Arbitrary-Shaped Building Boundary-Aware Detection with Pixel Aggregation Network

Xin Jiang, Xinchang Zhang, Qinchuan Xin, Xu Xi, and Pengcheng Zhang

***Abstract*—Large-scale building extraction is an essential work in the field of remote sensing image analysis. The high-resolution image extraction methods based on deep learning have achieved state-of-the-art performance. However, most of the previous work has focused on region accuracy rather than boundary quality. Aiming at the low accuracy problems and incomplete boundary of the building extraction method, we propose a predictive optimization architecture, BAPANet. Notably, the architecture consists of an encoder-decoder network and residual refinement modules responsible for prediction and refinement. The objective function optimizes the network in the form of three levels (pixel, feature map, and patch) by fusing three loss functions: binary cross-entropy (BCE), intersection over-union (IoU) and structural similarity (SSIM). The five public datasets' experimental results show that the extraction method in this paper has high region accuracy, and the boundary of buildings is clear and complete.**

***Index Terms*—Building extraction, high-resolution, boundary quality, structural similarity.**

I. Introduction

WITH the development of spatial data acquisition technology and increasing resources of databases, geospatial data acquisition methods show the characteristics of multi-platform, multi-sensor, and multi-angle. The issue of acquiring data rapidly and extracting information more effectively from different sources for analysis becomes essential. As an essential component of a city, buildings are widely used in many applications such as urban planning [1], cartography [2], civil-military emergency response [3]. The realization of automatic, intelligent, reliable, and accurate building extraction has application value for the acquisition and update of primary geographic data because of the complexity of remote sensing imaging mechanism, spectrum, texture, and contour of buildings.

Traditionally, based on the experiential feature to express "what is a building" and creating a corresponding feature set for automatic recognition and extraction of buildings. The commonly used index including spectrum [5-6], edge [7], shape [8], texture [9], shadow [10], height [11] and semantics [12]. However, these features change significantly with seasons, lighting, atmospheric conditions, sensor quality, observation scale, building style, and environment. As a result, empirical features often only deal with specific data rather than being genuinely automated. In recent years, with the rapid development of computer hardware technology, deep learning methods based on convolutional neural networks (CNNs) have shown great application potential in object detection [13], image segmentation [14], text recognition [15] and other fields. The advantage of CNNs that it introduces the concept of end-to-end learning, which automatically extracts the most descriptive and remarkable features of the dataset. The neural network is quite suitable in the remote sensing image for its good generalization ability and gradually substitutes the traditional artificially designed method.

Recently, the image semantic segmentation algorithm based on Fully Convolutional Networks (FCNs) is widely used in building extraction. FCNs does not use the full connection layer to construct a set of predictive feature vectors after multi-layer convolution and pooling operations. Instead, the deconvolution operation is to obtain a result with the same resolution as the input image. It prevents spatial information from being lost during the propagation process in the image. The semantic segmentation algorithm is prominent in the extraction of buildings. For example, Maggiori [16] and Huang [17] used FCNs and its variants to extract buildings. The SRI-Net proposed by Liu [18] accurately detects large buildings that are easy to be missed while maintaining the global morphology identical and local details. Zhang et al. [19] proposed adaptive segmentation and developed a multistage classifier to improve buildings' extraction accuracy further. Zhao [20] constructed a multi-scale pyramid based on multi-scale images to mine the spatial information. Following the idea of FCNs semantic segmentation, Badrinarayanan [21] proposed the SegNet network, which fuses the Encoder-Decoder structure and the features of skip connection, the generality of the model allowed it to get more accurate feature maps.Audebert [22] designed a multi-kernel convolution layer to improve the original SegNet network model, given the multi-scale features of imagery.Xu [23] and Chen [24] extracted image features based on ResNet

Manuscript received May 10, 2020; revised July 16, 2020; accepted August 10, 2020. This research was funded by National Key R & D Program of China (Grant No. 2018YFB2100702), National Natural Science Foundation of China (Grant Nos. 41875122, 41431178, 41801351 and 41671453), Natural Science Foundation of Guangdong Province: 2016A030311016, National Key R & D Program of China (grant nos. 2017YFA0604300 and 2017YFA0604400), Research Institute of Henan Spatio-Temporal Big Data Industrial Technology: 2017DJA001, Hunan Botong Information Co.,ltd.: BTZH2018001. Western Talent (grant no. 2018XBYJRC004) and Guangdong Top Young Talents of Science and Technology (grant no. 2017TQ04Z359). (Corresponding authors: Xinchang Zhang.)

X. Jiang is with the Southern University of Science and Technology, Shenzhen, 518055, China (e-mail: jiangx3@mail.sustech.edu.cn)

X. Zhang is with the Guangzhou University, Guangzhou 510275, China (e-mail: eeszxc@mail.sysu.edu.cn)

Q. Xin are with the Sun Yat-sen University, Guangzhou 510275, China (e-mail: xinqinchuan@mail.sysu.edu.cn)

X. Xi is with the Suzhou University of Science and Technology, Suzhou, 215000, China (e-mail: xixu2016sysu@outlook.com)

P. Zhang is with the Guangzhou Urban planing & design survey research institute, Guangzhou 510275, China. (e-mail: 1260142133@qq.com)

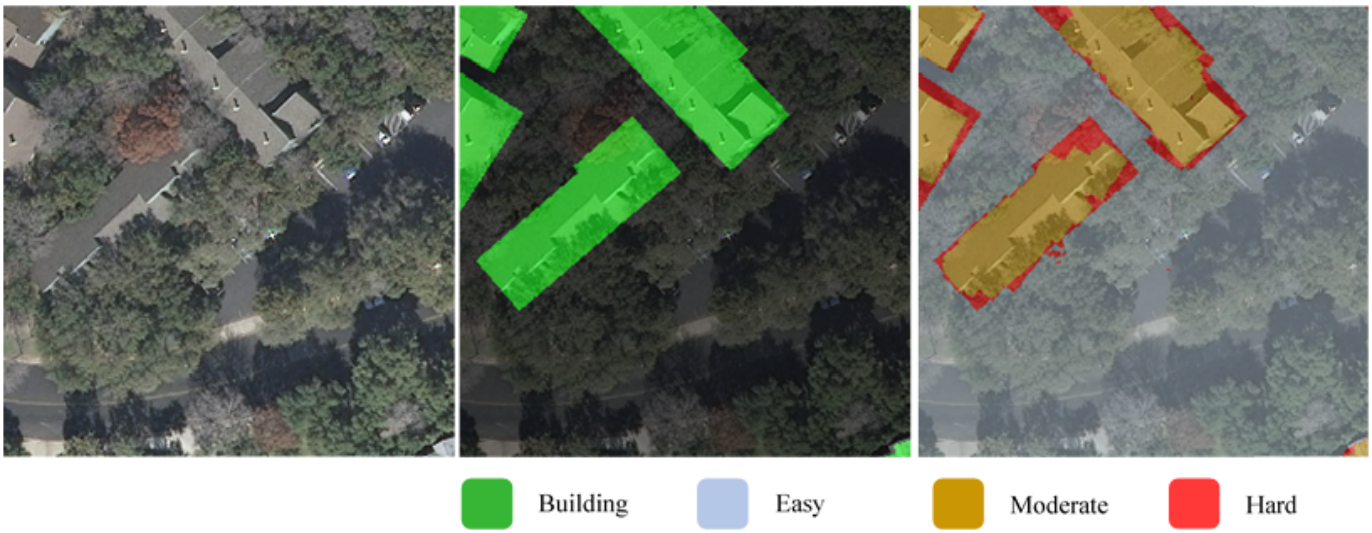


Fig.1. A "building" image from the Inria dataset (left) and its ground truth label image (middle). The difficulty level of pixels is visualized in the right image, where the pixels are divided into three sets, including the "easy", "moderate", and "extremely hard" sets.

[25] to improve segmentation accuracy. Huang [26] proposed a new post-processing framework for building detection to extend to complex environments. Yang [27] accurately established the United States building distribution map based on various training sets from different geographical regions. For the multi-source features, Sun [28], Maltezos [29], and Huang [30] proposed a deep neural network model that combines Lidar data with optical remote sensing data.

There are many complex fine-structure objects in the ISPRS-Potsdam dataset based on very high resolution. Volpi [31] proposed a labeling method for dense areas of buildings by learning the rules for sampling the original resolution from rough spatial features. Audebert [32] studied how deep convolutional networks fit the semantic labeling of very high-resolution and multi-scale remote sensing data. Liu [33] proposed an end-to-end self-cascaded convolutional neural network (ScasNet), which effectively marks the buildings from coarse to fine by correcting the potential residual. Marcos [34] proposed a CNN architecture of the Rotation Equivariant Vector Field Network (RotEqNet) based on the image's prior information for extracting feature types in any direction.

