Received 11 December 2022, accepted 19 December 2022, date of publication 21 December 2022,
date of current version 29 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3231362

RESEARCH ARTICLE

FMAM-Net: Fusion Multi-Scale Attention Mechanism Network for Building Segmentation in Remote Sensing Images

HUANRAN YE[1], RUN ZHOU[1], JIANHAO WANG[1], AND ZHILIANG HUANG[2], (Member, IEEE)
[1]School of Mechanical and Electrical Information, Yiwu Industrial & Commercial College, Jinhua, Zhejiang 322000, China
[2]College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

Corresponding author: Huanran Ye (huanranye@ywicc.edu.cn)

This work was supported in part by the Zhejiang Provincial Department of Education under Grant 2022JYTYB03, in part by the Zhejiang Federation of Social Sciences under Grant 2023B071, and in part by the Jinhua Science and Technology Association under Grant 2022JKX37.

ABSTRACT As the largest target in remote sensing images, buildings have important application value in urban planning and old city reconstruction. However, most networks have poor recognition ability on high resolution images, resulting in blurred boundaries in the segmented building maps. Then, the similarity between buildings and backgrounds will lead to inter-class indistinction. Finally, the diversity of buildings brings difficulties to segmentation, which requires the network to have better generalization ability. To address these problems, we propose Fusion Multi-scale Attention Mechanism Network (FMAM-Net). Firstly, we design Feature Refine Compensation Module(FRCM) to improve the boundary ambiguity problem, including Feature Refinement Module(FRM) and Feature Compensation Module(FCM). FRM utilizes the densely connected architecture to refine features and increase recognition capabilities. FCM introduces low-level features to make up for the lack of boundary information in high-level features. Secondly, to handle inter-class indistinction, we design Tandem Attention Module(TAM) and Parallel Attention Module(PAM). TAM is designed to sequentially filter some features from channels and spaces for adaptive feature refinement. PAM combines context information and uses high-level features to guide low-level features to select more distinguishable features. Finally, based on the binary cross entropy loss function, we add an evaluation index to reduce the error caused by determining the optimization direction only through cross entropy. On the Inria Aerial Image Labeling Dataset, FMAM-Net achieves mean IoU of 85.34%, which is 5.58% higher than AMUNet and 3.77 % higher than our baseline(U-Net ResNet-34). On the WHU Dataset, IoU reached the maximum value of 91.06% on FMAM-Net, 1.67% higher than SARB-UNet and 0.2% higher than MAP-Net. The visualization results show that FMAM-Net improves the fuzzy boundary of building segmentation and reduces the inter-class indistinction.

INDEX TERMS Remote sensing image, building segmentation, attention mechanism, feature refinement, encoder-decoder.

I. INTRODUCTION

With the rapid development of aerospace technology, a large number of satellites have been launched one after another, and the acquisition of high-resolution remote sensing images is also simpler [1]. Therefore, automatic semantic segmentation of these images is required to quickly and efficiently identify objects of interest, such as buildings [2], roads [3], and land cover [4], [5]. Among them, building extraction is the most critical task and is usually used for monitoring subtle changes in urban areas, urban planning and building demolition statistics. Moreover, the manual annotation task in the remote

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

sensing field is time-consuming and labor-intensive, which makes it urgent to design an automatic semantic segmentation model with high accuracy.

Early building segmentation used machine learning methods to achieve segmentation by carefully designing some feature descriptors. The work in [6] proposed fuzzy clustering techniques, as well as support vector machine (SVM) classifiers to deal with the problem of unsupervised image segmentation of a satellite image in a number of homogeneous regions. The work in [7] combined the results of a fuzzy output SVM classifier with spatial information extracted from image segmentation to improve segmentation accuracy. In [8], a modifed swarm optimization approach and OTSU's binary threshold method are helpful to improve the sensitivity, specificity and accuracy of high resolution satellite images. Although traditional machine learning has achieved certain results, there are also obvious shortcomings. For example, the poor generalization ability of manually selected features leads to inaccurate segmentation. In addition, unfavorable factors on remote sensing images, such as tree occlusion and shadows, can also lead to suboptimal results, as shown in Figure 1.

The deep learning algorithms avoid manual selection of features and has higher segmentation accuracy. The most used deep learning algorithms in the segmentation field mainly include: encoder-decoder networks, region selection networks and deepLab series networks. The encoder-decoder network [9], [10], [11], [12], [13] uses the encoder to extract features, the decoder to fuse features, and the skip connections to transfer features. The network architecture based on region selection [14], [15], [16], [17], [18] was first used for detection, and gradually evolved into segmentation with development, and achieved higher accuracy in the segmentation algorithm. The network based on the DeepLab series [19], [20], [21], [22] shows that the multi-scale information is combined to complete the segmentation task.

A. RELATED WORK

With the rapid improvement of GPU computing speed, deep learning has become a research hotspot for building segmentation. Among them, the encoder-decoder network is the most widely used and easiest to understand of segmentation network architecture. Some scholars are committed to improving the encoder and decoder to achieve the improvement of the feature extraction level. The U-Net ResNet-34 network proposed by [23] uses ResNet with better feature extraction ability to replace the encoder, and combines the knowledge of transfer learning to improve the accuracy from the source. In [24], selective spatial pyramid dilated (SSPD) network proposes enhanced encoder and the dual-stage decoder to recover the crucial multi-scale information better. The work in [25] builds the Strip Pooling module (SPM) and the Mixed Pooling module (MPM) in encoder to better extract the vacancy features. In [26], a single-side dual-branch network (SSDBN) based on an encoder–decoder structure uses an improved Res2Net model to capture multi-scale information and enhancing high-level semantic details.

