frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/feart.2023.1268628

OPEN ACCESS

EDITED BY
Bahareh Kalantar,
Riken, Japan

REVIEWED BY
Vahideh Saeidi,
Putra Malaysia University, Malaysia
Mohammed Oludare Idrees,
University of Abuja, Nigeria

*CORRESPONDENCE
Xueli Chang,
chang99@hbut.edu.cn

RECEIVED 28 July 2023
ACCEPTED 09 October 2023
PUBLISHED 19 October 2023

CITATION
Jin H, Fu W, Nie C, Yuan F and Chang X (2023), Extraction of building from remote sensing imagery base on multi-attention L-CAFSFM and MFFM.
Front. Earth Sci. 11:1268628.
doi: 10.3389/feart.2023.1268628

Extraction of building from remote sensing imagery base on multi-attention L-CAFSFM and MFFM

Huazhong Jin[1], Wenjun Fu[1], Chenhui Nie[2], Fuxiang Yuan[3] and Xueli Chang[1]*

[1]College of Computer Science, Hubei University of Technology, Wuhan, China, [2]Zhejiang Academy of Surveying and Mapping, Hangzhou, China, [3]College of Management, Jianghan Art Vocational College, Qianjiang, China

Building extraction from high-resolution remote sensing images is widely used in urban planning, land resource management, and other fields. However, the significant differences between categories in high-resolution images and the impact of imaging, such as atmospheric interference and lighting changes, make it difficult for high-resolution images to identify buildings. Therefore, detecting buildings from high-resolution remote sensing images is still challenging. In order to improve the accuracy of building extraction in high-resolution images, this paper proposes a building extraction method combining a bidirectional feature pyramid, location-channel attention feature serial fusion module (L-CAFSFM), and meticulous feature fusion module (MFFM). Firstly, richer and finer building features are extracted using the ResNeXt101 network and deformable convolution. L-CAFSFM combines feature maps from two adjacent levels and iteratively calculates them from high to low level, and from low to high level, to enhance the model's feature extraction ability at different scales and levels. Then, MFFM fuses the outputs from the two directions to obtain building features with different orientations and semantics. Finally, a dense conditional random field (Dense CRF) improves the correlation between pixels in the output map. Our method's precision, F-score, Recall, and IoU(Intersection over Union) on WHU Building datasets are 95.17%、94.83%、94.51% and 90.18%. Experimental results demonstrate that our proposed method has a more accurate effect in extracting building features from high-resolution image.

KEYWORDS

remote sensing image, building detection, building extraction, location-channel attention feature serial fusion module (L-CAFSFM), meticulous feature fusion module (MFFM)

1 Introduction

With the rapid development of sub-meter-level high-resolution earth observation satellites, it has become possible to obtain high-resolution images of the surface over a large area. Buildings are one of the most common and essential elements of the earth's surface (Cai, et al., 2021; Sheikh, et al., 2022; Yuan and Mohd Shafri, 2022; Zhang, et al., 2022). Building detection from remote sensing images has been widely used in urban development planning, land development and utilization, post-disaster damage assessment, and other fields. Factors such as the size, shape, texture difference, cloud occlusion, surface

material reflection, and shadow of buildings in remote sensing images can reduce the accuracy of building detection. Improving the accuracy of building detection has essential application value and practical significance for urban 3D modeling, map updating, disaster assessment, etc (Bauchet, et al., 2021; Chen and Sun, 2022; Fang, et al., 2022; Hou, et al., 2022; Yang, et al., 2022).

Traditional methods of extracting buildings from remote sensing images mainly rely on manually extracted features, such as brightness, texture, shape, and prior knowledge. (Lin and Zhang, 2017) proposed an object-based morphological building index (OBMBI) by comprehensively using image segmentation and graph-based mathematical morphology top-hat reconstruction technology, using image segmentation to obtain objects and establish topological relationship diagrams between objects. The feature function of the graph is created using the brightness value feature of the object, and a bidirectional mapping relationship between the pixels, objects, and graph nodes is established. Morphological building index maps are generated using top-hat reconstruction techniques. But this method is only suitable for remote sensing images with a resolution better than 1 m. (Ma, et al., 2019) proposed a new morphological attribute building index (MABI), which establishes morphological attribute filters (AFs) with the building features of the input images (Such as high local contrast, internal homogeneity, shape, and size) and is used for image segmentation to obtain building regions with high homogeneity. However, this method's segmentation standard deviation threshold must be manually set. (Zhang, et al., 2019) used a novel rotation uniform invariant local binary pattern algorithm to obtain low-density feature maps and use mean shifts to extract building edges. This method can accurately segment the boundary of simple buildings, but the detection effect could be better when there are many buildings and complex scenes. (Wang, et al., 2019) proposed the Adaptive Morphological Attribute Profile under Object Boundary Constraint (AMAP-OBC) method under the constraints of building boundaries and combined with Morphological Attribute Profiles (MAPs). In the preprocessing step, candidate object sets are extracted through MAPs. Secondly, the candidate object set is processed by AMAP-OBC to obtain the initial building set. Finally, the building sets are segmented using an adaptive threshold to obtain final building extraction results. However, the morphological property profile of this method is challenging to obtain in advance.

Because traditional building extraction methods for high-resolution remote sensing images often rely on low-level features, and the means of describing and representing building features are single and specific, it is challenging to extract the features of different types of buildings (Borba, et al., 2021; Xiao, et al., 2022). In fact, traditional building extraction methods are not universal and cannot meet the needs of most scenes. Assuming that the high-level semantic features of high-resolution remote sensing images are utilized, along with their low-level features (Xie, et al., 2020; Tian, et al., 2022), such as shape, texture, brightness, and contour, different levels of features are fused, which can improve the accuracy of building extraction.

In recent years, deep learning technology has led the continuous in-depth application and expansion of artificial intelligence in different industries and fields, especially in computer vision (Chen, et al., 2022a; Shi, et al., 2022; Wei, et al., 2022). The reason is that deep learning can manipulate data or symbols to form different levels of features to recognize patterns, model approximate functions in the data or symbols, and interpret and understand what people see. Deep learning, as a learning mechanism, writes rules for specific patterns or defines symbols for fuzzy concepts through self-supervised learning of large amounts of data. Therefore, image-building extraction can define and describe a pattern or concept from many images (Abdollahi, et al., 2020; Li, et al., 2020a; Saini, et al., 2021; Chen, et al., 2022b; Yan, et al., 2022; You, et al., 2022).

Some researchers have used deep learning methods to segment buildings in high-resolution remote sensing images, significantly improving detection accuracy. Yu, et al. (2021) proposed the Capsule Feature Pyramid Network (CapFPN), which utilizes the characteristics of the feature pyramid and fuses the features of the capsule network at different levels. CapFPN can extract features with high resolution and intrinsic solid semantics, effectively improving the extraction accuracy of pixel-level buildings. (Zhu, et al., 2021) used Multi Attending Path Neural Network (MAP-Net) to learn multi-scale features in feature space. They use an attention module to adaptively compress the features of each channel, which is used to fuse multi-scale features. Then, global dependency can be captured using a pyramid spatial pooling module to optimize discontinuous buildings. (Zhu, et al., 2018) used a Bidirectional Feature Pyramid Network (BFPN) to fuse feature maps of different scales and enhance the feature encoding ability. The above methods use the feature pyramid structure to extract and fuse multi-scale features. However, the feature information extracted by a single feature pyramid network must be more prosperous, and the model's ability to perceive features needs to be improved.