Although the above studies have achieved good segmentation accuracy, these methods may excessively compute the pixels that distinguish the boundaries of buildings and resulting in the misclassification of the edge pixel points that should clear and continuous boundaries, which makes the extraction results exist such issues as smooth edges and loss of information. In the classical semantic segmentation network, such as FCNs, Deeplab, the CNN generally downsamples the input image 16 times and then tries to upsample it back. In more detail, for Deeplabv3+, the model ends up being a 4x bilinear interpolation upsampling, which is very unfavorable for the prediction of the object edges. The edge prediction situation is not ideal for the image segmentation task mentioned in many previous works. For example, "Not All Pixels Are Equal: difficult-aware Semantic Segmentation via Deep Layer Cascade" [35] has made detailed statistics on semantic segmentation, in which the pixels are easy to lead to misjudged in classifying, the edge of the object as shown in red in Fig.1. To improve the boundary recognition ability of different categories, Chai [36] replaces the pixel-level optimization method with a distance map to get sufficient spatial context information. The result of the building edge is smoother than the post-processing using a conditional random field (CRF).

Similarly, Marmanis [37] proposed an end-to-end deep convolution neural network to improve boundary recognition for different semantic categories. Yuan [38] designed a deep convolutional network with a simple architecture that fuses multilayer activation function for pixel-by-pixel prediction. To enhance the building's expression ability, the author finally introduces the output result of the distance function. Li [39] proposed a novel two-step method to improve extracting buildings from remote sensing images. First, an improved model based on CRF is used to reduce edge misclassification and then further make full use of the building's saliency features to improve edge information expression. Although semantic segmentation models usually use loss functions to optimize network parameters, such as cross-entropy and intersection over-union, they are not sensitive enough to boundary misalignment. Even if the extracting boundary deviates from the valid 5-10 pixels, it will not significantly affect the value of the above loss function and common indicators of evaluation. In order to solve the problems mentioned above, we propose a corresponding detection model—BAPANet, a boundary-aware pixel aggregation network. The work-flow is displayed in Fig. 4. It dramatically improves the recognition ability of arbitrarily shaped buildings and optimizes the boundary quality. The contributions of this paper include:

(1) a new building shape recognition model based on boundary perception: BAPANet consists of an encoder-decoder and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3017934, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

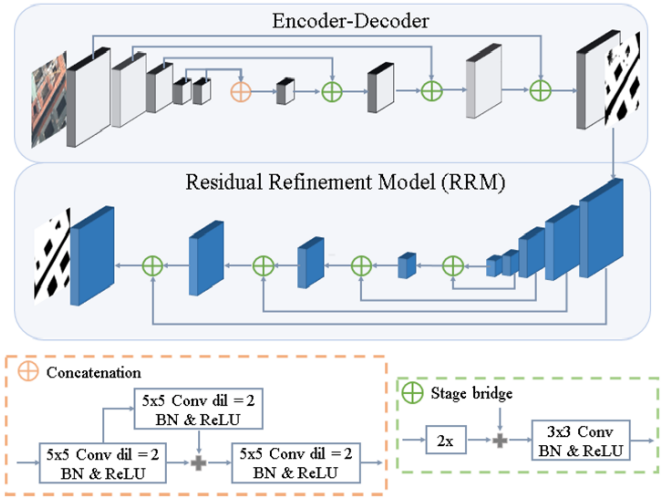


Fig.2. Overall architecture of BAPANet.

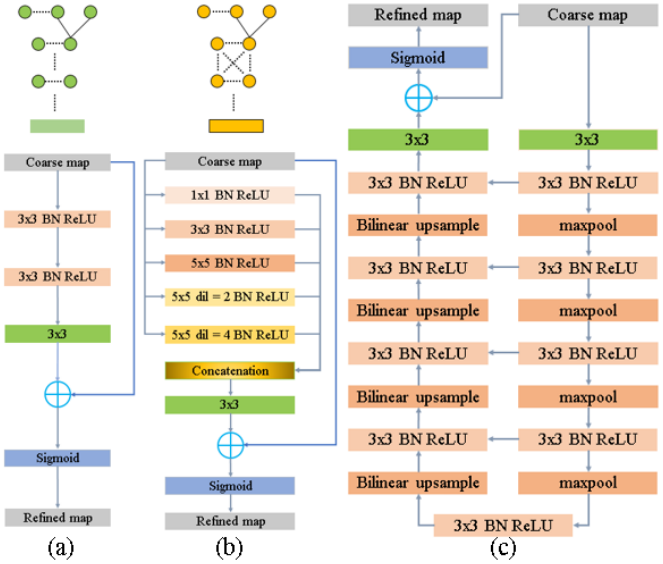


Fig.3. Comparison of different optimization modules. (a) Local boundary refinement module (LRM); (b) Multi-scale refinement module (MSRM); (c) We design the architecture of Encoder-Decoder residual refinement module (EDRRM).

a residual optimization module.

(2) data augmentation: image morphology transformation is uesd to improve the robustness and generalization of the deep learning model.

(3) a new objective optimization function: the hybrid loss function consists of binary cross-entropy (BCE), intersection over-union (IoU), and structural similarity (SSIM) loss, which supervises the training process of object detection on three levels: pixel level, feature map level, patch level.

II. METHODOLOGY

A. Overview of Network Architecture

We designed the building object recognition module—Encoder-Decoder network, which combines low-level detail and high-level global information. The encoder is composed of the first convolution layer from ResNet-34 and the basic residual module. Fig. 2 shows the overall architecture of BAPANet, which can be divided into two parts: the backbone network and a feature optimization module. We use the lightweight ResNet-34 as the backbone network to reduce the computational burden of the model and improve the efficiency. Merely using such a shallow backbone network does not have enough receptive field and cannot extract robust features. BAPANet adds a bridge structure between encoder and decoder of the backbone network to enhance the ability of feature expression. It is composed of three convolutional layers, and each convolutional layer consists of 512 hole convolutions, followed by a BN layer and a ReLU activation function layer. Each bridge's input is composed of the previous bridge and the up-sampling feature layer output by the corresponding encoder. The bridge has the following advantages: the module integrates features of different scales well, and the perceptive field of the features will increase.

B. Refine Module

The residual optimization module (Fig.3a) was first proposed for boundary optimization [40] based on local information. Islam et al. [41] and Deng et al. [42] optimize feature maps on multiple scales due to the relatively small area of the classical receptive field. Wang et al. [43] used the pyramid pooling module in [44] to piece together the pyramid pooling features of three scales. To avoid loss of detailed information due to the pooling operation, a multi-scale refinement module (MSRF, Fig.3b) uses hole convolution to obtain multi-level information. However, these modules are relatively shallow, and it is challenging to obtain higher-level information. To optimize the region's problem and boundary defects in the feature map, we designed a new residual optimization module. The residual refinement module (RRM) utilizes the encoder-decoder architecture-EDRRM (Fig.3c). The architecture is similar to the central network architecture, including encoder, bridge, and decoder. In this way, features of different depths fuse to combine low-level and high-level semantic information. We use the max-pooling layer in the encoder and then use bilinear interpolation in the decoder. The final output of our model is a feature map of the RRM module.

C. Data Augmentation Module

The fine information obtained from high-resolution imagery is better applied to image analysis and interpretation, bringing new challenges to image segmentation technology. With the deepening of the neural network, the optimization parameters will increase, which will easily lead to overfitting. Overfitting means that the neural network highly fits the distribution of training data, and it lacks generalization ability. There are many reasons for the over-fitting of the training network, and the most direct reasons are the small number of datasets and poor quality. Later Inception networks [45], VGG [46], and ResNet all used Scale Jittering, a scale-to-length-to-width enhancement transformation method. These

state-of-the-art studies have shown that data augmentation plays a crucial role in the final recognition performance and extensive deep network's generalization ability. Therefore, to avoid overfitting and the amount of data is small, data augmentation is necessary.Compared with original image recognition, classification, and segmentation datasets, the existing remote sensing image datasets are usually small in scale, and it is not elementary to train an excellent semantic segmentation model directly.Appropriate data augmentation operations such as rotation, scaling, and scale transformation improve the model's training accuracy and enhance the generalization ability of the model. We use image morphological transformation to increase the number of samples in the dataset and increase the dataset's diversity, mainly includes the following methods.

