Urban building extraction from high-resolution remote sensing imagery based on multi-scale recurrent conditional generative adversarial network

Zhen Wang a,b, Nan Xu c, Buhong Wang b, Yaohui Liu d and Shanwen Zhang b

aSchool of Information Engineer, Xijing University, Xi’an, Shaanxi, China; bSchool of Information and Navigation, Air Force Engineering University, Xi’an, Shaanxi, China; cSchool of Marine Science and Engineering, Nanjing Normal University, Nanjing, Jiangsu, China; dSchool of Surveying and Geo-Informatics Shandong Jianzhu University, Jinan, Shandong China

ABSTRACT

Urban building extraction from high-resolution remote sensing imagery is important for urban planning, population statistics, and disaster assessment. However, the high density and slight boundary differences of urban building regions pose a great challenge for accurate building extraction. Although existing building extraction methods have achieved better results in urban building extraction, there are still some problems, such as boundary information loss, poor extraction effect for dense regions, and serious interference by building shadows. To accurately extract building regions from high-resolution remote sensing images, in this study, we propose a practical method for building extraction based on convolutional neural networks (CNNs). Firstly, the multi-scale recurrent residual convolution is introduced into the generative network to extract the multi-scale and multi-resolution features of remote sensing images. Secondly, the attention gates skip connection (AGs) is used to enhance the information interaction between different scale features. Finally, the adversarial network with parallel architecture is used to decrease the difference between the extracted results and the ground truths. Moreover, the conditional information constraint is introduced in the training process to improve robustness and generalization ability of the proposed method. The qualitative and quantitative analyses are performed on IAILD and Massachusetts datasets. The experimental results show that the proposed method can accurately and effectively extract building regions from remote sensing images.

1. Introduction

With the development of remote sensing technology, the spatial resolution of remote sensing images is constantly improved, which makes it convenient to obtain the high-resolution remote sensing imagery. Building extraction from high-resolution remote sensing images plays an important role in urban planning, population estimation, disaster monitoring, and smart city construction (Lu et al. 2018; Song et al. 2020; Zhou et al. 2021). Traditional building extraction methods mainly use manual mapping, which exhibit the problems of low efficiency and high cost, and cannot meet the real-time requirement (Alshehhi et al. 2017; Hu et al. 2020). Benefiting from the progress of computer vision and pattern recognition, many methods for natural optical image processing have been widely applied in remote sensing building extraction tasks. However, different illumination conditions, image acquisition angles, and building materials inevitably pose greater challenges to the accurate extraction of buildings from remote sensing images.

Previous studies of building extraction from remote sensing images mainly relied on the image basic features, such as spectral, shape, contour, texture, color, shadow, etc. In this field, some effective methods have been proposed. Zheng and Wang (2015) proposed an object-based Markov random field (OMRF) model for building extraction, which establishes a weighted region adjacency graph by region size and edge feature information, and then uses OMRF with a region penalty term to complete accurate building regions extraction. Zhang et al. (2016) proposed a building extraction method based on saliency analysis, which extracts the multi-scale texture and edge features of remote sensing images by Fourier transform and adaptive wavelet. The experimental results show that the method can effectively extract different size building regions. Zhong, Gao, and Zhang (2016) proposed a multi-scale and multi-feature normalized cut framework for remote sensing building extraction, which uses the proposed framework to construct super-pixel multi-scale maps to obtain the multi-scale spatial
features, and then extracted the building regions by N-Cut model. Xie and Zhou (2017) completed the building regions extraction by using the extended multi-resolution segmentation (EMRS) and back-propagation (BP) networks, where EMRS is used for multi-scale spatial resolution feature representation, and BP network is used for classification of pixel point with different building regions. Liu et al. (2017) proposed a local competitive superpixels segmentation method, which can effectively fuse the spatial resolution and multi-scale features of remote sensing images, and complete the accurate extraction of building regions. Liu et al. (2018) obtained the spectral, texture, and shape features of remote sensing images by non-subsampled contourlet transform, and the region merging criteria is constructed using different features, finally used the global optimization strategy to complete the building region extraction. Li et al. (2017) proposed a building extraction method based on mixed sparse representation, which slices the remote sensing image into combinations of subgraph with different components, then used the sparse representation to express different subgraph features, and uses SVM to complete the building region extraction. Yin, Zhang, and Karim (2018) proposed a building region extraction method combining fuzzy region competition and Gaussian mixture model, which define the class attributes of building region pixels by Gaussian mixture model, and the fuzzy clustering is used to suppress noise interference and extract building region information. Su (2019) used the scale-variable region merging method to complete the building extraction of different regions, which used Fuzzy C-mean clustering to obtain the spatial contextual features, and then used the region merging with scale factor to obtain the building regions in remote sensing images. Zhang et al. (2020) proposed a building extraction method using saliency feature analysis and multi-scale spatial feature extraction, which applied guided filtering and metric analysis to obtain spatial feature information, and then extracts building regions by global contrast calculation and saliency analysis. The above building extraction methods based on basic features have achieved some results. However, the results still exhibit some problems, such as boundary information loss and incomplete shape structure because of the insufficient extraction of deep semantic features and global spatial features contained in remote sensing images.

In recently years, benefit from the powerful feature extraction and representation ability of convolution neural networks (CNNs), it has been widely used in image classification, target detection, image segmentation, and target tracking (Jiang and Chi 2019; Liu et al. 2020a; Hang et al. 2021). The CNNs is a deep feedforward neural network model with local connection and weight sharing, which has strong local invariance to input information and can automatically obtain different scale feature information. Therefore, the use of CNNs can solve the problem of manually designing feature extractors in conventional building extraction methods. Existing CNNs-based building extraction methods can be summarized as the following categories, 1) the first method is based on image classification task with CNNs, in which a fixed-size image is putted into a CNNs model and predicts the classes of one or several pixels; 2) the second method is called object-oriented CNNs semantic segmentation which combines image segmentation with neural network classification; 3) the third method is called semantic segmentation and is based on fully convolution neural networks (FCNs).

The most recent studies on building extraction exclusively utilized the FCNs-based methods. Fully convolution neural networks (FCNs) is an end-to-end deep learning network for image semantic segmentation, which replaces the fully connected layers with the convolution layers, so that the final feature map contains pixel position information. Based on the idea of FCNs, many building extraction methods have been proposed, such as Xu et al. (2018) proposed a building extraction method combining deep residual network and guided filtering, which uses residual network to extract the multi-scale resolution features of remote sensing images, and then performs pixel-wise segmentation of building regions by guided filtering. Hui et al. (2019) proposed a multi-task U-Net model for building extraction, which uses the multi-feature skip connection to obtain the spatial resolution of remote sensing images, and the multi-task learning is used to merge the regional structure feature information of buildings. Tan, Xiong, and Yan (2020) proposed a building extraction framework combining multi-branch convolution and graph model, which uses the multi-branch convolution to extract deep semantic feature information of remote sensing images, and uses the graph model to optimize the building extraction results.; Shao et al. (2020) proposed a refined
residual network for building extraction, which used the encoder-decoder prediction module to obtain the multi-scale features of remote sensing images, and the residual refinement module is used to optimize the building extraction results. Zhang et al. (2020) proposed a local-global dual-stream network (DS-Net) that can obtain the global context feature information for building region extraction, which uses a dual-stream complementary approach to exchange information between different features to obtain better building extraction results. Xie et al. (2020) used the multi-feature convolutional neural network (MFCNN) to extract multiple feature information of the building regions, and then the morphological filtering is used for building extraction from high resolution remote sensing images. Wang and Li (2020) proposed a multi-scale residual neural network with shape representation regularization for building extraction, which uses residual connection and multi-scale dilated convolution to extract multi-level feature information of remote sensing images, and uses the regularization with shape representation to reduce the over-fitting phenomenon in the model training process. ZZhu et al. (2020) based on the multi-scale semantic segmentation network D-LinkNet and Markov model to extract the building region, which extracts the general contour of building regions by D-LinkNet, and then use the conditional Markov model to further refine the extraction results. Xu et al. (2021) proposed a building extraction method by combining super-resolution module and semantic segmentation module, which firstly uses the super-resolution module to recover low-resolution building regions into high-resolution images, and the end-to-end semantic segmentation module is used to obtain the building regions. Liu et al. (2021) proposed a multi-scale U-shaped CNN model for building extraction, which used the multi-scale fusion U-shaped network (MFUN) to obtain the multi-scale spatial features information, the region proposal network (RPN) is used to location the building regions, and the building extraction results are optimized by edge-constrained multi-task network (EMU-CNN). Abdollahi, Pradhan, and Alamri (2021) proposed an end-to-end remote sensing road extraction model based on UNet network, which used two different UNet structure to obtain different scale feature information contained in remote sensing images, and the loss function of focal loss weighted by median frequency balancing (MFB_FL) is used to solve the training data imbalance problem. Chen et al. (2021) proposed a novel building extraction model DR-Net, which uses encoder-decoder structure as the backbone network, and the densely connected convolution neural network (DCNN) and residual network (ResNet) are used to obtain global features and local details features contained in remote sensing images. Although existing FCNs-based building extraction methods have achieved better results, they cannot fully extract the multi-scale feature information and spatial feature information contained in remote sensing images, so that the extracted results have different degrees of boundary blur and contour information loss phenomenon.

