ReA-Net: A Multiscale Region Attention Network With Neighborhood Consistency Supervision for Building Extraction From Remote Sensing Image

Shengjun Xu, Bowen Deng, Yuebo Meng, Guanghui Liu, and Jiuqiang Han

Abstract—Aimed at the challenge of low accuracy of building segmentation caused by poor continuity of remote-sensing-image regions and blurred boundaries, a remote sensing building semantic segmentation algorithm based on multiscale regional consistent attention supervision is proposed. First, based on the Unet encoder–decoder architecture, the proposed algorithm constructs the region attention network (ReA-Net), which employs a multiscale receptive field-guidance model to simultaneously focus on regional features and edge details of remote sensing image objects. Second, the self-attention mechanism is employed to establish the correlation representation of regional-level features of remote sensing images, and multiscale regional attention features of remote sensing images are obtained through weighted regional-level correlation mapping. Finally, to address the lack of spatial correlation constraints on the prediction of remote sensing images segmentation, a loss function with multiscale neighborhood consistency supervision is suggested to constrain the consistency of pixel label assignment related to a local region. Experimental results on WHU building dataset showed that intersection over union (IOU) reached 91.6%, precision reached 95.61%, recall reached 95.68%, and F1-score reached 95.64%; On the Massachusetts building dataset, IOU reached 74.77% and precision reached 83.93%, recall reached 87.53%, and F1-score reached 85.69%. Therefore, the proposed algorithm not only has a good segmentation effect but also has a strong robustness for remote sensing building image segmentation.

Index Terms—Building segmentation, neighborhood consistency supervision, region attention, remote sensing image, U-Net.

I. INTRODUCTION

BUILDINGS are the main ground object in urban and rural areas. Accurate building distribution information is of great significance for monitoring the change of urban areas, urban planning, and population estimation. The semantic segmentation of high-resolution remote sensing buildings is a major part of remote sensing earth observation technology. Its main task is to use the collected remote sensing images to extract the relevant characteristic information of buildings, and to classify the target represented by each pixel in the remote sensing images. However, compared with natural objects, such as water bodies and forests in remote sensing images, buildings are often affected by severe disturbances, such as illumination, season, unclear angle and boundary, and complex background information. These disturbances bring great challenges to the accurate segmentation of buildings in remote sensing images.

With the development of computer vision technology, more and more researchers have undertaken extensive studies on the semantic segmentation of high-resolution remote sensing buildings [2], [3], [4], [5], [6], and proposed a large quantity of remote sensing image semantic segmentation methods. These research methods are mainly divided into the following two parts: one is traditional machine-learning-based segmentation and the other deep-learning-based segmentation.

The segmentation method based on traditional machine learning mainly utilizes artificial features, such as shape, texture, color, spectrum, and spatial details to train the classifier. These methods mainly include local binary patterns (LBPs) [7], random forests (RFs) [8], K-nearest neighbors (KNN) [9], Gaussian maximum likelihood classifier [10], logistic regression [11], and Markov random field [12]. LBPs [7] are an image local texture feature extraction algorithm. In the domain of remote sensing image segmentation, the relationship between a single pixel and its neighbor pixels is used to describe the features of local texture structure. However, the feature dimension of local feature is too large after coding, so it is difficult to use the LBPs to express the texture features of remote sensing images in the practical applications. RF [8] is an ensemble learning model with decision tree as the basic classifier, which is composed of multiple decision trees. When the remote sensing image to be segmented is input, the final classification result of each pixel is determined by multiple decision trees voting. KNN [9] is a supervised classification method, which is based on the principle of utilizing distance metric to calculate the distance between unknown data and known data, and then determining
the similarity of different pixel semantic features of remote sensing images based on their distances, and selecting the nearest K distances as the basis for determining the class of unknown data. Markov random field [12] combines graph theory with Bayesian probability theory, and uses the spatial features of neighboring pixels as prior knowledge to establish a semantic segmentation model for remote sensing buildings. Although this method can effectively improve the continuity of the segmentation region, the traditional Markov random field is difficult to describe the complex statistical features of remote sensing images, meanwhile its inference is still an NP-hard problem. Although the segmentation methods based on traditional machine learning have achieved certain effects, the effectiveness of such methods strongly depends on the accurate description of the prior knowledge of the specific scene images. This problem often leads to the shortcomings of traditional machine learning-based image segmentation methods, such as poor generalization ability and complex design, and it is challenging to solve the problem of remote sensing image semantic segmentation in a real complex environment.

In recent years, deep learning has achieved satisfactory results in the domain of remote sensing image semantic segmentation due to its ability of strong generalization and self-learning characteristics [13], [14], [15], [16], [17], [18]. Deep-learning-based remote sensing image semantic segmentation networks primarily utilize encoder–decoder structure to extract the semantic feature of remote sensing images and restore the resolution [19], so as to achieve the segmentation of the target remote sensing objects. Shao et al. [20] suggested a remote sensing building semantic segmentation network (ReA-Net) based on multiscale feature fusion and residual refinement structure. The fusion of deep semantic features and the refinement of segmentation results are realized and refined by introducing atrous convolution module with distinct receptive fields and a residual refinement module. Chen et al. [21] introduced a dense Xception module based on DeepLabV3+Net and utilized a multiscale fusion approach to fuse shallow and deep features, which effectively enhanced the network’s ability to extract contextual information. Zeng et al. [22] employed PCA to downscale the input raw remote sensing data for problems, such as missing edge information of remote sensing images, and performed segmentation and contour extraction of remote sensing image buildings based on multitask learning, thus improving the segmentation result of building target contours. To enhance the ability of deep learning networks to characterize targets to be extracted from different remote sensing images, it is an effective strategy to add attention mechanism in the encoder–decoder module [23], which enriches the ability of deep learning networks to capture long-range dependencies with contextual information. Yang et al. [24] proposed a semantic segmentation method for high-resolution remote sensing images based on the attention fusion network, which obtains satisfactory segmentation results by fusing high-level semantic features and low-level semantic features of the encoder through a multipath attention fusion block module, and uses the decoder to recover the resolution of the fused features. Deng et al. [25] designed an attention-gate structure based on the encoder–decoder structure, which is based on the spatial pyramidal pooling structure of the atrous convolution and utilizes the atrous convolution to further expand the perceptual field while fusing the multilayer features of the decoder, and enhances the pixel-level feature information expression in the relevant regions through the attention-gate structure, thus improving the results of semantic segmentation of remote sensing images of buildings. Li et al. [26] proposed a multcore attention mechanism to extract global contextual information while utilizing the hierarchical aggregation strategy to fuse different levels of contextual information, which effectively exploits the feature information at each stage and enhances the network feature representation. To solve the problem of unclear global semantics at image-level in remote sensing images, Ding et al. [27] proposed a local attention module using the characteristics of local image blocks with clear semantics to achieve region-level attention weighting, which effectively enhanced the network’s ability to express region-level semantics.