1) Random folding: include three methods of horizontal, vertical and diagonal.

2) Random scaling: image random scaling at most 10%.

3) Random offset: the image is randomly offset by up to 10%.

4) Random stretch: the image along the vertical or horizontal direction randomly pull up to 10%.

After the above four transformations, the 256*256 part of the image center is intercepted, insufficient to add 0.

D. Objective Function

The problem of scale impacts is ubiquitous, and the objects of buildings of different sizes have different characteristics in the application of remote sensing due to the complexity of space features and scale dependence. The encoder layer of the CNN convolves and pools the original image to obtain feature maps of different sizes. The shallow network pays more attention to details, the high-level network pays more attention to semantic information, and the high-level semantic information accurately detects a target, so we use the feature map on the last convolutional layer to make predictions. The method exists in most deep networks, such as UNet, PSPNet, BiSeNet, which use the features of the last layer of the deep network for image segmentation. The advantage of this method is that it is fast and requires less memory. Its disadvantage is that we only focus on the features of the last layer in the deep network and ignore other layers' features. The detailed information improves the accuracy of segmentation to a certain extent. The design idea of deep-supervised encoding and decoding uses both low-level features and high-level features to make predictions at different layers at the same time. The remote sensing images may have several different sizes, to distinguish the different goals may require different features. For simple objects, we need shallow features to detect it. For complex objects, we need to use sophisticated features to identify it. The whole process is to first perform a deep convolution on the original image, and then make predictions on different feature layers. In the backbone network's decoder phase, six feature maps with different resolutions are output, while RRM only outputs the feature map of the last layer. The feature maps generated in seven different stages are added to the loss function for calculation simultaneously. This multi-layer and multi-loss design method help the network better converge on the one hand, and it will enable the network to pay attention to the significance map of different scales to obtain more robust semantic information. The objective function is defined as the sum of all outputs:

$$L_{total} = \sum_{m=1}^{M} a_m l^m$$

Among them, $l^m$is the loss of the m-th feature map, M is the number of output feature maps, and $a_m$is the weight of each layer's output loss. The target detection model is supervised by seven outputs, i.e., M = 7, which includes the outputs of six backbone networks and an RRM optimization module's output. We define l^m as the mixed loss function:

$$l^m = l^m_{bce} + l^m_{iou} + l^m_{ssim}$$

The l_{bce} loss checks each pixel and compares the prediction result of each pixel category with the label. The loss of the entire image is the average of the loss of each pixel. The pixel-level cross-entropy loss function is defined as follows:

$$l_{bce}(x) = -(ylogf(x) + (1-y)log(1-f(x)))$$

l_{iou} is the commonly used index in semantic segmentation. It is not only used to determine positive samples and negative samples but also has scale invariance. The l_{iou}at the feature map level is defined as follows:

$$l_{iou}(x) = -\frac{1}{C}\sum_{c=1}^{C} \frac{\sum\limits_{pixels} y_{true}y_{false}}{\sum\limits_{pixels}(y_{true}+y_{false}-y_{true}y_{false})}$$

C is the number of categories, BCE coefficient is a judgment index of segmentation effect, and its formula is equivalent to the intersection ratio of predicted result area and ground truth area, so it calculates the loss function by taking all pixels of a category as a whole. Besides, the IoU coefficient directly takes the segmentation effect evaluation index as a loss function to supervise the network and ignores many background pixels when calculating the intersection ratio, thus solving unbalanced positive and negative samples problems. If the above two kinds of loss functions are a kind of area matching metric to supervise the network learning target, we also use a boundary matching metric to supervise boundary loss. We use structural similarity to evaluate the predicted and real boundary pixels. From the perspective of image composition, the structural information is defined as the property of object structure independent of brightness and contrast, and models distortion as a combination of three different factors: brightness, contrast, and structure. The mean value was used as an estimate of brightness, the standard deviation as an estimate of contrast, and the covariance as a structural similarity measure. Structural similarity index:

$$SSIM(x) = -\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$$

SSIM is used to learn the structured information between the target and the ground truth and evaluate the picture quality. In simple terms, to calculate the structural similarity of the two images, we need a local window (N * N size), calculate the structural similarity loss in the window, slide in pixels, and finally take the average of structural similarity loss of all

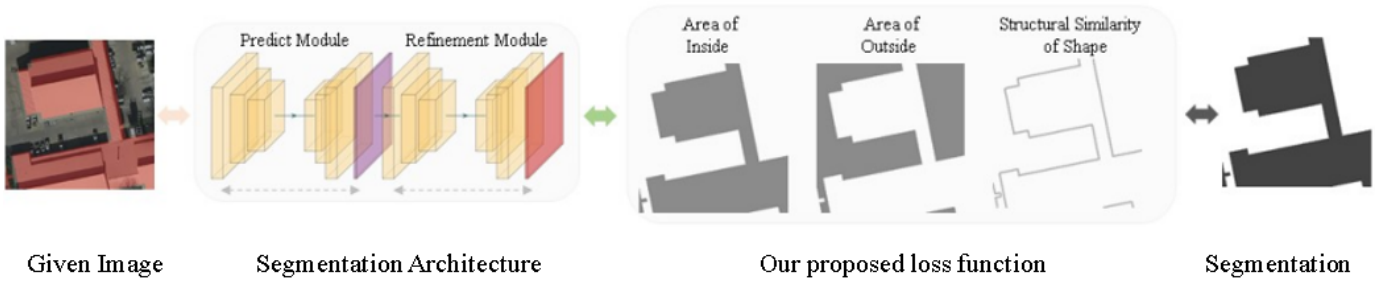


Fig.4. Overview of our proposed method which takes the object's area and shape of the boundaries into account during training.

windows. The specific calculation method is to represent the corresponding pixel points of the two pictures as x and y, where x = $\{x_i : i = 1, ..., N^2\}$ and y = $\{y_i : i = 1, ..., N^2\}$; μ_x and μ_y, σ_x and σ_y are the mean and variance of x and y, respectively, and σ_{xy} is the covariance of x and y. c_1=0.012 and c_2=0.032 to avoid a 0 denominator. The key to SSIM loss acting on patch-level is that it focuses on the boundary. In order to obtain high-quality region segmentation and clear boundary,the structural similarity loss function is introduced Structural similarity loss:

$$l_{ssim}(x) = 1 - SSIM$$

The l_{bce} predicts each pixel as an independent sample, while the l_{iou} looks at the final predicted output in a more complete. The l_{ssim} is a patch-level measure that considers the adjacent local region of each pixel point. The loss of the boundary will be higher than the weight assigned to the target's interior or elsewhere, and it will output clear object boundaries to solve the boundary blur. BAPANet simultaneously considered three loss functions l_{bce}, l_{iou},and l_{ssim} optimized the network in three levels (pixel, feature map, and patch) to segment the target area effectively.

III. Results

The subsequent GL-DenseUNet [47], DenseASPP [48], and Res2Net [49] all achieved high extraction accuracy in the field of image segmentation. However, with the increase of network depth and feature dimension, these methods may over-compute the pixels that distinguish the building's boundaries, resulting in the misclassification of edge pixels that should have clear and continuous boundaries. In this paper, the structural similarity loss function is introduced to make full use of buildings' edge characteristics and reduce the influence of the inability to extract the edge information of buildings in complex scenes, focusing on solving the problem of blurred boundaries while maintaining the region accuracy.

The dataset of this experiment comes from the institute national de recherche en Informatique et Automatique [50] (Inria), Mnih [51], and ISPRS-Potsdam [52]. This dataset contains remote sensing images from residential areas in different cities in the United States, Austria, and Germany, labeled as buildings and non-buildings. The Inria aerial image dataset released in 2018 has a spatial resolution of 0.3m and contains 180 images covering five cities, with 36 high-resolution remote sensing images. Images from Austin, Chicago, and Vienna were selected, among which 31 images were used for training, and five images were used for testing. The pixels of each image are 5000 × 5000, and the coverage area is about 2.25 km^2. The Massachusetts dataset contains 137 training images and 10 test images with three red, green, and blue bands, all of which are 1500 pixels long and wide, with a spatial resolution of 1 m, covering the surface area of about 340 km^2. The Potsdam dataset contains 38 patches, each consisting of a true orthophoto. We select three bands (red, green, blue) from the five channels (RGB+NIR+DSM) to experiment and train a neural network with adam optimizer.

Considering the computer performance, we take 128 pixels as the step size, crop the test image to 256 × 256 pixels, and remove the images without buildings to obtain 8899 training samples and 5332 test samples. The specific information is shown in Table I.