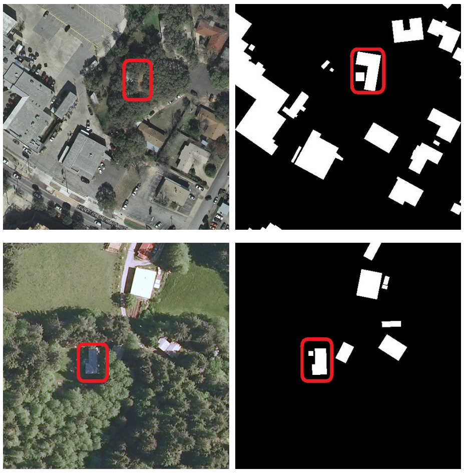


FIGURE 1. Tree occlusion and shadow in remote sensing images. In the red box on the above, the house was obscured by the green trees, making it difficult to distinguish. In the red box on the below, building obscured by shadow has a different appearance than other buildings and is easily misidentified as non-building.

Other scholars consider improving skip connections, so that skip connections can better refine features and reduce information loss. It is not difficult to find from the original U-Net model that skip connections are only used to transfer the encoder features to the decoder, and feature optimization is not performed during the transfer process. Therefore, UNet++ proposed by [27] connects the first four modules of the encoder to select more beneficial features from different levels of features, aiming to change the connectivity between the encoder and decoder. The work in [28] proposes UNet 3+, which makes use of full size jump connections. Full size jump connections combine low-level details and high-level semantics of feature maps from different scales. The Web-Net proposed by [29] consists of encoder, decoder, and node layers nested, making more complex improvements on skip connections. The node layer absorbs the feature maps of adjacent node layers and remote node layers along the horizontal and vertical directions, and adjusts the input layer features at different levels to the same size for further processing at each node layer. The work in [30] proposed channel transformer module to replace skip connection, and effectively connected to decoder features through multi-scale channel cross fusion sub module and multi-scale channel information guided fusion to solve semantic gap.

Attention model is another powerful tool for deep learning [31]. DFN [32] increases channel attention to select more discriminative features, which helps to reduce ambiguity. The work in [33] propose a hybrid first and second order attention network (HFSA) that explores both the global mean and the inner-product among different channels to achieves better building segmentation. In [34], through the self-attention module, the network will pay more attention to positions where there may be salient regions. In [35],

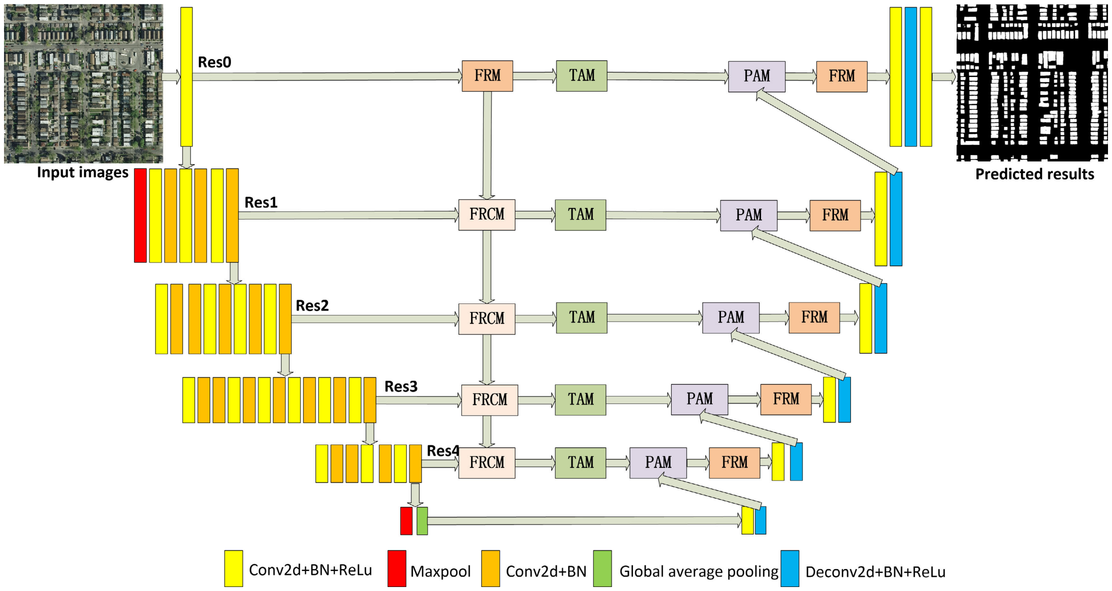


FIGURE 2. Architecture of Fusion Multi-scale Attention Mechanism Network. The left part of the figure is the encoder, the middle is the nested skip connections architectures and the right part is the decoder. Features at different stages have different information. The downward arrow represents feature extraction, the right arrow represents feature transfer, and the upward arrow represents feature upsampling. FRCM: Feature Refine Compensation Module. FRM: Feature Refinement Module. TAM: Tandem Attention Module. PAM:Parallel Attention Module.

MTPA-Net introduces channel attention module and location attention module respectively to capture long-term context information from spatial and channel dimensions. MHA-Net [36] design a separable convolution block attention module and an attention downsampling module as the basic modules with separable convolutions and channel attention. Therefore, we introduce both spatial and channel attention models to select features similar to MTPA-Net.

B. CONTRIBUTIONS

However, the current automatic segmentation tasks for buildings still have the following problems: (1) The poor recognition ability of most networks on high-resolution images is a major difficulty. (2) Similarity between buildings and backgrounds leads to inter-class indistinction is a major difficulty. (3) Diverse architectural styles and sizes require a more adaptable network model, which is a major difficulty. To solve the above problems, this paper proposes a novel Fusion Multi-scale Attention Mechanism Network (FMAM-Net). The main contributions of this work can be summarized as follows:

- Nesting architecture enables features to be refined and filtered in the horizontal direction, and features at different levels in the vertical direction are semantically compensated.
- The Feature Refine Compensation Module(FRCM) realizes feature refinement by reusing, and integrates low-level features into high-level features to compensate for the lack of spatial boundary information, which helps to obtain clear building contours.
- The Tandem Attention Module(TAM) scales attention map from channel and spatial dimensions back to the original feature map for adaptive feature refinement.
- The Parallel Attention Module(PAM) introduces the high-level features to guide the selection of low-level features is helpful to screen features with strong recognition ability.
- A new multi-restriction loss function introduces IoU as one of the terms based on the binary cross-entropy loss function, which helps to reduce errors caused by only a single restriction.

