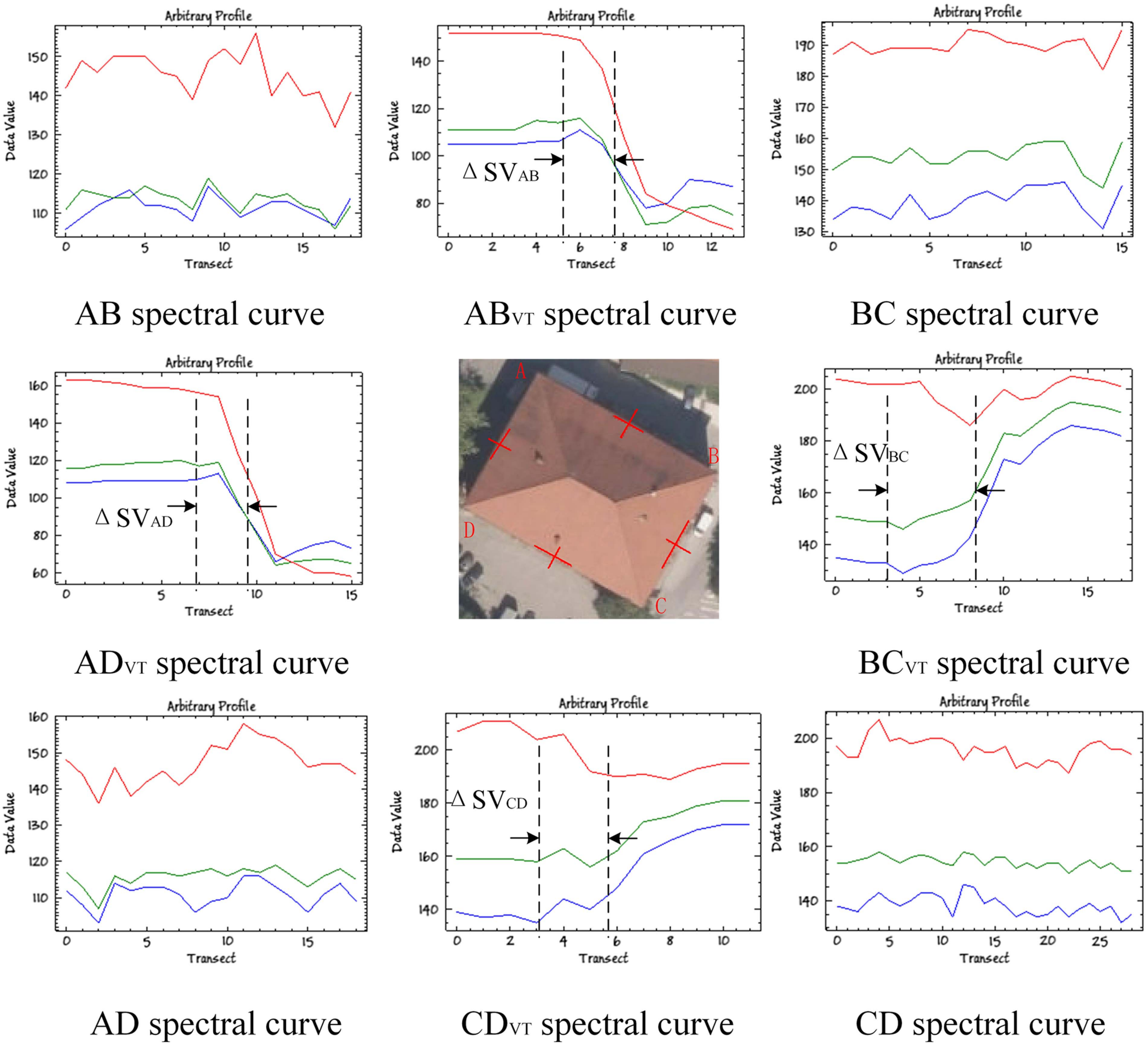
pattern perpendicular to the building rooftop edge direction, and finally the boundary loss function L_{BR} is established based on this boundary model. Equation is as follows:

$$L_{BH} = \frac{1}{N}\sum_{i=1}^{N}(|g_i^H - g_{i-1}^H| - |p_i^H - p_{i-1}^H|) \tag{7}$$

$$L_{BV} = \frac{1}{T}\sum_{i=1}^{T}(|g_i^V - g_{i-1}^V| - |p_i^V - p_{i-1}^V|) \tag{8}$$

$$L_{BR} = \alpha L_{BH} + \beta L_{BV}. \tag{9}$$

As shown in Table III, N denotes the number of edge pixels along the rooftop edge direction and T denotes the number of edge pixels perpendicular to the building rooftop edge direction, i.e., the pixel width of the edge transition zone in this boundary model. p_i denotes the predicted probability value of the ith pixel of the image, g_i denotes the true value of the ith pixel of the image. H and V indicate the pixel width along the building rooftop edge direction and perpendicular to the building rooftop edge direction, respectively. α, β is the hyperparameter that controls the effect of these two losses and sets the initial value $\alpha = \beta = 0.5$, which is generally related to the spatial resolution and boundary intensity of the image. Sobel operator weights the difference between the gray values of the four neighbors of each pixel in the image, up, down, left and right, to reach the extremes at the edges thus detecting the edges. The proposed method, however, effectively handles the gradient in all eight directions surrounding each pixel. The loss function is constructed based on the statistical feature pattern of pixel grayscale distribution in the edge transition region (as illustrated in Fig. 6). In addition, it adheres to the principle of geographic correlation by considering the spatial distribution of pixels. This integration combines the merits of knowledge-driven and data-driven approaches, thereby enhancing model interpretability.



AB spectral curve

AB_{VT} spectral curve

BC spectral curve

AD_{VT} spectral curve

BC_{VT} spectral curve

AD spectral curve

CD_{VT} spectral curve

CD spectral curve

Fig. 8. Profile with pixel values (with red rooftop).

Based on the above-mentioned established loss function, the mask loss L_{mask} of the generator can be expressed as follows:

$$L_{\text{mask}} = \frac{1}{m^2}\lambda\sum_{1}^{m^2}[-y * \log(\text{sigmoid}(x)) - (1-y) * \log(1-\text{sigmoid}(x))] + \eta L_{BR} \quad (10)$$

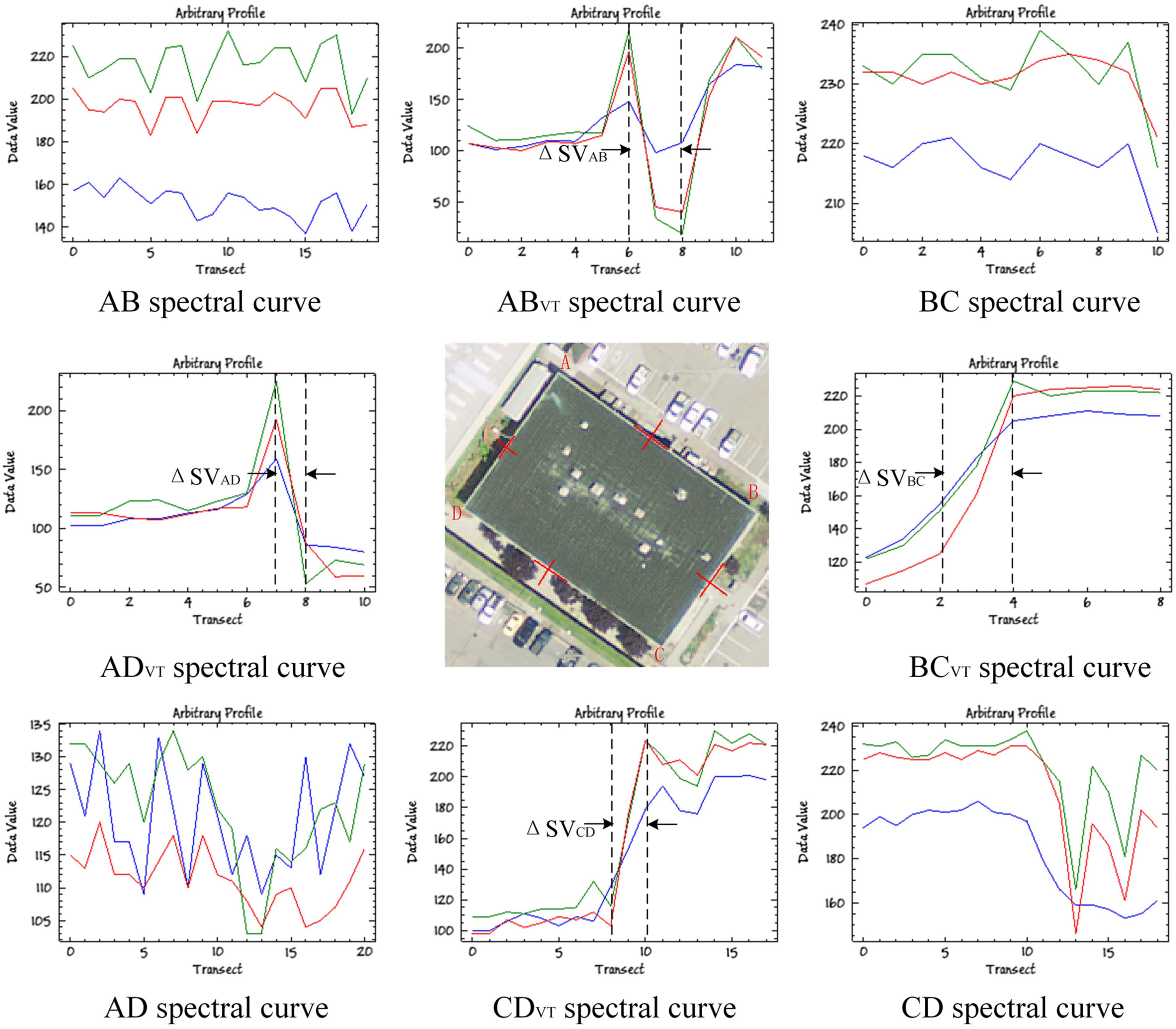
where the first part is the original mask loss (binary cross-entropy loss) of the generator MS-CNN [10], m^2 denotes the output of the mask branch with $m * m$ dimensions, and the second part, i.e., our newly added boundary loss function L_{BR} for optimal building rooftop contour generation. λ, η is the hyperparameter that constrains the two losses, hyperparameters are related to the geographical spatial–temporal heterogeneity of buildings and set the initial value $\lambda = 1$, $\eta = 0.5$.

IV. Experiments

A. Datasets

To verify the effectiveness of the method, we combine public building datasets and high-resolution remote sensing image data of China's building styles regions to construct a self-annotated building rooftop pedigree sample dataset BUCEA2.0. The public datasets include the WHU building dataset, the INRIA aerial image dataset and the Massachusetts building dataset. The resulting BUCEA2.0 dataset covers sensors, resolutions, cities and scenes from different environments.