In summary, the spatial attention mechanism has made impressive advances in the field of semantic segmentation of remote sensing buildings based on deep learning. However, the segmentation method based on the classical spatial attention mechanism focuses primarily on pixel-level spatial correlation of images, although this mechanism has the feature of enhancing strong correlation features and suppressing weak correlation features, due to the high resolution of remote sensing buildings images with large overall resolution and the small area to be segmented, therefore, directly applying the classical spatial attention mechanism to the semantic segmentation of remote sensing buildings images will lead to the following two problems: 1) The classical spatial attention mechanism aggregates the global statistical information of the image, which frequently leads to an exponential increase in the computational complexity of the network for high-resolution remote sensing images, resulting in an excessive occupation of computational resources; 2) The pixel-level attention mechanism is challenging to accurately capture the region-level spatial correlation of remote sensing images at the large scale, and therefore, lacks learning and supervision of the building regions and the coherence of building edges during the training and segmentation of the remote sensing image model, which leads to a large impact of interference information such as illumination, season, angle, and background on the accurate segmentation of buildings.

To solve the abovementioned problems, a multiscale region attention (MRA) network with neighborhood consistency supervision for remote sensing buildings is proposed. The proposed network constructs a multiscale regional attention module based on the Unet encoder–decoder structure and employs the multiscale neighborhood information of remote sensing images to build the multiscale region-level feature descriptors. The self-attention (SA) mechanism is operated to capture the intraregional correlation and the correlation between different local regions in remote sensing images, and enhance the ability of semantic features to pay attention to the correlation features of target regions at different scales. Based on the assumption of spatial consistency of local region pixel label assignment, a loss function for multiscale neighborhood consistency supervision is
The main contributions of this article are described as follows.

1) To solve the problem that the attention mechanism at the pixel level is difficult to effectively characterize the global semantic features of salient targets in remote sensing images, a multiscale regional attention module is devised to intensify the regional focusing capability of semantic features on building regions of remote sensing images.

2) Based on the spatial consistency assumption that pixels in local regions of remote sensing images tend to take the same labels, a loss function for multiscale neighborhood consistency supervision is proposed, and a neighborhood consistency supervision mechanism is established to constrain the regional consistency of pixel segmentation labels, which enhances the robustness of the network to noise, texture mutation and other disturbing information.

3) Fusing MRA and neighborhood consistency supervision mechanism, ReA-Net based on multiscale region consistency attention supervision is proposed, and a comparative study is carried out in two remote sensing building datasets to validate the effectiveness of the proposed method.

The rest of this article is organized as follows. Section II presents the details of the proposed building extraction network architecture, including the feature extraction module, the resolution recovery module, and the MRA module, and the proposed network loss function. Experimental results and demonstrations are reported and analyzed in Section III. Finally, Section IV concludes this article.

II. PROPOSED METHODS

Many remote sensing building extraction methods based on deep convolutional neural networks have been presented since the emergence and development of convolutional neural networks. Semantic segmentation algorithms based on the encoder–decoder structure of Unet frequently utilize the characteristics of residual connectivity and dense connectivity and have achieved numerous research achievements in the semantic segmentation of high-resolution remote sense buildings [28], [29], [30]. To enhance the feature representation capability of deep convolutional neural networks for remote sensing image segmentation, researchers commonly adopt the attention mechanism in the feature extraction stage to obtain more information about the target of remote sensing images and suppress the background, noise, and other interference feature of remote sensing images.

However, most of these methods are based on pixel-level attention mechanisms, which cannot sufficiently express the spatial correlation between local regions of remote sensing images and the spatial coherence of adjacent pixels within the region, thus, often leads to boundary discontinuity and mis-segmentation of intraregional patches in remote sensing image target segmentation results. To address the aforementioned challenges, we propose a multiscale attention network with neighborhood consistency supervision for remote sensing buildings, which is based on Unet encoder–decoder structure. The overall structure of ReA-Net network is shown in Fig. 1. An MRA module is constructed to enhance the regional concentration ability of semantic features on building regions in remote sensing images. Then, a neighborhood consistency supervision module is proposed, which effectively enhance the smoothness of the network to assign labels to pixels in local areas.

A. Encoder–Decoder Segmentation Network Based on ReA-Net

The proposed network is mainly composed of the following three parts: the feature extraction module of remote sensing image based on Unet encoder structure, the resolution recovery module based on the decoder, and the MRA module. The feature
extraction module of the remote sensing image extracts the texture, boundary, and deep semantic features of buildings in the remote sensing image, and inputs the extracted features into the resolution restoration module for resolution restoration. To enhance the ability of the network to represent the correlation between regions of remote sensing images, an MRA module is introduced into the resolution recovery module, and the region-level features in different scales are constructed by atrous convolution and pooling operation, and input into the SA module to obtain an enhanced map of region-level correlations of the feature map. Finally, the local weighting (LW) module is used to fuse the enhanced image of regional correlation with the input feature to achieve the presentation of regional correlation in remote sensing images. Each module is detailed as follows.

1) Remote Sensing Image Feature Extraction Module: The remote sensing image feature extraction module mainly consists of two stages. In the first stage, a convolution block (Conv1) is used to extract the low-level texture features of remote sensing images. The Conv1 mainly contains two convolution layers with convolution kernels size of \(3 \times 3\) and a padding of 1. Each convolutional layer is followed by a batch normalization (BN) layer and leaky corrected linear unit (LeakyReLU). The second stage consists of four residual blocks based on residual structure (RESConv1-RESConv4), in which each residual block contains a max pooling layer with a kernel size of \(2 \times 2\) for down-sampling operation, and at the same time, there are two convolution layers with a convolution kernel \(3 \times 3\) and the padding of 1. Each convolutional layer is followed by a BN with LeakyReLU, which is used to increase the network’s ability to extract deeper semantic features from remote sensing images. The parameters of the proposed encoder are shown in Table I, where \(k\) represents kernel size; IC represents number of input channels; OC represents number of output channels; Pool represents the detail of pooling layer; Output represents the size of outputs.

| Stage | IC | OC | Pool | Output |
|-------|----|----|------|--------|
| Input | /  | /  | /    | \(H \times W\) |
| 1(Conv1) | 3x3 | 3 | 64 | \(H \times W\) |
| 2(ResConv1) | 3x3 | 64 | 128 | Max \(2 \times 2\) \(H / 2 \times W / 2\) |
| 2(ResConv2) | 3x3 | 128 | 256 | Max \(2 \times 2\) \(H / 4 \times W / 4\) |
| 2(ResConv3) | 3x3 | 256 | 512 | Max \(2 \times 2\) \(H / 8 \times W / 8\) |
| 2(ResConv4) | 3x3 | 512 | 512 | Max \(2 \times 2\) \(H / 16 \times W / 16\) |