II. METHOD

A. NETWORK ARCHITECTURE

FMAM-Net is improved on the basis of U-Net ResNet-34. The overall architecture is shown in Figure 2. FMAM-Net makes different improvements in the horizontal and vertical directions of the encoder, decoder, and skip connections. We replaces the original convolutional layer in the encoder with ResNet-34, which possesses better feature extraction effect to improve the segmentation accuracy from the source. We embeds Feature Refine Compensation Module(FRCM) and Tandem Attention Module(TAM) in the skip connections of each layer. In the horizontal direction, the extracted features are first further refined by FRCM. This is mainly due to the dense network carefully designed in this paper, which contains multiple reused modules to maximize the utilization of features. In TAM, the features are sequentially fed to the channel attention module and the spatial attention module in series, so that the model knows "what" and "where" to focus on. In the vertical direction, low-level features are

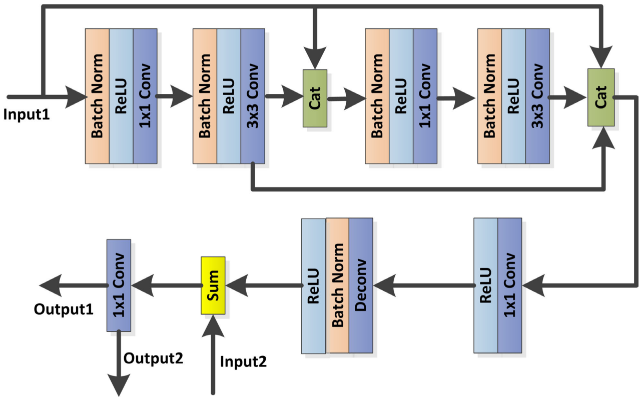


FIGURE 3. Architecture of the Feature Refine Compensation Module(FRCM). The upper half of the graph is Feature Refinement Module(FRM),and the lower half of the graph is Feature Compensation Module(FCM). Input1 represents the features of the current stage,and Input2 represents the features of the lower stage. Output1 represents the features output to the next FRCM, and Output2 represents the features output to the decoder.

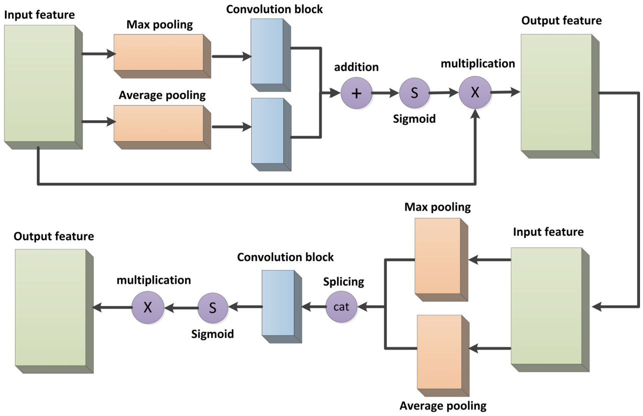


FIGURE 4. Architecture of Tandem Attention Module.

integrated into high-level features by resizing, the benefit is that low-level features have better spatial information. In the decoder, we first introduces a convolution to arrange features, which are more conducive to subsequent deconvolution for upsampling. Then, we pay more attention to feature selection, using Parallel Attention Module(PAM) to select more discriminative features from channel and space dimensions.

B. FEATURE REFINE COMPENSATION MODULE

As an important part of skip connections, Figure 3 shows the workflow of the Feature Refine Compensation Module(FRCM). When the buildings have similar appearances with the background, it is easy to confuse the two categories. For example, buildings are covered by trees with the same color as vegetation, and building construction materials are basically the same as pavement materials. Therefore, we design FRCM to enhance recognition capabilities and expand the distinction between classes, and solve the blurred and irregular contours of the building maps. This is similar to the edge spatial attention block in [37]: (1) both modules use reuse mechanism in design to further refine features; (2) In order to obtain more edge information, the both modules obtain different dimensional features in different levels respectively for fusion. FRCM includes two sub-module: Feature Refinement Module(FRM) and Feature Compensation Module(FCM).

FRM is inspired by DenseNet[38], which can enhance the recognition ability and refine the features. Therefore, the output features of each level of the encoder are all refined through the FRCM, as shown in Figure 2. DenseNet reuses features for many times to achieve feature extraction and refinement, and FRM also introduces this idea and makes corresponding improvements. After testing, it is found that using more than two bottleneck layers not only makes the effect worse, but also reduces the algorithm speed. Therefore, FRM design two bottleneck layers in skip connection, and the output can be calculated by the following formula:

$$x_4 = H([x_1 + x_2 + x_3]) \tag{1}$$

where x_1 is the input feature, x_2 represents the output after the first bottleneck layer, x_3 represents the output after the second bottleneck layer, x_4 represents the combined output, and H represents the non-linear transformation. The first part of the FRM is the basic bottleneck layer, which is used to refine the features. 1 × 1 convolution is used to reduce the number of parameters, while 3 × 3 convolution is used to extract features. This is followed by another basic bottleneck layer with consistent features, and the features of each bottleneck layer have been reused to enhance the recognition ability.

Considering the features of different levels include different information(such as low-level features have more complete spatial information, while high-level features have more obvious semantic features), FCM introduces low-level features to supervise and learn the boundary, as shown in the lower half of Figure 3. FCM is inspired by U-Net's decoder. In the U-Net decoder, the low-level features passed by skip connections have precise spatial information, which largely bridges the semantic gap. We also use low-level spatial information to make up for segmentation defects on the boundary. The difference is that this operation is done on skip connections. FCM includes two inputs, such as Input1 from the FRM and Input2 from the previous stage FCM. Because these two inputs come from different stages, their feature map size and number of channels are different. First, a 1 × 1 convolutional layer is used to unify the number of channels of the two inputs. The subsequent deconvolution layer upsamples the output to ensure the same size as Input2. Then high-level (Input1) and low-level (Input2) features are merged. Finally, the feature map is restored to its original size through the 1 × 1 convolutional layer. Output1 is used for the next FRCM and Output2 is applied for decoder. With this design, we can make full use of the boundary information to obtain a sharper boundary.