1) Public Building Datasets for Building Extraction. WHU buildings dataset: This dataset was presented by Ji et al. [1]. The dataset has a large variation in image features including different illumination and atmospheres, sensors, scales, and building structures. The dataset is abundant in diverse data sources, including QuickBird, Worldview series, IKONOS, ZY-3, etc., and the spatial resolution of the image's ranges from 0.3 to 2.5 m.



AB spectral curve

AB$_{VT}$ spectral curve

BC spectral curve

AD$_{VT}$ spectral curve

BC$_{VT}$ spectral curve

AD spectral curve

CD$_{VT}$ spectral curve

CD spectral curve

Fig. 9. Profile with pixel values (with gray rooftop).

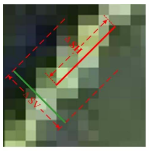
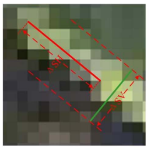
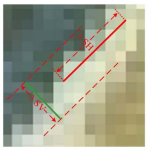
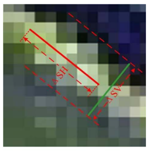
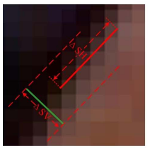
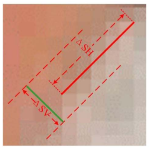
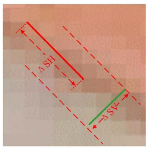
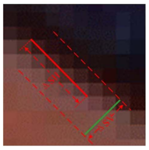
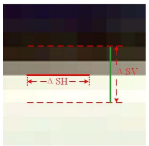
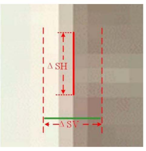
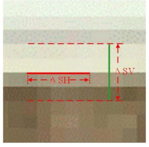
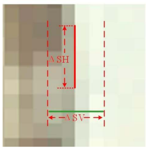
TABLE I
STATISTICS OF PIXEL WIDTH CORRESPONDING TO THE BOUNDARY TRANSITION AREA

Roof color	Band	ΔSH_{AB}	ΔSH_{BC}	ΔSH_{CD}	ΔSH_{AD}
White	R	12	12	5	9
	G	10	7	6	4
	B	14	6	8	8
	AH	13	8	6	7
Red	R	26	13	16	22
	G	16	14	6	11
	B	13	14	13	13
	AH	18	14	12	15
Gray	R	25	5	10	15
	G	30	10	15	31
	B	20	6	15	24
	AH	25	7	13	23

TABLE II
STATISTICS OF THE RANGE OF PIXEL VALUE CHANGES IN THE BOUNDARY TRANSITION AREA: ΔSV

Roof color	Band	ΔSV_{AB}	ΔSV_{BC}	ΔSV_{CD}	ΔSV_{AD}
White	R	160	30	70	45
	G	160	30	85	45
	B	170	35	90	55
	AV	163	32	82	48
Red	R	65	15	15	85
	G	45	35	20	55
	B	30	40	20	45
	AV	47	30	18	62
Gray	R	150	100	120	100
	G	200	75	105	170
	B	50	40	50	80
	AV	133	72	92	117

TABLE III
STATISTICS OF THE PIXEL WIDTH CORRESPONDING TO THE BOUNDARY TRANSITION AREA

Data source	World view-3		Spatial resolution	0.3m
Example of an edge transition zone				
T	8.48	9.9	7.07	7.07
Example of an edge transition zone				
T	8.48	9.9	7.07	7.07
Example of an edge transition zone				
T	5	5	8	8

The dataset contains 204 512 × 512 RGB images, providing aerial images and corresponding ground-truth images. Examples are shown in Fig. 10(a) and (b).

INRIA aerial image dataset: This dataset, presented by Maggiori et al. [53], has a total area of 810 km, of which 405 km are used for the training set and the test set contains 805 km. The images cover different urban buildings, ranging from dense urban villages to high mountain towns.

Massachusetts buildings dataset: This dataset was presented by Volodymyr [54]. He presented the Massachusetts road dataset and the Massachusetts building dataset in Chapter 6 of his Ph.D. thesis. One of the building datasets consists of 151 aerial images of the Boston, USA, divided into a training set of 137 images, a validation dataset of four images, and a test set of ten images. Each image has 1500 × 1500 pixels, and each image covers an area of 2.25 km^2. Therefore, the whole dataset covers about 340 km^2. The dataset contains example image data, as shown in Fig. 10(c) and (d) in the following. In this study, 10% (125) of the images in the training set, cropped to 512 × 512 pixels, were selected to construct a genealogical sample library of building rooftop.

2) Self-Annotated Genealogical Sample Dataset BUCEA 2.0: A highly accurate sample pool is one of the most important factors influencing the recognition results of deep

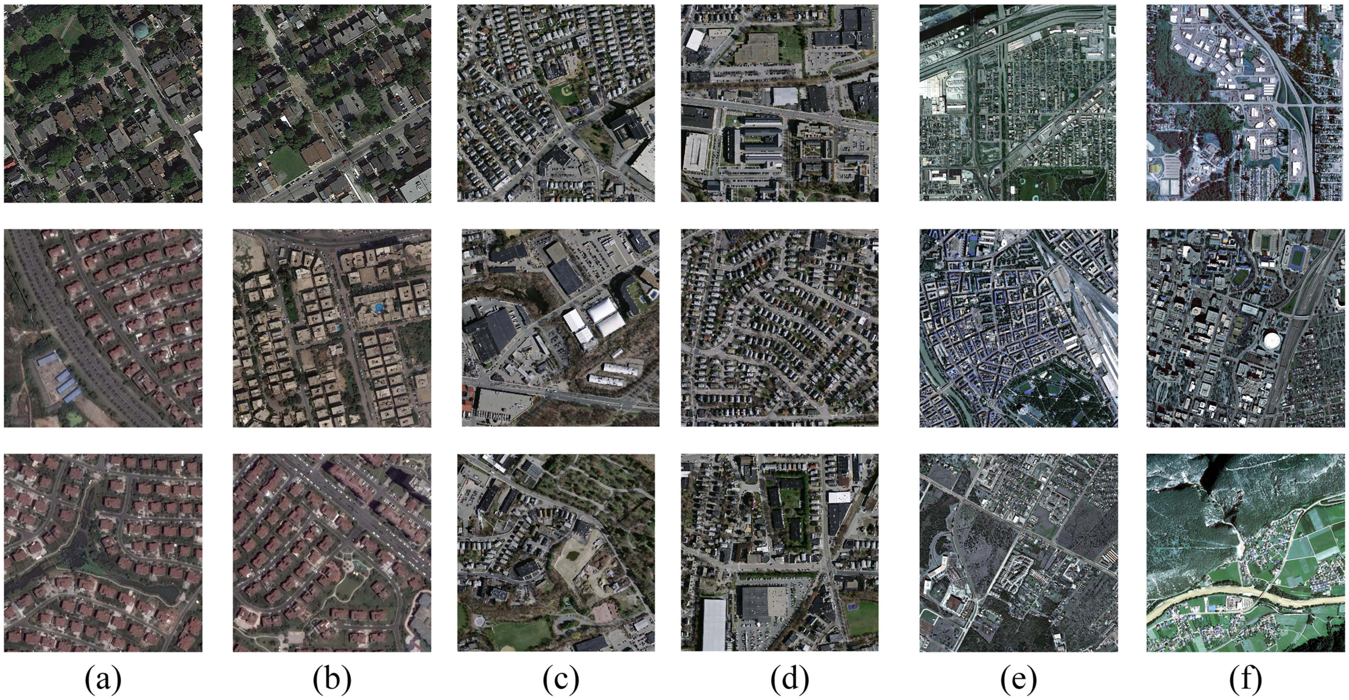


Fig. 10. Overview of the dataset from (a) and (b) WHU dataset; (c) and (d) Massachusetts dataset; (e) and (f) INRIA building dataset.

learning network models. The spatial differences in solar radiation, land and sea location, and altitude cause the natural environment and biomes to differentiate geographically, resulting in the geographical spatial–temporal heterogeneity of buildings as a vehicle for human settlement. The spatial distribution of buildings and building types are not sufficiently characterized by the commonly used sample libraries, which cannot reflect their geographical spatial–temporal heterogeneity, resulting in models with high extraction accuracy in some areas or a certain type of buildings, but seriously lacking in generalization ability. At the same time, most of the current building extraction methods only focus on the first level of building classification, and the lack of corresponding classification label data when facing the demand for accurate extraction of different building rooftop types greatly hinders the engineering application of the model. Therefore, this study builds BUCEA2.0, a Tupu sample library with coupled geographic environment factors, based on Tupu theory and geographic knowledge, such as the laws of geography.

As shown in Fig. 11, this study uses high-resolution remote sensing images of representative cities at the global scale as the base data, selecting places such as Milan and Wuhan in the Asia-European plate, Chicago and Santiago in the American plate, Cairo in the African plate, and Christchurch in the Indian Ocean plate, respectively. In accordance with the law of geographic heterogeneity, BUCEA 2.0 was built on the basis of BUCEA 1.0 [10] and guided by the intrinsic mechanism of the sample genealogical features of the geographically coupled buildings [11] to more efficiently migrate the model with both geographic coupling and generalizability. The geometric and spectral features of building rooftops are extracted from the remote sensing images of these representative cities to form a subset of the spectral samples, which in turn form a global sample library of Tupu.

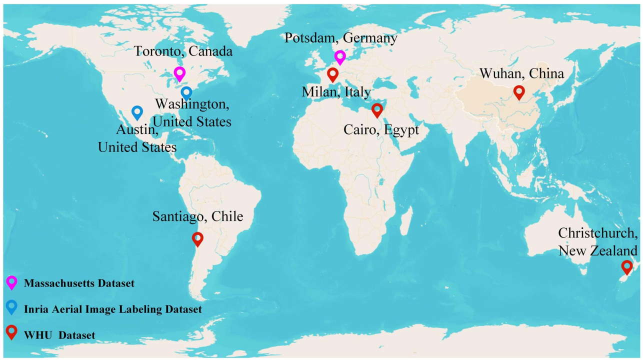


Fig. 11. Global distribution of representative cities in the self-labeled sample dataset.