2) Resolution Recovery Module Based on the Decoder: The decoder-based resolution recovery module focuses on the resolution recovery of extracted remote sensing image feature maps using a progressive up-sampling strategy, and realize the preclassification of dense pixels. The structure consists of four up-sampling blocks (Upsample1- Upsample4), four multiscale region attention (MRA1-MRA4) modules, and one convolution block (Outconv), which are divided into the following five stages. From stage 1 to stage 4, each stage consists of an up-sampling module and an MRA module. The up-sampling module completes the progressive spatial resolution recovery of the feature map, and to eliminate the checkerboard artifacts caused by deconvolution [31], and each up-sampling block contains a bilinear up-sampling layer and two convolutional layers with a convolutional kernel size of \(3 \times 3\), padded with 1, each followed by a BN and a LeakyReLU. Meanwhile, in the feature fusion phase of up-sampling, the MRA module is employed to fuse distinct multiscale region level correlation enhancement maps, which improves the network’s ability to reflect larger-scale spatial correlations between regions in remote sensing images. The MRA block is mainly composed of average pooling layers and convolutional layers. The multiscale regions are extracted by the atrous convolution layers with kernel size \(3 \times 3\) and different atrous rate \([1, 3, 5, 7]\). The fifth phase involved a convolutional block that implements the classification of the remote sensing image’s output feature map. Table II shows the parameters of the proposed decoder structure, where \(k\) represents kernel size; IC represents number of input channels; OC represents number of output channels; Scale represents the scale factor of up-sample layer; Output represents the size of outputs.

| Stage | IC | OC | Scale | Output |
|-------|----|----|-------|--------|
| Input | 512 | /  | /     | \(H / 16 \times W / 16\) |
| 1(Upsample1 + MRA1) | 3x3 | 512 | 512 | 2 \(H / 8 \times W / 8\) |
| 2(Upsample1 + MRA2) | 3x3 | 512 | 256 | 2 \(H / 4 \times W / 4\) |
| 3(Upsample1 + MRA3) | 3x3 | 256 | 128 | 2 \(H / 2 \times W / 2\) |
| 4(Upsample1 + MRA4) | 3x3 | 128 | 64  | 2 \(H / W\) |
| 5(Outconv) | 3x3 | 64  | 2    | \(H / W\) |
sensing images, severe interference factors, such as illumination, season, angle, and unclear boundary of buildings often result in uneven segmentation regions and the unclear edge of buildings. To alleviate the aforementioned issues, we propose an MRA module. The overall structure is illustrated in Fig. 2.

Let $F_{\text{in}} \in \mathbb{R}^{C \times H_f \times W_f}$ denotes the input MRA feature map, and $F_{\text{out}} \in \mathbb{R}^{C \times H_f \times W_f}$ represents the output MRA feature map, where $W_f, H_f, C$ are the height, width, and channel number of the input MRA feature map, respectively. The proposed MRA is mainly composed of the multiscale neighbor extraction module, the region embedding (RE) module, the SA module and the LW module. The MRA process is described as follows:

First, to obtain the multiscale neighborhood feature $A \in \mathbb{R}^{4C \times H_f \times W_f}$, the multiscale neighborhood information of the input remote sensing building feature map $F_{\text{in}}$ is extracted utilizing atrous Conv1 with different atrous rates, and then the neighborhood features extracted by convolution layers of the different atrous rate are stitched together. The processes are shown in the following equation:

$$F_{\text{in},d} = \text{ReLU}(\text{BN}(\text{Conv}_{d,k,\text{pad}}(F_{\text{in}})))$$

$$A = \text{Concat}(F_{\text{in},1}, F_{\text{in},3}, F_{\text{in},5}, F_{\text{in},7})$$

where $\text{Conv}_{d,k,\text{pad}}(\cdot)$ represents the multilayer atrous convolution layer, $d \in [1, 3, 5, 7]$ is parameter atrous rate, $k = 3$ is the size of convolution kernel, and pad $\in [1, 3, 5, 7]$ is the padding; BN(·) stands for the BN layer; ReLU(·) stands for the leaky linear rectifier layer; Concat(·) stands for the concatenation operation.

Second, we input feature map $A$ incorporated multiscale neighborhood information into the RE module. To eliminate the redundancy of multiscale features and improve the presentation of remote sensing image features, the feature map $A$ is dimensionally selected by utilizing a convolution layer with a kernel size of one to obtain the feature map $B \in \mathbb{R}^{C \times H_f \times W_f}$. Simultaneously, to gather regional feature information for all regions, feature map $B$ is subjected to a regional average pooling operation to create a multiscale region-level descriptor $C \in \mathbb{R}^{C \times H_p \times W_p}$ of the input remote sensing images

$$C = \text{Avgpool}_{k,s}(\text{BN}(\text{Conv}_{d,k,\text{pad}}(A)))$$

where $\text{Conv}_{d,k,\text{pad}}(\cdot)$ represents the dimension reduction convolution layer, $d = 1$ is parameter atrous rate, $k = 1$ is the convolution kernel size, pad = 0 is the padding, and the dimension reduction rate is 0.25; BN(·) stands for the BN layer; Avgpool$$_{k,s}(\cdot)$ stands for the average pooling layer with convolution kernel $k = 4$, and stride $s = 4$; let $H_p = 0.25H_f$ and $W_p = 0.25W_f$.

Furthermore, to capture correlations between distinct feature regions of the remote sensing image, an SA method is applied to characterize the correlations between all-region descriptors in the region-level feature map. First, the region-level features are input into the SA module, and the feature map $C$ is encoded and reshaped three times by using convolutional layer with kernel size of one to produce three encoded feature matrices: $V \in \mathbb{R}^{M \times C}$, $G \in \mathbb{R}^{C \times M}$, and $I \in \mathbb{R}^{C \times M}$. Then, the spatial attention scores matrix $Z$ is obtained by multiplying the two feature matrices $V$ and $G$ through sigmoid(·) activation function. Finally, the feature matrix $Z$ and the spatial attention matrix $I$ are multiplied and reshaped to obtain the feature enhancement map $K \in \mathbb{R}^{C \times H_p \times W_p}$ with inter-regional correlation. The calculation process is shown in the following equation:

$$V = \text{Reshape} (\text{Conv}_{d,k,\text{pad}}(C))$$

$$G = \text{Reshape} (\text{Conv}_{d,k,\text{pad}}(C))$$

$$I = \text{Reshape} (\text{Conv}_{d,k,\text{pad}}(C))$$

$$K = \frac{\exp(V_i \otimes G_i)}{\sum_{i=1}^{M} \exp(V_i \otimes G_i)} \cdot I$$

where $\text{Conv}_{d,k,\text{pad}}(\cdot)$ represents the convolution layer, $d = 1$ is the parameter atrous rate, $k = 1$ is the convolution kernel size, and pad = 0 is the padding; $\otimes$ stands for multiplying the
corresponding elements; \( V_i, G_i \) represents the \( i \)th value of \( V, G \), respectively; let \( M = W_p \times H_p \).

Then, the inter-regional correlation feature enhancement map \( K \in \mathbb{R}^{C \times H_p \times W_p} \) is input into the LW module, furthermore the correlation weighted map of region-level is projected into the original feature map \( F_{in} \). After up-sampling the enhanced map \( K \), the region relevance weighted map \( H \) is obtained. Finally, the region attention feature map \( Q \in \mathbb{R}^{C \times H_f \times W_f} \) is constructed by multiplying \( H \) by \( F_{in} \) pixel by pixel. The compute process is as follows:

\[
H = \text{Upsample}_{\text{scale}, \text{mode}} (K) \tag{6}
\]

\[
Q = H \otimes K \tag{7}
\]

where \( \text{Upsample}_{\text{scale}, \text{mode}}(\cdot) \) is the up-sampling layer with parameter scaling scale = 4, and the up-sampling mode parameter mode set as the nearest sampling method.