C. TANDEM ATTENTION MODULE

The features output by the FRCM will be sequentially fed into the Tandem Attention Module(TAM), in order to make the features focus on things and locations of interest[39], as shown in the middle of Figure 2. It is easy to confuse the two categories when buildings have a similar appearance to the background. Especially when they are adjacent at the boundary, the contours are fuzzy and irregular. Therefore, TAM is designed to expand the distinction between classes, which addresses the blurred and irregular contours of building segmentation. The workflow of TAM is shown in Figure 4.

The design principle of the channel attention block is to redistribute the weights according to the importance of each channel, as shown in the upper part of Figure 4. First, the input features are passed through the max pooling layer and the average pooling layer to generate two vectors with 1×1. Since only the elements in the channel are concerned, both pooling layers compress the input features into the channel statistics. Subsequently, both vectors are fed to the convolution block to further generate the channel attention map. The convolutional block consists of two convolutional layers and a ReLU function. In order to reduce parameters and reduce operation consumption, the output channel of the first convolutional layer is reduced to 1/16 of the input channel. The number of output channels of the second convolutional layer is restored to its original size. In order to summarize the features, a sum operation is performed to merge the two channel attention maps, and then the sigmoid activation function is used to output the score map of the channel. To perform element-wise multiplication between the input features and the channel score map, the formula can be expressed as follows:

$$y_p = e_p \times x_p \tag{2}$$

where e_p denotes the score map, x_p denotes the input features and y_p is the rescaled input features.

The principle of the spatial attention block is to compress the channel while retaining the complete feature map to achieve the effect of paying attention to the spatial information. The detailed structure is shown in the lower half of Figure 4. First, max pooling and average pooling are performed on the channel dimension to generate two feature maps with a channel number of 1. Then stitch the two feature maps and apply a convolutional layer to generate a spatial attention map while reducing the number of channels. The classification probability of each pixel is normalized to [0, 1] by the sigmoid activation function. According to the probability score map of the features, the input features are rescaled to selectively enhance the features of interest and achieve the effect of solving the blurring of building boundaries.

D. PARALLEL ATTENTION MODULE

To handle the inter-class indistinction well, Parallel Attention Module(PAM) is designed in the decoder. In the encoder-decoder architecture, most scholars focus on improving the encoder and skip connections. The decoder still uses simple upsampling or deconvolution, which largely ignores contextual information, resulting in inter-class indistinction in the results. Therefore, we propose PAM that low-level features and high-level features are introduced in the decoder. High-level features are used to guide low-level feature selection, and more discriminative features are selected from both spatial and channel dimensions. PAM consists of two parts: the channel attention module and the spatial attention module, as shown in Figure 5.

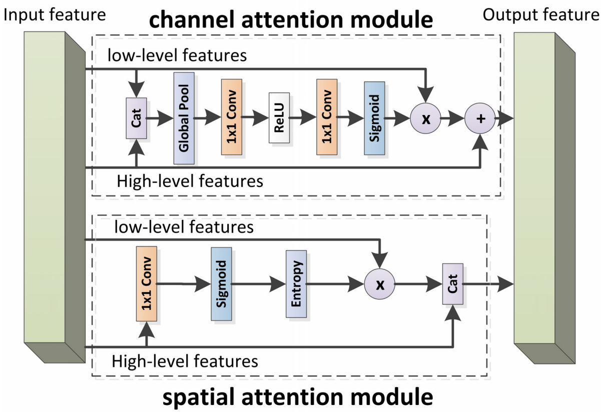


FIGURE 5. Architecture of parallel attention module.

The channel attention module aims to change the weights of features in each channel to enhance the consistency of features, as shown in the upper part of Figure 5. First, high-level features and low-level features are combined in the channel dimension to form a new feature map, which is conducive to the effective use of features. In order to achieve the effect of only focusing on the features in the channel, the global average pooling is used to compress the input features x into the channel statistics s, and the c-th channel of s can be calculated by the following formula:

$$s_c(x_c) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} x_c(i, j) \tag{3}$$

where x_c is the c-th channel of the input feature x, and $M \times N$ denote spatial dimensions of x_c. In order to restore the number of feature map channels, the first 1×1 convolution is used to restore the number of merged channels to the original size, and then the features is activated by RELU function. The second 1×1 convolution is used to generate the feature map, and then the feature score map is generated by the sigmoid function. Multiply operation rescales the score map back to the low-level features, which use the score map of the high-level features to guide the low-level features to select effective features with greater weights on the channel. Finally, the selected low-level features and high-level features are summed by addition operation to output.

The detailed structure of the spatial attention mechanism is shown in the lower half of Figure 5. The high-level features are first passed through a 1×1 convolutional layer for the purpose of dimensionality reduction on the number of channels to focus on the spatial features. The score map is normalized to [0, 1] by sigmoid function. Then, the entropy score map is calculated element-wise for the score map and multiplied with the low-level features to assign the weights

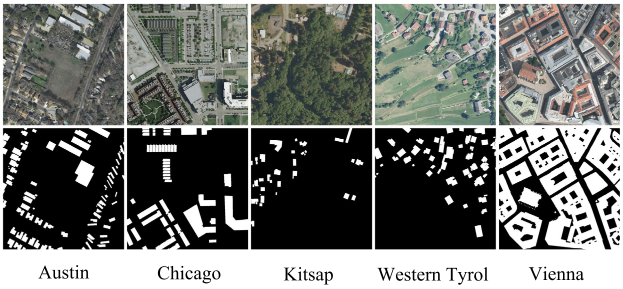


FIGURE 6. Inria aerial image labeling dataset.