To solve the problem that the existing methods cannot fully extract the building region in remote sensing images, in this study, we propose an urban building extraction method based on multi-scale recurrent conditional generative adversarial network (MSRGAN), which can extract the multi-scale features and global spatial features contained in remote sensing images, and obtain more accurate building extraction results. The propose method consists of generative network and adversarial network, where the generative network is an end-to-end encoder-decoder structure, and the adversarial network is a parallel symmetric structure. Specially, the generative network uses recurrent residual convolution to obtain multi-scale feature information and spatial resolution information of remote sensing images, and enhances the interaction between different convolution features by attention gates skips connection operations; the adversarial network optimizes the difference between the building extraction results of the generative network and the ground truth, which further guides the model training. Moreover, to improve the model training speed and robustness, we propose a novel adversarial training strategy, which introduces conditional information in the training process, making the model training process more stable. The main contributions of our work are summarized as follows,

1) To solve the existing problems of CNN-based remote sensing building extraction methods, we construct a novel generative adversarial network framework for building extraction, in which the generative network can obtain the preliminary building extraction results, and the adversarial network fine-tunes the generative network to output more accurate extraction results.
2) To alleviate the interference of non-building regions features in CNNs feature extraction process, we propose a recurrent residual convolution module, which can directly extract the features of building at different scales and suppress background feature information interference in the convolution process.

3) To enhance the interaction transmit between different convolution layer features, we propose an attention gates skip connection mechanism, which can enhance the model nonlinear learning ability, and effectively fuse the building regions features at different scales and spatial resolutions.

4) To assess the performance of the proposed method, we perform a wide range of comparisons between different building extraction methods. The experiments on the IAILD and Massachusetts dataset show that our method is far more efficient than other state-of-the-art methods.

The remainder of this work is organized in the following manner. Section 2 and Section 3 describe the related works and proposed approach, respectively. Section 4 presents the experimental results and details analysis, which lead to the conclusion in Section 5.

2. Related work

2.1 Residual network for building extraction

To improve the feature extraction ability of traditional CNNs, He et al. (2016) proposed a residual network model, which uses residual connection to make CNNs have a deep network structure, thus the model can obtain multi-scale features of input data. In the task of extracting building from remote sensing images, many methods based on residual network have been proposed. Zhang, Liu, and Wang (2015) combined residual learning and U-Net proposed a lightweight neural network model for building extraction, which uses residual learning to accelerate the model training process, and uses skip connection for information interaction between different features. Ererapu et al. (2019) proposed a dense refinement residual network for building extraction, which uses dense refinement modules with different scales to extract multi-scale and multi-resolution features of remote sensing images, and transmission features by multiple skip connection operations. To improve the generalization ability of CNNs model, Meng et al. (2019) proposed a multi-path residual network model, which uses multiple residual functions to widen the model structure to obtain the multi-scale features of remote sensing images. Liu et al. (2020b) proposed an end-to-end building extraction model based on residual network, which uses generalized residual learning to obtain deep semantic features of remote sensing images, and reduces the model computational parameters by deep separable convolution. Ding and Bruzzone (2021) proposed a direction-aware residual network for building and road extraction, which makes full use of the feature extraction ability of the residual network, and uses the building and roads as direction-aware constraints to obtain better extraction results. Although the building extraction method based on residual network can obtain the multi-scale features of remote sensing images, it ignores the transfer and interaction between different feature information. To enable the model to fully obtain the multi-scale features information, we propose a novel recurrent residual convolution module, which enhances the feature information transfer and interaction by a recurrent neural network and attention mechanism.

2.2 GAN for building extraction

The traditional CNN model ignores the interrelationship between pixels in the process of semantic segmentation, while GAN can extract the interrelationship between pixels by adversarial training and obtain better semantic segmentation results. Based on the advantages of GAN, many remote sensing extraction methods have been proposed. Lin et al. (2017) proposed a multilayer feature matching generative adversarial network (MFM-GAN) framework for remote sensing image semantic segmentation, which uses feature matching to generate pseudo-labels to expand the original dataset, and then uses feature fusion to obtain the global feature information, so achieve better semantic segmentation results. Pan et al. (2019) proposed a building extraction method combining attention mechanism and generative adversarial network (AM-GAN), which obtains multi-scale features by spatial and channel attention mechanism, and uses the GAN to accurately extract the building region from the remote sensing images. To address the problem of large intra-class variations in remote sensing images, Sun et al. (2021) proposed an orthogonal generative adversarial
network (O-GAN), which uses generative network to obtain multi-scale feature vectors of remote sensing images, and then uses discriminative network to extract the building regions. To adaptively obtain the multi-level feature information of remote sensing images, Liang, Bao, and Shen (2021) proposed a multi-scale adaptive feature fusion generative adversarial network (MSAFF-GAN), which introduced the scale attention module in the generative network to obtain the deep semantic feature information, and uses the adaptive training strategies to optimize the model training process. Guo, Xia, and Luo (2021) proposed a GAN framework with similarity loss constraint for remote sensing images semantic segmentation, which uses the gated self-attention mechanism to suppress background information interference, and constructs a generative network with pyramid residual structure to obtain better semantic segmentation results. Existing GAN models exhibit some problems of insufficient feature extraction and over-fitting in the remote sensing images segmentation, hence we introduce the conditional information constraints in the GAN to speed up the model training and improve the robustness.

3. Methodology
The proposed MSR-cGAN includes the generative network and the adversarial network, where the generative network is used to generate building extraction results, and the adversarial network is used to correct the difference between extraction results and ground truths, the structure of MSR-cGAN is shown in Figure 1.

The generative network is an encoder-decoder architecture, where the encoder module is used to extract the remote sensing image multi-scale features, and the decoder module is used to restore the image resolution size. Specifically, the encoder module includes five recurrent residual convolution modules, each of which contains $3 \times 3$ convolution, $1 \times 1$ convolution, and max-pooling operations; the decoder module includes four deconvolution modules, each of which contains deconvolution and up-sampling operations; the attention gates skip connection is used for feature transfer between encoder module and decoder module.

The adversarial network is a dual-branch parallel structure, and each branch contains two down-sampling modules and three convolution modules.
Specifically, the input of adversarial network is the predicted results of generative network, original remote sensing image, and ground truths; the down-sampling module contains down-sampling layer, GroupNorm layer, and LeakyReLU; the convolution module contains $3 \times 3$ convolution layer, GroupNorm layer, and LeakyReLU operation; the output result of each branch uses $1 \times 1$ convolution layer for feature fusion, and uses L1_Loss to calculate the error between predicted result and ground truths.

### 3.2 Generative network

The generation network in MSR-cGAN is constructed on the basis of U-Net, and to make the generation network have better adaptive capabilities, the end-to-end network structure of U-Net is retained. Due to U-Net has deep layers and complex structure, direct processing of remote sensing images will cause gradient disappearance, gradient explosion, and over-fitting problems (Ronneberger, Fischer, and Brox 2015). Moreover, because of the building region in the remote sensing images with complex background interference, U-Net cannot suppress such interference information. The proposed generation network is an encoder-decoder structure, in which the encoder module fully extracts the multi-scale feature information of the input image, and the decoder module alleviates the background noise interference to restore the target area segmentation results. The structure of generation network is shown in Figure 2.