The attention mechanism is widely used in the field of image processing, inspired by the research on human attention. In image analysis, the attention mechanism can focus on important feature information with high weight and ignore irrelevant information with low weight. (Guo, et al., 2020) proposed a U-Net building extraction method with attention modules and multiple losses. It can improve the model's sensitivity through the attention module and suppress the background influence of irrelevant feature areas. However, as a fully supervised method, it relies on many manual label samples. (Das and Chand, 2021) proposed Attention Building Net (ABNet), which utilizes a convolutional attention module with a channel and spatial attention mechanism to focus on essential features selectively. Building boundaries can be accurately extracted because it improves the overall feature representation. However, this method needs to pay attention to the correlation between features of different levels and scales, resulting in poor detection in complex scenes. (Cai and Chen, 2021) designed a downsampling module combining separable convolution and channel attention to extract features from the input graph. However, single-channel attention only pays attention to the channel information of features and needs help to obtain good feature space information. (Li, et al., 2020b) used a convolutional neural network to extract pixel-level building shadows and used conditional random fields (CRF) as post-processing optimization experimental results, which achieved good results. However, CRF needs to use the correlation between pixels; there is still room for improvement.

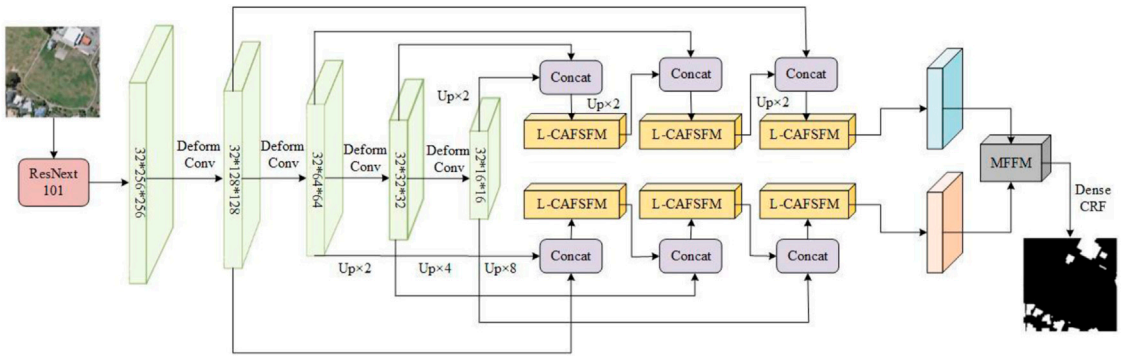

FIGURE 1
The overall structure of the network. Deform Conv denotes deformable convolution, Concat denotes concatenating, and Up denotes upsampling.

The above scholars have provided different methods to improve the model network and attention mechanism. However, some methods still need to be improved, such as insufficient extraction of features from a single model, insufficient attentional fusion, and failure to consider correlations between features in neighboring hierarchies. In response to these issues, this paper designs the location-channel attention feature serial fusion module (L-CAFSFM) and the meticulous feature fusion module (MFFM) in the bidirectional feature pyramid network. First, based on the ResNeXt101 network, combined with deformable convolution, a group of feature maps with different levels and resolutions is generated. Then, the L-CAFSFM is used to iteratively calculate the two adjacent feature maps from low level to high level and from high level to low level. MFFM is used to fuse the output of two directions. Finally, Dense Conditional Random Field (Dense CRF) is applied to optimize the results and output the prediction image.

2 Methodology

In this section, we will elaborate on our method. First, the overall network structure diagram is introduced. Then, the Location-Channel Attention Feature Serial Fusion module (L-CAFSFM), Meticulous Feature Fusion module (MFFM), and loss function are introduced in detail.

2.1 Model overview

Traditional neural network models want to improve accuracy by deepening or widening the network. However, with the increase of super parameters (such as the number of channels and convolution size), the difficulty of network design and computational expense will increase. ResNext deep neural network can improve the accuracy without increasing the complexity of the parameters and also reduce the number of super parameters. Based on VGG/ResNets' duplicate layer strategy and split transform merge strategy, the ResNext101 network proposes an aggregate transformations method, which uses a parallel stack of blocks with the same topology structure to replace the original ResNets' three-layer convolutional block. The model's accuracy is improved without significantly increasing the number of parameters. At the same time, because of the same topology, the super parameters are reduced accordingly.

Traditional convolution kernels' size is usually fixed (e.g., 3×3, 5×5, 7×7). They have poor adaptability to changes in unknown objects and weak generalization ability. Deformable convolution introduces a learnable offset in the receptive field so that the receptive field becomes a polygon, which is no longer limited to a square, and can extract more accurate features at different levels. Therefore, we use the ResNeXt101 network (Xie, et al., 2017) combined with deformable convolution (Dai, et al., 2017) to extract feature maps of different scales and levels of the input image.

The network structure diagram proposed in this paper is shown in Figure 1.

High-level features contain rich semantic information about buildings, and low-level feature maps have fine local detail features of buildings. Inspired by the article (Zhu, et al., 2021), we concatenate adjacent high-level features with low-level features and input them into the L-CAFSFM for computation. Two different directions are used to calculate semantic information iteratively: one is from high-level to low-level, and the other is the opposite, to obtain multi-scale information in different directions and levels. Then, MFFM fuses the outputs from both directions. Finally, dense conditional random field (Dense CRF) improves the correlation of each pixel in the output image to obtain a building prediction map.

2.2 Location-channel attention feature serial fusion

Attention mechanisms can focus on the more critical information of the current task in a large amount of information, reduce attention to other information, and even filter out irrelevant information to improve the efficiency and accuracy of task processing. The attention mechanism is widely applied in computer vision fields such as image segmentation and classification and is roughly divided into three categories: spatial, channel, and spatial-channel hybrid attention. SE (Squeeze-and-Excitation) attention (Jie, et al., 2017) is typical channel attention, which only considers the internal channel information and ignores