Finally, to incorporate the global context semantic information and the spatial correlation of local regions of the remote sensing image, we use pixel-by-pixel addition to fuse the MRA feature map \( Q \) and the global attention feature map \( T \in \mathbb{R}^{C \times H_f \times W_f} \) from preceding steps. Then, the fusion result is pixel by pixel multiplied by the activation function \( \text{Sigmoid}(\cdot) \), and the original feature \( F_{in} \) to construct multiscale regional attention feature output of MRA \( F_{out} \). The compute process is as follows:

\[
F_{out} = \text{Sigmoid} \left( (Q \oplus T) \otimes F_{in} \right). \tag{8}
\]

### B. Neighborhood Consistency Supervision Loss Function

Let \( X = \{x_1, x_2, \ldots, x_N\} \) denote the set of remote sensing image training sets, where \( x_i \in X \) denotes the \( i \)th remote sensing training image, \( N \) is the number of remote sensing training images in a single input. Let \( Y = \{y_1, y_2, \ldots, y_N\} \) is the set of labels corresponding to the remote sensing image training set one by one, and \( L = \{\ell_1, \ell_2, \ldots, \ell_N\} \) denotes the output set of predicted label maps of ReA-Net. Let \( S = \{s \mid 1 \leq s \leq B \times R \} \) denote a finite set of lattice points defined on a predicted label map \( \ell_i = \{\ell_i^s \mid s \in S\}, i \in [1, N] \), where \( \ell_i^s \) represents the set of neighborhood nodes of node \( s \) and \( B, R \) is the width and height of the input training image, respectively, and \( U = \{u_1, u_2, \ldots, u_N\} \) denotes the output set of the predicted label probability maps of ReA-Net, \( \mathcal{W} = \{w_1, w_2, \ldots, w_N\} \) denotes the set of neighborhood consistency penalty weights for the output of ReA-Net segmentation result \( \mathcal{U} \).

To improve the smoothing of remote sensing image segmentation results, a multiscale neighborhood consistency supervision loss function is proposed based on the assumption that adjacent pixels within a local region tend to be assigned to the same labels. The strategy of the local region consistency constraint will result in better smoothing within the segmented local region. The purpose of establishing the neighborhood consistency supervision loss is to quantitatively evaluate the error between the ReA-Net segmentation results and the true labels for remote sensing images. The proposed neighborhood consistency loss structure shows in Fig. 3.

The proposed neighborhood consistency supervision loss function is defined as

\[
\text{Loss}_{\text{lc}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{B \times R} \left[ w_{i,j} \times \left[ y_{i,j} \log (u_{i,j}) \right] \right. \\
\left. + (1 - y_{i,j}) \log \left( 1 - u_{i,j} \right) \right] \tag{9}
\]

where \( y_{i,j} \in Y \) represents the \( j \)th label of the \( i \)th remote sensing training image, \( u_{i,j} \in \mathcal{U} \) represents the \( j \)th probability of prediction label of the \( i \)th remote sensing training image. \( w_{i,j} \in \mathcal{W} \) represents the \( j \)th neighborhood consistency penalty weight of the \( i \)th remote sensing training image. The weight \( w_{i,j} \) is defined as follows:

\[
w_{i,j} = \frac{1}{|\partial j|} \sum_{s \in \partial j \setminus j} \left[ \ell_{s,i} - \ell_{i,j} \right]^2 + \left( y_{i,s} - \ell_{i,j} \right)^2 \frac{|s - j|}{|s - j|} \tag{10}
\]

where \( \ell_{i,j} \) represents the \( j \)th label predicted value of the \( i \)th predicted label map \( \ell_i \); \( \ell_{i,s} \) represents the \( s \)th label predicted value of the \( i \)th predicted label map \( \ell_i \); \( |\partial j| \) denotes the set of the neighborhood lattice point of the \( j \)th node, excepts for the node \( j \)th itself; \(|\partial j|\) is the potential of the neighborhood lattice point of the \( j \)th node; \( y_{i,j} \) is the predicted label value corresponding to \( \ell_{i,j} \) one-to-one; The denominator term \(|s - j|\) denotes the Euclidean distance of spatial location between the \( j \)th label and its neighbor label. The farther the spatial location between the center label
and its neighbor label, the smaller the weight $w_{i,j}$; The molecular term $(\ell_{i,x} - \ell_{i,j})^2 + (y_{i,s} - \ell_{i,x})^2$ is the sum of the errors of the prediction label $\ell_{i,j}$ of the center in local region and its neighborhood prediction label $\ell_{i,s}$ and the training label $y_{i,s}$ of the neighborhood, respectively. The larger the error value, the larger the weight $w_{i,j}$.

Substituting (10) into (9), the proposed neighborhood consistency loss function is shown in the following equation:

$$
\text{Loss}_{nc} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{B \times R} \left\{ w_{i,j} \times \left[ y_{i,j} \log(u_{i,j}) + (1 - y_{i,j}) \log(1 - u_{i,j}) \right] \right\} 
+ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{B \times R} \left\{ \frac{1}{|\mathcal{N}(j)|} \sum_{s \in \mathcal{N}(j)} \left[ \frac{(\ell_{i,s} - \ell_{i,j})^2 + (y_{i,s} - \ell_{i,j})^2}{|s-j|} \right] \right\}
\times \left[ y_{i,j} \log(u_{i,j}) + (1 - y_{i,j}) \log(1 - u_{i,j}) \right] \right\}
$$

(11)

where $L_{\text{Smooth}}(y, \ell)$ represents the consistency constraint term of pixel label allocation process in the local regions of remote sensing image $x$, which makes the prediction labels $\ell$in the local regions of remote sensing image $x$ tend to be assigned the same label, that is, the constraint segmentation result is smoother. $L_{\text{Data}}(y, u)$ represents the likelihood penalty term of the training label map $y$ and the prediction label probability map $u$ of the output remote sensing image, which constrains the pixel of the remote sensing image to be assigned a label according to the maximum likelihood probability.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Settings

The experimental platform’s workstation is configured with an Intel Xeon E5 2650 processor, 376GB of memory, with 4 NVIDIA 2080Ti 11G graphics cards. PyTorch 1.8, NVIDIA’s CUDA11.2 GPU runtime platform, and the cuDNN8.0 deep learning GPU acceleration package are all used in the deep learning framework.

During the proposed ReA-Net training stage, the size of the input training remote sensing images is set as $512 \times 512$ or $256 \times 256$. To augment the remote sensing image set, random horizontal flip, and normalized operation strategies are employed. The network’s training parameter batch size is set to 12, the total number of learning rounds is 151, the starting learning rate is $1 \times 10^{-3}$, and the AdamW optimizer [47] is utilized for optimization in the experiment. The cosine learning rate adjustment approach is employed in the network training process, with a 20-round adjustment cycle and a minimum learning rate of $1 \times 10^{-6}$.