To solve the problem that the building extraction accuracy decreases due to the increase of network layers, the recurrent residual convolution module is introduced into the generation network to enhance the feature information propagation and feature map reusability to further improve the model performance. Moreover, to integrate feature information more effectively, the attention gates skip connection (AGs) is added on the basis of original UNet skip connection, which can ensure that the feature information weight of the remote sensing building region is increased while reducing the feature information weight of the background region. As shown in Figure 2, the generative network includes encoder module, AGs skip connection, and decoder module. Specially, the encoder module consists of four sets of down-sampling layers and convolution layers, each down-sampling layer

---

**Figure 2.** The structure of generative network, including convolution, max-pooling, up-sampling, deconvolution, recurrent residual convolution, and attention gates skip connection (AGs).
contains two parallel channels, the size of convolution layer is $3 \times 3$, the feature channel of initial convolution layer is 64, and its convolution operation is recurrent residual convolution; the decoder module is composed of four sets of up-sampling layers and convolution layers, its structure, and convolution layer parameters are the same as the encoder module, and the last layer of decoder module is $1 \times 1$ convolution layer to output the building extraction results; the AGs skip connection operation is set between the encoder module and decoder module, which can fuse image feature information and reduce the impact influence of background feature on the building region.

### 3.2.1 Recurrent residual convolution module
To extract the multi-scale feature and enhance feature transfer, recurrent convolution (Liang and Hu 2015; Alom et al. 2018) improves the model feature representation ability for different target regions by extracting the temporal correlation. Inspired by previous studies on recurrent convolution and residual network (He et al. 2016), we propose a recurrent residual convolution module (RRCM) for remote sensing building image feature extraction, which can effectively extract multi-scale feature and enhance information transmission between different features. The residual learning is proposed on the basis of CNNs, and the basic idea is to assume that the input feature map is $x$ and the function map of the network is $H(x)$, then the residual map output of the network fitting is $H(x) - x = F(x)$, where $F(x)$ represents the overall residual network. Using the residual mapping method can solve the model degradation problem of the neural network as the number of layers increases, and can avoid the gradient disappearance and gradient explosion phenomenon. The recurrent residual convolution module accumulates the convolution feature map of previous layer and current layer, and the result is used as the input of the next layer. Assuming that feature map $X_l$ is the output pixel value obtained by linear transformation and nonlinear activation function at the $l-1$ network layer, the output of recurrent residual convolution module can be expressed as the following,

$$ Y_l = H_l(W_l^T \ast X_l + W_{l-1}^T \ast X_{l-1}, b) $$

where, $H(\cdot)$ represents the nonlinear network mapping of the $l$–th layer; $W_{l-1}$ and $W_l$ are the weight matrix of the input feature in the $l-1$ convolution layer and $l$ convolution layer; $b$ represents the bias term.

The recurrent residual convolution structure can strengthen the features transmission, merges the different layer features, and efficiently uses the output feature map, which is beneficial to the feature extraction of the building region. The use of recurrent residual convolution can learn multi-scale features of different receptive field, and the residual learning method can further the feature use efficiency. The recurrent residual convolution module structure is shown in Figure 3, which includes three feature extraction modules composed of convolution, batch normalization, and ReLU activation functions. The input feature map is performance recurrent convolution and merges in the feature extraction module, and then input to the $1 \times 1$ convolution layer for compression, which can solve the problem of feature information and network parameter redundancy caused by the convolution operation. Finally, the residual structure is used to output results, which can solve the performance degradation caused by the large number of network layers, and can accelerate the network convergence speed.

### 3.2.2 Attention gates skip connection
The shape, size, and contour of the remote sensing building region are complex and diverse. Therefore, the pixel enhancement of the remote sensing image
can better eliminate the background noise interference during the building extraction process. In UNet model, the skip connection between encoder and decoder network can enhance the feature information transfer. This method of copy and crop corresponding hierarchical feature maps can realize the information integration in the sampling process, so that the network can better learn the relationship between the coarse-grained location information and global information. The attention mechanism can make the model focus on feature extraction of target region and suppress background region information interference. To solve the problem of original skip connection, inspired by Attention U-Net (Oktay et al. 2018), we introduce AGs into the skip connection operation, reducing the parameter redundancy, and suppressing the background complex feature response.

\[ x^{(l)j}_{ij} = a^{(l)j}_{ij} \cdot x^{(l)j}_{ij} \]  

(2)

where, \( x^{(l)j}_{ij} \) represents the feature mapping of the input; \( l \) represents the number of network layers of the feature input; \( i \) represents the channel scale; \( j \) represents the pixel space size; \( a^{(l)j}_{ij} \in [0,1] \) is the attention coefficient corresponding to the input feature.

In remote sensing image building extraction tasks, there are many building categories, so AGs skip connection is a multi-dimensional operation, and each AGs needs to focus more on the building region. To obtain the focus position and context information on each channel pixel, the additive attention mechanism is introduced in the AGs skip connection to select feature channels and improve the model feature mapping ability. The formal description of the AGs skip connection operation that introduces the additive attention mechanism is as follows,

\[ v^{(l)j} = W_Q^T x^{(l)j}; k^{(l)j} = W_K^T q^{(l)j} \]  

(3)

\[ y^{(l)j} = \lambda(F(v^{(l)j}) + k^{(l)j}, b_x), b_y) \]  

where, \( a^{(l)j} \) represents the attention selection channel vector; \( W_Q \) is the weight vector of the input \( x^{(l)j} \); \( W_K \) represents the weight vector of the down-sampling selection channel vector \( q^{(l)j} \); \( \lambda \) represents the convolution function; \( b_x \) and \( b_y \) represent the bias term; \( F(\cdot) \) represents the Parametric Rectified Linear Unit (PReLU) activation function; \( G(\cdot) \) represents the sigmoid activation function.

The attention gates skip connection structure is shown in Figure 4, the remote sensing image outputs a sparse matrix after the up-sampling pooling layer, and it is decoded by different deconvolution operations. The attention mechanism is used to analyze the information of different channels, and get the matching attention coefficients. The shallow convolution layer of the AGs module is used to extract the contour and morphological feature of the remote sensing target area and the deep convolution layer is used to extract the texture features, so that the generation network pays more attention to the target area feature, and weakens the interference of background noise on the segmentation results. The AGs unit transfers the feature information of decoder network by skip connection in the encoder network, and directly cascades the feature information of the encoder network to its corresponding decoder network, then uses the \( 1 \times 1 \) convolution layer to perform linear change, which effectively restores the feature information loss caused by down-sampling.

### 3.3 Adversarial network

In the MSRCGAN model, the adversarial network is used to discriminate the error between the generation network prediction result and the ground trues.

![Figure 4. The architecture of attention gates (AGs) skip connection.](image-url)
In the remote sensing image building extraction task, the optimization goal of generation network is to minimize the generation loss, making it difficult for the adversarial network to distinguish the difference between the extraction result and ground truths; the optimization goal of adversarial network is to minimize the discriminative loss and maximize the difference between the prediction result and ground truths. In the model training process, the adversarial network will gradually improve the discriminative ability and guide the generation network for training. Therefore, the adversarial network is equivalent to a trainable loss function, which can calculate the difference between the generated result and the real data from the deep data distribution, so the adversarial network can well supervise the generation network training. The adversarial network of MSR-cGAN is an encoder-decoder structure similar to the generation network, which makes it easier to train, and can avoid the training collapse of the generation network caused by the strong discriminative ability of the adversarial network. Moreover, in order to prevent over-fitting of the adversarial network, the constructed model uses a simple down-sampling pooling layer and convolution layer, and the structure of adversarial network is shown in Figure 5.

The input of adversarial network is an image pair composed of the prediction result output by generation network, ground truths, and original remote sensing image. In the deep image classification network, the output features of the encoder structure will be flattened into feature vectors, then the full connection layer will be used to output the classification results, so it can only input images of a fixed size. Because the adversarial network structure is a two-class model, to processing multi-scale images and stabilize the training process of the adversarial network, we use the Patch-GAN method in Pix2Pix for model training. The constructed adversarial network removes the feature flattening operation and full connection layer, directly allows the network to output a single-channel feature map, and then uses the discriminative feature of generated data and real data for feature matching. In this way, the model can better train the generation network and enhance its performance. The commonly used feature matching loss functions are the regularization loss L1 function and L2 function. Compared with the L2 loss function, the L1 loss function has a built-in features selection and has more uncertain solutions, so it has better robustness. Therefore, the L1 loss function is used in the MSR-cGAN adversarial network to perform feature matching on the discriminative feature of generated data and real data.