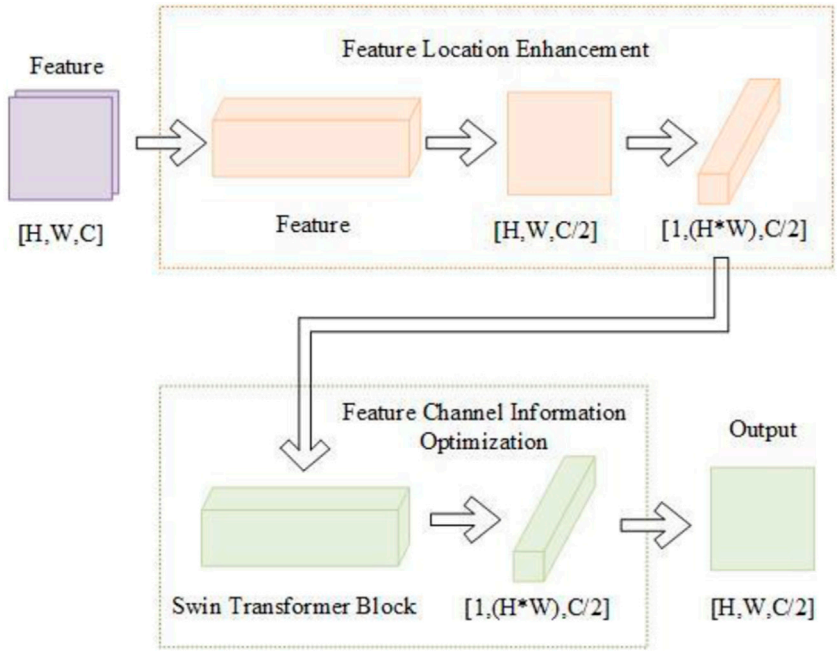

FIGURE 2
The structure of L-CAFSFM.

the importance of location information. BAM(Bottleneck Attention Module) (Park, et al., 2018) and CBAM(Convolutional Block Attention Module) (Woo, et al., 2018) try to introduce location information by global pooling on channels, but they can only capture local information instead of long-range dependent information. The self-attention is an improvement of the attention mechanism, which reduces the dependence on external information and is better at capturing the internal correlation of features. However, when using the self-attention mechanism to encode the information about the current position, the model will excessively focus on its own position.

Given the above attention mechanism problems, we introduce coordinate attention and Swin Transformer Block to build the Location-Channel Attention Feature Serial Fusion module (L-CAFSFM). Coordinate attention can improve the ability to obtain location information and channel information and reduce the loss of channel information and location information caused by downsampling operations (Hou, et al., 2021). Swin Transformer Block can capture the internal correlation of location and channel information and improve the network's sensitivity to information (Liu, et al., 2021). The L-CAFSFM is shown in Figure 2.

L-CAFSFM can be divided into feature location enhancement and feature channel information optimization.

2.2.1 Feature location enhancement

In the feature location enhancement step, the Coordinate Attention Block calculates the input feature map. Coordinate attention captures feature details across channels and includes feature orientation and location information. It enables the model to locate and identify target regions more accurately and enhances the model's feature expression capabilities. Coordinate attention can take an arbitrary $X = [x_1, x_2, \ldots, x_C] \in R^{C\times H\times W}$ as input and transform it into $Y = [y_1, y_2, \ldots, y_C] \in R^{C\times H\times W}$, which is output with the same size and the same channel as X.

Coordinate attention encodes channel relationships and long-term dependencies with precise location information and can be divided into coordinate information embedding and attention generation. The structure of coordinate attention is shown in Figure 3.

In the coordinate information embedding module, global average pooling is performed respectively on the horizontal and vertical directions of the input feature map so that the attention module can capture the interaction information in different directions and different spaces. Specifically, for the input feature X, the pooling kernels of size $(1, W)$ and $(H, 1)$ are used to encode along the vertical and horizontal directions, respectively, so the output of the c-*th* channel at height h is:

$$Y_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i) \tag{1}$$

the output of the c-*th* channel at height h is:

$$Y_c^w(w) = \frac{1}{W}\sum_{0\le i<W} x_c(w,i) \tag{2}$$

By aggregating features along two spatial directions, a pair of direction-aware feature maps can be obtained, enabling the attention module to capture long-term dependencies along one spatial direction and preserve precise location information along the other. It helps the network to more accurately locate the target of interest.

In the coordinate attention generation module, a better global receptive field can be obtained after the above transformation, and the precise location information of the feature can be encoded. In order to capture the information between channels simultaneously,

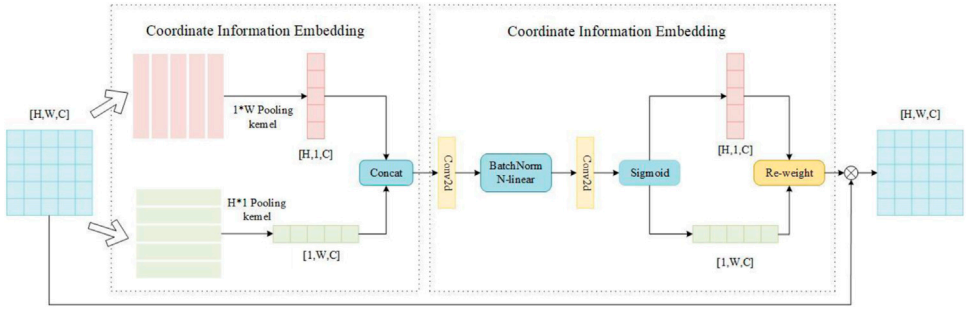

FIGURE 3
The structure of coordinate attention.

the two outputs of the module in the previous step are concatenated and computed through the convolution transformation function.

$$F = Nl\left(Concat\left(\left[Y^h, Y^w\right]\right)\right) \tag{3}$$

where *Concat* (·) is a concatenation operation along the horizontal and vertical directions, and *Nl* (·) is a nonlinear activation function. F is the intermediate feature vector that encodes the spatial information in the horizontal and vertical directions. Then it is decomposed into two separate vectors F^h and F^w along the horizontal and vertical directions, which are activated by convolution transformation and Sigmoid function, respectively. The results are:

$$S^h = Sig\left(C_h\left(F^h\right)\right) \tag{4}$$

$$S^w = Sig\left(C_w\left(F^w\right)\right) \tag{5}$$

where $C_h(\cdot)$ and $C_w(\cdot)$ represent convolution operations on F^h and F^w. The output T of the last Coordinate Attention Block is:

$$T = X \times R\left(S^h\right) \times R\left(S^w\right) \tag{6}$$

where $R(\cdot)$ represents the Re-Weight operation, that is, restore $S^h \in R^{C\times 1\times W}$ and $S^w \in R^{C\times H\times 1}$ to $C\times H\times W$ size. The output $T \in R^{C\times H\times W}$ has the same size and dimension as the input X. We compress T into $T' \in R^{C\times 1\times (H\times W)}$. T' is a row vector with dimension C and size $[1, (H \times W)]$. T' is used as the output of this stage.

2.2.2 Feature channel information optimization

In optimizing feature channel information, we choose Swin Transformer Block in the article (Liu, et al., 2021) to calculate the output of the previous step. Swin Transformer Block consists of a self-attention based on a sliding window and a self-attention without a sliding window. Multi-head attention is added to self-attention, which can extract feature information from multiple dimensions. Constraining attention computation within a window through multiple windows and using Shifted Window to link multiple windows make it easier to capture fine local features. Therefore, this paper designs a feature optimization module based on Swin Transformer Block, as shown in Figure 4.