Table I
INFORMATION ON ALL DATASET IMAGES FOR FOUR CITIES

Location	Resolution m	Area km^2	Building quantity	Percent -age(%)
Austin	0.3	81	20449	14.93
Chicago	0.3	81	39673	24.57
Vienna	0.3	81	6459	46.62
Massachusetts	1.0	330	209907	12.03
Postdam	0.05	3.42	1770	26.28

This experiment is based on the Pytorch deep neural network framework, using a single 11GB GTX 1080Ti graphics card to complete the model training. After the model training, the model's prediction results in five different regions on the test set are given. To verify the effectiveness of the proposed method, the results are compared with the four classic neural network models of SegNet, FRRN-B, FC-DenseNet, and Deeplabv3+. These models correspond to the most classical network structure at present. Among them, SegNet uses VGG16, FRRN-B uses a dual-stream structure to combine multi-scale context information with pixel-level precision. FC-DenseNet corresponds to a dense connection mode, which is a network structure related to multi-branch structure. Deeplabv3+ uses ResNet as a feature extractor and uses atrous spatial pyramid pooling technology to form a faster and more powerful Encoder-Decoder network for semantic segmentation.Fig.5 shows the results obtained by using five

Table II
REGION ACCURACY EVALUATION RESULTS

Model	Austin		Chicago		Massachusetts		Vienna		ISPRS-Potsdam	
	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1
SegNet	71.79	88.51	73.61	85.75	77.33	91.74	57.53	72.68	82.05	90.19
FRRN-B	73.98	89.95	73.02	85.29	79.07	92.69	60.00	74.92	**84.34**	**91.57**
FC-DenseNet	75.18	90.46	75.30	86.76	80.11	92.99	62.12	76.72	78.65	88.45
Deeplabv3+	71.84	88.73	71.25	84.24	73.70	89.87	66.15	80.26	74.36	85.61
BAPANet	**78.05**	**92.00**	**83.17**	**91.72**	**82.54**	**94.36**	**86.35**	**93.86**	83.05	90.93

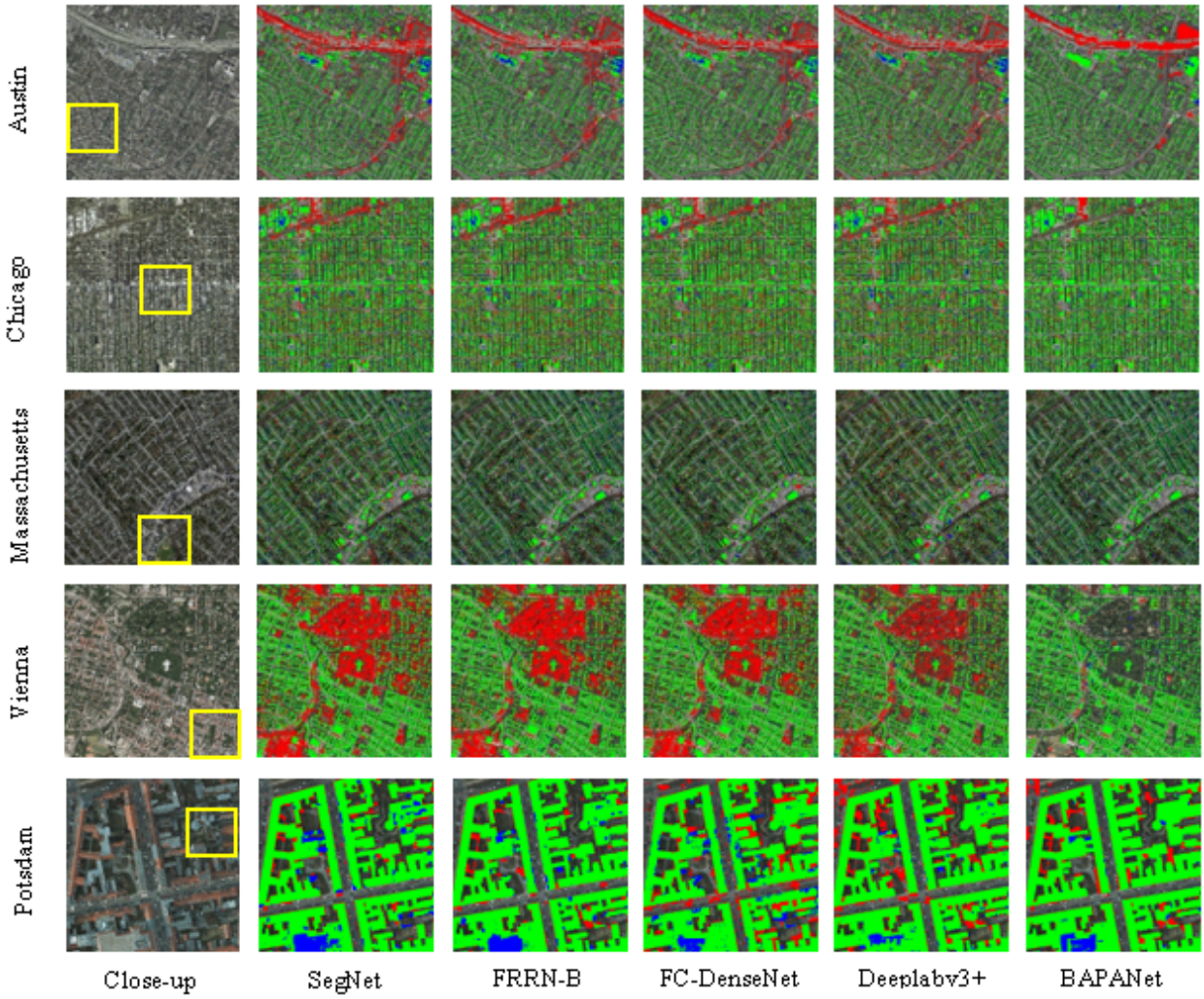


Fig.5. Images of the original true color composite image are displayed and compared the prediction results in five regions under different deep learning methods. The false positive (FP), true positive (TP), and false negative (FN) are marked in red, green, and blue, respectively. The yellow rectangles in the original images are enlarged for close-up inspection in Fig. 6.

methods to extract buildings in the image. The buildings in Austin are all small residential buildings, and all networks are well classified, among which FC-DenseNet and BAPANet have higher accuracy. However, the five models have a higher false detection rate for more full roads. Fig.6 shows the local extraction results of the building. It is seen from the figure that most buildings in Chicago are rectangular and regular, but the buildings have more shadows.BAPANet has done the best in the integrity of buildings, while the other four models have identified partial shadows, the integrity of buildings is reduced. All the five models extract the Massachusetts buildings with fewer building types and more straightforward structures; however, the types of features and structure in the Vienna area are complex; there are many irregular buildings. In the very high-resolution ISPRS-Potsdam 2D semantic markup dataset, we found that the FRRN achieves the best building extraction results, and our proposed method is suboptimal. Although FC-DenseNet based on dense connection structure performed second and achieved better segmentation results in Austin, Chicago, and Massachusetts, it performed poorly in finer ISPRS-Potsdam. Compared with the other four methods,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3017934, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