As shown in Figure 5, the adversarial network uses two down-sampling pooling layers with the same structure. Each down-sampling layer uses a 4 × 4 pooling kernel with a stride size of 2, and padding operation is performed on the input features before the down-sampling operation. In this way, the down-sampling pooling feature can be effectively merged, and the output feature can be sparsity, to avoid the problem of detail feature information loss caused by multiple down-sampling operations. Moreover, the adversarial network uses three convolution modules with the same structure, and each convolution module uses a convolution kernel with a size of 3 × 3 and stride of 1 for feature extraction, the padding operation with size of 2 is performed on the input features before the convolution operation. The use of the constructed convolution module for feature

![Figure 5. The architecture of adversarial network.](image-url)
extraction can effectively fuse multi-scale feature without changing the feature size. The output layer of the adversarial network is a single convolution layer with a kernel size of 1 × 1 and stride size of 1, which can ensure that the output feature layer will not be down-sampling, and the output coefficient feature can be better matched with the L1 loss function.

The adversarial network in MSR-cGAN uses Leaky Rectified Linear Unit (LeakyReLU) as the activation function, which is different from the ReLU function, when the input feature tensor value is less than zero, LeakyReLU will output a leak value, so it can avoid the neuron inactivation phenomenon. The formal description of LeakyReLU activation function is as follows,

\[
f(x) = \begin{cases} 
  x, & x > 0 \\
  \varepsilon x, & x \leq 0 
\end{cases}
\]

where, \(\varepsilon\) represents the leak value, and the leak values in adversarial network are set to 0.2.

### 3.4 Loss function

The proposed MSR-cGAN performs model training by adversarial learning approach. GANs include generation network and adversarial network, while in MSR-cGAN, the generation network is used to generate the predicted labeling map to spoof the adversarial network, and the adversarial network is used to distinguish the different between the predicted labeling maps generated by the generation network and the ground truths. Due to the large difference between the target area and the background area in the remote sensing image, to make the model training process more stable, the introduction of conditional information between the generation network and the adversarial network can stabilize the model training process, and make the generate results more closer to the ground truths (Mirza and Osindero 2014). Assume that the \(x\) represents ground truths, \(y\) represents the input image, \(z\) represents the generate result, and \(x\) represents random noise. MSR-cGAN learns the mapping from conditional information \(x\) to random noise \(z\) to output result \(y\). Moreover, the addition of conditional information makes the generated network results have ground truths information. The formal description of MSR-cGAN is as follows,

\[
GAN(G, D) = E_{xy}[\log D(x, y)] + E_{xz}[\log(1 - D(x, G(x, z)))] 
\]

where, \(G\) represents the generation network, \(D\) represents the adversarial network, \(E(\cdot)\) represents the expected value of distribution function, \(D(x, y)\) represents the probability that the input image is ground truths, \(D(x, G(x, z))\) represents the probability that the input image is generate image.

The L1 distance loss function is introduced into the adversarial loss function, which can better improve the model learning process. The L1 loss function is expressed as follows,

\[
L_{L1}(G) = E_{xy}[||y - G(x, z)||_1] 
\]

Combining the MSR-cGAN loss function with the L1 loss function, the final loss functions of generation network and adversarial network are as follows,

\[
G^* = \arg \min_G L_{gen}(G, D) = \arg \min_G \max_D L_{MSR-cGAN}(G, D) + \lambda L_{L1}(G) 
\]

\[
D^* = \arg \min_D L_{dis}(G, D) = \arg \min_D \max_G L_{MSR-cGAN}(G, D) 
\]

where, \(G^*\) represents the loss function of generation network, \(D^*\) represents the loss function of adversarial network, \(\lambda\) represents penalty coefficient.

### 4. Experiments

#### 4.1 Datasets

To verify the effectiveness of the proposed MSR-cGAN, two publicly available building extraction datasets IAILD (Inria Aerial Image Labeling) dataset and Massachusetts building dataset are used in this study. These two building extraction datasets contain multiple category building characteristics, such as size, contour, shape, distribution area, and spatial resolution, which can effectively verify the generalization ability of the proposed method.

The IAILD dataset contains 810 km² remote sensing images with a spatial resolution of 0.3 m, of which 405 km² buildings have been manually labeled. The dataset is comprised of multiple orthographic aerial red-green-blue (RGB) bands images in Austin, Chicago, Kitsap, Western Tyrol, and Vienna five areas, and the building styles of each area are significantly different, some samples are shown in
Figure 6. IAILD dataset contains 180 remote sensing images with a resolution of 5,000 × 5,000 pixels, to speed up the model training and improve the generalization ability, the dataset images are cropped into 4,500 remote sensing with a resolution of 1,000 × 1,000 pixels. In the experiment, we divided 4,500 dataset images into train dataset, validation dataset, and test dataset. Specially, the train dataset used for model training contains 3,150 images, the verification dataset for optimizing model performance contains 900 images, and the test dataset contains 450 images.

The Massachusetts remote sensing building image dataset covers a 500 km² surface area, from suburban areas to rural areas, which included ground object such as roads, oceans, buildings, tree, bridges, and vehicles. The remote sensing image is comprised of red-green-blue (RGB) bands, the image size in the dataset is 1,500 × 1,500 pixels, and the spatial sampling resolution of each pixel is 1.2 m, some dataset samples are shown in Figure 7. The original Massachusetts dataset contains 1,171 images, and to improve the model generalization ability, a two-stage data expansion method is used to expand the dataset. Specially, in the first stage, four fixed-position regions of original remote sensing are extracted, and then 25 images are randomly cropped using a sliding window with size of 300 × 300; in the second stage, the randomly cropped image is flipped and translated in the horizontal direction to obtain the expended dataset. The expanded dataset contains 5,540 remote sensing images, using 3,878 images as the train dataset, 1,108 images as the validation dataset, and 554 images as the test dataset.

4.2 Evaluation metrics

To compare the algorithm performance more fairly, different quantification indicators are used for the IAILD dataset and the Massachusetts dataset. Due to the dataset image only contains the building and non-building region, it can be regarded as pixel two classification problems, so the intersection-over-union (IoU), recall, precision, and F1_measure are used as evaluation metrics. The calculation of different evaluation metrics are as follows,

1) IoU is commonly used to evaluate image semantic segmentation result by measuring the similarity between the predicted result and the ground truth. IoU is calculated as Eq. (11):

\[
\text{IoU} = \frac{TP}{FN + TP + FP}
\] (11)

2) Recall represents the ratio of pixels classified as buildings to truth building pixels. The calculation of recall is shown in Eq. (12).

\[
\text{recall} = \frac{TP}{TP + FN}
\] (12)

3) Precision represents the ratio of pixels that are correctly classified as buildings to all pixels that are classified as building. Precision is calculated as Eq. (13).

\[
\text{precision} = \frac{TP}{TP + FP}
\] (13)

4) F1_measure is a comprehensive evaluation metric of recall and precision. The calculation of F1_measure is shown in Eq. (14).

Figure 6. Sample image from the IAILD dataset. (a) Original image. (b) Ground truth.
where, $TP$ represents the overlap between the building region in ground truth and the building region extracted by algorithm; $FP$ represents the overlap between the background region in ground truth and the building region extracted by algorithm; $TN$ represents the overlap between the background region in ground truth and the background region extracted by algorithm; $FN$ represents the overlap between the building region in ground truth and the background region extracted by the algorithm.

4.3 Implementation details

The proposed MSR-cGAN is constructed by PyTorch framework on a PC with Ubuntu 18.04 operation system, and the NVIDIA GeForce RTX 3090 GPU with 24GB memory is used for acceleration. In the training phase, the random gradient descent (SGD) is used to optimize the weight. The batch size is set to 24, and the epoch is set to 200. In the first 150 epochs, the learning rate is fixed to $10^{-4}$, and then the learning rate is automatically adjusted by exponential decay. The alternate training method is used to train MSR-cGAN model. The parameters of the generation network and adversarial network are initialized by normally distributed random variables. The initial learning rate is set to $10^{-3}$ and divided by 2 every 15 epochs, and the batch size is set to 64. The alternate training method is used to train MSR-cGAN model. The pseudo code of alternate training is shown in Table 1.