The main body of the module consists of two Swin Transformer Blocks. W-MSA (Window-based Multi-head Self Attention) is a multi-head self-attention without a sliding window. Different from the global self-attention calculation, W-MSA can calculate self-attention. It reduces the amount of computation without causing a lot of memory consumption.

Because self-attention is calculated in multiple windows, the information between windows cannot interact, and the effect of global modeling cannot be achieved. To solve this problem, the article (Liu, et al., 2021) proposes the SW-SAM module. SW-SAM (Shifted Window-based Multi-head Self Attention) is a multi-head self-attention with a sliding window. By sliding the window in the feature map, the information of different windows is collected, and the communication between the windows is realized to establish a global model. MLP is a multilayer perceptron that performs nonlinear classification of features. This module calculates the output of the previous step: as the input dimension $Z \in R^{C\times 1\times (H\times W)}$, after two Swin Transformer Block calculations, the output $Z' \in R^{C\times 1\times (H\times W)}$ is the same as the input dimension, and Z' is a row vector with C channels and size $[1, (H \times W)]$. Finally, Z' is restored to a feature map $M \in R^{C\times H\times W}$ according to the number of channels C and the size of $[C, H, W]$ as the output of L-CAFSFM.

As shown in Figure 5, the characteristic diagram is not processed by the L-CAFSFM and is calculated by the L-CAFSFM.

2.3 Meticulous feature fusion module

After the iterative calculation of the L-CAFSFM by the bidirectional feature pyramid network, fusing the outputs of two opposite paths is necessary. Inspired by the article (Zhu, et al., 2018), we propose the Meticulous Feature Fusion module (MFFM). The module structure is shown in Figure 6.

For the outputs $M_i \in R^{C\times H\times W}$ and $M_j \in R^{C\times H\times W}$ of the two opposite paths, use a 1×1 convolution operation to connect the two outputs to obtain the meticulous feature L_1.

$$L_1 = Concat\left[F_c\left(M_i\right), F_c\left(M_j\right)\right] \tag{7}$$

where *Concat* (·) represents the connection operation, and $F_c(\cdot)$ is the 1×1 convolution operation.

After connecting M_i and M_j, the channel of the feature map is $2C$, and the size is $[H, W]$. Input it into this paper's improved SE Attention (Fusion-SE), and then use the Sigmoid operation. The output is a meticulous feature L_2 with channel two and size $[H, W]$. It is:

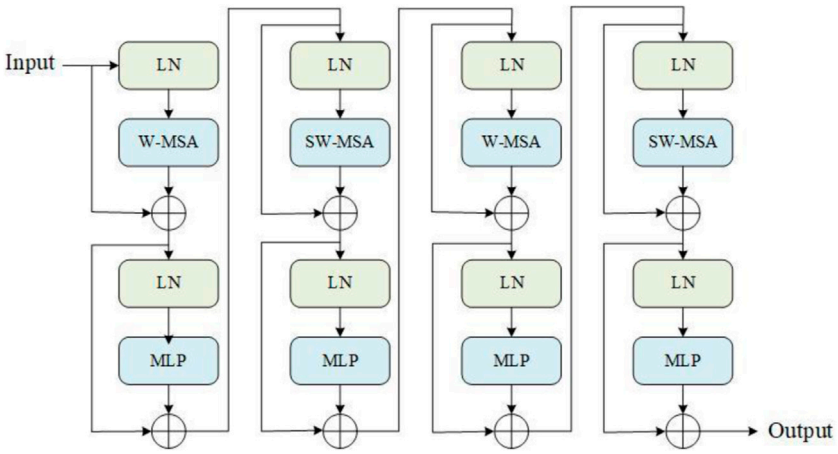

FIGURE 4
Structure of feature optimization module.

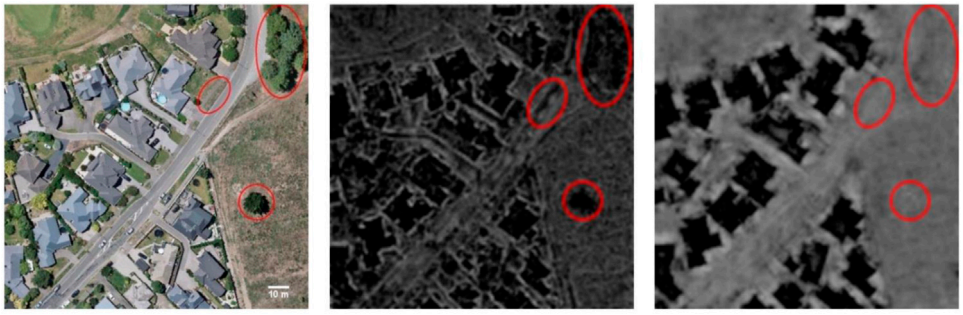

FIGURE 5
Figure 1 is the input image. Figure 2 shows the feature map the L-CAFSFM still needs to calculate. The distinction between the building area and the trees and roads in the red circle needs to be more apparent, and the overall brightness value of the feature map is close to the building area. Figure 3 shows the feature map calculated by L-CAFSFM. In Figure 3, the building area's features differ from other features, suppressing the other features. It shows that after L-CAFSFM calculation, the model can extract more accurate and rich building features.

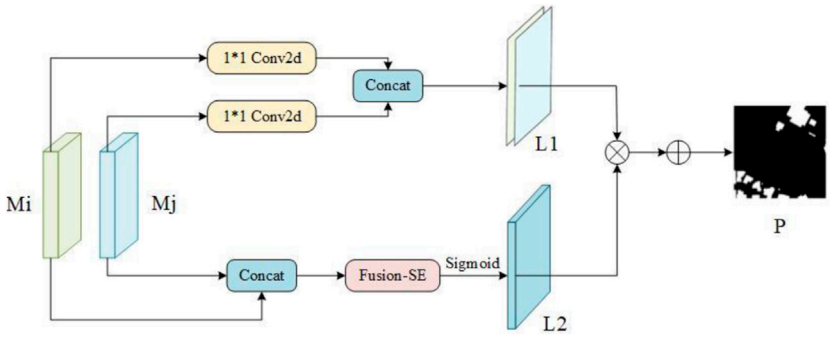

FIGURE 6
The structure of MFFM.