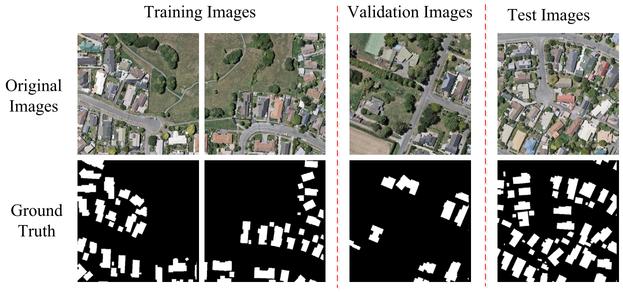


FIGURE 7. Several training, validation and testing sample images in the WHU Dataset.

of the entropy score map to the low-level features. After that, the high-level features are summed with the weighted low-level features for further processing. It is worth noting that the entropy score map has a strong relationship with building boundaries in the building extraction task. Therefore, the spatial attention mechanism can bring benefits for building boundary segmentation, especially in combination with the overall binary cross-entropy loss to train segmentation network.

III. EXPERIMENTS

A. DATASETS

We evaluate our approach on two public datasets: Inria Aerial Image Labeling Dataset [40] and WHU Dataset [41].

The Inria Aerial Image Labeling dataset mainly contains five urban settlements from Austin, Chicago, Kissap County, West Tyrol, and Vienna, and each image is 5000 × 5000 pixels in size. To ensure that the image size meets the training requirements, the images are cropped to 1024 × 1024 pixels(2880 samples). At the same time, the first five images of each region are selected to create a test set(400 samples), and the rest of the images are used as the training set(2480 samples), which is consistent with other literatures. As shown in Figure 6, we can see that the image quality in Inria is high and the architectural styles are diverse.

The WHU Dataset consists 8189 tiles of 512 × 512 pixel. According to the needs of training, the dataset is divided into training set, validation set and test set, which are composed of 4736 images, 1036 images and 2416 images respectively. As shown in Figure 7, we can see that the style of the buildings is basically similar but the background is more complex.

B. METRICS

This paper chooses the similarity coefficient (Dice Coeffient) as the evaluation index, also known as the intersection over union (IoU), which can effectively evaluate the closeness of the two distributions, thus overcoming the influence of the imbalance phenomenon. This is consistent with most algorithms that use Inria as a dataset, and the specific calculation formula can be expressed as:

$$IoU = 2 \times \frac{|A \bigcap B| + C}{|A \bigcup B| + C} \tag{4}$$

where A denotes the ground truth and B means the predicted result. C is a constant 1e-15. In addition, the overall accuracy(Acc.) is also the criterion for evaluation, which is the same metric as in other literatures. Acc. is the proportion of the correctly labeled pixels and its formula is as follows:

$$Acc. = \frac{tp + tn}{tp + tn + fp + fn} \tag{5}$$

where tp denotes the number of true positive pixels, fp denotes the number of false positive pixels, tn denotes the number of true negative pixels, and fn denotes the number of false negative pixels.

For the WHU Dataset, the precision and recall are usually used to evaluate the segmentation performance. They can be calculated as:

$$precision = \frac{tp}{tp + fp} \tag{6}$$

$$recall = \frac{tp}{tp + fn} \tag{7}$$

C. LOSS FUNCTION

In our experiments, our labels are only two types: buildings and non-buildings. We can consider the semantic segmentation of buildings as a binary classification problem of pixels [42]. Therefore, we choose binary sigmoid cross entropy loss as the main body of the loss function, which can be written as:

$$H = -\frac{1}{n}\sum_{i=1}^{n}(y_i log\hat{y}_i + (1 - y_i)log(1 - \hat{y}_i)) \tag{8}$$

where n is the number of images. y_i is the ground truth, and $\hat{y}_i$ is the actual output of the network. However, the direction of the gradient is only determined by the cross entropy is inaccuracy. We add the evaluation indicator to make our training direction more clear. Moreover, the dual restriction has a better experimental effect than the single restriction, and can balance the loss and evaluation indicator. Therefore, we can construct the loss function, shown as following:

$$L = (1 - W) \times H - W \times log(IoU) \tag{9}$$

where H is the binary sigmoid cross entropy loss mentioned above. W represents the best weight after experiment.

TABLE 1. **Detailed performance comparison of components in our model. TAM: Tandem Attention Module. PAM:Parallel Attention Module. FRCM: Feature Refine Compensation Module.**

Method	Mean IoU(%)
(1)U-Net ResNet-34 (baseline)	81.57
(2)U-Net ResNet-34+Loss	82.31
(3)U-Net ResNet-34+Loss+TAM	83.36
(4)U-Net ResNet-34+Loss+TAM+PAM	84.19
(5)U-Net ResNet-34+Loss+TAM+PAM+FRCM	**85.34**

IV. RESULTS AND DISCUSSION

A. ABLATION EXPERIMENT

In this section, we will gradually decompose each part of the model and carry out experiments to analyze the results from both qualitative and quantitative perspectives.