4.4 Hyper-Parameters analysis

To analyze the rationality and effectiveness of using L1 distance loss in the model loss function, we compared the effects of different loss functions on the model training process and building extraction precision. The compared loss functions include L1 distance loss, L2 distance loss, and cross entropy loss. The influence of different loss functions on the model performance is show in Figure 1, where Figure 8(a) shows the influence of different loss functions on the model training process, from which it can be seen that the L1 loss can make the model obtain a relatively stable training process, and it obtains the lowest loss value under 200 epochs. Figure 8(b) shows the influence of different loss functions on the building extraction accuracy of the model, from which it can be seen that with the increase of epochs, the building extraction precision of the model gradually increases, while the L1 loss has the fastest growth rate and enables the model to obtain the optimal precision at 180 epochs. The reason for using L1 loss can make the model obtain better results is that L1 loss is more robust to outliers, while there are a lot of noise data in the training process of generative adversarial network. Therefore, introducing L1 loss can make the model obtain better training effect.
Table 1. The pseudo code of alternate training in MSR-cGAN.

Algorithm: MSR-cGAN Training Process
1: **Input**: Training Dataset (original images, ground truths)
2: **Output**: Trained Generator model
3: for *i in* range (Epoch):
4: D.trainable = False
5: for *j in* range (Steps_per_Epoch):
6: real_image, real_ground_truths = next(Data_augmentation)
7: G_x_batch = concatenate(real_image, real_ground_truths)
8: G_y_batch: all = 1
9: GAN.train_on_batch(G_x_batch, G_y_batch)
10: D.trainable = True
11: for *j in* range(Steps_per_Epoch):
12: real_image, real_ground_truths = next(Data_augmentation)
13: real_input = concatenate(real_image, real_ground_truths)
14: fake_input = concatenate(real_image, G(real_image))
15: D_x_batch = concatenate(real_input, fake_input)
16: D_y_batch: real_input = 1, fake_input = 0
17: D.train_on_batch(D_x_batch, D_y_batch)

4.5 Comparison to state-of-the-Art methods

4.5.1 Inria aerial image labeling dataset

To evaluate the algorithm performance on the Inria Aerial Image Labeling Dataset (IAILD), we compare the proposed MSR-cGAN with other state-of-the-art method, which includes the TreeUNet (Yue et al. 2019), and BRR-Net (Shao et al. 2020). Moreover, we compare the GAN-based remote sensing image building extraction method including Generative Adversarial Network with Spatial and Channel Attention Mechanisms (GAN-SCA) (Pan et al. 2019), Bayesian Segmentation Network (BAS-Net) (Xiong et al. 2020), and NAS-Net (Xu et al.). Specifically, TreeUNet uses residual neural network for multiscale feature extraction, and constructs the model weight parameters adaptively by Tree-Cutting algorithm. BRR-Net is a classical FCNs structure, which uses the dilated convolution to strengthen the model feature extraction ability, and uses the residual refinement module to refine the building extraction results. GAN-SCA introduces spatial and channel attention mechanisms on the basis of GAN, the generation network introduces spatial and channel two attention mechanisms, which increases the feature extraction ability for remote sensing images, and the adversarial network introduces a channel attention mechanism to enhance the model discriminate ability. Based on the Bayesian theory, BAS-Net inputs the segmentation results of FCNs as prior knowledge into GAN for remote sensing image segmentation, which effectively avoids the over-fitting problem during model training. NAS-Net consists of feature resolution module and semantic segmentation module, where the

![Figure 8](image.png)

**Figure 8.** The influence of different loss functions on the model performance. (a) Training process comparison. (b) Extraction accuracy comparison.
feature resolution module is used to obtain the deep semantic feature information and the semantic segmentation module is used to output the extraction results of different region building.

Table 2 and Figure 10 show the building extraction results of different methods on the IAILD dataset. Due to the relatively simple model structure of TreeUNet and GAN-SCA, which cannot fully extract the remote sensing building regional features, so the F1_measure is 0.8146 and 0.8257, respectively, and it can be seen from the visualization results in Figure 10 that large areas of buildings in remote sensing image have not been accurately extracted. The precision, recall, and F1_measure of BRR-Net are 0.9202, 0.8643, and

| Methods      | TreeUNet | GAN-SCA | BRR-Net | BAS-Net | NAS-Net | MSR-cGAN |
|--------------|----------|---------|---------|---------|---------|----------|
| Precision    | 0.8574   | 0.8673  | 0.9202  | 0.9318  | 0.9425  | **0.9541** |
| Recall       | 0.7914   | 0.8025  | 0.8643  | 0.8724  | 0.8835  | **0.8972** |
| IoU          | 0.7652   | 0.7738  | 0.8246  | 0.8357  | 0.8462  | **0.8615** |
| F1_measure   | 0.8146   | 0.8257  | 0.8913  | 0.9011  | 0.9126  | **0.9247** |

Figure 9. The influence of conditional information on the model performance. (a) Training process comparison. (b) Extraction accuracy comparison.

Figure 10. Building extraction results of different methods in IAILD dataset. Yellow boxes represent the best performance; Red boxes represent a relative worse performance.
respectively. Due to the introduction of attention mechanism, the multi-scale features of building area can be better extracted, and it can be seen from the visualization results that BRR-Net can extract a large area of buildings region, but the result of extraction on the buildings boundary is poor. NAS-Net can extract the contour and boundary of the building region better, but the extraction effect of dense building area is poor, and the precision, recall, and F1_measure are 0.9245, 0.8835, and 0.9126, respectively. The proposed MSR-cGAN is superior to the compared methods in quantitative indicators, from the Table 2 it can be seen that the precision, recall, and F1_measure are 0.9541, 0.8972, and 0.9247, respectively. The visualization results in Figure 8 shows that MSR-cGAN accurately extracts the building region of the remote sensing image, and can accurately complete the extraction of building in dense region. Figure 11 shows the ROC curve and PR curve of different methods, which can be seen that the performance of TreeUNet and GAN-SCA are obvious inferior to BRR-Net, BAS-Net, and NAS-Net, while the results of MSR-cGAN are better than other methods, shows the effectiveness of the proposed method.

4.5.2 Massachusetts Dataset
We further verify the performance of the proposed algorithm MSR-cGAN on the Massachusetts dataset. Different from IAILD dataset, the Massachusetts dataset contains more challenging scenarios. Table 3 shows the quantitative results of different methods on the Massachusetts dataset. Figure 12 shows the visualization results of different methods on Massachusetts dataset. The building extraction results of proposed MSR-cGAN are better than that of compared algorithms, and compared with the GAN-based optimal model NAS-Net, the evaluation indicators precision, recall, and F1_measure are higher by 0.0189, 0.0191, and 0.0155, respectively. It can be seen from the visualization results in Figure 12 that the TreeUNet, GAN-SCA, and BRR-Net can extract the contour information of the building region, but the boundary details of the extract results are relatively rough and cannot completely segment the building region. The building extract results of BAS-Net and NAS-Net are better than the other methods, where the BAS-Net can completely extract the building region, and NAS-Net can better optimize the boundary details of the building extract results. However, neither BAS-Net nor NAS-Net can accurately segment buildings in dense area. The proposed MSR-cGAN is not only able to extract the building region completely, but also can better extract the boundary detail information. Figure 13 shows the ROC curve and PR curve of different methods on the Massachusetts dataset. It can be seen from Figure 13 that the performance difference between TreeUNet, GAN-SCA, and BRR-Net is small, while the performance of BAS-Net, NAS-Net, and MSR-cGAN is significantly better than that of TreeUNet and GAN-SCA, of which MSR-cGAN has the better than performance.

4.6 Failure cases analysis
Although the proposed MSR-cGAN achieved better building extraction results on different datasets, it still cannot obtain perfect results for very challenging scenarios. In Figure 14, we show some failure cases of our MSR-cGAN. Firstly, for small area building region,
our method cannot achieve perfect building extraction results. As shown in the first row of Figure 14, the proportion of building region in the image is very small, and our method cannot effectively obtain the key feature information contained in the building region, so that the building extraction results have error with the ground truths. The main reason is that to relieve the GPU computational budget, we cropped and scaled the original image, so that the small area building region feature information is lost in the feature extraction process. A direct solution to this problem could be exploiting more computation resources to enable the training and testing of large-size inputs. Secondly, for the building region with complex background interference, our method fails to achieve accurate building region extraction. As shown in the second row of Figure 14, the building region and background region have high texture similarity, which makes it difficult to accurately extraction the building region feature information in the feature extraction process, so that our method fails to obtain perfect results. To address this issue, more attempts should be made to develop better learning strategies for the understanding of semantic feature information in the task of building extraction.