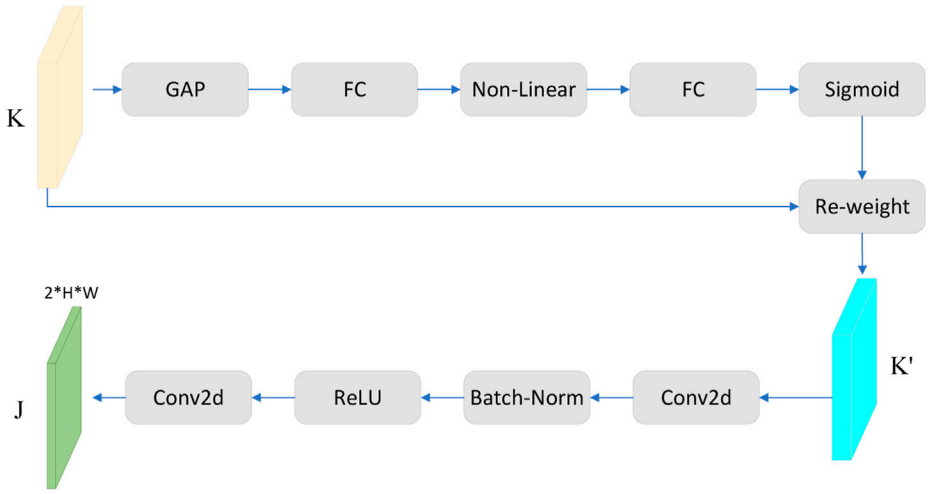

FIGURE 7
The structure of fusion-SE attention.

FIGURE 8
Figure 1 is the label image, Figure 2 shows the extraction results without Dense CRF optimization, and Figure 3 shows the extraction results with Dense CRF optimization. Comparing objects of the red circle box in Figure 2 and Figure 3, it is clear that the extracted edges of the buildings are clearer and more regular after the Dense CRF optimization of the predicted maps.

$$L_2 = \phi\{F_{SE}[Concat(M_i, M_j)]\} \tag{8}$$

where $\phi(\cdot)$ represents the sigmoid operation, and $F_{SE}(\cdot)$ represents the Fusion-SE attention. It is the Fusion-SE attention module diagram, as shown in Figure 7.

In Figure 7, GAP stands for global average pooling. After inputting $K \in R^{C\times H\times W}$ into the SE attention module, the output of $K' \in R^{C\times H\times W}$ is consistent with the input dimension and size. After convolution transformation, batch normalization and linear activation, the output is $J \in R^{2\times H\times W}$. The fusion-SE attention module reduces the dimension C of the input K to 2, which is convenient for calculation with the meticulous feature L_1 of the previous step.

The final output of building a prediction map is:

$$P = Sum(L_1 \times L_2) \tag{9}$$

Dense CRF optimizes the prediction map to improve the correlation between different pixels in the prediction map. It can be seen from Figure 7 that after Dense CRF optimization of the prediction map, the building edge details have been further optimized, which is closer to the label map. Thus, the final building detection map of our method is obtained, as shown in Figure 8.

3 Experiments and results

3.1 Dataset and experimental settings

In order to evaluate the method proposed in this paper, we selected the public building dataset (WHU) of Wuhan University. WHU is an aerial image dataset. This dataset consists of aerial images obtained in April 2012 and covers an area of 20.5 km^2 12,796 buildings. The aerial image data comes from the New Zealand Land Information Service website, with a ground resolution of 0.3 m after low sampling, selected from approximately 22,000 buildings in Christchurch. This dataset contains 8,188 remote sensing images and has a resolution of 512 ×

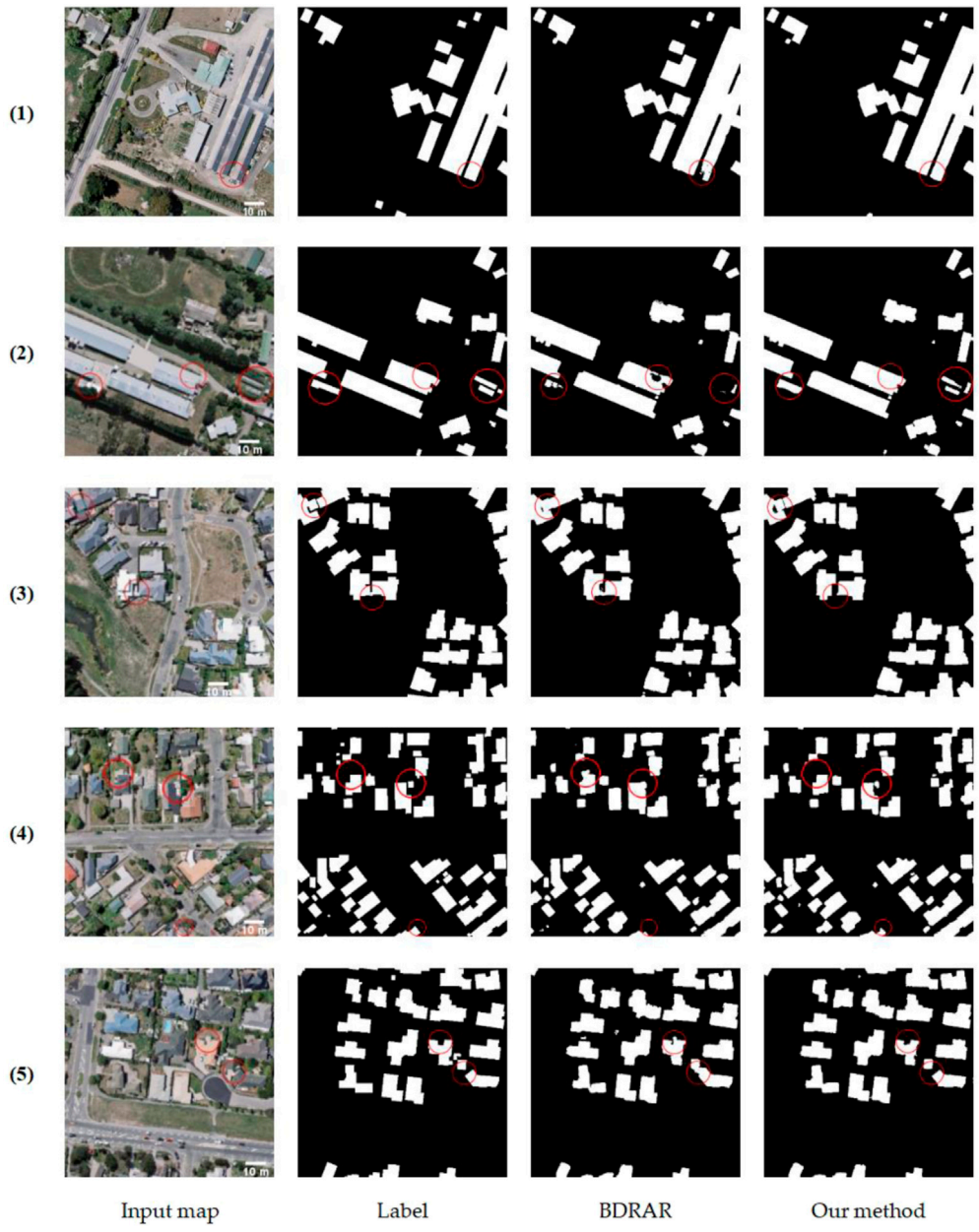

FIGURE 9
Comparison of the accuracy of our method and BDRAR in extracting building shape. The parts circled in red in the picture show significant differences in comparison. In the above figure, the first column is the input map, the second column is the label map, the third column is the experimental result map of the BDRAR model (Zhu, et al., 2018), and the fourth column is the experimental result map of this paper.