In addition, this study illustrates the geographical coupling of the building rooftop dataset and the idea of constructing the Tupu sample database of this article, using the geographical distribution of typical building landscapes in China as an example. China is a vast country, spanning different climatic zones from north to south, such as cold temperate, middle temperate, warm temperate, subtropical, and tropical. The north-eastern plains (in the case of Changchun) and the plains of northern China have a temperate monsoon climate, with high temperatures and rain in summer and cold, dry winters; the middle and lower reaches of the Yangtze River and the hills of the south-east (in the case of Fuzhou) have a subtropical monsoon climate,

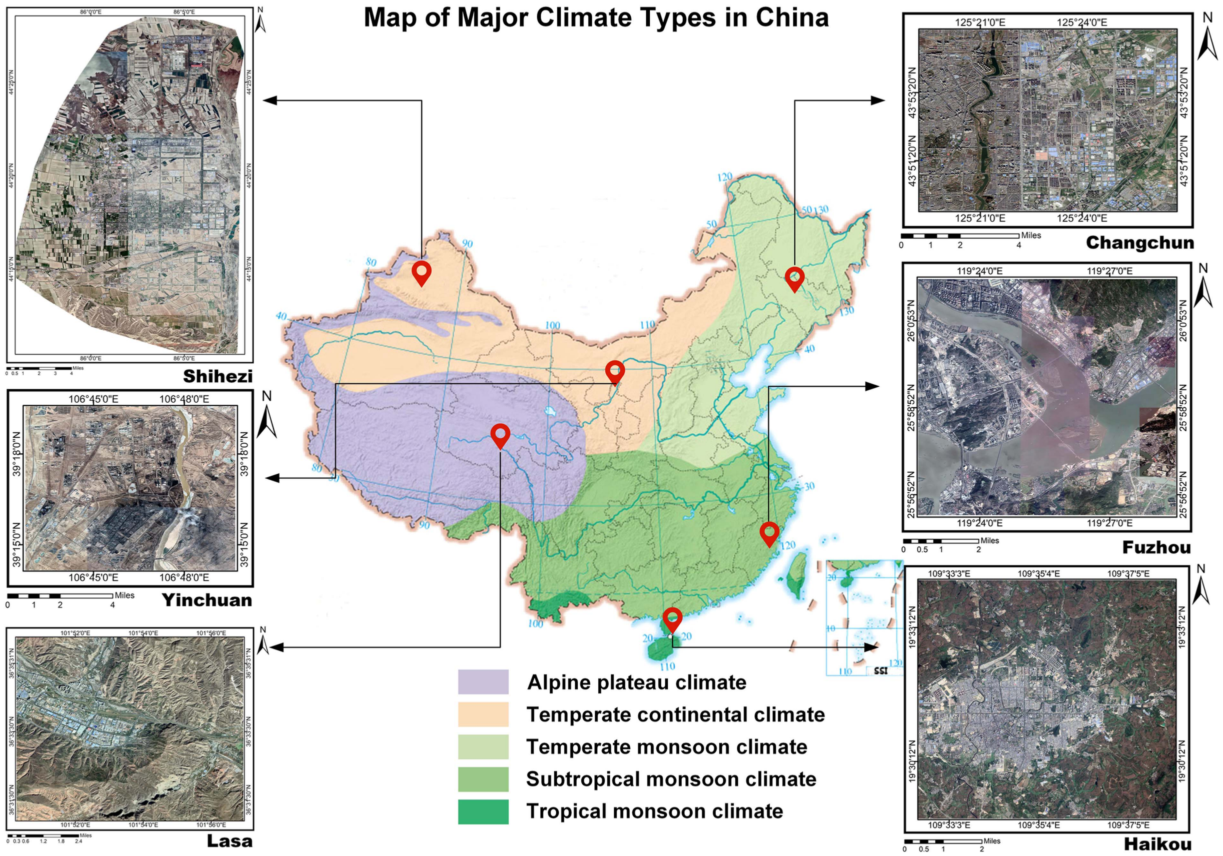


Fig. 12. Chinese geographically coupling pedigree building rooftop sample dataset range distribution (standard map source: http://bzdt.ch.mnr.gov.cn/).

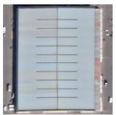
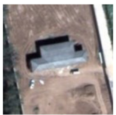
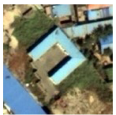
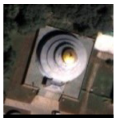
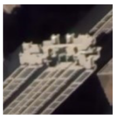
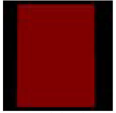
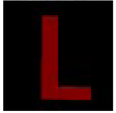
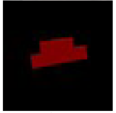
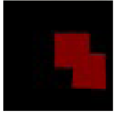
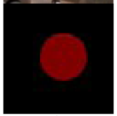
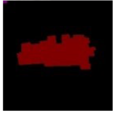
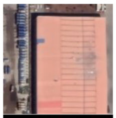
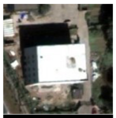
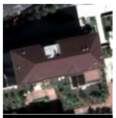
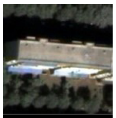
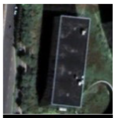
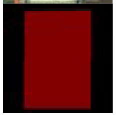
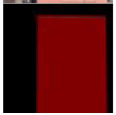
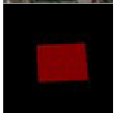
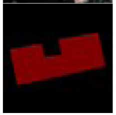
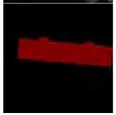
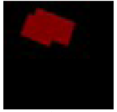
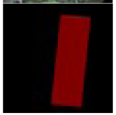
with high temperatures and rain in summer, mild winters and sloping roofs; southern Taiwan, Hainan Island (in the case of Haikou), and southern Yunnan have a tropical monsoon climate, with no winter and high temperatures and rain all year round; the northwest region (Shihezi and Yinchuan, for example) is far from the sea, with a dry climate and scarce precipitation, and has a temperate continental climate; the Qinghai–Tibet Plateau (Lasa, for example) is at a high altitude, covered with snow and ice during the winter half of the year, and cool and pleasant during the summer half of the year, a typical plateau climate. As shown in Fig. 12, according to the law of geographic heterogeneity, we select high-resolution remote sensing images of six typical areas of architectural landscape to create a sample dataset of building rooftops, so that geometric and spectral features can be characterized in diversity. The building styles are differentiated with regional climate and other environmental changes, reflecting the geographical spatial and temporal heterogeneity of buildings. Therefore, the construction of a spectral sample dataset for these regions based on Tupu theory fully considers the law of geographic heterogeneity and satisfies the need for geographical coupling of the sample dataset.

As shown in Table IV, according to the style and feature type of building rooftops from remote sensing images of the study area, building rooftops can be classified according to geometric features as follows: rectangle, character shape (H, L, T, U, Z, and other geometric shapes and variants), circle, combination shape, etc.; according to the building rooftop spectral characteristics can be classified into the following categories by visual color: red, yellow, green, blue, dark gray, white, gray, brown, etc. We used the image annotation tool Labelme to extract the feature-generating attributes and mask information of various types of target buildings, and then transformed them into COCO dataset format to construct a geographically coupled spectral high-resolution remote sensing building rooftop sample dataset. BUCEA 2.0 sample library [55] contains 1400 images, BUCEA 2.0 sample library contains 1400 images, 800 images are 512 × 512 pixels, 600 images are 512 × 512 pixels.

B. Experimental Procedure

1) Method Implementation: All experiments in this article were implemented in the framework of Pytorch 0.4.1 powered by CUDA 9.0 on Ubuntu 16.04 OS and run on a single GPU of NVIDIA GeForce GTX 1060. This experiment trains the network by using backpropagation from the original MS-CNN losses and a newly added boundary loss function. The generators and discriminators are trained alternately. The generator is first frozen and the discriminator is trained. As the discriminator is trained to try effectively discerning real data from false data, it must learn how to identify defects in the generator. Similarly, the discriminator is frozen while the generator is trained. As the

TABLE IV
EXAMPLE FEATURES OF THE PEDIGREE SAMPLE LIBRARY

Geometric features	Rectangle	H-building	L-building	T-building	U-building	Z-building	Oval	Combined
Imagery								
Mask								
Count	25525	28	1921	608	290	139	62	3903
Percentage	78.60%	0.10%	5.90%	1.90%	0.90%	0.40%	0.20%	12.00%
Spectral features	Blue	Red	White	Brown	Yellow	Grey	Green	Dark gray
Imagery								
Mask								
Count	8289	5850	2856	3334	442	4690	581	4720
Percentage	27.00%	19.00%	9.30%	10.80%	1.40%	15.30%	1.90%	15.30%

training progresses, both the generator network and the discriminator network exhibit enhanced capabilities, ultimately enabling the generator to generate predictions that closely approximate the ground truth. In all experiments, we used a weight decay of 0.0001 and a momentum of 0.9 with SGD optimizer. In addition, we used the same learning rate of 0.001 for both the generator and the discriminator.

In order to quantitatively evaluate the performance of MSBR-GNet, this study evaluated performances of the model in two aspects. The mask metrics are used to evaluate the generated building rooftop masks. The boundary metrics are used to evaluate the quality of the extracted building rooftop boundaries.

2) Evaluation Metrics:

1) *Mask metrics:* The evaluation of the example segmentation in this study incorporates several commonly employed metrics, including precision, recall, F1-score, and IoU. Precision indicates the percentage of correctly predicted areas as buildings, while recall indicates the proportion of correctly predicted buildings in the building ground truth. F1-score is the sum of precision and recall The F1-score is the weighted average of precision and recall; IoU represents the ratio of the intersection area between building prediction and ground truth to the joint area. The following equation can be used to express the evaluation metrics:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{12}$$

$$F_1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \tag{13}$$

$$\text{IoU} = \frac{|P_p \bigcap P_t|}{|P_p \bigcup P_t|} = \frac{\text{TP}}{\text{TP + FN + FP}} \tag{14}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Here, P_p denotes the set of pixels predicted to be buildings and P_t denotes the ground-truth set. $|\bullet|$ denotes the function to calculate the number of pixels in the set.