B. Experimental Dataset and Evaluation Metrics

The experiments employed two remote sensing building image datasets, the aerial imagery dataset [34] and the Massachusetts building dataset [35], which is a subdataset of the WHU building dataset, to evaluate the effectiveness of the proposed algorithm. The aerial imagery dataset consists of Aerial and Satellite Imagery of Christchurch, New Zealand, comprising approximately 22,000 individual buildings. The original ground resolution of the image is 0.075 m. The dataset contains total of 8187 remote sensing images with $512 \times 512$ resolution after cropping, including 4735 training images, 1036 evaluation images, and 2416 test images. The Massachusetts building dataset is constituted of 151 aerial imageries from the Boston area, each with $1500 \times 1500$ pixels and 137 training images, four evaluation images, and ten test images. To facilitate training and evaluation, the experiment used the edge overlap method to divide each original image into 36 subgraphs with the resolution of $256 \times 256$. After the pruning, 5436 training images, 144 evaluation images, and 360 test images are established.

To verify the efficiency of the proposed algorithm in this article, evaluation metrics such as precision [36], recall [36], intersection over union (IoU) [36], and F1-Score [37] are used for a comparison study of the method’s effectiveness, which are defined as follows:

- **Precision**
  \[
  \text{Precision} = \frac{TP}{TP + FP} \tag{12}
  \]

- **Recall**
  \[
  \text{Recall} = \frac{TP}{TP + FN} \tag{13}
  \]

- **IoU**
  \[
  \text{IoU} = \frac{TP}{TP + FN + FP} \tag{14}
  \]

- **F1-Score**
  \[
  F1 - \text{Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{15}
  \]

where $TP$ is the total number of pixels correctly segmented by building category in segmentation result; $TN$ is the total number of pixels correctly segmented by background category in segmentation result; $FP$ is the total number of pixels incorrectly segmented by building category in segmentation result, and $FN$ is the total number of pixels incorrectly segmented by background category in segmentation result.

To evaluate the complexity of the proposed model, multiply-accumulate operations (MACs) [38] and the size of parameters (Params) are utilized as a metric

\[
\text{MACs} = C_{\text{out}} \times H_{\text{in}} \times W_{\text{in}} \times (C_{\text{in}} \times K \times K \times \text{bias}) \tag{16}
\]

where $H_{\text{in}}$ and $W_{\text{in}}$ are height, width of input feature map, respectively; $C_{\text{in}}$ and $C_{\text{out}}$ are the number of channels of the input feature map and output feature map, respectively; $K$ is the height and width of kernel; bias is the bias of kernel.

C. Ablation Study

An ablation study is performed on the aerial imagery dataset to evaluate the effectiveness of each module of the proposed ReA-Net. precision, recall, IoU, F1-Score, MACs, and Params
metrics of several variant networks, were compared in the ablation study. All alternative models had the same related parameter settings and training strategies, and the semantic segmentation network from the literature [39] was used as the baseline network. Table III shows the configuration and description of each network version, and Table III also shows the comparison results of ablation experiments (the input network image resolution is set as $512 \times 512$).

It can be seen from the analysis of Table IV.

1) Compared with the Baseline network, the performance of the network is improved more significantly by adding the regional attention module, while the computational complexity and the size of model parameters are slightly increased. Baseline + RA1 improved over Baseline by 2.35% (IoU), 0.65% (Precision), 1.48% (Recall), and 1.33% (F1-Score); Baseline + RA1 + RA2 improved over Baseline by 2.53% (IoU), 0.73% (Precision), 1.68% (Recall), and 1.43% (F1-Score), respectively; Baseline + RA1 + RA2 + RA3 improved over Baseline by 2.70% (IoU), 0.77% (Precision), 1.69% (Recall), and 1.46% (F1-Score), respectively; Baseline + RA1 + RA2 + RA3 + RA4 improved over Baseline by 2.76% (IoU), 0.83% (Precision), 2.02% (Recall), and 1.66% (F1-Score). All these effects indicate that the addition of the region attention module enhances the attention to relevance and consistency of local regions, and improves the proposed network’s ability to focus on region and boundary information of the segmentation target building (see Section III-D Visualization Experiments for details).

2) Comparing the performance of the Baseline + RA1 + RA2 + RA3 + RA4 and the Base-line + RA1 + RA2 + RA3 + RA4 + MS with that of baseline + RA1 + RA2 + RA3 + RA4 + MS + RL, the latter has the region consistency constraint among pixels in remote sensing images by introducing the proposed neighborhood consistency loss function, which enhances the network’s ability to focus on building regions and boundaries (see Section III-D for details of the visualization experiments). Thus, the latter have more accurate segmentation results in the boundaries of the segmentation buildings. Compared with the baseline + RA1 + RA2 + RA3 + RA4 + MS, the latter’s IoU, precision, and F1-Score improve by 0.37%, 0.34%, and 0.1%, respectively.

Fig. 4 shows the variation of F1-score and IoU at different neighborhood sizes (NS). It can be seen that without neighborhood consistency supervision loss (NS = 0), the network performance on F1-score and IoU is poorer reaching 95.53% and 91.4%, respectively. After introducing neighborhood consistency supervision loss with NS = 2, the network performance on F1-score and IOU has significantly improved to 95.58% and 91.5%, respectively. The best network performance was achieved at NS = 3, with 95.64% and 91.62%, respectively. The ablation experiments presented in Fig. 4 show that increasing the NS size after NS = 3 does not significantly improve the performance of the network on F1-score and IOU, while increasing the NS size implies a significant increase in the computational complexity of the network. Therefore, through comprehensive consideration, NS = 3 is chosen in this article for the subsequent experiments.

D. Visualization

To further prove the effectiveness of the proposed ReA-Net, the feature maps of the RA1 and RA3 phases of the ReA-Net are visualized based on the aerial imagery dataset, and the visualization results are shown in Fig. 5. To demonstrate the effectiveness of the proposed region consistency supervision loss function, the penalty weights $\omega$ from the training is visualized and analyzed, and the results are shown in Fig. 6.

The selected test images contain scattered groups of buildings as shown in the first and third rows of Fig. 5, and large regions of buildings as shown in the second and fourth rows of Fig. 5, respectively. Scattered building clusters tend to suffer from a lack of attention to the regions, and edges of small targets at the larger scale of the remote sensing image, resulting in frequent undersegmentation of scattered small target regions. For large-region buildings, oversegmentation tends to occur when the color of the background regions is similar to that of the building regions to be segmented, and because the building regions to be segmented are large, the problem of poor consistency of
pixel classification in the local regions tends to occur due to noise, texture, and other interference factors, resulting in regionally mis-segmentation and blurred edges in the segmentation results.