### Table 3. Comparison results of different methods in Massachusetts dataset, including TreeUNet, GAN-SCA, BRR-Net, BAS-Net, NAS-Net, and MSR-cGAN. The bold represents optimal performance.

| Methods  | TreeUNet | GAN-SCA | BRR-Net | BAS-Net | NAS-Net | MSR-cGAN |
|----------|----------|---------|---------|---------|---------|----------|
| Precision| 0.8315   | 0.8412  | 0.8972  | 0.9152  | 0.9238  | **0.9427** |
| Recall   | 0.7752   | 0.7824  | 0.8417  | 0.8274  | 0.8623  | **0.8814** |
| IoU      | 0.7413   | 0.7531  | 0.7936  | 0.8146  | 0.8273  | **0.8541** |
| F1_measure| 0.7826   | 0.8014  | 0.8692  | 0.8845  | 0.8972  | **0.9127** |

**Figure 12.** Building extraction results of different methods in Massachusetts dataset. Yellow boxes represent the best performance; Red boxes represent a relative worse performance.
5. Conclusions

In this article, we propose a practical framework for building extraction based on CNNs. This study further develops the application of CNNs for remote sensing image processing. The propose method consists of the generative network and adversarial network, where the generative network uses the recurrent residual convolutional module to extract the multi-scale and multi-resolution features of remote sensing building regions, and the attention gates skip connection is used to enhance the information transfer and interaction between different scale features. The adversarial network is a parallel encoder-decoder structure, which guides the model to optimize training by calculating the error between the predicted result and the ground truth. Moreover, the conditional information constraint is introduced into the model training process to alleviate the over-fitting problem and improve the building extraction accuracy. Experimental results on IAILD and Massachusetts building datasets show that MSR-cGAN is significantly superior over the other state-of-the-art methods in terms of building extraction accuracy. In our future work, we will introduce unsupervised learning in the propose method, so that the model can achieve better extraction results in unlabeled datasets.

Acknowledgements

All authors would sincerely thank the reviewers and editors for their beneficial, careful, and detailed comments and suggestions for improving the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China under Grant numbers 62172338; Jiangsu Innovation and Entrepreneurship Program under Grant numbers 184080H10827; Neural Science Foundation of Shanxi Province under Grant 2021JQ-879.
ORCID

Zhen Wang http://orcid.org/0000-0002-5765-0827

Data Availability Statement

The source code and dataset used to support the findings of this study have been in the https://github.com/darkseed-arch/MSR-cGAN

References

Abdollahi, A., B. Pradhan, and A. Alami. 2021. "RoadVecNet: A New Approach for Simultaneous Road Network Segmentation and Vectorization from Aerial and Google Earth Imagery in a Complex Urban et-up." GIScience & Remote Sensing 58 (7): 1151–1174. doi:10.1080/15481603.2021.1972713.

Alom, M. Z., M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari. 2018. "Recent Recurrent Convolutional Neural Network Based on U-Net (RU-Net) for Medical Image Segmentation." arXiv preprint arXiv: 1802.06955

Alshehhi, R., P. R. Marup, W. L. Woon, and M. D. Mura. 2017. "Simultaneous Extraction of Roads and Buildings in Remote Sensing Imagery with Convolutional Neural Networks." ISPRS Journal of Photogrammetry and Remote Sensing 130: 139–149. doi:10.1016/j.isprsjprs.2017.05.002.

Chen, M., J. Wu, L. Liu, W. Zhao, F. Tian, Q. Shen, B. Zhao, et al. 2021. "DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image." Remote Sensing 13 (2): 294–312. doi:10.3390/rs13020294.

Diakogiannis, F. I., F. Waldner, P. Caccetta, and C. Wu. 2020. "ResUNet-A: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data." ISPRS Journal of Photogrammetry and Remote Sensing 12 (3): 94–114. doi:10.1016/j.isprsjprs.2020.01.013.

Ding, L., and L. Bruzzzone. 2021. "DiResNet: Direction-Aware Residual Network for Road Extraction in VHR Remote Sensing Images." IEEE Geoscience Remote Sensing 13 (2): 1–12. doi:10.1109/TGRS.2020.3034011.

Eerapu, K. K., B. Ashwath, S. Lal, A. V. Narasimha, and A. V. Narasimha Dhan. 2019. "Dense Refinement Residual Network for Road Extraction from Aerial Imagery Data." IEEE Access 7 (2): 151764–151782. doi:10.1109/ACCESS.2019.2928882.

Guo, D., Y. Xia, and X. Luo. 2021. "Self-Supervised GANs with Similarity Loss for Remote Sensing Image Scene Classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2): 2508–2521. doi:10.1109/JSTARS.2021.3056883.

Hang, R., Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya. 2021. "Hyperspectral Image Classification with Attention-Aided CNNs." IEEE Geoscience Remote Sensing 59 (3): 2281–2293. doi:10.1109/TGRS.2020.3007921.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, pp. 770–778, Las Vegas, NV, USA. doi: 10.1109/CVPR.2016.90.

Hu, Q., L. Zhen, Y. Mao, X. Zhou, and G. Zhou. 2020. "Automated Building Extraction Using Satellite Remote Sensing Imagery." Automation in Construction 123: 103–113. doi:10.1016/j.autcon.2020.103509.

Hui, J., M. Du, X. Ye, Q. Qin, and J. Sui. 2019. "Effective Building Extraction from High-Resolution Remote Sensing Images with Multitask Driven Deep Neural Network." IEEE Geoscience Remote Sensing Letter 16 (5): 786–790. doi:10.1109/LGRS.2018.2880986.

Jiang, Y., and Z. Chi. 2019. "A CNN Model for Semantic Person Part Segmentation with Capacity Optimization." IEEE Transactions on Image Processing 28 (5): 2465–2478. doi:10.1109/TIP.2018.2886785.

Li, F., L. Xin, Y. Guo, J. Gao, and X. Jia. 2017. "A Framework of Mixed Sparse Representations for Remote Sensing Images." IEEE Geoscience Remote Sensing 55 (2): 1210–1221. doi:10.1109/TGRS.2017.2752750.

Liang, H., W. Bao, and X. Shen. 2021. "Adaptive Weighting Feature Fusion Approach Based on Generative Adversarial Network for Hyperspectral Image Classification." Remote Sensing 13 (2): 198–208. doi:10.3390/rs13020198.

Lin, D., K. Fu, Y. Wang, G. Xu, and X. Sun. 2017. "MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification." IEEE Geoscience Remote Sensing Letter 14 (11): 2092–2096. doi:10.1109/LGRS.2017.2752750.

Liu, J., Z. Tang, Y. Cui, and G. Wu. 2017. "Local Competition-Based Superpixel Segmentation Algorithm in Remote Sensing." Sensors 17 (6): 1364–1375. doi:10.3390/s17061364.

Liu, D., L. Han, X. Ning, and Y. Zhu. 2018. "A Segmentation Method for High Spatial Resolution Remote Sensing Images Based on the Fusion of Multi-Features." IEEE Geoscience Remote Sensing Letter 15 (8): 1274–1278. doi:10.1109/LGRS.2018.2829807.

Liu, C., P. Liu, W. Zhao, and X. Tang. 2020a. "Visual Tracking by Structurally Optimizing Pre-Trained CNN." IEEE Transactions on Circuits and Systems for Video Technology 30 (9): 3153–3166. doi:10.1109/TCSVT.2019.2938038.

Liu, J., S. Wang, X. Hou, and W. Song. 2020b. "A Deep Residual Learning Serial Segmentation Network for Extracting Buildings from Remote Sensing Imagery." International Journal of Remote Sensing 41 (14): 5573–5587. doi:10.1080/01431161.2020.1734251.

Liu, Y., D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu. 2021. "Multiscale U-Shaped CNN Building Instance Extraction Framework with Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery." IEEE Geoscience Remote Sensing 59 (7): 6106–6120. doi:10.1109/TGRS.2020.3022410.