512 pixels, covering residential, factory, urban, rural, etc. buildings in the area. The training set size is 4,736 images, the validation set size is 1,036 images, and the test set size is 2,416.

The training platform used in this paper has a 24G GPU and an 8-core CPU. Train the model using the training set from the WHU dataset. The training batch size is 4, the number of training times is 15w, and the initial learning rate is 0.005. The entire training network is optimized using Stochastic Gradient Descent (SGD) with a weight decay 0.0005. We use a deep supervision method (Lee, et al., 2014) to set a branch classifier on the output of each L-CAFSFM to supervise the quality of the output, thus facilitating the dissemination of helpful information. The loss of the L-CAFSFM in each direction in the bidirectional feature pyramid network is calculated, and the total loss is the sum of the losses in the two directions.

3.2 Comparative test

We use the test set in the WHU dataset to test our model. The test set has 2,416 remote sensing images and has a resolution of 512 × 512 pixels. The experiment in this paper is compared with that

in the original paper (Liu, et al., 2021), and some results are shown in Figure 9.

As we can see from the first and second rows of Figure 9. The results of the BDRAR model showed some missing inspections, cavities, or missing corners, resulting in an incomplete building shape. The shapes of the buildings detected in this paper are relatively regular and complete. In the third row, the BDRAR model identifies multiple buildings with close distances into one, while this paper can clearly display the boundaries of multiple buildings. It can be seen from the fourth and fifth rows that when BDRAR model distinguishes buildings and open spaces in front of doors, it mistakenly detects open spaces as buildings, and some buildings are not detected under the shelter of trees. This method can distinguish buildings from their adjacent open spaces and can still identify buildings in the case of tree interference.

TABLE 1 Comparison of experimental results.

	Precision (%)	F-score (%)	Recall (%)	IoU (%)
BOMSC-Net	95.14	94.80	94.50	90.15
BMFR-Net	94.31	93.95	94.42	89.32
RSR-Net	94.92	-	92.63	88.32
B-FGC-Net	95.03	94.76	94.49	90.04
STT	-	94.13	-	89.01
SRI-Net	**95.21**	94.23	93.28	89.09
DR-Net	-	93.80	-	88.30
Ours	95.17	**94.83**	**94.51**	**90.18**

The bolded text in table is meant to highlight the maximum values for each of the evaluation indicators.

4 Discussion

In order to quantitatively analyze the detection effect of our method and other excellent networks, we select the building detection network or semantic segmentation network in recent 3 years as the comparison network of this paper, namely, BOMSC-Net (Zhou, et al., 2022), BMFR-Net (Ran, et al., 2021), STT (Chen, et al., 2021a), SRI-Net (Liu, et al., 2019), DR-Net (Chen, et al., 2021b), RSR-Net (Huang, et al., 2022)and B-FGC-Net (Wang, et al., 2022).

4.1 Evaluation and comparisons

BOMSC-Net proposes a Multi-Scale Context Awareness Module (MSCAM) and a Direction Feature Optimization Module (DOM) by combining boundary optimization and multi-scale context awareness for problems such as tree and shadow occlusion and complex building roof materials. BMFR-Net combines Continuous Atrous Convolution Pyramid (CACP) module and Multi-scale Output Fusion Constraint (MOFC) for building detection. The two-way path conversion module is proposed in the Self-Service Terminal (SST) network, which can learn the long-term dependence features in space and channel dimensions and obtain more accurate building features. Spatial Residual Inception Network (SRI-Net) introduces deeply separable convolution and convolution decomposition, significantly reducing the number of model parameters while retaining global morphological features and local details. It makes the model lighter and more accurate in extracting building features. Dual-Rotation Network (DR-Net) combines densely connected convolutional neural network (DCNN) and residual network (ResNet) structures to extract buildings. RSR-Net improves the model's performance by introducing the SE attention module to reduce the noise effect of shallow features in feature fusion. B-FGC-Net optimizes network training by introducing residual learning and spatial attention units, highlighting the spatial information representation of features.

As mentioned above, the networks use feature pyramids and attention mechanism fusion methods, which are close to our method, so they are selected as the comparative experimental method in this paper. The experimental evaluation indicators are Precision, F-score, Recall, and IoU. Four indicators judge the ability of the network model from different aspects. Precision and Recall can measure the ability of the network to distinguish between false and correct targets. IoU and F-score can evaluate the overall accuracy of the network model.

The above four indicators data are from the original paper for fairness. "-" indicates that the data of this indicator is not provided in the article. The bolded indicator data indicates that the indicator is the highest, and the underline indicates that the indicator is the second. Quantitative evaluations are shown in the table below.

As shown in Table 1, our method's Precision, IoU, Recall, and F-score are 95.17%, 90.18%, 94.51%, and 94.83%, respectively. The accuracy of our method is lower than that of SRI-Net, indicating that the ability to detect the correct target is slightly lower than that of SRI-Net. However, the remaining three indicators in this paper are higher than that, indicating that our method is better than SRI-Net in distinguishing between correct and incorrect targets. Except for SRI-Net, the four indicators in this paper are higher than the above networks, which is enough to prove the advantages of our method in building detection. Therefore, our proposed method can more accurately detect buildings in remote sensing images.

4.2 Ablation studies

In order to verify the effectiveness of the L-CAFSFM and MFFM in this paper, ablation experiments are designed in this paper. The experimental models are divided into BDRAR (Hou, et al., 2022), the network with only the L-CAFSFM, the network with only the MFF module, and the network in this paper. For fairness, all training and testing data configurations are the same. This paper evaluates the four network models from two aspects of qualitative analysis and quantitative calculation. Qualitative analysis and results are shown in Figure 10.

In the above Figure, it can be seen from the second column that the missed detection rate of the BDRAR model is high. In the third column, it can be seen that after adding only the MFF module, the missed detection area decreases. It can be seen from the fourth column that after only adding the L-CAFSFM, the missed detection area is greatly reduced, but at the same time, the false detection area also increases. In the fifth column, that after adding L-CAFSFM and MFFM, compared with using the BDRAR model, the false detection area is slightly increased, but the missed detection area is greatly reduced.