2) *Boundary metrics:* The Hausdorff distance (HD) and structural similarity measure (SSIM) metrics are commonly used in evaluating object detection performance. However, it is important to note that the boundary IoU metric exhibits higher sensitivity toward large object boundary errors without excessively penalizing small objects. Nevertheless, it should be acknowledged that this metric may not be universally applicable when assessing building rooftops. Therefore, this evaluation metric is not used in this article [56]. In addition, four new evaluation metrics are introduced in this study to impose constraints on the building rooftop boundaries, which are AD, circumference deviation (CD), geometrical center distance (GCD), and main direction deviation (MDD). The following equation can be used to express the evaluation index:

$$\begin{aligned} d_H(X,Y) &= \max\{d_{XY}, d_{YX}\} \\ &= \max\{\text{maxmin}(x,y), \text{maxmin}(x,y)\} \end{aligned} \tag{15}$$

where X and Y denote the ground-truth and prediction maps, respectively. It can be interpreted as the maximum value of the shortest distance from a point in a point set to another point set. To eliminate the effect of very small outlier sets, the HD is multiplied by 95% to obtain the final evaluation index (95% HD). When the distance is small, the predicted result exhibits a shape that closely resembles the form of the actual label

$$S(X,Y) = F(l(X,Y), c(X,Y), s(X,Y)) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{16}$$

where $l(X,Y)$, $c(X,Y)$, $s(X,Y)$ denote the luminance function, contrast function, and structure function, respectively. μ, σ, and σ_{xy} represent the mean, variance, and covariance, respectively. To avoid dividing by zero, C_1 and C_2 denote the constants 6.50 and 58.52. The range of SSIM is (–1,1). The value of SSIM is equal to 1 when the two images are identical

$$\text{AD} = 1 - \log\left(\frac{\sum_i^n S_X}{\sum_i^n S_Y}\right) \tag{17}$$

$$\text{CD} = 1 - \log\left(\frac{\sum_i^n C_X}{\sum_i^n C_Y}\right) \tag{18}$$

$$\text{GCD} = 1 - \log\left(\frac{\sum_i^n \sqrt{(x_g)^2 + (y_g)^2}}{\sum_i^n \sqrt{(x_p)^2 + (y_p)^2}}\right) \tag{19}$$

$$\text{MDD} = |\theta_X - \theta_Y| \tag{20}$$

where X and Y denote the ground-truth and predicted result, respectively. S_X denotes the true area of the building rooftop ground and S_X denotes the area of the building rooftop split mask; C_X denotes the true parameter of the building rooftop ground and C_Y denotes the parameter of the building rooftop split mask; (x_g, y_g) denotes the geometric centroid of the true value of the building rooftop ground and (x_p, y_p) denotes the geometric centroid of the predicted value of the building rooftop; AD, CD, and GCD all quantify the proximity between the predicted result and the ground truth, with a higher closeness to 1 indicating a closer alignment between them. θ_x indicates the principal directional angle of the ground-truth value of the building rooftop, and θ_y indicates the principal directional angle of the predicted value of the building rooftop, where the principal directional angle of the predicted value of the building rooftop is the principal directional angle of the predicted mask after it is turned into the smallest external rectangle, as shown in Fig. 13.

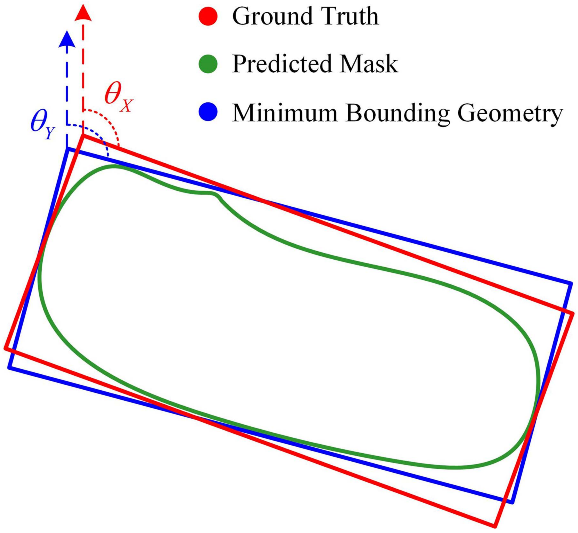


Fig. 13. Schematic diagram of MDD.

C. Result

The results of the building rooftop test identification on remote sensing imagery from three different cities are shown in Fig. 14. (a) is Bellingham, a city in Washington State, USA, which is characterized by less variation in the spectral characteristics of building rooftops which predominantly exhibit cool tones. The majority of buildings are single-family houses with a large number of small and medium-sized structures; (b) is Tyrol, a region of the Alps straddling western Austria and northern Italy, characterized by significant variations in the spectral properties of building rooftops and a higher prevalence of large contiguous structures; (c) is the south-western Austrian city of Innsbruck, a region where Gothic-style buildings are lined up and much of the old town appearance is preserved. As can be seen in Fig. 14, MSBR-GNet is able to extract a large range of building rooftops from different geographical areas, and the contour patterns of large and small buildings remain intact and have prominent edge features. The comparative analysis of the local detail information of the rooftop will be described in the "Discussion" section.

V. Discussion

A. Comparison Experiments

To further analyze and compare the strengths and limitations of MSBR-GNet, as well as to validate the generalization capability of the model, we present building prediction results and local detail comparison plots on three publicly available datasets, respectively. In this article, the MSBR-GNet method is compared with five mainstream SOTA building extraction methods, including mask R-CNN [23], MS-CNN [10], MAP-Net [45], STT [57], and CBR-Net [58]. Mask R-CNN is the most popular building instance extraction method, which adds a network branch to the original R-CNN for predicting segmentation masks on each RoI. MS-CNN is an improved residual block based on mask R-CNN, which introduces multiscale parameters to improve the model's ability to identify multiscale building symbiosis. The MAP-Net is an innovative neural network that employs multiple attending paths to learn spatial localization-preserved multiscale features. These features are extracted through a multiparallel path, where each stage progressively generates high-level semantic features with a fixed resolution. STT uses the transformer for efficient building extraction and introduces a new module, the sparse token sampler, by which buildings are represented as a set of "sparse" feature vectors in their feature space, greatly reducing computational complexity. CBR-Net introduces a boundary refinement module that reflects building predictions by sensing the orientation of each pixel in an optical remote sensing image to the center of the nearest object to which it may belong, for the purpose of boundary optimization. These methods have achieved SOTA performance in either semantic

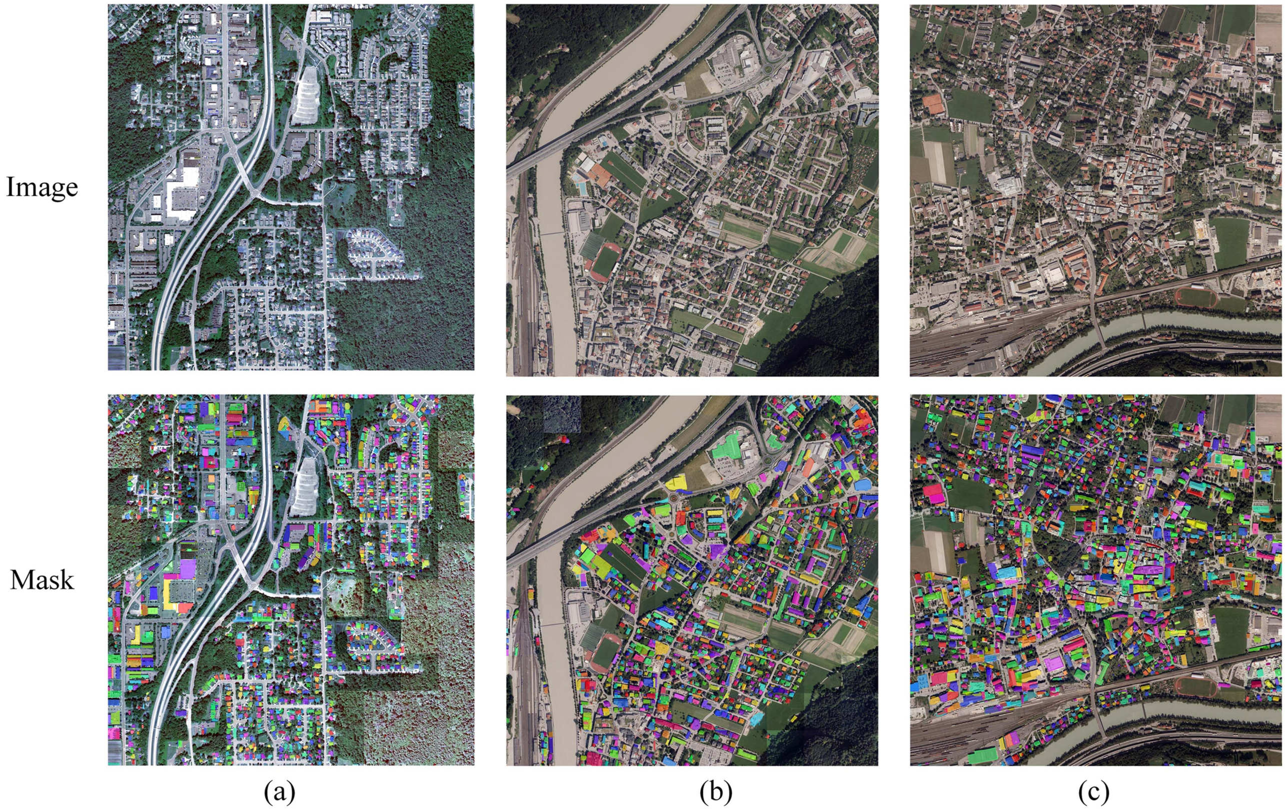


Fig. 14. Test results from remote sensing images of three different cities. (a) Bellingham. (b) Tyrol. (c) Innsbruck.

TABLE V
QUANTITATIVE COMPARISON OF EXTRACTION ACCURACY WITH SOTA METHODS ON INRIA AERIAL IMAGE DATASET

Method	Precision	Recall	F1-score	IoU
Mask R-CNN	0.8238	0.7809	0.8018	0.7679
MS-CNN	0.9228	0.8458	0.8826	0.8243
MAP-Net	0.9141	**0.9056**	0.9099	0.8346
STT	0.8974	0.9024	0.8999	0.818
CBR-Net	0.9068	0.8783	0.8923	0.8055
MSBR-GNet	**0.9275**	0.9012	**0.9142**	**0.8584**

Bold values indicate the best value for the corresponding indicator.