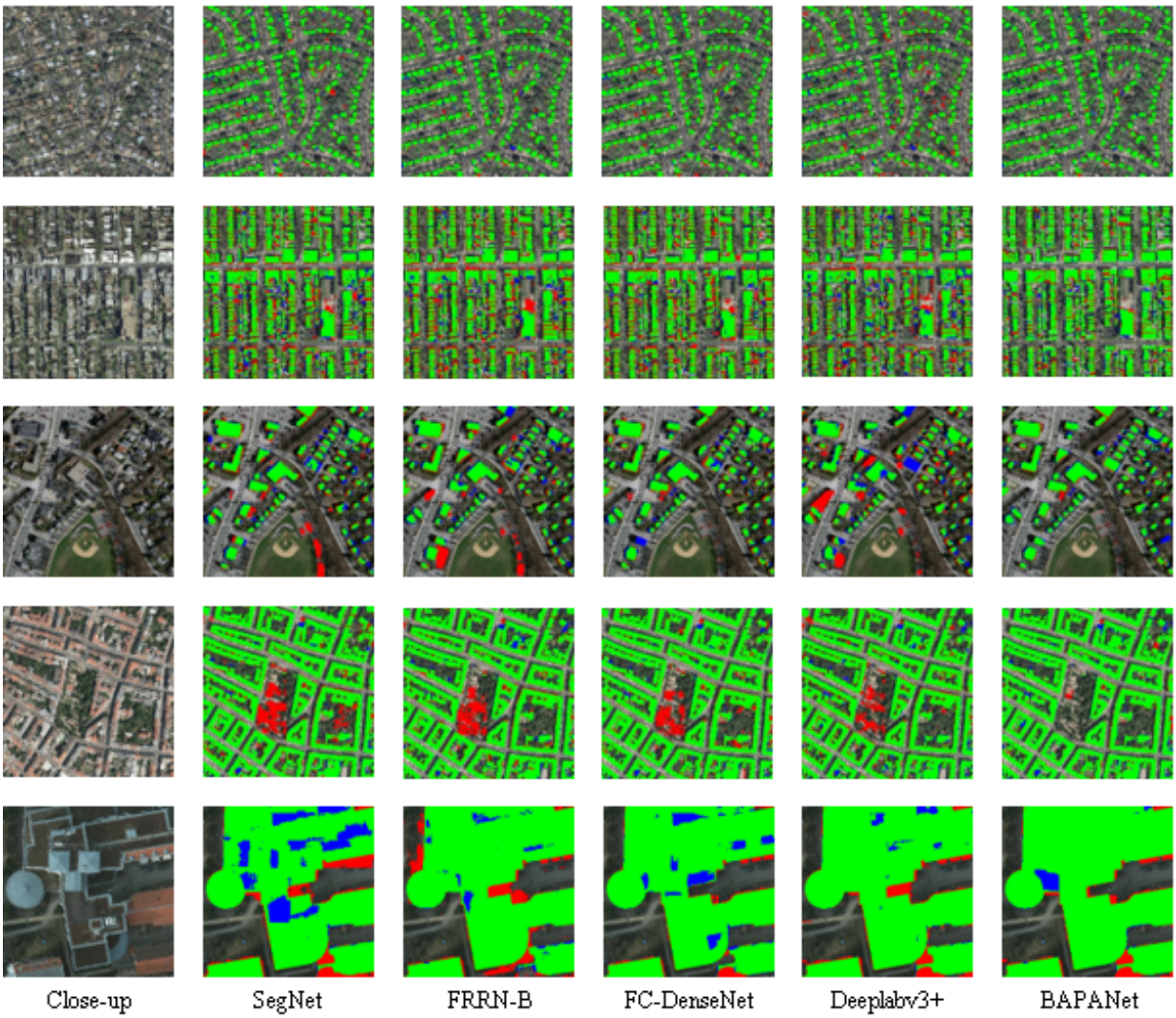


Fig. 6. Local extraction results. A close-up view of the original true-color composite image and classification results is displayed across five regions. The images are the subset from the yellow rectangles marked in Fig. 4. False negative (FN), true positive (TP) and false positive (FP) were marked as blue, green, and red, respectively.

BAPANet can better distinguish buildings and backgrounds in densely populated areas. While reducing the false detection rate, the edge details of buildings, and the contours of large irregular buildings can be better extracted.

To evaluate the segmentation accuracy of the building on datasets, four evaluation indexes, including accuracy, intersection over-union, F-measure, and mean absolute error (MAE), were used to evaluate the accuracy of five network models of SegNet, FRRN-B, FC-DenseNet, Deeplabv3+, and BAPANet. The first three indexes are used to evaluate the accuracy of region segmentation, and MAE is used to evaluate the performance of different loss functions on extracting building boundaries. MAE [53] represents the average absolute difference per pixel between the predicted probability map and its ground truth mask. Given a prediction result map, MAE is defined as:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{h=1}^{H} |R(j,i) - P(j,i)|$$

Where R and P are ground truth and its probability image, respectively, W and H represent the image's width and height, and (j, i) denotes the pixel coordinates. The results are shown in Table 2. It can be seen that compared with the classic image segmentation model, BAPANet has been improved on all three indicators. Although the buildings in these three regions are relatively small and simple in structure, Deeplabv3+ has the worst performance. The model ends up being a 4x bilinear interpolation upsampling, which is very unfavorable to predicting the edge of the object, leading to the severe absence of small buildings. On the contrary, the performance of SegNet with a simple structure is more stable. Deeplabv3+ got a suboptimal result for the Vienna region with many irregular buildings in a variety of image features and intricate structures. In contrast, SegNet, with a simple structure, performed the worst. Synthesizing the evaluation results of the five regions, we find that the method in this paper can extract the buildings in Austin, Chicago, and Massachusetts with low-resolution

structures and low-resolution buildings and the large and irregular buildings in Vienna. Although the proposed network for high-resolution images with more complex structures does not have the best overall accuracy, its ability to identify buildings is better than other models, as shown in Fig.5.

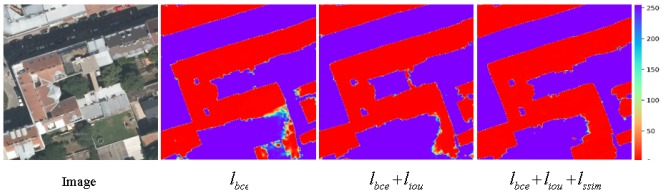


Fig.7. Illustration of the impact of different loss functions. Red and blue indicate the possibility of pixels with buildings and backgrounds, respectively.

We illustrate the effects of three different loss functions in Fig.6. These heat maps show the loss of each pixel. These three cols correspond to l_{bce}, $l_{bce}+l_{iou}$,and $l_{bce}+l_{iou}+l_{ssim}$, respectively. l_{bce} is pixel-by-pixel, which helps convergence of all pixels. l_{ssim} is a patch level measure that takes into account the local neighborhood of each pixel. We use the l_{bce} to keep it relatively smooth for all pixels while using the to focus more on foreground targets, while SSIM is used to constrain the loss near the boundary to predict the structure of the original image better. To quantitatively evaluate the edge significance of the segmentation object, a series of F-measures are generated by binarization of each prediction result with different threshold values. Fig. 7 shows the likelihood of the predicted outcome. We can get the result of building segmentation by binarization.

The MAE results in Table IV show that our proposed SSIM-based hybrid loss function will improve performance, especially for boundary quality. Also, combining with the results in Fig.8, data enhancement can improve the accuracy of building regional segmentation and extract the edge information of buildings accurately. In summary, although it can be seen that each network achieve better recognition results, the BAPANet model proposed in this paper achieves the most accurate extraction for both Vienna region with complex feature types and Massachusetts region with a small area, which preserves the clear boundary and integrity of the building, and solves the problem that the edge contour of the extraction result is too smooth.

TABLE III
COMPARISON OF THE EFFICIENCY OF DIFFERENT NETWORK MODELS

Model	Inference (ms)	Model size (MB)	FPS
SegNet	82	419	43
FRRN-B	120	297	27
FC-DenseNet	173	106	17
Deeplabv3+	53	235	73
BAPANet	263	332	11

In addition, deep learning methods are suitable for building extraction tasks with a large amount of accurately labelled data. It should be pointed out that this method tends to overfit when the amount of training data is small, which is also the disadvantage of data-driven deep learning methods. It is crucial to study small and efficient CNN models in these scenarios, and models as large and complex as Deeplabv3+ and Res2Net are challenging to apply directly. First, the models are too large and face the problem of insufficient memory. Second, some scenarios require low latency or fast response time. As shown in Fig. 8, with the model's training, the overall accuracy of BAPANet in these five datasets is significantly higher than that of the other four methods. However, the results in Table III show that the proposed network is at the lowest level, both in terms of the time used to train the model and the prediction's timeliness. Therefore, it is an important research direction to reduce the size of the model and improve the speed and low latency of the model.