As shown in Fig. 5, RA1 in ReA-Net directly processes the deep-level features, and since RA1 holds the larger regional perceptual field, it is capable of extracting richer semantic information from remote sensing images. Therefore, compared with the original Unet, RA1 in ReA-Net can extract multiscale region-level features of remote sensing images, expand the perceptual field of the network to the target building region, and establish the correlation representation between region descriptors in the region-level feature map employing the SA mechanism, which effectively captures the intraregional correlation and the correlation between different local regions in remote sensing images. Consequently improves the saliency of RA1 in extracting the features of the building regions to be segmented, and decreases the missing features of the building regions caused by the interference of the background regions of the deep feature map in the resolution recovery stage, enhancing the ability of the network to characterize the correlation between different regions of the remote sensing image and the ability to focus on the building regions to be segmented, so that the proposed network not only suppresses the features of the background regions of the remote sensing image but also effectively enhances the features of the building regions. RA3 in ReA-Net processes deeper features, and because RA3 has a smaller local field of perception, it is more capable of extracting rich spatial detail information of remote sensing images. Compared to the original Unet, RA3 in ReA-Net can focus precisely on the shape and edge features of the building regions to be segmented, while suppress background regions interference. Therefore, as shown in Fig. 5, the visualization results demonstrate that the proposed region attention mechanism can better capture the region and edge information of the buildings to be segmented for a variety of building distribution cases.

Fig. 6 depicts a heatmap of the penalty weights $w$ in the training stage, with blue regions representing small weights and red regions representing significant weights. The high-weight activation area is predominantly dispersed on the edges of the buildings and the inaccurate segmentation regions, as can be observed from the visualization results in Fig. 6. The suggested strategy improves the loss of building edges and inaccurate segmentation regions utilizing dot product weighting in the processing of training, and enhances the network’s attention to such problems. Therefore, the proposed regional consistency supervision loss function can effectively improve the edge extraction ability of the buildings to be segmented, and at the same time strengthen the network’s ability to constrain the image consistency of remote sensing.
In summary, the proposed region attention mechanism and region consistency supervision loss cannot only strengthen the network’s attention to the region to be segmented under different typical building distributions, but also further suppress the salience of the background region of the remote sensing image, so as to effectively extract the regional features of the building to be segmented and the information of the building edge features, and improve the accuracy of building segmentation result.

As illustrated in Fig. 7(a), compared with FCN, PSPNet, and Res-Unet, the proposed ReA-Net not only effectively extracts the spatial detail information of buildings, but also makes usage of the constraint of spatial consistency function in the local regions of remote sensing images, so that the ReA-Net can effectively solve the problem of inaccuracy segmentation caused by building adhesion, which more clearly segmented the adhesion of buildings in the image, and the segmentation edges are clearer. For the problem of very similar building color and background
Fig. 7. Qualitative comparison results of the aerial imagery dataset.

In summary, it can be seen from the comparative experimental results of the aerial imagery dataset, compared with the comparison networks, the proposed ReA-Net cannot only obtain higher quality segmentation results for the remote sensing building image segmentation in complex scenes, but also has strong robustness to the interference of challenging problems with adhesion, light-shadow interference, complex foreground-background color interference, and small targets.

Quantitative analysis. The quantitative analysis of the segmentation results of different networks on the aerial imagery dataset is shown in Table V. The metrics employed in the quantitative analysis are precision, recall, IoU, F1-score, MACs, and Params. The proposed ReA-Net on the aerial imagery dataset precision is 95.61%, recall is 95.68%, IoU is 91.6%, MACs is 235.16 G, and Params is 22.14 M. Rows 1 to 12 are the semantic segmentation quantification of the comparisons algorithms, such as FCN [40], SegNet [41], DeeplabV3 [42], PSPNet [43], Unet [39], Res-Unet [44], HR-Net [45], Deng et al. [25], DR-Net [21], Chen et al. [46], RSR-Net[48], and DS-Net [49], respectively. Compared with FCN, SegNet, DeeplabV3, PSPNet, Unet, Res-Unet, HR-Net, Deng et al., DR-Net, Chen et al., RSR-Net, and DS-Net, the experiments results showed that the precision of ReA-Net is improved by 4.86%, 3.13%, 2.1%, 1.15%, 1.26%, 1.13%, 2.94%, 0.64%, 1.31%, 2.36%, 2.39%, and 0.76%,
TABLE V

| Network       | IoU(%) | Precision(%) | Recall(%) | F1-score(%) | MACs(%) | Params(%) |
|---------------|--------|--------------|-----------|-------------|---------|----------|
| FCN[40]       | 82.46  | 90.75        | 90.00     | 90.39       | 160.86G | 15.31M   |
| SegNet[41]    | 85.20  | 92.48        | 91.55     | 92.00       | 160.56G | 29.44M   |
| DeeplabV3[42]| 87.42  | 93.51        | 93.06     | 93.20       | 64.61G  | 15.31M   |
| PSPNet[43]    | 88.68  | 94.46        | 93.55     | 94.00       | 201.37G | 53.58M   |
| Unet[39]      | 88.26  | 94.35        | 93.66     | 93.77       | 160.36G | 17.26M   |
| Res-Unet[44]  | 88.05  | 94.48        | 92.84     | 93.65       | 323.93G | 13.04M   |
| HR-Net[45]    | 86.51  | 92.67        | 93.66     | 93.16       | 90.78G  | 29.53M   |
| Deng et al.[25]| 90.29 | 94.94        | 94.81     | 94.90       | /       | /        |
| DR-Net[21]    | 88.3   | 94.3         | 94.3      | 93.8        | /       | /        |
| Chen et al.[46]| 89.39 | 93.25        | 95.56     | 94.4        | /       | /        |
| RSR-Net[48]   | 88.32  | 93.22        | 92.25     | 92.73       | 3.0G    | 2.90M    |
| DS-Net[49]    | 90.4   | 94.85        | 95.06     | 94.96       | /       | /        |
| Ours          | 91.6   | 95.61        | 95.68     | 95.64       | 235.16G | 22.14M   |

The bold entities indicate the best result of the comparison methods.

respectively; the Recall of ReA-Net is improved by 5.68%, 4.13%, 2.13%, 2.02%, 0.87%, 1.38%, 0.12%, 3.43%, and 0.62%, respectively; the IoU of ReA-Net is improved by 9.14%, 6.4%, 4.18%, 2.92%, 3.34%, 3.55%, 5.09%, 1.3%, 3.3%, 2.21%, 3.28% and 1.20%, respectively; the F1-score of ReA-Net is improved by 5.25%, 3.64%, 2.44%, 1.64%, 1.87%, 1.99%, 2.48%, 0.73%, 1.84%, 1.24%, 2.91%, and 0.68%, respectively. The computational complexity of ReA-Net is higher than that of FCN, SegNet, DeeplabV3, PSPNet, Unet, HR-Net, DS-Net, but smaller than Res-Net; the overall size of parameters of ReA-Net is higher than that of FCN, DeeplabV3, Unet, Res-Unet, DS-Net, but smaller than SegNet, PSPNet, and HR-Net.

In summary, it can be seen from the qualitative evaluation that the remote sensing building images in complex scenes not only have good segmentation results, but also has satisfactory robustness to various strong interference factors that often exist in remote sensing images. From the quantitative evaluation, the comparison results of the segmentation experiments on the aerial imagery dataset show that, although the computational complexity and the parameters size of the proposed ReA-Net algorithm are slightly higher or the same as those of the comparison algorithms, the proposed algorithm is more effective in terms of the more important effectiveness indicators, such as precision, recall, IoU, and F1-score. The proposed algorithm obtained the best result in the quantitative evaluation. Therefore, compared with the state-of-the-art networks, the proposed ReA-Net can effectively achieve more accurate segmentation results of remote sensing building images in complex scenes.