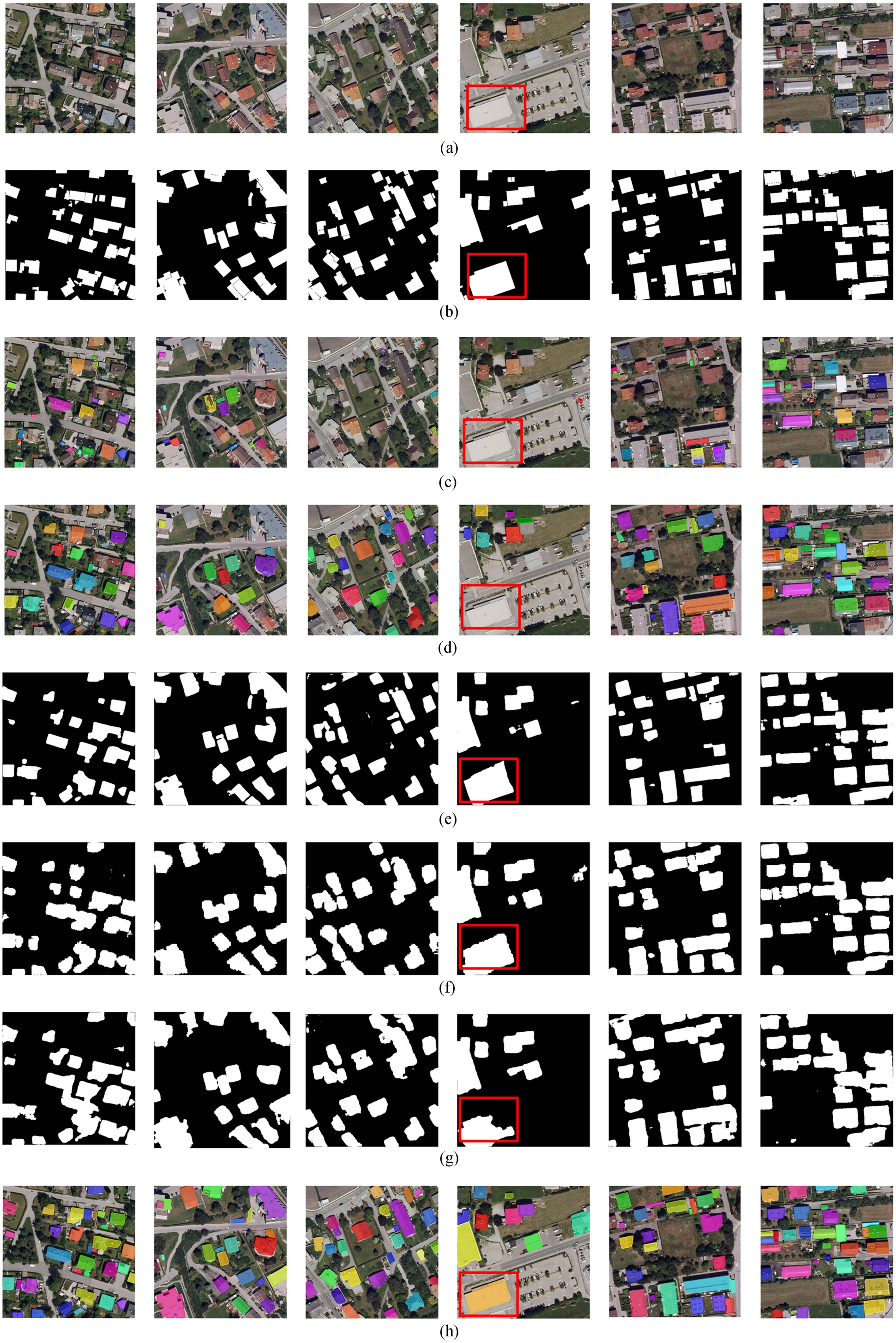
segmentation or building extraction tasks. Among them, mask R-CNN, MS-CNN, and MSBR-GNet belong to the instance segmentation methods, while we added experiments with three semantic segmentation models (MAP-Net, STT, and CBR-Net) to compare the two classes of methods. All methods use the same experimental setup and the same self-labeled building rooftop dataset BUCEA 2.0 mentioned in 4.1 as the training set, and are tested on three publicly available building datasets to compare the model recognition effectiveness and generalization capability.

1) Evaluation With the INRIA Aerial Image Dataset: The extraction accuracy comparison results on the INRIA aerial image dataset are presented in Table V. MSBR-GNet achieved the best results in precision, F1-score, and IoU metrics, with F1-score and IoU reaching 91.42% and 85.54%, respectively. In terms of precision, the proposed method outperformed mask R-CNN, MS-CNN, MAP-Net, STT, and CBR-Net by 10.37%, 0.47%, 1.34%, 3.01%, and 2.07%, respectively. It indicates that the method can effectively improve the accuracy of building rooftop extraction. The index in recall is slightly lower than MAP-Net and STT (0.0044, 0.0012) due to the higher resolution of the INRIA aerial remote sensing image dataset and the influence of the "same spectrum with different objects" phenomenon in remote sensing images, which confuses the spectral features of buildings with the surrounding features and increases the false detection rate of building rooftops.

As shown in Table VI, the comparison results of the boundary quality on the INRIA aerial imagery dataset. The MSBR-GNet model exhibited superior performance across all boundary evaluation metrics, with a minimum HD_{95} value of 80.59 and a maximum SSIM value of 0.9245, indicating its effectiveness in enhancing the accuracy of boundary extraction. Due to the small and scattered buildings in the INRIA dataset, the values of all indicators tested in all models on this dataset are relatively high.

As shown in Fig. 15, a further look at the results of the tests in the INRIA aerial dataset in detail. MSBR-GNet was able to ensure the integrity of the large building outline extraction, as shown by the red box in the fourth column in Fig. 15. In the INRIA aerial imagery dataset, the buildings in the area are relatively scattered and not contiguous, which has less impact on the extraction of building rooftops.

2) Evaluation With the WHU Building Dataset: As shown in Table VII, the comparison results of extraction accuracy



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 15. Visualized results of building instance extraction using the INRIA aerial image dataset. (a) Original remote sensing imagery. (b) Ground truth. (c) Mask R-CNN. (d) MS-CNN. (e) MAP-Net. (f) STT. (g) CBR-Net. (h) Proposed model.

TABLE VI
QUANTITATIVE COMPARISON OF BOUNDARY QUALITY WITH SOTA METHODS ON INRIA AERIAL IMAGE DATASET

Method	HD_{95}	SSIM	AD	CD	GCD	MDD
Mask R-CNN	84.13	0.8744	0.8344	0.701	0.8145	2.35°
MS-CNN	83.51	0.8833	0.9023	0.688	0.8232	2.19°
MAP-Net	83.24	0.8747	0.8336	0.625	0.7424	5.67°
STT	99.65	0.8953	0.8293	0.6452	0.7528	5.58°
CBR-Net	97.58	0.9037	0.8925	0.6895	0.7815	3.15°
MSBR-GNet	**80.59**	**0.9245**	**0.9065**	**0.708**	**0.8564**	**1.03°**

Bold values indicate the best value for the corresponding indicator.

TABLE VII
QUANTITATIVE COMPARISON OF EXTRACTION ACCURACY WITH SOTA METHODS ON WHU DATASET

Method	Precision	Recall	F1-score	IoU
Mask R-CNN	0.8238	0.7809	0.8018	0.7679
MS-CNN	0.8564	0.8051	0.83	0.7756
MAP-Net	0.9043	0.8362	0.8689	0.7964
STT	0.8962	0.8435	0.8691	0.7855
CBR-Net	0.9058	0.8389	0.8711	0.8098
MSBR-GNet	**0.9228**	**0.8458**	**0.8826**	**0.8243**

Bold values indicate the best value for the corresponding indicator.

on the WHU building dataset. MSBR-GNet achieved the best results in all four metrics, with the F1-score and IoU reaching 88.26% and 82.43%, respectively. In terms of precision, the proposed method outperformed mask R-CNN, MS-CNN, MAP-Net, STT, and CBR-Net by 9.90%, 6.64%, 1.85%, 2.66%, and 1.70%, respectively. In terms of recall, the proposed method outperformed mask R-CNN, MS-CNN, MAP-Net, STT, and CBR-Net by 6.49%, 4.07%, 0.96%, 0.23%, and 0.69% respectively. MSBR-GNet was effective in improving the accuracy of building rooftop extraction. The overall metrics tested on this dataset were slightly lower than the INRIA dataset. In addition to differences in satellite sensors, atmospheric conditions, panchromatic and multispectral fusion algorithms in the imagery, atmospheric, and radiometric corrections, and seasonal variations, the ability to generalize the model was also influenced by seasonal fluctuations. The densely distributed buildings, primarily consisting of factory structures and exhibiting a contiguous pattern, also exerted an impact on the overall evaluation metrics.

The comparison results of the boundary extraction quality on the WHU satellite image dataset are shown in Table VIII. On both HD_{95} and SSIM, the proposed model MSBR-GNet achieved the best results with 185.46 and 0.9277, respectively, the overall metrics in the WHU dataset are comparatively low when compared to those in the INRIA dataset, primarily due to the varying resolution of building images and their overall darker tone, which unfavorably impacts the predicted results after being input into the model.

The results of building extraction on the WHU building dataset are shown in Fig. 16. As shown in the red boxed part of the second column of Fig. 16, dense building area adhesion occurred in the STT and CBR-Net recognition results, and building omission occurred in mask R-CNN and MAP-Net, while our proposed MSBR-GNet model can ensure the integrity of building contour extraction while guaranteeing a certain detection rate. In the remote sensing images of the WHU dataset, the features around the building have similar spectral, textural, shape, and color characteristics to the plant, which increasing the incidence of wrong and missed detection of the building rooftop. As shown in the red box in the fifth column of Fig. 16, even in complex scenes containing tree shading, mislabeled or buildings of different sizes and styles, MSBR-GNet can still extract building rooftop results with a more complete boundary shape than other models.

3) Evaluation With the Massachusetts Building Dataset: As shown in Table IX, the comparison results of extraction accuracy on the Massachusetts building dataset. MSBR-GNet achieved the best results in precision, F1-score, and IoU metrics, with F1-score and IoU reaching 87.15% and 83.44%, respectively. In terms of precision, the proposed method outperformed mask R-CNN, MS-CNN, MAP-Net, STT, and CBR-Net by 12.33%, 5.33%, 1.01%, 4.41%, and 4.96%, respectively. In terms of recall, MSBR-GNet was significantly higher than comparable methods, and MSBR-GNet was effective in extracting building rooftops in dense areas. The MAP-Net method had higher precision values, but lower recall values, indicating many missed detections.

The results of the boundary quality evaluation for building extraction in the Massachusetts dataset are shown in Table X. The HD_{95} values are generally high for all methods due to the presence of many complex buildings in the Massachusetts dataset, however MSBR-GNet still achieved a minimum value of 254.26, which indicates that our model is effective for complete and regular extraction of complex buildings.

A qualitative evaluation of the test results in the Massachusetts building dataset is shown in Fig. 17. Both mask R-CNN and MAP-Net show missed detection of large-scale and intricate buildings. The MS-CNN and CBR-Net also show a certain degree of distortion in their recognition results, compared to MSBR-GNet, which achieves regularized extraction of large-scale and intricate buildings. In addition, as shown in the red box in the fifth column of Fig. 17, MSBR-GNet still able to achieve a certain degree of completeness while achieving regularized extraction of dense building areas.