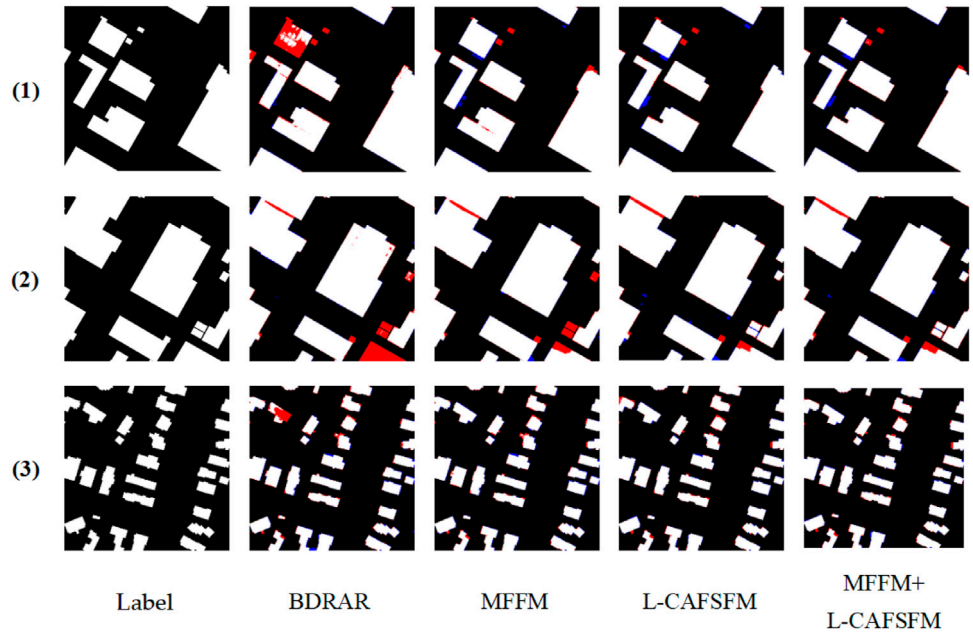

FIGURE 10
Results of ablation experiments. The first column is the label map, the second column is the resulting map of the BDRAR model, the third column is the resulting map of only MFFM, the fourth column is the resulting map of only the L-CAFSFM and the fifth column is the experimental results of the method. The red part in the Figure is the missed detection area, the blue part is the false detection area, and the other parts are the same as the label map, which is the positive detection area.

TABLE 2 Results of ablation experiments.

	Precision (%)	F-score (%)	Recall (%)	IoU (%)
BDRAR	93.91	93.67	93.42	88.09
Only L-CAFSFM	94.94	94.32	93.70	89.25
Only MFFM	94.64	94.13	94.13	89.37
L-CAFSFM+ MFFM	**95.17**	**94.83**	**94.51**	**90.18**

The bolded text in table is meant to highlight the maximum values for each of the evaluation indicators.

The quantitative assessment of the ablation experiments is shown in Table 2.

It can be seen from Table 2 that after adding only the L-CAFSFM, the prediction accuracy and F-score are greatly improved, and the recall rate and IoU are only slightly improved. After only adding the MFF, all four indicators are improved, but the recall rate and IOU are greatly improved. After adding L-CAFSFM and MFFM, the four indicators have been greatly improved, consistent with the conclusions in the above experimental results. Therefore, it can be proved that L-CAFSFM and MFFM proposed in this paper have sound effects.

5 Conclusion

In this article, we propose a building extraction method of combining L-CAFSFM, MFFM with a bidirectional feature pyramid network. L-CAFSFM calculates and fuses the feature maps of two adjacent levels to extract finer building details. The bidirectional feature pyramid network iteratively calculates L-CAFSFM and gradually learns multi-level feature information from two different directions to obtain fine features at different levels. MFFM integrates outputs from two directions to complement building feature information. The results are optimized using a dense conditional random field. Through the above improvements, the ability of the method to obtain rich and specific spatial features is further enhanced. Our method achieves state-of-the-art performance compared to other advanced models on the WHU dataset. The combination of L-CAFSFM and MFFM still has excellent potential to be applied in the field of computer vision, and we will continue to learn how to better apply L-CAFSFM and MFFM to building detection in the future.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://study.rsgis.whu.edu.cn/pages/download/.

Author contributions

WF: Conceptualization, Methodology, Validation, Writing–review and editing. FY: Conceptualization, Formal analysis, Methodology, Writing–review and editing. XC: Conceptualization, Formal analysis, Methodology, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdollahi, A., Pradhan, B., Gite, S., and Alamri, A. (2020). Building footprint extraction from high resolution aerial images using generative adversarial network (gan) architecture. *IEEE Access* 8, 209517–209527. doi:10.1109/ACCESS.2020.3038225

Bauchet, J.-P., Mapurisa, W., Gobbin, A., Tripodi, S., Tarabalka, Y., Duan, L., et al. (2021). Rooftops or footprints? Reliable building footprint extraction from high-resolution satellite images. *IEEE Int. Geoscience Remote Sens. Symposium*, 274–277. doi:10.1109/IGARSS47720.2021.9554755

Borba, P., Diniz, F. D. C., Silva, N. C. D., and Bias, E. d. S. (2021). Building footprint extraction using deep learning semantic segmentation techniques: experiments and results. *IEEE Int. Geoscience Remote Sens. Symposium*, 4708–4711. doi:10.1109/IGARSS47720.2021.9553855

Cai, J., and Chen, Y. (2021). MHA-net: multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 14, 5807–5817. doi:10.1109/JSTARS.2021.3084805

Cai, Y., Chen, D., Tang, Y., Zhang, J., and Gao, Y. (2021). "Multi-scale building instance extraction framework in high resolution remote sensing imagery based on feature pyramid object-aware convolution neural network," in Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, July 2021, 2779–2782. doi:10.1109/IGARSS47720.2021.9554016

Chen, H., and Sun, W. (2022). "Building extraction from remote sensing images with conditional generative adversarial networks," in Proceedings of the 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, July 2022, 655–658. doi:10.1109/ICSIP55141.2022.9886096

Chen, J., Zhang, D., Wu, Y., Chen, Y., and Yan, X. (2022a). A context feature enhancement network for building extraction from high-resolution remote sensing imagery. *Remote Sens.* 14, 2276. doi:10.3390/rs14092276

Chen, K., Zou, Z., and Shi, Z. (2021a). Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* 13 (21), 4441. doi:10.3390/rs13214441

Chen, M., Wu, J., Liu, L., Zhao, W., Du, R., Shen, Q., et al. (2021b). DR-Net: an improved network for building extraction from high resolution remote sensing image. *Remote Sens.* 13 (2), 294. doi:10.3390/rs13020294

Chen, S., Shi, W., Zhou, M., Zhang, M., and Xuan, Z. (2022b). CGSANet: a contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 15, 1526–1542. doi:10.1109/JSTARS.2021.3139017

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). Deformable convolutional networks. Avaialble at: https://arxiv.org/abs/1703.06211.

Das, P., and Chand, S. (2021). "AttentionBuildNet for building extraction from aerial imagery," in Proceedings of the International Conference on Computing, Communication, and Intelligent Systems, Greater Noida, India, February 2021, 576–580. doi:10.1109/ICCCIS51004.2021.9397178

Fang, F., Zheng, D., Li, S., Liu, Y., Zeng, L., Zhang, J., et al. (2022). Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 15, 1629–1642. doi:10.1109/JSTARS.2022.3144176

Guo, M., Liu, H., Xu, L., and Huang, Y. (2020). Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* 12 (9), 1400. doi:10.3390/rs12091400

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, June 2021, 13708–13717. doi:10.1109/CVPR46437.2021.01350

Hou, X., Wang, P., and An, W. (2022). "Multi-scale residual network for building extraction from satellite remote sensing images," in Proceedings of the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, July 2022, 1348–1351. doi:10.1109/IGARSS46834.2022.9883509

Huang, H., Chen, Y., and Wang, R. (2022). A lightweight network for building extraction from remote sensing images. *IEEE Trans. Geoscience Remote Sens.* 60 (5614812), 1–12. doi:10.1109/TGRS.2021.3131331

Jie, H., Li, S., and Gang, S. (2017). "Squeeze-and-Excitation networks," in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.