Lu, X., B. Wang, X. Zheng, and X. Li. 2018. "Exploring Models and Data for Remote Sensing Image Caption Generation." IEEE Transactions on Geoscience and Remote Sensing 56 (4): 2183–2195. doi:10.1109/TGRS.2017.2776321.
Maji, D., P. Sigedar, and M. Singh. 2021. "Attention Res-UNet with Guided Decoder for Semantic Segmentation of Brain Tumors." Biomedical Signal Processing and Control 9 (1): 103077–103092. doi:10.1016/j.bspc.2021.103077.

Meng, Z., L. Li, X. Tang, Z. Feng, L. Jiao, and M. Liang. 2019. “Multipath Residual Network for Spectral-Spatial Hyperspectral Image Classification.” Remote Sensing 11 (6): 1896–1903. doi:10.3390/rs11061896.

Mirza, M., and S. Osindero. 2014. “Conditional Generative Adversarial Nets.” arXiv preprint arXiv:1411.1784.

Oktay, O., J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, and D. Rueckert. 2018. “Attention U-Net: Learning Where to Look for the Pancreas.” arXiv preprint arXiv:1804.03999.

Pan, X., Y. Fan, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren. 2019. “Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms.” Remote Sensing 11 (8): 917–924. doi:10.3390/rs11080917.

Ronneberger, O., P. Fischer, and T. Brox. 2015. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” Medical Image Computing and Computer-Assisted Intervention 234–241. doi:10.1007/978-3-319-24574-4_28.

Shao, Z., P. Tang, Z. Wang, N. Saleem, S. Yarn, and C. Sonnma. 2020. “BBRNet: A Fully Convolutional Neural Network for Automatic Building Extraction from High-Resolution Remote Sensing Images.” Remote Sensing 12 (6): 1050–1062. doi:10.3390/rs12061050.

Song, W., W. Song, H. Gu, and F. Li. 2020. “Progress in the Remote Sensing Monitoring of the Ecological Environment in Mining Areas.” International Journal of Environmental Research and Public Health 17 (6): 1846–1852. doi:10.3390/ijerph17061846.

Su, T. 2019. “Scale-Variable Region-Merging for High Resolution Remote Sensing Image Segmentation.” ISPRS Journal of Photogrammetry and Remote Sensing 147: 319–334. doi:10.1016/j.isprsjprs.2018.12.003.

Sun, S., L. Mu, L. Wang, P. Liu, X. Liu, and Y. Zhang. 2021. “Semantic Segmentation for Buildings of Large Intra-Class Variation in Remote Sensing Images with O-GAN.” Remote Sensing 13 (3): 475–483. doi:10.3390/rs13030475.

Tan, Y., S. Xiong, and P. Yan. 2020. “Multi-Branch Convolutional Neural Network for Built-up Area Extraction from Remote Sensing Image.” Neurocomputing 39 (6): 358–374. doi:10.1016/j.neucom.2018.09.106.

Wang, C., and L. Li. 2020. “Multi-Scale Residual Deep Network for Semantic Segmentation of Buildings with Regularize of Shape Representation.” Remote Sensing 12 (18): 2932–2943. doi:10.3390/rs12182932.

Xie, J., and J. Zhou. 2017. “Classification of Urban Building Type from High Spatial Resolution Remote Sensing Imagery Using Extended MRS and Soft BP Network.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (8): 3515–3528. doi:10.1109/JSTARS.2017.2686422.

Xie, Y., J. Zhu, Y. Cao, D. Feng, M. Hu, W. Li, Y. Zhang, and L. Fu. 2020. “Refined Extraction of Building Outlines from High-Resolution Remote Sensing Imagery Based on a Multi-Feature Convolutional Neural Network and Morphological Filtering.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2): 1842–1855. doi:10.1109/JSTARS.2020.2991391.

Xiong, D., C. He, X. Liu, and M. Liao. 2020. “An End-To-End Bayesian Segmentation Network Based on a Generative Adversarial Network for Remote Sensing Images.” Remote Sensing 12 (2): 216–225. doi:10.3390/rs12020216.

Xu, Y., L. Wu, Z. Xie, and Z. Chen. 2018. “Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters.” Remote Sensing 10 (1): 144–152. doi:10.3390/rs10010144.

Xu, P., H. Tang, J. Ge, and L. Feng. 2021. “ESPC_NASUNet: An End-to-End Super-Resolution Semantic Segmentation Network for Mapping Buildings from Remote Sensing Images.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (3): 5421–5435. doi:10.1109/JSTARS.2021.3079459.

Yin, S., Y. Zhang, and S. Karim. 2018. “Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model.” IEEE Access 6 (3): 26069–26080. doi:10.1109/ACCESS.2018.2834960.

Yue, K., L. Yang, R. Li, W. Hu, F. Zhang, and W. Li. 2019. “TreeUNet: Adaptive Tree Convolutional Neural Networks for Subdecimeter Aerial Image Segmentation.” ISPRS Journal of Photogrammetry and Remote Sensing 156: 1–13. doi:10.1016/j.isprsjprs.2019.07.007.

Zhang, Z., Q. Liu, and Y. Wang. 2015. “Road Extraction by Deep Residual U-Net.” IEEE Geoscience Remote Sensing 15 (5): 749–753. doi:10.1109/LGRS.2018.2802944.

Zhang, L., A. Li, Z. Zhang, and K. Yang. 2016. “Global and Local Salience Analysis for the Extraction of Residential Areas in High-Spatial-Resolution Remote Sensing Image.” IEEE Geoscience Remote Sensing 54 (7): 3750–3763. doi:10.1109/TGRS.2016.2527044.

Zhang, H., Y. Liao, H. Yang, G. Yang, and L. Zhang. 2020. “A Local-Global Dual-Stream Network for Building Extraction from Very-High-Resolution Remote Sensing Images.” IEEE Transactions on Neural Networks and Learning Systems 12 (1): 1–15. doi:10.1109/TNNLS.2020.3041646.

Zhang, L., Y. Sun, and J. Zhang. 2020. “Pan-Sharpening Based on Common Saliency Feature Analysis and Multiscale Spatial Information Extraction for Multiple Remote Sensing Images.” International Journal of Remote Sensing 41 (8): 3095–3118. doi:10.1080/01431161.2019.1698784.

Zheng, C., and L. Wang. 2015. “Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model with Regional Penalties.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8 (5): 1924–1935. doi:10.1109/JSTARS.2014.2361756.

Zhong, Y., R. Gao, and L. Zhang. 2016. “Multiscale and Multi-Feature Normalized Cut Segmentation for High Spatial Resolution Remote Sensing Imagery.” IEEE Geoscience Remote Sensing 54 (10): 6061–6075. doi:10.1109/TGRS.2016.2580643.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. “Learning Deep Features for Discriminative Localization.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, pp. 2921–2929, Las Vegas, NV, USA. doi: 10.1109/CVPR.2016.319.
Zhou, Z. M., M. R. Siddiquee, N. Tajbakhsh, and J. Liang. 2020. “UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation.” *IEEE Transactions on Medical Imaging* 39 (6): 1856–1867. doi:10.1109/TMI.2019.2959609.

Zhou, X., X. Wang, Y. Zhou, Q. Lin, J. Zhao, and X. Meng. 2021. “RSIMS: Large-Scale Heterogeneous Remote Sensing Images Management System.” *Remote Sensing* 13 (9): 1815–1824. doi:10.3390/rs13091815.

Zhu, Q., Z. Li, Y. Zhang, and Q. Guan. 2020. “Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields.” *Remote Sensing* 12 (23): 3983–3992. doi:10.3390/rs12233983.

### Appendix

**Ablation Study**

We evaluate whether the RRCM and AGs operation can bring benefits to the generative network performance, so the base semantic segmentation network U-Net and the improved U-Net network (U-Net-RRCM, U-Net-AGs, our generative network) are compared. In the experiment, firstly, we compare the training process of different methods to analyze whether the RCM and AGs modules can optimize the training process of the model; then, we visualize the feature maps and heat maps of different methods, and analyze the influence of RRSC and AGs on the feature extraction ability; finally, the extraction results of different methods are quantitatively compared using multiple evaluation indicators. In order to further demonstrate the extraction performance of different methods, 1,250 images in the IAILD dataset and Massachusetts dataset are randomly selected as the train dataset, 755 images as the validation dataset, and 425 images as the test dataset. As shown in Figure A1, the range of optimal epoch for training U-Net is 145–165, the range of optimal epoch for training U-Net-RRCM is 105–125, the range of optimal epoch for training U-Net-AGs is 110–130, and the range of optimal epoch for training the proposed generative network is 100–120. The convergence speed of U-Net-RRCM and U-Net-AGs are faster than that of U-Net, which shows that the introduction of RRCM and AGs can speed up the training speed and optimization efficiency of the model. The proposed generative network introduces RRCM and AGs, and the final loss value is 0.252, which is better than other compared methods.