B. Ablation Experiments

1) Model Ablation: In this section, the ablation study will test the validity of two key components of the model. Without loss of generality, the ablation experiments and analysis are carried

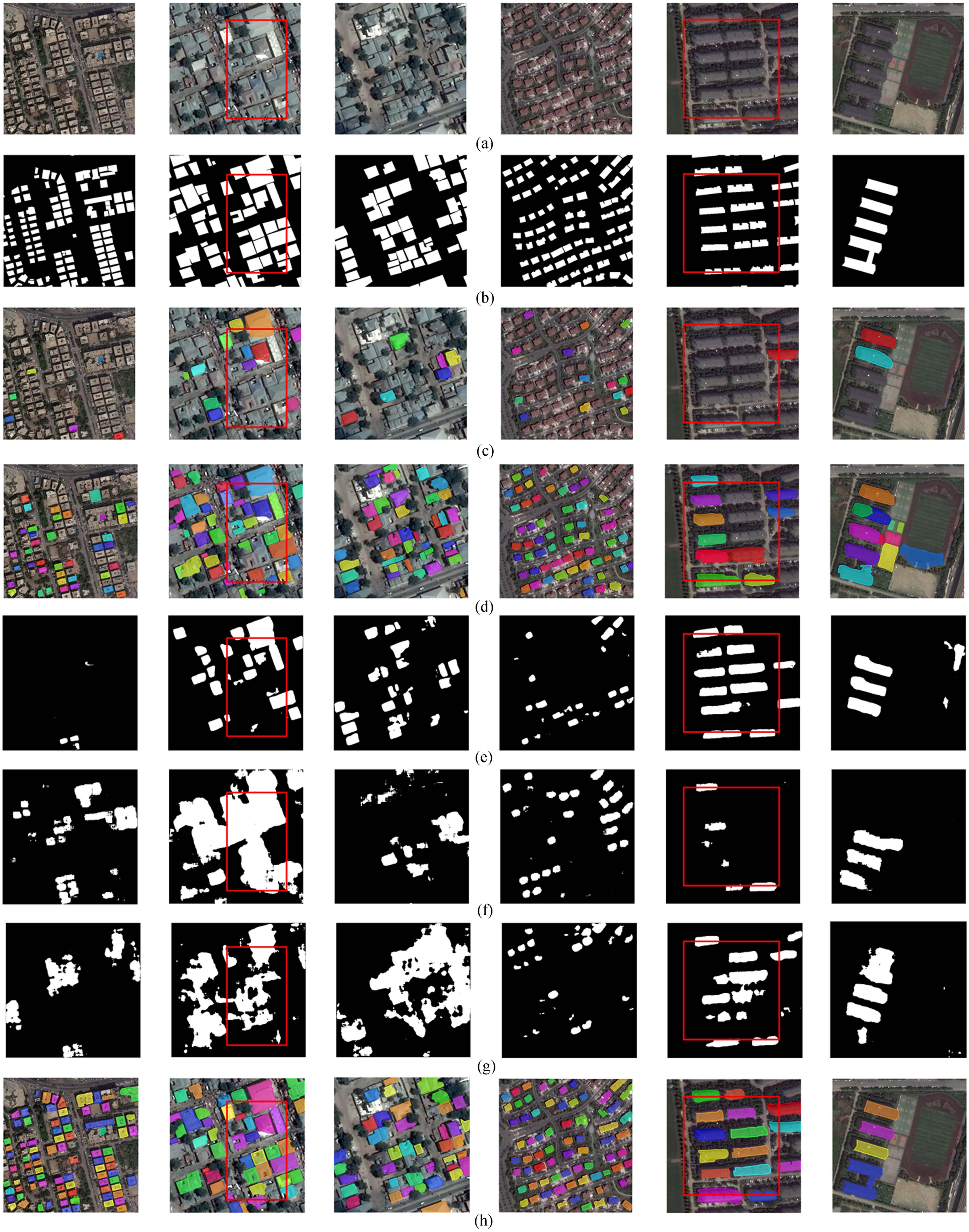


Fig. 16. Visualized results of building instance extraction using the WHU building dataset. (a) Original remote sensing imagery. (b) Ground truth. (c) Mask R-CNN. (d) MS-CNN. (e) MAP-Net. (f) STT. (g) CBR-Net. (h) Proposed model.

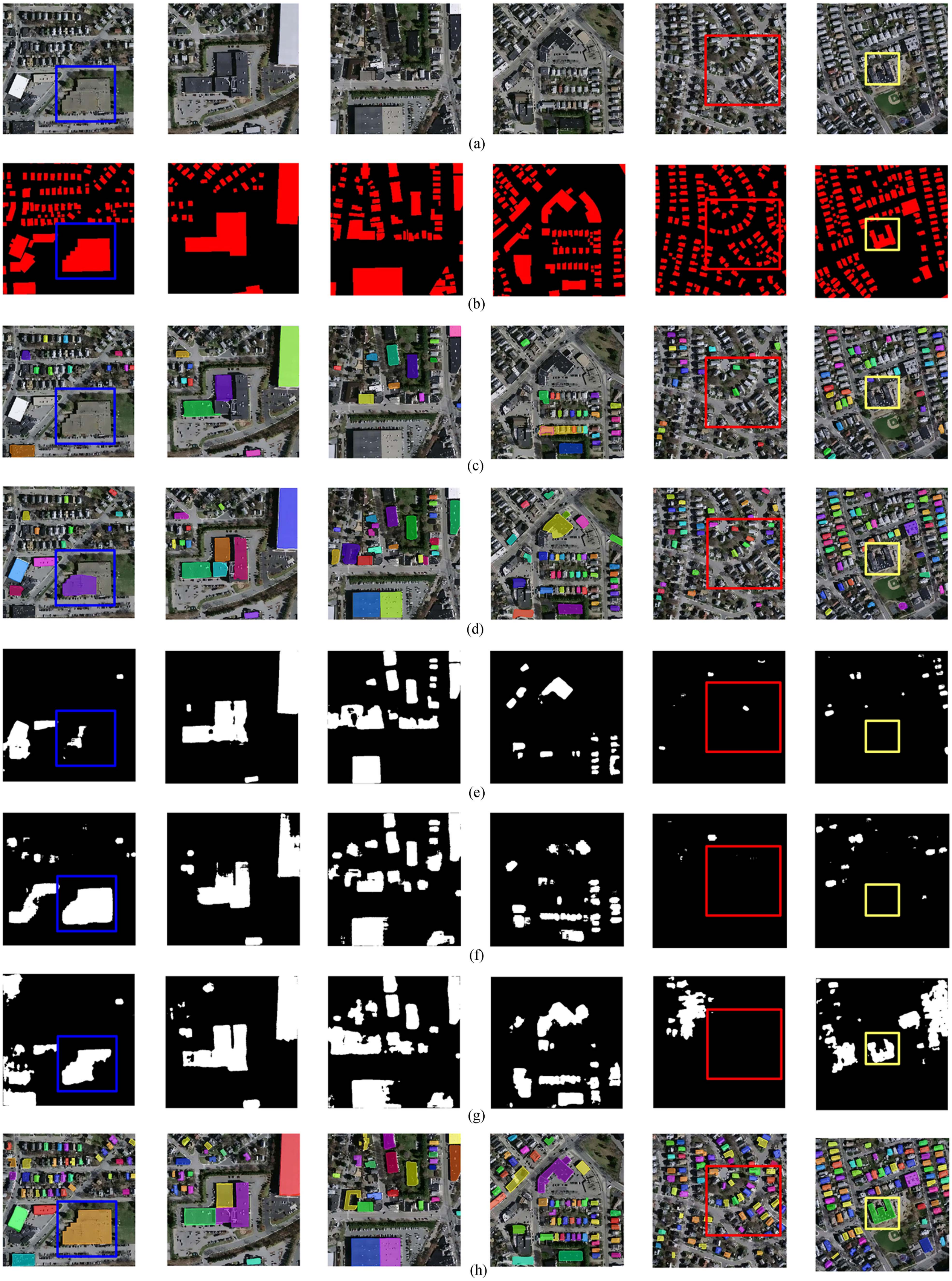
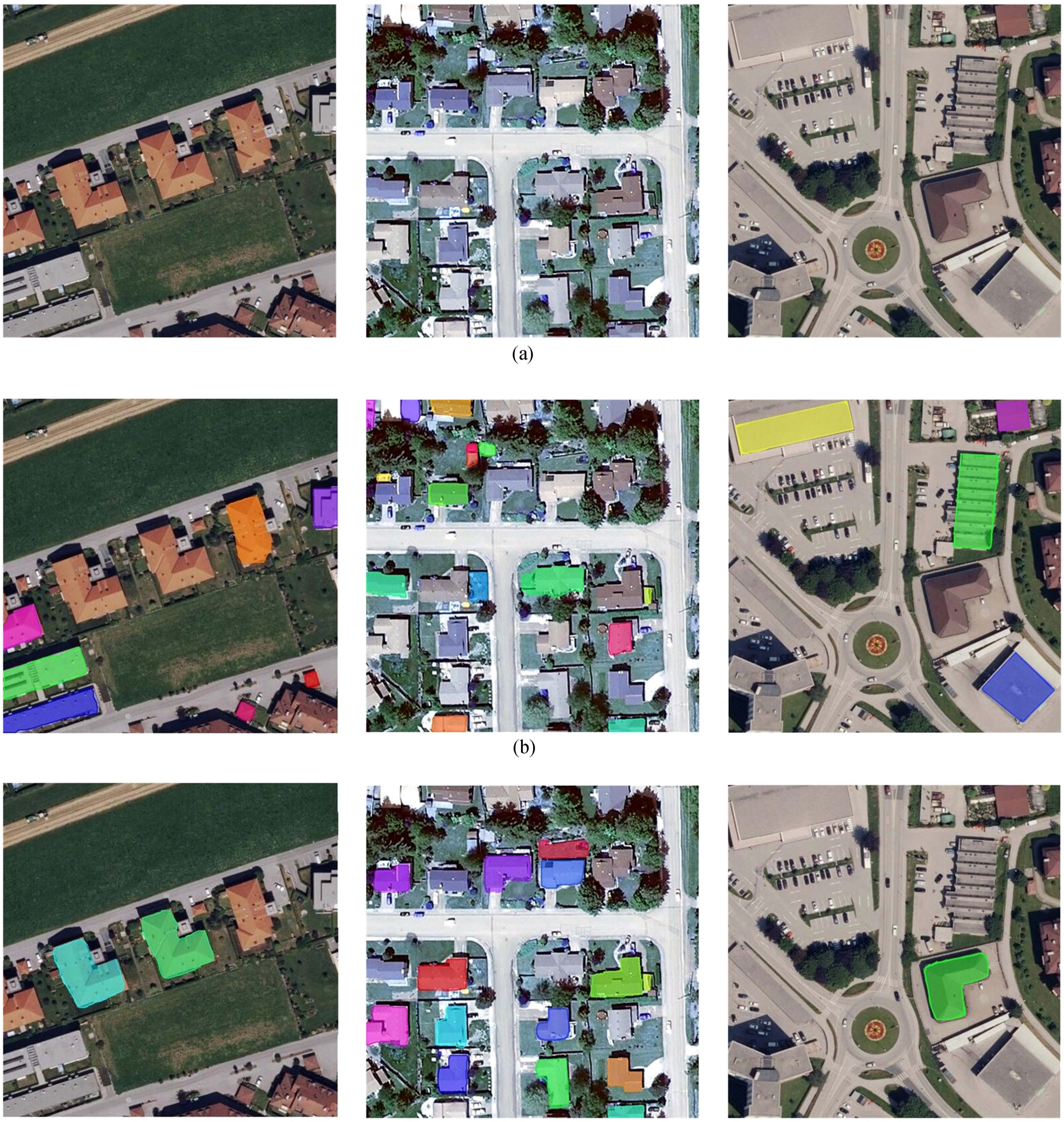


Fig. 17. Visualized results of building instance extraction using the Massachusetts building dataset. (a) Original remote sensing imagery. (b) Ground truth. (c) Mask R-CNN. (d) MS-CNN. (e) MAP-Net. (f) STT. (g) CBR-Net. (h) Proposed model.