E. Massachusetts Dataset Experimental Results and Analysis

To further demonstrate the generalizability of the proposed ReA-Net, we also conducted a comparison experiment between ReA-Net and the state-of-the-art segmentation networks on the Massachusetts dataset. The comparison networks include FCN [40], SegNet [41], DeeplabV3 [42], PSPNet [43], Unet [39], Res-Unet [44], and HR-Net [45]. The comparison of partial segmentation results of various comparison algorithms is shown in Fig. 8. From left to right in Fig. 8, the first column is the input test remote sensing building image, and the second to eighth columns are the semantic segmentation results of FCN [40], SegNet [41], and DeeplabV3 [42], PSPNet [43], Unet [39], Res-Unet [44], and HR-Net [45], respectively. The ninth column is the semantic segmentation result of the proposed ReA-Net, and the ground truth is shown in the last column.

Qualitative analysis. The overall imaging quality of the Massachusetts building dataset is poor. There are the following four mainly cases:

1) the case where the buildings are small and sticky, as shown in Fig. 8(a);
2) the case where the features, such as insignificant building boundaries and colors due to lighting and shadows, as shown in Fig. 8(b);
3) the case where small building are densely distributed and have blurred images, as shown in Fig. 8(c);
4) the case where the foreground and background colors are relatively similar, as shown in Fig. 8(d).

The segmentation comparison results of the different networks are shown in Fig. 8. The proposed ReA-Net obtained
the best segmentation results compared to the comparative networks, such as FCN, PSPNet, and Res-Unet. As shown in the segmentation comparison results marked by the red box in Fig. 8, the proposed network not only obtains a smoother building segmentation region, but also a clearer contour of the building segmentation, which is closer to the ground truth map. This is because the proposed MRA mechanism and region consistency supervision strategy can better enhance the network’s ability to extract semantic features from the remote sensing images, and by constraining and supervising the building regions and boundaries in the remote sensing images, it can make the network enrich the degree of attention to the local regions and contours of buildings, which improves the segmentation result of the proposed network.

Quantitative analysis. The quantitative analysis results of different networks on the Massachusetts dataset are shown in Table VI. The indicators used in the quantitative analysis are precision, recall, IoU, F1-score, MACs, and Params. It can be seen from Table VI that the proposed ReA-Net algorithm on the Massachusetts dataset has a precision of 83.93%, a recall of 87.53%, an IoU of 74.77%, an MACs of 58.79 G, and a Params of 22.14 M. In Table VI, rows 1 to 10 are FCN [40], SegNet [41], DeeplabV3 [42], PSPNet [43], Unet [39], Res-Unet [44], HR-Net [45], Chen et al. [46], BRRNet [20], and DS-Net [49] for semantic segmentation quantification, respectively. The experiments comparison results show that, compared with FCN, SegNet, DeeplabV3, PSPNet, Unet, Res-Unet, HR-Net, Chen et al., BRRNet, and DS-Net, the precision of the proposed algorithm has improved by 6.37%, 17.87%, 5.49%, 4.17%, 3.57%, 6.96%, 4.42%, 0.59%, and −3.84%, respectively, and the recall of the proposed algorithm has increased by 7.11%, 4.78%, 4.84%, 5.29%, 3.13%, 4.95%, 5.25%, 1.38%, and 5.29%, respectively;
the IoU of the proposed algorithm has improved by 9.53%, 16.7%, 7.39%, 6.73%, 4.79%, 8.56%, 6.88%, 1.28%, 0.31%, and 0.98%, respectively; the F1-score of the proposed algorithm has improved by 6.73%, 12.22%, 5.18%, 4.69%, 3.35%, 6.02%, 4.82%, 0.97%, 0.33%, and 0.78%, respectively.

To sum up, compared with the comparison networks, it can be seen from the qualitative comparison that the proposed algorithm has the best segmentation results on the Massachusetts dataset, and can perform higher quality segmentation for remote sensing images with poor imaging quality. From the quantitative metrics, the proposed ReA-Net algorithm has obtained the best results on the aerial imagery dataset and the Massachusetts building dataset. Therefore, the comparative experimental results on the two datasets show that the proposed algorithm has satisfactory generalizability.

IV. CONCLUSION

Aimed at the challenge of low building segmentation accuracy caused by poor continuity of remote sensing image regions and blurred boundaries in remote sensing images, a semantics segmentation algorithm for remote sensing buildings based on MRA and neighborhood consistent supervision is proposed. The proposed algorithm employs the encoder to extract features, such as texture, boundary, and deep semantic of buildings in remote sensing images, and restores a resolution of extracted feature images through the progressive up-sampling strategy of the decoder. The proposed multiscale neighborhood extraction fusion strategy can effectively enhance the correlation and consistency ability of higher order neighborhoods, and then enhances the network’s ability to extract region and boundary feature information of the target to be segmented. Furthermore, by introducing the region consistency supervised loss, the network’s attention to region smoothness is strengthened, and the network’s sensitivity to building boundaries and the accuracy of pixel classification are enhanced. Furthermore, the effectiveness and robustness of the proposed algorithm are demonstrated by quantitative experiments, qualitative experiments, and ablation experiments on the publicly available datasets the aerial imagery dataset and the Massachusetts dataset.

REFERENCES

[1] I. Kotaridis and M. Lazaridou, “Remote sensing image segmentation advances: A meta-analysis,” ISPRS J. Photogrammetry Remote Sens., vol. 173, pp. 309–322, Mar. 2021.

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

[3] M. Irandi and P. Saeedi, “Three-dimensional polygonal building model estimation from single satellite images,” IEEE Trans. Geosci. Remote Sens., vol. 50, no. 6, pp. 2254–2272, Jun. 2012.

[4] W. Zhao, C. Persello, and A. Stein, “Building instance segmentation and boundary regularization from high-resolution remote sensing images,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2020, pp. 3916–3919.

[5] X. Li, X. Yao, and Y. Fang, “Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 10, pp. 3680–3687, Oct. 2018.

[6] J. Chen et al., “DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, no. 93, pp. 1194–1206, Nov. 2021.

[7] L. Huang, C. Chen, W. Li, and Q. Du, “Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors,” Remote Sens., vol. 8, no. 6, Jun. 2016, Art. no. 483.

[8] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, “Investigation of the random forest framework for classification of hyperspectral data,” IEEE Trans. Geosci. Remote Sens., vol. 43, no. 3, pp. 492–501, Mar. 2005.

[9] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[10] J. Edirwicckrema and S. Khorram, “Hierarchical maximum-likelihood classification for improved accuracies,” IEEE Trans. Geosci. Remote Sens., vol. 35, no. 4, pp. 810–816, Jul. 1997.