The feature map is a visual result of feature extraction from input image by convolution operation, which can intuitively display the feature information extracted by different convolution layers. The class activation mapping (Zhou et al. 2016) is

![Figure A1](image_url). The curve of loss value in training process. (a) U-Net training process. (b) U-Net-RRCM training process. (c) U-Net-AGs training process. (d) Generative network training process.
used as visualization method for feature map. From Figure A2 that the original U-Net can only extract the shallow information, and its extract feature ability to extract boundary and contour features is poor, so the deep semantic information on the building region cannot be obtained. Compared with the original U-Net, the feature extraction ability of U-Net-RRCM for remote sensing images has been greatly improved, and it can be seen from the experimental results that the U-Net-RRCM enhances the convolution layer feature transfer ability, so it can better extract the edge and contour feature of the target area. U-Net-AGs can better extract the texture and detailed features of non-target area, resulting in features redundant phenomenon. The proposed generation network can not only fully extract the boundary and contour features of remote sensing images, but also extract texture and detail features, indicating that the introduction of RRMC and AGs can greatly improve the model feature extraction ability.

The heat map represents the segmentation probability map output by the model, which can indicate how much the model pays attention to the remote sensing image building region. Figure A3 is the visualization result of the heat map, where the heat map represents the segmentation probability map output by the model. It can be seen from Figure A3 that the heat maps of U-Net covers a small area, indicating that the model cannot accurately extract the remote sensing building area. U-Net-RRCM introduces a recurrent residual convolution module, which is more sensitive to the building area feature, and its heat map covers most of the building area in the remote sensing image. The heat map of U-Net-AGs covers a large of the building area, because the AGs operation can enhance the module feature reuse ability, the shallow features and deep features can be better fused. The proposed generation network is superior to the compared methods in the coverage area of the heat map, which can be seen from the experimental results that the heat map of the generation network not only covers the building area, but also can automatically filter the non-building area. On the basis of the obtained heat map results, the receiver-operating characteristic (ROC) curve and precision recall (PR) curve can be obtained. The area under the ROC curve (AU-ROC) and area under the PR curve (AU-PR) are commonly used quantitative indicators for segmentation result evaluation. The ROC curve and PR curve of different methods are shown in Figure A4, from which it can be seen that the proposed generation network can achieve better results, indicating that it is suitable for solving the problem of remote sensing building extraction.

Table A1 and Figure A5 are the quantization and visualization results of different methods on the INRIA and Massachusetts dataset. It can be seen that the building extraction effect of U-Net is poor, the precision, recall, F1_measure are 0.8265, 0.7843 and 0.8048, respectively, and it can be seen from Figure

![Figure A2](image-url)  
**Figure A2.** Visualization comparison between the output feature maps of the U-Net, U-Net-RRCM, U-Net-AGs, and our generative network. The feature map illustrates the feature extraction effect of the different models on the remote sensing image building region.
that the building region cannot be accurately extracted. U-Net-RRCM and U-Net-AGs have greatly improved the performance of the original U-Net, and it can be seen from the experimental results that the F1 measure is 0.0474 and 0.0639 higher than the original U-Net, respectively. It can be seen from the visualization results in Figure A5 that U-Net-RRCM and U-Net-AGs can extract large-area building region better, but the extraction effect for small-area and densely-area building is relatively poor. The proposed generation network is superior to the compared methods in all evaluation indicators, and the results in Figure A5 show that the generation network can accurately extract the building region.

**Complex Analysis**

To further assess the practicality and complexity of the proposed method, we randomly selected 5,250 images from IAILD and Massachusetts as the training dataset, 1,635 images as the test dataset, and 920 images as the validation dataset to compare and analyze the training time, single-image test time, memory space.
usage during training, and model computational parameters for different methods. The complexity analysis results of different methods are shown in Table A2. From Table A2, the complexity analysis indicator of the proposed method is optimal, the training time of 5.75 h, and the calculation parameters of 65,783,621, indicating that the method has a strong practicality. The memory space used during model training is 7.32GB, indicating that the proposed method can be train and test on multiple computing devices. Further, the complex analysis result show that our method has improvement space, it needs to compress the calculation parameters while ensuring building extraction accuracy.

**Robustness Analysis**

To better illustrate that the proposed generative network is more suitable for remote sensing building region extraction, we compared four improved UNet architecture with the proposed generative network. The improved UNet structures compared include ResUNet-a (Diakogiannis et al. 2020), Attention Res-UNet (Maji et al., 2021), UNet++ (Zhou et al. 2020). Firstly, to demonstrate that the proposed generative network has better feature extraction effect and suppression ability to feature information of non-building region, we first compare the feature masks of building region extracted by different UNet structures, and the visualization results are shown in Figure A6. As can be seen from Figure A6, since the proposed generative network uses the recurrent convolution and attention mechanism, it obtains more building feature information, such as texture, contour, and details. However, other improved UNet structures have the phenomenon of feature information loss, which only extracts the contour feature, and has a relatively worse extraction effect on the detail feature. Secondly, in terms of the specific effect of building extraction, we conducted a comparative analysis. Table A3 and Figure 21 show the quantitative analysis and visualization comparisons of different improved UNet architecture and MSR-cGAN, from which it can be seen that the proposed generative network can accurately extracts the building region and obtain the optimal evaluation metrics, while other methods have problems, such as incomplete extraction of building region and boundary contour information loss.

**Table A1.** Quantitative result of different methods, including U-Net, U-Net-RRCM, U-Net-AGs, and our generative network. The bold represents optimal performance.

| Methods          | Precision | Recall | IoU   | F1_measure |
|------------------|-----------|--------|-------|------------|
| U-Net            | 0.8265    | 0.7843 | 0.7893| 0.8048     |
| U-Net-RRCM       | 0.8963    | 0.8124 | 0.8524| 0.8522     |
| U-Net-AGs        | 0.9189    | 0.8238 | 0.8665| 0.8687     |
| Generative Network | 0.9348    | 0.8546 | 0.8894| 0.8928     |

*Figure A5a.* Building extraction results of U-Net, U-Net-RRCM, U-Net-AGs, and our generation network. Yellow boxes represent the best performance; Red boxes represent a relative worse performance.
**Figure A6.** Building extraction results of ResUNet-a, attention ResUNet, UNet++, and MSR-cGAN.

**Table A2.** The complexity analysis results of different methods, including TreeUNet, GAN-SCA, BRR-Net, BAS-Net, NAS-Net, and MSR-cGAN. The bold represents optimal performance.

| Model      | Memory Space | Training Time | Testing time | Parameters   |
|------------|--------------|---------------|--------------|--------------|
| TreeUNet   | 11.23 GB     | 8.63 h        | 7.41 s       | 97,365,274   |
| GAN-SCA    | 10.12 GB     | 7.18 h        | 6.26 s       | 91,318,432   |
| BRR-Net    | 9.73 GB      | 6.56 h        | 5.85 s       | 86,241,856   |
| BAS-Net    | 8.26 GB      | 6.85 h        | 3.52 s       | 85,473,268   |
| NAS-Net    | 8.52 GB      | 6.21 h        | 3.14 s       | 71,714,165   |
| MSR-cGAN   | **7.32 GB**  | **5.75 h**    | **2.13 s**   | **65,783,621** |

**Table A3.** Evaluation metrics of ResUNet-a, Attention ResUNet, UNet++, and MSR-cGAN. The bold represents optimal performance.

| Methods     | Recall  | Precision | IoU   | F1_measure |
|-------------|---------|-----------|-------|------------|
| ResUNet-a   | 0.8206  | 0.8673    | 0.7852| 0.8257     |
| Attention ResNet | 0.7332  | 0.8931    | 0.7982| 0.8662     |
| UNet++      | 0.8415  | 0.9280    | 0.8146| 0.8795     |
| MSR-cGAN    | **0.8975** | **0.9541** | **0.8815** | **0.9278** |

**Figure A5b.** The feature masks visualization results of ResUNet-a, Attention ResUNet, UNet++, and MSR-cGAN.