1) QUANTITATIVE ANALYSIS

We choose U-Net ResNet-34 as our baseline. In order to solve the problem of de novo training of the feature extraction network, U-Net ResNet-34 uses the pre-trained ResNet-34 to replace the original convolutional layer to extract features. The evaluation index IoU reaches 81.57%, as shown in Table 1(1).

a: ABLATION FOR LOSS

We use the improved loss function while ensuring the consistency of the overall architecture, dataset and evaluation indicators. But the performance of the above structure reaches 82.31%, is 0.74% higher than U-Net ResNet-34, as shown in Table 1(2). We can observe that the improved loss function is the key of performance improvement. The addition of evaluation indicator compensates for the error of the single restriction.

b: ABLATION FOR TAM

Subsequently, we introduce Tandem Attention Module to solve the fuzzy and irregular contours of building segmentation. The features are scaled back to the original feature map from the channel and spatial dimensions in sequence through the Tandem Attention Module for adaptive feature refinement. As shown in Table 1(3),evaluation index reaches 83.36%, an increase of 1.05%.

c: ABLATION FOR PAM

To address the inter-class indistinction, we embed Parallel Attention Module in the decoder to select more distinguishable features from two dimensions. PAM improves the performance from 83.36% to 84.19%, as shown in Table 1(4). Obviously, the high-level features in the PAM guide the selection of low-level features to effectively avoid inter-class indistinction.

d: ABLATION FOR FRCM

Finally, we add Feature Refine Compensation Module with feature refinement function and feature compensation function to form the final Fusion Multi-scale Attention Mechanism Network. The mean IoU also reaches the maximum value of 85.34% as show in Table 1(5). On the one hand, the dense connection architecture refines the features in FRM. On the other hand, only in FCM, low-level features are introduced into high-level features to compensate spatial information, which helps to segment the building contour.

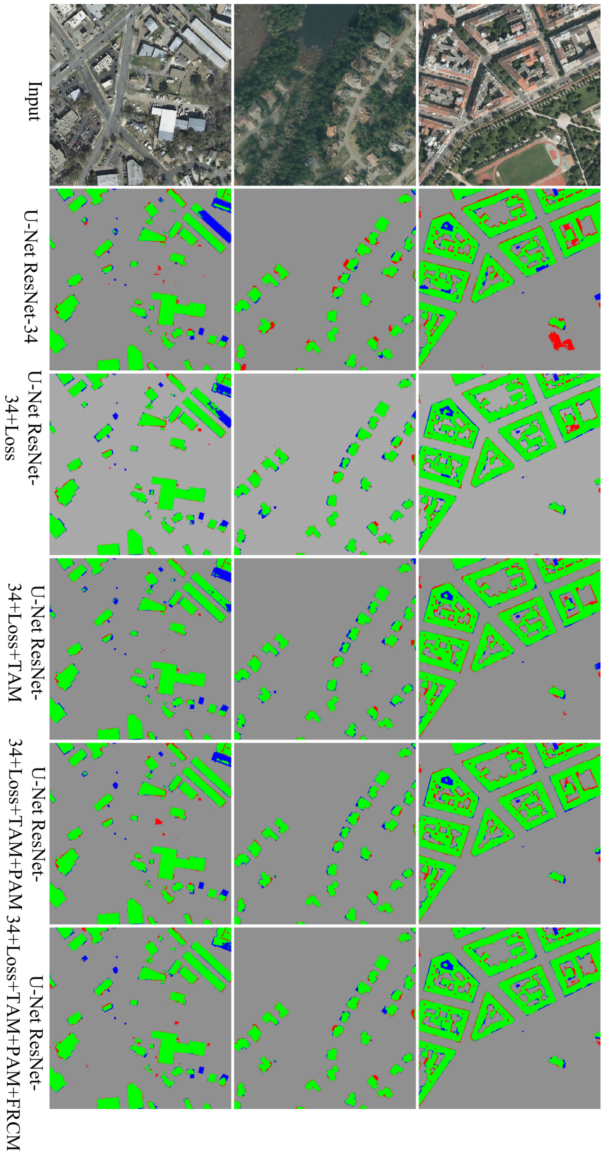


FIGURE 8. **Building segmentation results of ablation experiments. Green: True positive pixels. Gray: true negative pixels. Red: false positive pixels. Blue: false negative pixels.**

2) QUALITATIVE ANALYSIS

In order to see the improvement effect of each module more intuitively, Figure 8 shows the visualization results of five modules in the ablation experiment. As can be seen from the first column, U-Net ResNet-34 mistook inclined buildings in the upper right corner as background(Blue pixels). But this mis-segmentation has been improved in U-Net

TABLE 2. Evaluation results on the Inria Aerial Image Labeling Dataset. The "-" indicates that the method does not present this result in its paper.

Method	Metric	Austin	Chicago	Kitsap	West Tyrol	Vienna	Overall
E-D-Net	IoU	81.85	78.46	77.64	73.76	79.89	79.78
	Acc.	94.78	98.23	98.10	93.25	98.68	96.66
AMUNet	IoU	84.43	81.22	68.13	79.97	85.05	79.76
	Acc.	97.29	96.45	93.83	98.83	97.28	96.73
U-Net ResNet-34(baseline)	IoU	-	-	-	-	-	81.57
	Acc.	-	-	-	-	-	-
FMAM-Net(Proposed)	IoU	**87.92**	**83.17**	**85.58**	**84.44**	**87.13**	**85.34**
	Acc.	97.25	97.07	96.59	97.90	96.29	**97.02**

ResNet-34+Loss, and it has been correctly segmented(Green pixels) in U-Net ResNet-34+Loss+TAM+PAM. In U-Net ResNet-34+Loss+TAM+PAM+FRCM, addition of FRCM reduces the red and blue error pixels.

In the second column, the trees next to the building are more heavily occluded, so U-Net ResNet-34 mistakenly identifies the background as a building. Although U-Net ResNet-34+Loss eliminates a large number of wrong segmentations, there are still a small number of wrong pixels. While the wrong pixels are reduced in U-Net ResNet-34+Loss+TAM+PAM, thanks to the feature screening ability of TAM and PAM.

In the third column, although U-Net ResNet-34+Loss+ TAM +PAM can effectively reduce wrong pixel segmentation compared to U-Net ResNet-34, there are still some misjudgments. The addition of FRCM can better refine the feature function and reduce the possibility of inter-class indistinction, so there are fewer wrong pixels in the segmentation results of FMAM-Net.

B. COMPARATIVE EXPERIMENT

In this section, the best performance of FMAM-Net is compared with the relevant latest models to verify its validity and accuracy. The experiment was analyzed from both quantitative and qualitative perspectives on two datasets.