TABLE IV
BOUNDARY LOSS EVALUATION RESULTS

MAE	Configuration			
	Baseline+l_{bce}	Baseline+$l_{bce}+l_{iou}$	Baseline+$l_{bce}+l_{iou}+l_{ssim}$	Baseline+$l_{bce}+l_{iou}+l_{ssim}+augmentation$
Austin	0.0459	0.0452	0.0450	**0.0428**
Chicago	0.0743	0.0711	0.0738	**0.0665**
Massachusetts	0.0626	0.0606	0.0601	**0.0567**
Vienna	0.0749	0.0749	0.0732	**0.0623**
ISPRS-Potsdam	0.0585	0.0557	0.0485	**0.0427**

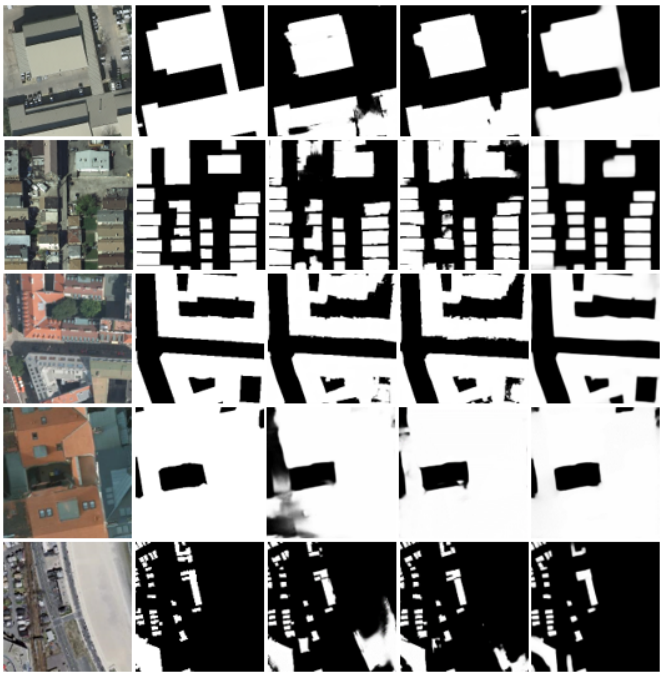


Fig.9. Qualitative comparison of the proposed method with different loss function. From top to bottom, the segmentation results of Austin, Chicago, Massachusetts, Vienna, and Potsdam.

In Table IV, we show that under the MAE indicator, we compare the quantitative evaluation results of different loss

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2020.3017934, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

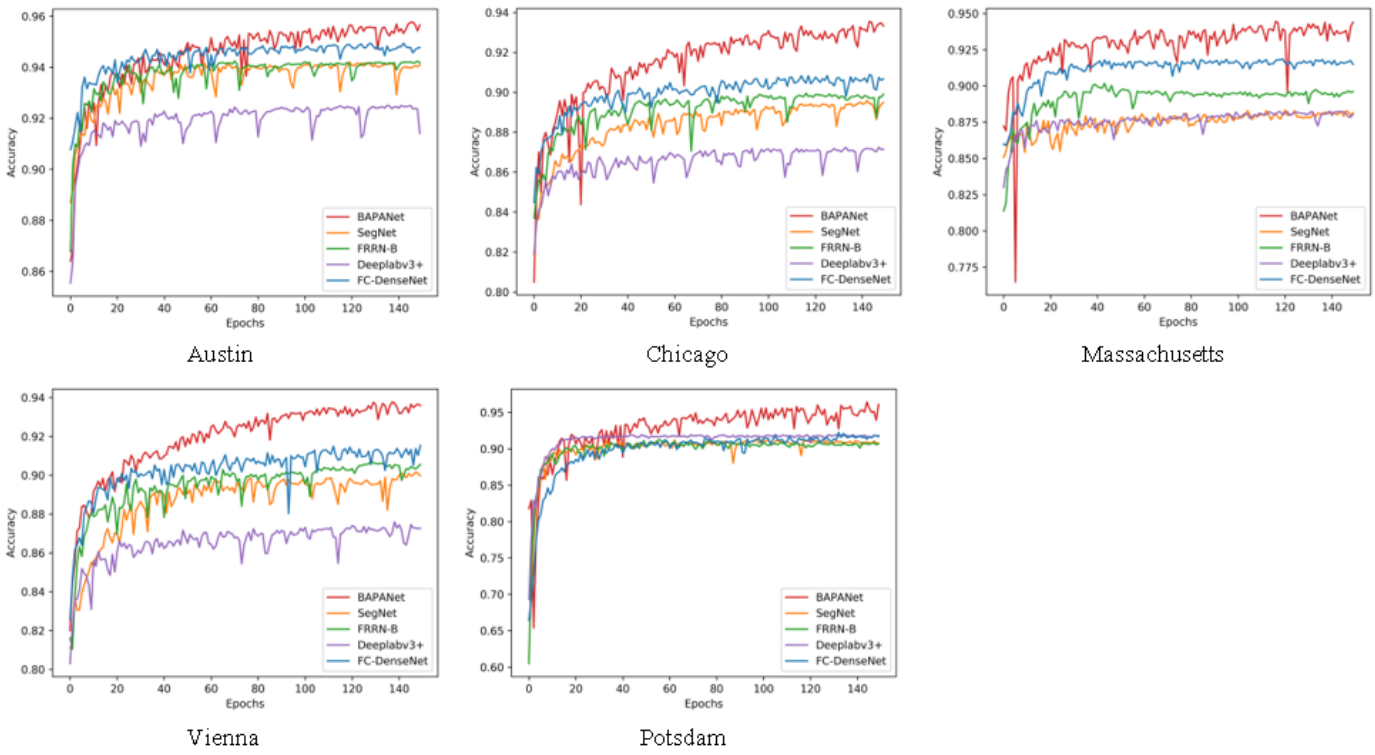


Fig.8. Plots showing the accuracy of the five models while training the datasets with increasing iterations.

functions. To the best of our knowledge, the backbone network using only the Resnet-34 module without any post-processing methods (such as CRF).It can be seen that our method can detect buildings in various situations well. It is also worth noting that due to the effect of contour loss, our results have obvious boundaries and more significant areas for precise positioning. In Fig 9, compared with the baseline BCE coefficient result, adding the IoU metric to the loss function eliminate background interference, while SSIM loss gets a better boundary.

IV. DISCUSSION

The current semantic segmentation algorithms focus on the accuracy of the building area and ignore the building's boundary quality. Moreover, the deep neural network model may over-calculate the pixel points that distinguish the boundaries of buildings, resulting in the misclassification of the edge pixels that should have clear and continuous boundaries, making the extraction results have problems such as smooth edges and loss of detailed information. To solve the above problems, we propose a building boundary perception detection network with arbitrary shape. The model is divided into two stages: the first stage is the prediction network, which is used to generate rough prediction results; the second stage is an excellent network that follows the prediction network to refine further the rough results obtained in the previous step to obtain a more accurate result. The network structure in these two phases is roughly the same, and both are classic Encoder-Decoder networks. Finally, four indexes are used to comprehensively evaluate the accuracy of the model region segmentation and the quality of the building boundary.

The quantitative evaluation results show that this paper's method reduces the impact of image noise while retaining the precise boundaries and integrity of the building and solves the problem of excessively smooth edge contours of the extracted results to a certain extent. Moreover, we comprehensively evaluate the model performance before and after adding the RRF module. The evaluation results show in Table V.

TABLE V
COMPARISON OF THE EFFICIENCY OF DIFFERENT LOSS FUNCTION WITH REFINE MODULE

Model	Inference (ms)	Model size (MB)	FPS
$Baseline + l_{bce}$	241	-	-
$Baseline + l_{bce} + l_{iou}$	246	-	-
$Baseline + l_{bce} + l_{iou} + l_{ssim}$	249	330	11
$Baseline + l_{bce} + l_{iou} + l_{ssim} + RRF(BAPANET)$	263	332	11

Where "Inference" and "FPS" represent the time required for one iteration and the number of frames predicted per second, to evaluate further the image of the building boundary recognition by the RRF module, we get the MAE results of the three regions Vienna, Massachusetts, and Potsdam, as shown in Table VI. The results of building edge detection in Fig. 10. The yellow, the blue, and green lines delineate the actual building contour and the prediction results of adding RRF and

no RRF module. We found that adding RRF modules can improve the building's edge detection performance.

To further confirm the availability of BAPANET, we also provide several loss functions that optimize the boundaries of objects and detect small objects, such as BF1 [54] metric, Hausdorff distance loss [55], and Lovasz-Softmax [56]. We proposed that the model obtained the best results from the evaluation results in Table VII.