Lee, C. Y., Xie, S., and Gallagher, P. (2014). Deeply-supervised nets. Avaialble at: https://arxiv.org/abs/1409.5185.

Li, J., Wang, C., Zhang, H., Wu, F., Li, L., and Gong, L. (2020a). "Automatic extraction of built-up areas for cities in China from GF-3 images based on improved residual U-Net network," in Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, October 2020, 4399–4402. doi:10.1109/IGARSS39084.2020.9324329

Li, Q., Shi, Y., Huang, X., and Zhu, X. X. (2020b). Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF. *IEEE Trans. Geoscience Remote Sens.* 58 (11), 7502–7519. doi:10.1109/TGRS.2020.2973720

Lin, X., and Zhang, J. (2017). Object-based morphological building index for building extraction from high resolution remote sensing imagery. *Acta Geod. Cartogr. Sinica* 46 (6), 724–733. doi:10.11947/j.AGCS.2017.20170068

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., et al. (2019). Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.* 11 (7), 830. doi:10.3390/rs11070830

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., and Zhang, Z. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, October 2021, 9992–10002. doi:10.1109/ICCV48922.2021.00986

Ma, W., Wan, Y., Li, J., Zhu, S., and Wang, M. (2019). An automatic morphological attribute building extraction approach for satellite high spatial resolution imagery. *Remote Sens.* 11 (3), 337. doi:10.3390/rs11030337

Park, J., Woo, S., and Lee, J. Y. (2018). BAM: bottleneck attention module. Avaialble at: https://arxiv.org/abs/1807.06514.

Ran, S., Gao, X., Yang, Y., Li, S., Zhang, G., and Wang, P. (2021). Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images. *Remote Sens.* 13 (14), 2794. doi:10.3390/rs13142794

Saini, A., Dixit, M., Prajapati, A., and Kushwaha, I. (2021). "Specific structure building extraction from high resolution satellite image," in Proceedings of the International Conference on Advances in Computing, Communication Control and Networking, Greater Noida, India, December 2021, 486–489. doi:10.1109/ICAC3N53548.2021.9725471

Sheikh, M. A. A., Maity, T., and Kole, A. (2022). IRU-net: an efficient end-to-end network for automatic building extraction from remote sensing images. *IEEE Access* 10, 37811–37828. doi:10.1109/ACCESS.2022.3164401

Shi, X., Huang, H., Pu, C., Yang, Y., and Xue, J. (2022). CSA-UNet: channel-spatial attention-based encoder–decoder network for rural blue-roofed building extraction from UAV imagery. *IEEE Geoscience Remote Sens. Lett.* 15, 1–5. doi:10.1109/LGRS.2022.3197319

Tian, Q., Zhao, Y., Li, Y., Chen, J., Chen, X., and Qin, K. (2022). Multiscale building extraction with refined attention pyramid networks. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/LGRS.2021.3075436

Wang, C., Shen, Y., Liu, H., Zhao, K., Xing, H., and Qiu, X. (2019). Building extraction from high–resolution remote sensing images by adaptive morphological attribute profile under object boundary constraint. *Sensors* 19 (17), 3737. doi:10.3390/s19173737

Wang, Y., Zeng, X., Liao, X., and Zhuang, D. (2022). B-FGC-Net: a building extraction network from high resolution remote sensing imagery. *Remote Sens.* 14, 269. doi:10.3390/rs14020269

Wei, R., Fan, B., Wang, Y., Zhou, A., and Zhao, Z. (2022). MBNet: multi-branch network for extraction of rural homesteads based on aerial images. *Remote Sens.* 14, 2443. doi:10.3390/rs14102443

Woo, S., Park, J., Lee, J., and Kweon, S. (2018). CBAM:Convolutional block attention module. Avaialble at: https://arxiv.org/abs/1807.06521.

Xiao, X., Guo, W., Chen, R., Hui, Y., Wang, J., and Zhao, H. (2022). A Swin transformer-based encoding booster integrated in U-shaped network for building extraction. *Remote Sens.* 14, 2611. doi:10.3390/rs14112611

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 2017, 5987–5995. doi:10.1109/CVPR.2017.634

Xie, Y., Zhu, J., Cao, Y., Feng, D., Hu, M., Li, W., et al. (2020). Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 13, 1842–1855. doi:10.1109/JSTARS.2020.2991391

Yan, X., Shen, L., Wang, J., Wang, Y., Li, Z., and Xu, Z. (2022). PANet: pixelwise affinity network for weakly supervised building extraction from high-resolution remote sensing images. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/LGRS.2022.3205309

Yang, B., Huang, Y., Su, X., and Guo, H. (2022). MAEANet: multiscale attention and edge-aware siamese network for building change detection in high-resolution remote sensing images. *Remote Sens.* 14, 4895. doi:10.3390/rs14194895

You, D., Wang, S., Wang, F., Zhou, Y., Wang, Z., Wang, J., et al. (2022). EfficientUNet+: a building extraction method for emergency shelters based on deep learning. *Remote Sens.* 14, 2207. doi:10.3390/rs14092207

Yu, Y., Ren, Y., Guan, H., Li, D., Yu, C., Jin, S., et al. (2021). Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery. *IEEE Geoscience Remote Sens. Lett.* 18 (5), 895–899. doi:10.1109/lgrs.2020.2986380

Yuan, Q., and Mohd Shafri, H. (2022). Multi-modal feature fusion network with adaptive center point detector for building instance extraction. *Remote Sens.* 14 (19), 4920. doi:10.3390/rs14194920

Zhang, L., Dong, R., Yuan, S., and Fu, H. (2022). "Srbuildingseg-E2: an integrated model for end-to-end higher-resolution building extraction," in Proceedings of the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, July 2022, 1356–1359. doi:10.1109/IGARSS46834.2022.9883295

Zhang, L., Zhong, B., and Yang, A. (2019). "Building change detection using object-oriented LBP feature map in very high spatial resolution imagery," in Proceedings of the 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), Shanghai, China, August 2019.

Zhou, Y., Chen, Z., Wang, B., Li, S., Liu, H., Xu, D., et al. (2022). BOMSC-net: boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. *IEEE Trans. Geoscience Remote Sens.* 60, 1–17. doi:10.1109/TGRS.2022.3152575

Zhu, L., Deng, Z., and Hu, X. (2018). *Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection*. Cham: Springer.

Zhu, Q., Liao, C., Hu, H., Mei, X., and Li, H. (2021). MAP-net: multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Trans. Geoscience Remote Sens.* 59 (7), 6169–6181. doi:10.1109/TGRS.2020.3026051