1) COMPARATIVE EXPERIMENT ON THE INRIA AERIAL DATASET

FMAM-Net is compared with state-of-the-art methods on the Inria Aerial Image Labeling Dataset, including E-D-Net [43], AMUNet [44] and U-Net ResNet-34. All the segmentation models are based on the encoder-decoder architecture. Calculate the overall accuracy and mean IoU in the test set, as shown in Table 2. From a vertical comparison, FMAM-Net achieves higher IoU than other methods in all five cities. FMAM-Net has better generalization ability whether in residential area or mountainous area. From the average metrics of five cities, FMAM-Net has 5.56% more IoU value than E-D-Net and 5.58% more IoU value than AMUNet. Compared with our baseline U-Net ResNet-34, the IoU indicator improves by 3.77%. Furthermore, FMAM-Net is also 0.36% higher than E-D-Net and 0.29% higher than AMUNet on the overall accuracy score. From the perspective of quantitative analysis, it can be seen that FMAM-Net is indeed superior to the contrasting methods.

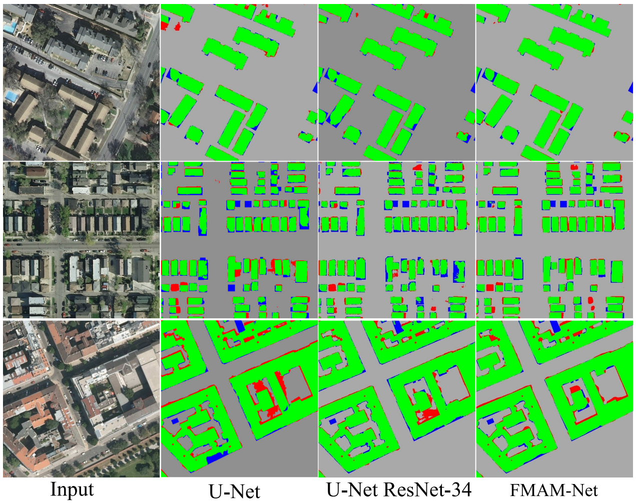


FIGURE 9. Building segmentation results of comparative experiments on the Inria Aerial Dataset. Green: True positive pixels. Gray: true negative pixels. Red: false positive pixels. Blue: false negative pixels.

For qualitative analysis, the segmentation results of FMAM-Net and U-Net ResNet34 are visualized, as shown in Figure 9. In the first row, it can be observed that U-Net ResNet-34 has more mis-segmented pixels in pool and vegetation shaded area. But in the FMAM-Net algorithm, the model can make accurate judgments about pools and shadows. In the second row, the similarity between the road and the roof leads to a more confusing segmentation. Therefore, U-Net ResNet-34 infers a large number of buildings as non-buildings, while FMAM-Net greatly reduces erroneous pixels by selecting discriminative channels. In addition, as shown in the third row, the shapes and colors of buildings are varied, causing the network to make wrong judgments. Compared with the baseline, FMAM-Net can obtain relatively accurate results. The overall outline is sharper, and there are significantly fewer erroneous pixels at the border. The experimental results show that FMAM-Net pays more attention to boundary prediction and inter-class indistinction.

2) COMPARATIVE EXPERIMENT ON THE WHU DATASET

FMAM-Net is compared with state-of-the-art methods on the WHU Dataset, including U-Net, SARB-UNet [45],

TABLE 3. Building segmentation results of comparative experiments on the WHU Dataset. The "-" indicates that the method does not present this result in its paper.

Method	Acc.	IoU	Pre.	Rec.
U-Net	98.45	86.80	92.54	92.15
SARB-UNet	-	89.39	93.25	**95.56**
MDAU-Net	-	-	**95.68**	90.63
MAP-Net	-	90.86	95.62	94.81
FMAM-Net(Proposed)	**98.84**	**91.06**	95.04	94.35

MDAU-Net[46] and MAP-Net [47], as shown in Table 3. First, the overall accuracy of U-Net has reached a high value, so FMAM-Net is improved by 0.39% compared with U-Net, which is not a big improvement. Then, IoU reached the maximum value of 91.06% on FMAM-Net, 1.67% higher than 89.39% on SARB-UNet and 0.2% higher than 90.86% on MAP-Net. The precision of FMAM-Net reached 95.04%, although not exceeding 95.68% of MDAU-Net, but 1.79% higher than 93.25% of SARB-UNet. The recall also did not exceed SARB-UNet algorithm, but it was 3.72% higher than MDAU-Net. In general, FMAM-Net has good performance in all evaluation indicators and has strong robustness.

Subsequently, we visualize some random prediction results for further analysis, as shown in Figure 10. In the first row, the variety of building shapes in the original image adds difficulty to the segmentation. Therefore, in U-Net and U-Net ResNet-34, there are pixels with segmentation errors at the edge of the building. While in FMAM-Net, this phenomenon is significantly reduced. In the second row, buildings and pavements have a similar appearance and it is easy to identify pavements as buildings, as shown in U-Net and U-Net ResNet-34.There is no similar situation in the segmentation results of FMAM-Net. In the third row, the container has a similar appearance and shape to the building, so a large number of false positive pixels appear in the segmentation results of U-Net and U-Net ResNet-34. FMAM-Net has solved a lot of problems of false segmentation, but it is still not completely eliminated.

C. LIMITATIONS DISCUSSION

We compare the state-of-the-art methods on datasets, and the results show that our method achieves better performance. This proves that our method not only solves the problems raised in this paper to a large extent, but also has strong generalization ability. But our method also has limitations. Firstly, compared with the remote sensing image with 1024 × 1024 pixels, the receptive field of small target area after feature extraction is very small. This leads to many wrong pixels in the segmentation result of small target area, as shown in the yellow box in Figure 11. Secondly, the appearance of the shadow area is not obvious due to the influence of light, and the segmentation error probability is very high, as shown in the purple box in Figure 11. Thirdly, our network model is more complex. Through calculation, we know that our FLOPs on the Inria Dataset reach 76.75GB, which requires more training time.