TABLE VI
COMPARISON OF THE MAE METRIC WITH REFINE MODULE

Model	MAE		
	Vienna	Massachusetts	Potsdam
Baseline+l_{bce} + $l_{iou} + l_{ssim}$	0.0686	0.0604	0.0456
Baseline+l_{bce} + l_{iou} + l_{ssim} + $RRF(BAPANET)$	0.0667	0.0583	0.0455

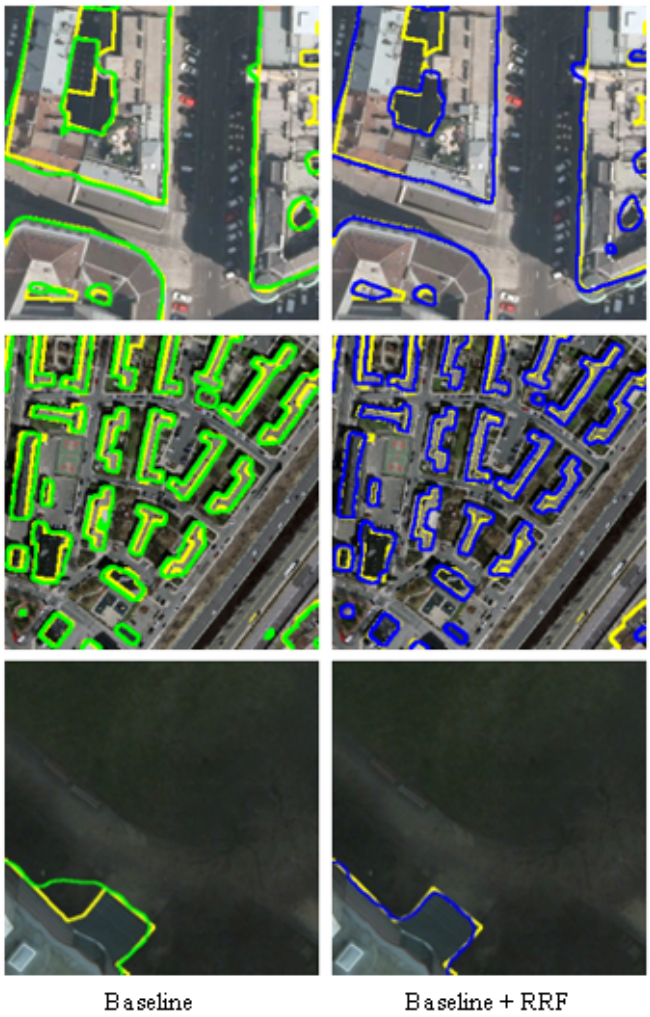


Fig.10. Example: Three building area where blue lines delineate the building extraction results with RRF module, and yellow lines denote the ground truth.

Because of the complexity and diversity of target information in high-resolution imagery, it is challenging for different segment objects. In general, the more considerable the amount of data, the easier it is for the model to learn representative features. From Table IV and Fig.12, we will also find that the model with enhanced data dramatically improves the accuracy of the building. Besides, the BAPANet model also achieved the most accurate extraction of large buildings with irregular shapes. The ResNet-34 structure based on Encoder-Decoder can strengthen features and further improve the ability to learn features. The mixed loss function of binary cross-entropy, intersection the effect of over-smoothing of the sample boundary in the building extraction task to a certain extent have a certain universality.

TABLE VII
COMPARISON OF THE EFFICIENCY OF DIFFERENT NETWORK MODELS

Loss function	Vienna		
	OA	mIoU	F1
$L_{bce+iou+ssim}$	**94.28**	**88.53**	**94.26**
L_{BF}	90.76	82.25	90.74
L_{HDDT}	90.74	82.20	90.73
L_{lovasz}	91.38	83.25	91.37

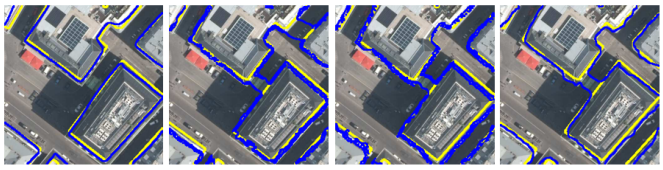


Fig.11. Example: Four different loss function where blue lines delineate the building extraction results, and yellow lines denote the ground truth.

The traditional method is to start with the edge line features of the building and perform a series of analysis and processing on the image's edge line features to extract the building.For example, Andrea [57] proposed a building extraction algorithm based on the analysis and merging of image edge segments; Chungan [58] proposed applying prior knowledge to extract rectangular elements with regular geometric shapes in remote sensing images. Although the traditional building extraction methods have achieved excellent results in specific application backgrounds, they cannot effectively integrate the context-dependent relationships of building edge line features for building extraction. Although the building extraction results based on the fusion of high-resolution imagery and deep neural networks perform well, this success is mostly due to the emergence of new neural network structures, such as ResNet, Inception, DenseNet. Designing a high-performance neural network requires a lot of expertise and trial and error, and the cost is exceptionally high, limiting the application of neural networks on many problems. There is still development potential for further research on improving the extraction accuracy of buildings by automatically designing high-performance network structures based on the sample set.

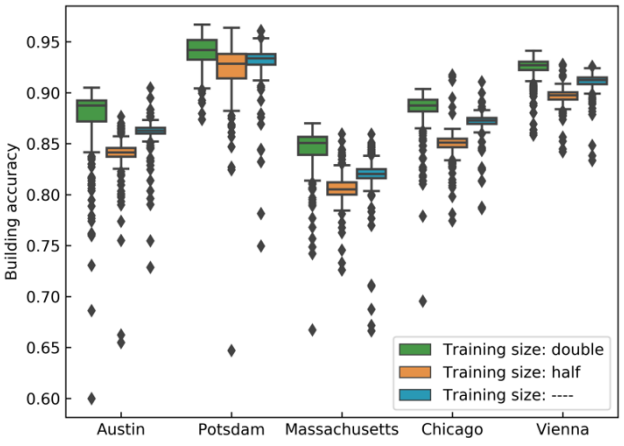


Fig.12. Statistical results of building accuracy of five test sets with different training data sizes. The orange, green, and cyan colors in the above figure indicate that the training data set is doubled, halved, and remains unchanged.

V. CONCLUSION

This paper proposes an original end-to-end boundary perception model-BAPANet, and a mixed loss function to mitigate the impact of overly smooth sample boundaries in building extraction tasks. The proposed BAPANet is a predictive optimization architecture consisting of two components: a predictive network and an optimization module. Combined with the mixing loss function, BAPANet will extract large and irregular buildings accurately. Experimental results on five datasets show that the network architecture is superior to the other four optimal methods in both region and boundary perception metrics. Besides, our proposed model is modular, and it can easily be extended or adapted to other tasks by replacing predictive networks or optimization modules.

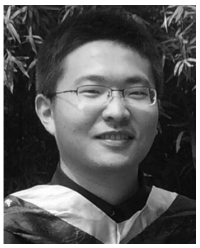
ACKNOWLEDGMENT

For the current study, X. Jiang and X. Zhang designed the study and conducted the analysis. Q. Xin, and their team provided computing resources. X. X and P. Zhang provided the guidance of using these data. All the authors contributed to the drafting of this paper. We are grateful to E. Maggiori et al. for providing the Inria Aerial Image Labeling Benchmark, and grateful to Volodymyr Mnih for providing the Massachusetts buildings dataset. We are also grateful to F. Rottensteiner et al. for providing the ISPRS Potsdam dataset.