[11] G. M. Foody and A. Mathur, “A relative evaluation of multiclass image classification by support vector machines,” IEEE Trans. Geosci. Remote Sens., vol. 42, no. 6, pp. 1335–1343, Jun. 2004.

[12] I. Griniás, C. Panagiotaíkis, and G. Tziritás, “MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images,” ISPRS J. Photogrammetry Remote Sens., vol. 122, pp. 145–166, Dec. 2016.

[13] A. Abdollahi and B. Pradhan, “Integrating semantic edges and segmentation information for building extraction from aerial images using UNet,” Mach. Learn. Appl., vol. 6, Dec. 2021, Art. no. 100194.

[14] K. Zhao, J. Kang, J. Jung, and G. Sohn, “Building extraction from satellite images using mask R-CNN with building resolution regularization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2018, pp. 247–251.

[15] S. Sherstha and L. Vannesci, “Improved fully convolutional network with conditional random fields for building extraction,” Remote Sens., vol. 10, no. 7, Jul. 2018, Art. no. 1355.

[16] Z. Zhang and Y. Wang, “IoRNet: A common neural network for road and building extraction,” Remote Sens., vol. 11, no. 6, Mar. 2019, Art. no. 696.

[17] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, “Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 4287–4306, May 2021.

[18] W. Boonpook, Y. Tan, and B. Xu, “Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry,” Int. J. Remote Sens., vol. 42, no. 1, pp. 1–19, Sep. 2021.

[19] Y. Wang, H. Lv, R. Deng, and S. Zhuang, “A comprehensive survey of optical remote sensing image segmentation methods,” Can. J. Remote Sens., vol. 46, no. 5, pp. 501–531, Aug. 2020.

[20] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yuan, and C. Somali, “BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images,” Remote Sens., vol. 12, no. 6, Mar. 2020, Art. no. 1050.

[21] M. Chen et al., “DR-Net: An improved network for building extraction from high resolution remote sensing image,” Remote Sens., vol. 13, no. 2, Jan. 2021, Art. no. 294.

[22] Y. Zeng, Y. Guo, and J. Li, “Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning,” Neural Comput. Appl., vol. 34, no. 4, pp. 2691–2706, Feb. 2022.

[23] S. Ghaifarian, J. Valente, M. Van Der Voort, and B. Tekinerdogan, “Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review,” Remote Sens., vol. 13, no. 15, Jul. 2021, Art. no. 2965.

[24] X. Yang et al., “An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery,” ISPRS J. Photogrammetry Remote Sens., vol. 177, pp. 238–262, Jul. 2021.

[25] W. Deng, Q. Shi, and J. Li, “Attention-gate-based encoder–decoder network for automatic building extraction,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 24, no. 2, pp. 2611–2620, Feb. 2021.

[26] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, “Multistage attention residual network for semantic segmentation of fine-resolution remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 19, Mar. 2021, Art. no. 8009205.

[27] L. Ding, H. Tang, and L. Bruzzone, “LA-Net: Local attention embedding to improve the semantic segmentation of remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 1, pp. 426–435, May 2021.

[28] Z. Tong, Y. Li, Y. Li, K. Fan, Y. Si, and L. He, “New network based on unet++ and densenet for building extraction from high resolution satellite imagery,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2021, pp. 2268–2271.

[29] M. S. Dey, U. Chaudhuri, B. Banerjee, and A. Bhattacharya, “Dual-path U-Net for road and building segmentation from satellite images,” IEEE Geosci. Remote Sens. Lett., vol. 19, Sep. 2021, Art. no. 3004005.

[30] S. Liu, H. Ye, K. Jin, and H. Cheng, “CT-UNet: Context-Transfer-unet for building segmentation in remote sensing images,” Neural Process. Lett., vol. 53, no. 6, pp. 4257–4277, Aug. 2021.
W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1874–1883.

Z. Zheng, S. Du, H. Taubenböck, and X. Zhang, “Remote sensing techniques in the investigation of aeolian sand dunes: A review of recent advances,” Remote Sens. Environ., vol. 271, Mar. 2022, Art. no. 112913.

X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” Expert Syst. Appl., vol. 169, May 2021, Art. no. 114417.

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.

V. Mnih, Machine Learning for Aerial Image Labeling. Toronto, ON, USA: Univ. Toronto, 2013.

Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, “Building extraction from multi-source remote sensing images via deep deconvolution neural networks,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2016, pp. 1835–1838.

Y. Zang, C. Wang, Y. Yu, L. Luo, K. Yang, and J. Li, “Joint enhancing filtering for road network extraction,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 3, pp. 1511–1525, Mar. 2017.

V. Camus, C. Enz, and M. Verhelst, “Survey of precision-scalable multiply-accumulate units for neural-network processing,” in Proc. IEEE Int. Conf. Artif. Intell. Circuits Syst., 2019, pp. 57–61.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput.-Assist. Intervention, 2015, pp. 234–241.

H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1520–1528.

V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

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

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2881–2890.

Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 749–753, May 2018.

J. Shi et al., “High-resolution representations for labeling pixels and regions,” 2019. [Online]. Available: arxiv.org/abs/1904.04514

Z. Chen, D. Li, W. Fan, H. Guan, C. Wang, and J. Li, “Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images,” Remote Sens., vol. 13, no. 13, Jun. 2021, Art. no. 2524.

I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: https://openreview.net/forum/?id=Bkg6RCvY7

H. Huang, Y. Chen, and R. Wang, “A lightweight network for building extraction from remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, Nov. 2021, Art. no. 5614812.

H. Guo, X. Su, S. Tang, B. Du, and L. Zhang, “Scale-robust deep-supervision network for mapping building footprints from high-resolution remote sensing images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, no. 762, pp. 10091–10100, Sep. 2021.

Shengjun Xu received the Ph.D. degree in engineering from Xi’an Jiaotong University, Xi’an, China, in 2013. He is currently an Associate Professor with the School of Information and Control Engineering, Xi’an University of Architecture and Technology, Xi’an, China. His research interests include image processing and artificial intelligence and automation.

Bowen Deng received the B.E. degree from the Xi’an University of Architecture and Technology, Xi’an, China, in 2020, where he is currently working toward the master’s degree in control theory and control engineering.

His research interests include computer vision, deep learning, and remote sensing image processing.

 Yuebo Meng received Ph.D. degree in engineering from Xi’an Jiaotong University, Xi’an, China, in 2014.

She is currently an Associate Professor with the School of Information and Control Engineering, Xi’an University of Architecture and Technology, Xi’an, China. Her research interests include machine learning and intelligent building technology.

Guanghui Liu received the Ph.D. degree in engineering from the Xi’an University of Architecture and Technology, Xi’an, China, in 2014. He is currently an Associate Professor with the School of Information and Control Engineering, Xi’an University of Architecture and Technology. His research interests include machine learning and intelligent building technology.

Jiupiang Han received the B.E. degree in automatic control from Xi’an Jiaotong University, Xi’an, China, in 1977. He is currently a Professor with the School of Electronic Science and Engineering, Xi’an Jiaotong University. His research interests include intelligent perception theory for visual recognition, artificial intelligence and automated systems, and intelligent control theory for model simulation.