TABLE VIII
QUANTITATIVE COMPARISON OF BOUNDARY QUALITY WITH SOTA METHODS ON WHU DATASET

Method	HD$_{95}$	SSIM	AD	CD	GCD	MDD
Mask R-CNN	321.65	0.6785	0.6254	0.649	0.7965	3.35°
MS-CNN	297.54	0.7562	0.6333	0.688	0.8142	3.59°
MAP-Net	190.24	0.9136	0.7859	0.705	0.7444	3.15°
STT	248.25	0.7933	0.5883	0.561	0.5256	4.28°
CBR-Net	283.33	0.7705	0.6014	0.543	0.5078	4.56°
MSBR-GNet	**185.46**	**0.9277**	**0.8043**	**0.6954**	**0.8324**	**2.03°**

Bold values indicate the best value for the corresponding indicator.


(a)

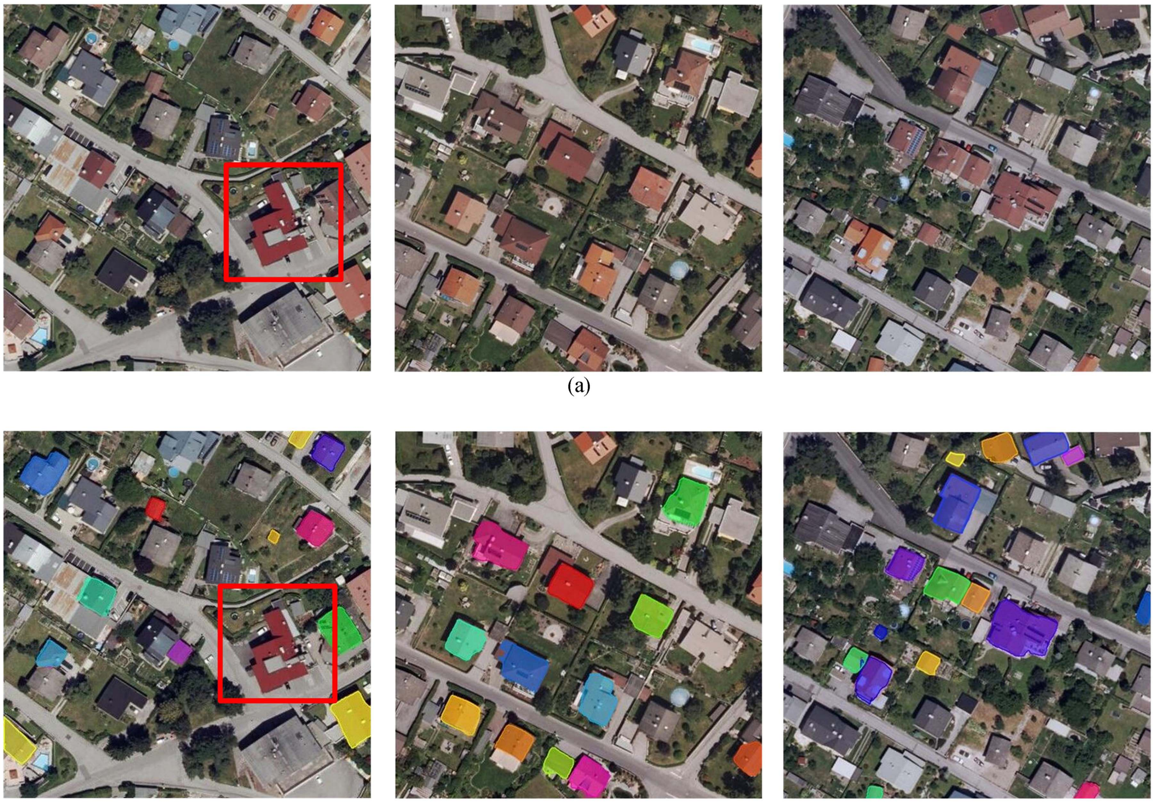
(b)

(c)

Fig. 18. Comparison of sample ablation results based on a subset of geometric features. (a) is the original image. (b) Test results for model training using only a subset of rectangular building rooftop samples. (c) Test results for model training using only a subset of L-shaped building rooftop samples. (a) Original image. (b) Only rectangle image. (c) Only L roof.

out in the INRIA dataset based on the analysis of the above-mentioned experimental results. The well-known mask R-CNN is also used as the baseline model, and then the components of each module are added gradually.

As shown in Table XI, in terms of precision, the fusion of the multiscale parameter residual blocks improved the baseline from 82.38% to 87.93%. In particular, the recall value improved from 78.09% to 86.25%. This percentage improvement indicates


(a)

(b)

Fig. 19. Comparison of sample ablation results based on a subset of spectral features. (a) is the original image. (b) Test results of model training using only a subset of red building rooftop samples. (a) Original image. (b) Only red.

TABLE IX
QUANTITATIVE COMPARISON OF EXTRACTION ACCURACY WITH SOTA METHODS ON MASSACHUSETTS DATASET

Method	Precision	Recall	F1-score	IoU
Mask R-CNN	0.7556	0.7368	0.7511	0.7152
MS-CNN	0.8256	0.7792	0.8017	0.7563
MAP-Net	0.8688	0.7559	0.8084	0.7233
STT	0.8348	0.762	0.7967	0.7364
CBR-Net	0.8293	0.852	0.8405	0.8029
MSBR-GNet	**0.8789**	**0.8643**	**0.8715**	**0.8344**

Bold values indicate the best value for the corresponding indicator.

that the introduction of a multiscale residual block is effective in achieving multiscale symbiosis recognition of building rooftops.

As shown in Table XII, the addition of BR loss increases the SSIM from 87.44% to 90.44% and the HD$_{95}$ value is reduced by 120.85, which means that BR loss helps to extract finer building rooftop contours. The majority of the extraction accuracy metrics exhibited a slight improvement compared to the baseline model results, but with a smaller magnitude. This observation further underscores the effectiveness of BR loss in mitigating irregularities during building rooftop extraction and achieving an optimal level of segmentation precision. Ultimately MSBR-GNet precision, F1-score and IoU all achieved optimal scores of 92.75%, 89.18%, and 85.84% respectively, and further improvements of 10.37%, 8.10%, and 11.31% over the baseline, while the boundary evaluation metrics also achieved optimal scores.

2) Sample Ablation: The construction of the self-labeled pedigree sample dataset BUCEA 2.0 was described in Experiments, where subsets of pedigree samples were separately constructed based on the geometric and spectral features of building rooftops in remote sensing imagery, thereby forming a comprehensive Tupu sample library. Based on the geometric and spectral features of the rooftops, we extracted 1–2 sample subsets corresponding to each of them to train the model schematically. The feasibility of the sample set organization form oriented toward the interpretability of rooftop features and the effectiveness of the sample set construction method are verified through ablation, in accordance with the intrinsic mechanism of pedigreed features.

As shown in Figs. 18 and 19, we extracted a subset of rectangular and L-shaped building rooftops and red building rooftops from the Tupu sample library for model training experiments, respectively. From the prediction results obtained, it can be seen that by training the sample subsets with different geometric and spectral features separately, the building rooftops with the required geometric and spectral features can be extracted separately from the remote sensing images, which verifies the feature interpretability of the construction of the Tupu sample library (set) in this article. At present, due to the limitation of the number of sample subsets, the building rooftops matching a certain spectral feature (e.g., the part in the first column of

TABLE X
QUANTITATIVE COMPARISON OF BOUNDARY QUALITY WITH SOTA METHODS ON MASSACHUSETTS DATASET

Method	HD_{95}	SSIM	AD	CD	GCD	MDD
Mask R-CNN	386.57	0.6825	0.7059	0.6012	0.8665	4.15°
MS-CNN	304.24	0.7591	0.7244	0.6225	0.8677	2.19°
MAP-Net	280.6	0.8154	0.734	0.6344	0.8852	2.25°
STT	310.53	0.6824	0.7535	0.6239	0.8757	3.55°
CBR-Net	298.72	0.8431	0.7765	0.6558	0.8898	3.82°
MSBR-GNet	**254.26**	**0.8524**	**0.794**	**0.685**	**0.897**	**1.53°**

Bold values indicate the best value for the corresponding indicator.

TABLE XI
QUANTITATIVE EVALUATION OF EXTRACTION ACCURACY IN ABLATION EXPERIMENTS

Methods	Precision (%)	Recall (%)	F1-score (%)	IoU (%)
Basic network	82.38	78.09	81.08	74.53
+Scale parameter	87.93 (+5.55)	**86.25** (+8.16)	87.08 (+6.00)	82.34 (+7.81)
+BR loss	83.56 (+1.18)	80.29 (+2.20)	81.89 (+0.81)	78.44 (+3.91)
MSBR-GNet	**92.75** (+10.37)	85.89 (+7.80)	**89.18** (+8.10)	**85.84** (+11.31)

Bold values indicate the best value for the corresponding indicator.

TABLE XII
QUANTITATIVE EVALUATION OF BOUNDARY QUALITY IN ABLATION EXPERIMENTS

Methods	AD (%)	CD (%)	GCD (%)	MDD	HD_{95}	SSIM (%)
Basic network	83.44	70.1	81.45	2.35°	321.13	87.44
+Scale parameter	83.56 (+0.12)	68.8 (–1.30)	82.32 (+0.87)	2.29° (–0.06°)	297.51 (–23.62)	88.33 (+0.89)
+BR loss	88.43 (+4.99)	77.56 (+7.46)	85.23 (+3.78)	1.11° (–1.24°)	200.28 (–120.85)	90.44 (+3.00)
MSBR-GNet	**90.65** (+7.21)	**80.80** (+10.7)	**85.64** (+4.16)	**1.03°** (–1.32°)	**180.59** (–140.54)	**92.45** (+5.01)

Bold values indicate the best value for the corresponding indicator.

red box in Fig. 19) are not marked in the sample due to their complex geometric features, and thus are not detected, which can be considered for optimization in the subsequent sample subset construction work.

VI. CONCLUSION

This article proposes a generative optimization model MSBR-GNet for regularized extraction of building rooftop boundary in aerial and satellite imagery, guided by interpretable statistical pattern in the boundary spatial and spectral domains. The method utilizes a generative network structure and incorporates an innovative boundary loss function into the generator's mask loss to constrain the regularized generation of the building roof mask boundary. The MSBR-GNet generative network structure can help improve the accuracy of the instance segmentation model by providing additional supervision, where the adversarial loss encourages the generator network to produce outputs closer to the ground-truth segmentation mask and generating more accurate segmentation masks. The boundary loss function can guide the generation of regularized mask boundary from both the spatial and spectral domains.

Experiments conducted on public datasets demonstrate the validity of the proposed MSBR-GNet. The module ablation study conducted in this article also shows that the boundary loss function in the method has good performance for building rooftop contours regularized generation, improving the accuracy of building extraction and performs well in multiscale symbiosis identification. In addition, this article constructs a set of sample library of building roof genealogical Tupu for the interpretability of geological laws, and verifies the feasibility and validity of the Tupu sample library construction method through sample set ablation experiments.