REFERENCES

[1] Lin, Chungan, and Ramakant Nevatia. "Building detection and description from a single intensity image." Comput. Vis. Image Underst., vol. 72, no. 2, pp. 101-121, 1998.
[2] Tally Jr, Robert T. "In the Deserts of Cartography: Building, Dwelling, Mapping." The Map and the Territory. Springer, Cham, pp. 599-608, 2018.
[3] Hamer, Melinda J. Morton, et al. "Economic community of West African states disaster preparedness tabletop exercise: building regional capacity to enhance health security." Dis. Med. Public Health Prep., vol. 13, no. 3, pp. 400-404, 2019.
[4] Sohn, Gunho, and Ian Dowman. "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction." ISPRS-J. Photogramm. Remote Sens., vol. 62, no. 1, pp. 43-63, 2007.
[5] Zhang, Yun. "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering." ISPRS-J. Photogramm. Remote Sens., vol. 54, no. 1, pp. 50-60, 1999.
[6] Sirmacek, Beril, and Cem Unsalan. "Building detection from aerial images using invariant color features and shadow information." 2008 23rd International Symposium on Computer and Information Sciences. IEEE, 2008.
[7] Ferraioli, Giampaolo. "Multichannel InSAR building edge detection." IEEE Trans. Geosci. Remote Sensing., vol. 48, no. 3, pp. 1224-1231, 2009.
[8] Dunaeva, Aleksandra Valer'evna, and Fiodor Andreevich Kornilov. "Specific shape building detection from aerial imagery in infrared range." Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya" Vychislitelnaya Matematika i Informatika"., vol. 6, no. 3, pp. 84-100, 2017.
[9] Awrangjeb, Mohammad, Chunsun Zhang, and Clive S. Fraser. "Improved building detection using texture information." International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences., vol. 38, pp. 143-148, 2011.
[10] Liow, Yuh-Tay, and Theo Pavlidis. "Use of shadows for extracting buildings in aerial images." Computer Vision, Graphics, and Image Processing., vol. 49, no. 3, pp. 242-277, 1990.
[11] dos Santos Galvanin, Edinéia Aparecida, and Aluir Porfírio Dal Poz. "Extraction of building roof contours from LiDAR data using a Markov-random-field-based approach." IEEE Trans. Geosci. Remote Sensing., vol. 50, no. 3, pp. 981-987, 2011.
[12] Zhong, Chen, et al. "Building change detection for high-resolution remotely sensed images based on a semantic dependency." 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2015.
[13] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
[14] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[15] Liu, Yuliang, et al. "Omnidirectional Scene Text Detection with Sequential-free Box Discretization." arXiv preprint arXiv:1906.02371, 2019.
[16] Maggiori, Emmanuel, et al. "Convolutional neural networks for large-scale remote-sensing image classification." IEEE Trans. Geosci. Remote Sensing., vol. 55, no. 2, pp. 645-657, 2016.
[17] Huang, Zuming, et al. "Building extraction from multi-source remote sensing images via deep deconvolution neural networks." 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2016.
[18] Liu, Penghua, et al. "Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network." Remote Sens., vol. 11, no. 7, pp. 830, 2019.
[19] Zhang, Xiuyuan, and Shihong Du. "Learning selfhood scales for urban land cover mapping with very-high-resolution satellite images." Remote Sens. Environ., vol. 178, pp. 172-190, 2016.
[20] Zhao, Wenzhi, and Shihong Du. "Learning multiscale and deep representations for classifying remotely sensed imagery." ISPRS-J. Photogramm. Remote Sens., vol. 113, pp. 155-165, 2016.
[21] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481-2495, 2017.
[22] Audebert, Nicolas, Bertrand Le Saux, and Sébastien Lefèvre. "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks." Asian conference on computer vision. Springer, Cham, pp. 180-196, 2016.
[23] Xu, Yongyang, et al. "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters." Remote Sens., vol. 10, no. 1, pp. 144, 2018.
[24] Chen, Qi, et al. "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings." arXiv preprint arXiv:1807.09532, 2018.
[25] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[26] Huang, Xin, et al. "A New Building Extraction Postprocessing Framework for High-Spatial-Resolution Remote-Sensing Imagery." IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., vol. 10, no. 2, pp. 654-668, 2017.
[27] Yang, Hsiuhan Lexie, et al. "Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States." IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., vol. 11, no. 8, pp. 2600-2614, 2018.

[28] Sun, Ying, et al. "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data." ISPRS-J. Photogramm. Remote Sens., vol. 143, pp. 3-14, 2018.
[29] Maltezos, Evangelos, et al. "Building Extraction From LiDAR Data Applying Deep Convolutional Neural Networks." IEEE Geosci. Remote Sens. Lett., vol. 16, no. 1, pp. 155-159, 2019.
[30] Huang, Jianfeng, et al. "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network." ISPRS-J. Photogramm. Remote Sens., vol. 151, pp. 91-105, 2019.
[31] Volpi, Michele, and Devis Tuia. "Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks." IEEE Trans. Geosci. Remote Sensing., vol. 55, no. 2, pp. 881-893, 2017.
[32] Audebert, Nicolas, Bertrand Le Saux, and Sebastien Lefevre. "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks." ISPRS-J. Photogramm. Remote Sens., vol. 140, pp. 20-32, 2018.
[33] Liu, Yongcheng, et al. "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network." ISPRS-J. Photogramm. Remote Sens., vol. 145, pp. 78-95, 2018.
[34] Marcos, Diego, et al. "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models." ISPRS-J. Photogramm. Remote Sens., vol. 145, pp. 96-107, 2018.
[35] Li, Xiaoxiao, et al. "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[36] Chai, Dengfeng, Shawn Newsam, and Jingfeng Huang. "Aerial image semantic segmentation using DCNN predicted distance maps." ISPRS-J. Photogramm. Remote Sens., vol. 161, pp. 309-322, 2018.
[37] Marmanis, Dimitrios, et al. "Classification with an edge: improving semantic image segmentation with boundary detection." ISPRS-J. Photogramm. Remote Sens., vol. 135, pp. 158-172, 2018.
[38] Yuan, Jiangye. "Learning Building Extraction in Aerial Scenes with Convolutional Networks." IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 11, pp. 2793-2798, 2018.
[39] Li, Er, et al. "Building Extraction from Remotely Sensed Images by Integrating Saliency Cue." IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., vol. 10, no. 3, pp. 906-919, 2017.
[40] Peng, Chao, et al. "Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[41] Islam, Md Amirul, et al. "Salient Object Detection using a Context-Aware Refinement Network." BMVC. 2017.
[42] Deng, Zijun, et al. "R3Net: Recurrent residual refinement network for saliency detection." Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI Press, 2018.
[43] Wang, Tiantian, et al. "A stagewise refinement model for detecting salient objects in images." Proceedings of the IEEE International Conference on Computer Vision. 2017.
[44] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904-1916, 2015.
[45] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[46] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.
[47] Xu, Yongyang, et al. "Road extraction from high-resolution remote sensing imagery using deep learning." Remote Sens., vol. 10, no. 9, pp. 1461, 2018.
[48] Yang, Maoke, et al. "Denseaspp for semantic segmentation in street scenes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[49] Gao, Shanghua, et al. "Res2net: A new multi-scale backbone architecture." IEEE Trans. Pattern Anal. Mach. Intell. 2019.
[50] Maggiori, Emmanuel, et al. "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark." 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2017.
[51] Mnih V. Machine learning for aerial image labeling[D]. Toronto:University of Toronto, 2013.
[52] Rottensteiner, Franz, et al. "The ISPRS benchmark on urban object classification and 3D building reconstruction." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences., vol. 3, no. 1, pp. 293-298, 2012.
[53] Perazzi, Federico, et al. "Saliency filters: Contrast based filtering for salient region detection." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.
[54] Csurka, Gabriela, et al. "What is a good evaluation measure for semantic segmentation?." BMVC. Vol, 27, 2013.
[55] Karimi, Davood, and Septimiu E. Salcudean. "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks." IEEE Trans. Med. Imaging., vol. 39, no. 2, pp. 499-513, 2019.
[56] Berman, Maxim, Amal Rannen Triki, and Matthew B. Blaschko. "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[57] Laliberte, Andrea S., and Albert Rango. "Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery." IEEE Trans. Geosci. Remote Sensing., vol. 47, no. 3, pp. 761-770, 2009.
[58] Lin, Chungan, and Ramakant Nevatia. "Building detection and description from a single intensity image." Comput. Vis. Image Underst., vol. 72, no. 2, pp. 101-121, 2019.



Xin Jiang received the B.S. degree in geographic information system from Hohai University, Nanjing, China, in 2017, and the master degree in geography, from Sun Yat-Sen University, Guangzhou, China, in 2020.

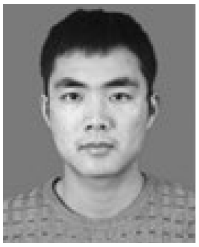
He is currently a research assistant with the School of Environmental Science & Engineering, Southern University of Science and Technology, Shenzhen.His research interests include machine learning, mathematical modeling, dynamic monitoring of surface water, and multi-source remote sensing data fusion.Mr. Jiang received the First Prize of China Undergraduate Mathematical Contest in Modelling in 2015, and the Honorable Mention of Mathematical Contest in Modeling (MCM/ICM) in 2016.

Xinchang Zhang received the B.S. degree in cartography from the Wuhan Institute of Surveying and Mapping, Wuhan, China, in 1982, the M.S. degree in cartography from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1994, and the Ph.D. degree in resources and environmental sciences from Wuhan University, Wuhan, China, in 2004.

He is currently a Professor with the School of Geographical Sciences, Guangzhou University, Guangzhou, China, and a Chair Professor with Henan University. His research interests include spatial database updating, spatial data integration, and smart city

Qinchuan Xin received the B.S. degree in physics from Peking University, Beijing, China, in 2005, and the Ph.D. degree in geography from Boston University, Boston, MA, USA, in 2012.

He is currently a Research Associate with the State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography and the Research Center for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi, China. His research interests include ecological remote sensing, terrestrial ecological modeling, global climate change, and ecosystem feedback.



Xu Xi received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2012, and the M.S. degree from Liaoning Normal University, Dalian, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Remote Sensing and Geographic Information Systems, Sun Yat-sen University, Guangzhou, China.

His research interests include geospatial data security and digital watermarking.

Pengcheng Zhang works at Guangzhou Urban planing & design survey research institute. His research interests include spatial database updating, smart city, and application of remote sensing (RS) technology.