In our future research, we aim to enhance the accuracy of building extraction from aerial imagery by incorporating semantic constraints and will further expand our sample library based on geological laws, with a specific focus on addressing the inherent contradiction between geographical coupling of samples and the model's universal applicability.

ACKNOWLEDGMENT

The authors would like to sincerely thank the reviewers and associate editors for their helpful comments and suggestions on improving this manuscript, and also like to thank Ms. Chen Shuhua for her suggestions on the content of this manuscript.

References

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

[2] J. Liu, J. Zhang, F. Xu, Z. Huang, and Y. Li, "Adaptive algorithm for automated polygonal approximation of high spatial resolution remote sensing imagery segmentation contours," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1099–1106, Feb. 2014.

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

[4] S. Lorenz, P. Ghamisi, M. Kirsch, R. Jackisch, B. Rasti, and R. Gloaguen, "Feature extraction for hyperspectral mineral domain mapping: A test of conventional and innovative methods," *Remote Sens. Environ.*, vol. 252, 2021, Art. no. 112129. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425720305022

[5] X. Lin and J. Zhang, "Object-based morphological building index for building extraction from high resolution remote sensing imagery," *Acta Geodaetica et Cartographica Sinica*, vol. 46, no. 6, 2017, Art. no. 724. [Online]. Available: http://xb.chinasmp.com/EN/abstract/article_7013.shtml

[6] Y. Ding, F. Feng, J. Li, Y. Hu, and W. Cui, "Right-angle buildings extraction from high-resolution aerial image based on multi-stars constraint segmentation and regularization," *Acta Geodaetica et Cartographica Sinica*, vol. 47, no. 12, 2018, Art. no. 1630. [Online]. Available: http://xb.chinasmp.com/EN/abstract/article_7363.shtml

[7] Q. H. Le, K. Youcef-Toumi, D. Tsetserukou, and A. Jahanian, "Instance semantic segmentation benefits from generative adversarial networks," *Comput. Sci. - Comput. Vis. Pattern Recognit.*, 2020.

[8] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[9] J. Gong, Y. Xu, X. Hu, L. Jiang, and M. Zhang, "Status analysis and research of sample database for intelligent interpretation of remote sensing image," *Acta Geodaetica et Cartographica Sinica*, vol. 50, no. 8, 2021, Art. no. 1013. [Online]. Available: http://xb.chinasmp.com/EN/abstract/article_12575.shtml

[10] Y. Liu, J. Liu, X. Ning, and J. Li, "MS-CNN: Multiscale recognition of building rooftops from high spatial resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 1, pp. 270–298, 2022, doi: 10.1080/01431161.2021.2018146.

[11] S. Chen, T. Yue, and H. Li, "Studies on geo-informatic Tupu and its application," *Geographical Res.*, vol. 19, no. 4, 2000, Art. no. 337. [Online]. Available: https://www.dlyj.ac.cn/EN/abstract/article_11300.shtml

[12] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425709000625

[13] X. Lin, J. Zhang, Z. Liu, and J. Shen, "Semi-automatic extraction of ribbon roads from high resolution remotely sensed imagery by T-shaped template matching," in *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, vol. 7147. Bellingham, WA, USA: SPIE, 2008, doi: 10.1117/12.813220.

[14] E. A. Wentz, D. Nelson, A. Rahman, W. L. Stefanov, and S. S. Roy, "Expert system classification of urban land use/cover for Delhi, India," *Int. J. Remote Sens.*, vol. 29, pp. 4405–4427, 2008, doi: 10.1080/01431160801905497.

[15] X. Wang and X. Pin, "The research on the way of knowledge-based classification in vegetation extraction," *Bull. Surveying Mapping*, pp. 48–50, 2004.

[16] Y. Hou, "A rule-based land cover classification method for the Heihe River Basin," *Acta Geographica Sinica*, vol. 66, pp. 549–561, 2011.

[17] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 257–272, Jan. 2013.

[18] V. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 67, pp. 93–104, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271611001304

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[24] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4969–4978.

[25] K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, and Q. Tian, "Corner proposal network for anchor-free, two-stage object detection," in *Proc. Eur. Conf. Comput. Vis.*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 399–416.

[26] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," *Comput. Sci. - Comput. Vis. Pattern Recognit.*, 2021.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[28] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 21–37.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[30] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9156–9165.

[31] X. Chen, R. Girshick, K. He, and P. Dollar, "Tensormask: A foundation for dense object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2061–2069.

[32] Q. Zhao et al., "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 33, pp. 9259–9266. [Online]. Available: https://api.semanticscholar.org/CorpusID:53283428

[33] E. Xie et al., "PolarMask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12190–12199.

[34] J. Liu, M. Du, and Z. Mao, "Scale computation on high spatial resolution remotely sensed imagery multi-scale segmentation," *Int. J. Remote Sens.*, vol. 38, no. 18, pp. 5186–5214, 2017, doi: 10.1080/01431161.2017.1325536.

[35] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 1511–1518.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[38] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 73–80.

[39] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Neural Inf. Process. Syst.*, 2013, vol. 26.

[40] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017 pp. 5998–6008.

[41] G. Liasis and S. Stavrou, "Building extraction in satellite images using active contours and colour features," *Int. J. Remote Sens.*, vol. 37, no. 5, pp. 1127–1153, 2016, doi: 10.1080/01431161.2016.1148283.

[42] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 242–2424.

[43] Y. Jin, W. Xu, C. Zhang, X. Luo, and H. Jia, "Boundary-aware refined network for automatic building extraction in very high-resolution urban aerial images," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 692. [Online]. Available: https://www.mdpi.com/2072-4292/13/4/692

[44] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu, "Building footprint generation through convolutional neural networks with attraction field representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[45] 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. 2021.

[46] A. Kirillov, Y. Wu, K. He, and R. Girshick, “PointRend: Image segmentation as rendering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9796–9805.

[47] S. Wei, T. Zhang, S. Ji, M. Luo, and J. Gong, “BuildMapper: A fully learnable framework for vectorized building contour extraction,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 87–104, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271623000217

[48] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, “BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images,” *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 1050. [Online]. Available: https://www.mdpi.com/2072-4292/12/6/1050

[49] Y. Zhou et al., “BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[50] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone, “Adversarial shape learning for building extraction in VHR remote sensing images,” *IEEE Trans. Image Process.*, vol. 31, pp. 678–690, 2022.

[51] Q. Li, S. Zorzi, Y. Shi, F. Fraundorfer, and X. X. Zhu, “RegGAN: An end-to-end network for building footprint generation with boundary regularization,” *Remote Sens.*, vol. 14, no. 8, 2022, Art. no. 1835. [Online]. Available: https://www.mdpi.com/2072-4292/14/8/1835

[52] J. Gong, L. Huan, and X. Zheng, “Deep learning interpretability analysis methods in image interpretation,” *Acta Geodaetica et Cartographica Sinica*, vol. 51, no. 6, 2022, Art. no. 873

[53] 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.

[54] V. Mnih, “Machine learning for aerial image labeling,” 2013.

[55] J. Liu, Y. Liu, X. Ning, M. Wang, and X. Wang, “Bucea2.0 Tupu sample library of high-resolution remote sensing building,” 2023. [Online]. Available: https://www.researchgate.net/publication/369890582_BUCEA20_Tupu_Sample_Library_of_High-Resolution_Remote_Sensing_Building

[56] A. Bokhovkin and E. Burnaev, “Boundary loss for remote sensing imagery semantic segmentation,” in *Proc. Adv. Neural Netw.–ISNN*, H. Lu, H. Tang, and Z. Wang, 2019, pp. 388–401.

[57] K. Chen, Z. Zou, and Z. Shi, “Building extraction from remote sensing images with sparse token transformers,” *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4441. [Online]. Available: https://www.mdpi.com/2072-4292/13/21/4441

[58] H. Guo, B. Du, L. Zhang, and X. Su, “A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 240–252, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621002975

Liu Jianhua received the B.S. degree in geographical information systems from Northwestern University, Xi’an, China, and the Ph.D. degree in communication and information systems from Fuzhou University, Fuzhou, China, in 2003 and 2011, respectively.

He is the Director of “Mobile Geospatial Big Data Cloud Service Innovation Team,” which is one of Beijing undergraduate entrepreneurial groups. In 2013, he was selected as one of the youth teacher’s development projects of Beijing Municipal Education Commission. In 2016, he was selected as one of the outstanding youth Researcher’s programs of Beijing University of Civil Engineering and Architecture, Beijing, China. In 2018, he was supported by the outstanding youth teacher development plan of Beijing Municipal Education Commission. Since 1999, he has been engaged in teaching and scientific research of geographic information science (GIS) and remote sensing (RS) application technology for twenty years. His research interests include GIS and RS application technology. At present, the area of research involves GIS (mobile GIS, LBS, mobile GIS, and Internet of Things, GIS and smart cities, GIS+BIM +CIM building map and information technology, mobile phone indoor positioning and navigation), and RS (high spatial resolution remote sensing image analysis, recognition and understanding, object-based image analysis GEOBIA and deep learning, GIS and RS integration). Team website.https://www.dxkjs.com/

Ning Xiaohe received the B.E. degree in surveying and mapping engineering from Qinghai University, Xining, China, in 2019. She is currently working toward the M.E. degree in photogrammetry and remote sensing with the Beijing University of Civil Engineering and Architecture, Beijing, China.

Her research interests include deep learning, remote sensing mapping, urban remote sensing, spatial data mining, and remote sensing applications.

Wang Mengchen received the B.E. degree in remote sensing science and technology in 2020 from the Beijing University of Civil Engineering and Architecture, Beijing, China, where she is currently working toward the M.E. degree in architectural heritage and conservation.

Her research interests include deep learning, remote sensing mapping, urban remote sensing, spatial data mining, and remote sensing applications.

Wang Xinyu received the B.S. degree in geographic information science from Central South University, Changsha, China, in 2021. He is currently working toward the M.E. degree in cartography and geographical information engineering with the Beijing University of Civil Engineering and Architecture, Beijing, China.

His research interests include deep learning, remote sensing mapping, urban remote sensing, spatial data mining, and remote sensing applications.

Liu Yuan received the B.E. degree in surveying and mapping engineering from the Henan University of Engineering, Zhengzhou, China, in 2016, and the M.E. degree in photogrammetry and remote sensing from the Beijing University of Civil Engineering and Architecture, Beijing, China.

Her research interests include remote sensing image analysis and recognition, urban remote sensing, spatial data mining, and deep learning.

Chen Xiaoyou is currently working toward the B.E. degree in remote sensing science and technology with the Beijing University of Civil Engineering and Architecture, Beijing, China.

Her research interests include deep learning, remote sensing mapping, and remote sensing applications.

Zeng Shiyi is currently working toward the B.E. degree in remote sensing science and technology with the Beijing University of Civil Engineering and Architecture, Beijing, China.

Her research interests include deep learning, remote sensing mapping, and remote sensing applications