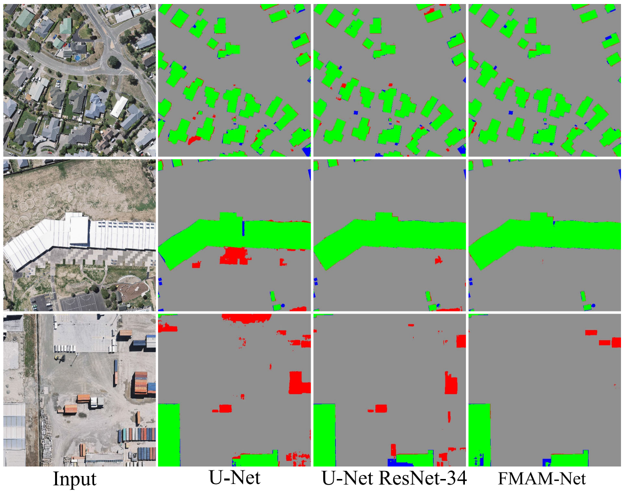


FIGURE 10. Building segmentation results of comparative experiments. Green: True positive pixels. Gray: true negative pixels. Red: false positive pixels. Blue: false negative pixels.

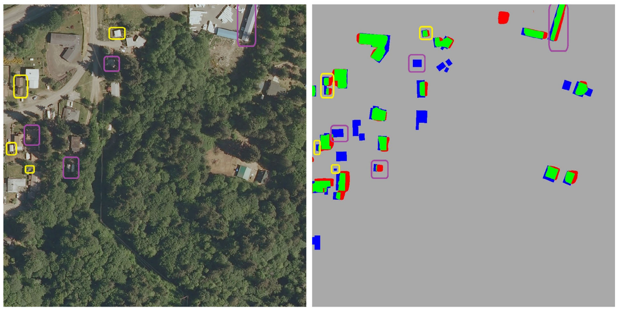


FIGURE 11. Yellow box represents the small target area, and the purple box represents the shadow area. Green: True positive pixels. Gray: true negative pixels. Red: false positive pixels. Blue: false negative pixels.

V. CONCLUSION

In this paper, we propose a novel Fused Multi-scale Attention Mechanism Network for building remote sensing image segmentation. In view of the current problem of less training data sets, this paper applies transfer learning to reduce the demand for data volume. To solve the problem of inaccurate contour segmentation, this paper introduces low-level features into high-level features to compensate for boundary information destroyed in feature extraction. Furthermore, we utilize dense connections to reuse features, which helps to redefine features. We also introduce both tandem and parallel attention mechanisms to focuses on features of interest. The spatial attention mechanism aims to obtain a score map from high-level features to guide low-level feature selection. The channel attention mechanism generates a score map over channels to guide feature fusion. The experiments were conducted on the Inria Dataset for buildings as well as the WHU Dataset. Compared with the state-of-the-art methods on both the datasets, the proposed FMAM-Net achieved higher overall accuracy (97.02%) and IoU (85.34%) for the Inria Dataset, higher overall accuracy (98.84%), IoU (91.06%) for the WHU Dataset. However, the precision and recall do not exceed the state-of-the-art methods.

Although the segmentation result of building contour has been effectively improved, the segmentation of small target area is still a major difficulty. In the future research, considering that the hole convolution is better for small target segmentation, we try to use hole convolution to design a novel segmentation network to solve the small target problem. At the same time, we can get inspiration from the special processing of small targets by target detection algorithm.

REFERENCES

[1] W. Liu, Y. Shu, and X. Tang, "Remote sensing image segmentation using dual attention mechanism Deeplabv3+ algorithm," *Trop. Geogr*, vol. 40, pp. 303–313, Jan. 2020.

[2] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, 2018.

[3] L. Gao, W. Shi, Z. Miao, and Z. Lv, "Method based on edge constraint and fast marching for road centerline extraction from very high-resolution remote sensing images," *Remote Sens.*, vol. 10, no. 6, p. 900, Jun. 2018.

[4] Z. Xue, B. Liu, A. Yu, X. Yu, P. Zhang, and X. Tan, "Self-supervised feature representation and few-shot land cover classification of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[5] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[6] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, "SVMeFC: SVM ensemble fuzzy clustering for satellite image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 1, pp. 52–55, Jan. 2012.

[7] S. K. Mylonas, D. G. Stavrakoudis, and J. B. Theocharis, "GeneSIS: A GA-based fuzzy segmentation algorithm for remote sensing images," *Knowl. Based Syst.*, vol. 54, pp. 86–102, Dec. 2013.

[8] S. Manju and K. Helenprabha, "A structured support vector machine for hyperspectral satellite image segmentation and classification based on modified swarm optimization approach," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 3659–3668, 2021.

[9] W. Feng, H. Sui, L. Hua, C. Xu, G. Ma, and W. Huang, "Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder–decoder architecture and historical land use vector map," *Int. J. Remote Sens.*, vol. 41, no. 17, pp. 6595–6617, Sep. 2020.

[10] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder–decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.

[11] S. Hu, "Segmentation of aerial image with multi-scale feature and attention model," in *Artificial Intelligence in China*. Singapore: Springer, 2020, pp. 58–66.

[12] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[13] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sens.*, vol. 11, no. 15, p. 1774, Jul. 2019.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Oct. 2017, pp. 2961–2969.

[18] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[23] A. Adiba, H. Hajji, and M. Maatouk, "Transfer learning and U-Net for buildings segmentation," in *Proc. New Challenges Data Sci., Acts 2nd Conf. Moroccan Classification Soc.*, Mar. 2019, pp. 1–6.

[24] H. Jing, X. Sun, Z. Wang, K. Chen, W. Diao, and K. Fu, "Fine building segmentation in high-resolution SAR images via selective pyramid dilated network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6608–6623, 2021.

[25] X. Zhang, Y. Yang, Z. Li, X. Ning, Y. Qin, and W. Cai, "An improved encoder–decoder network based on strip pool method applied to segmentation of farmland vacancy field," *Entropy*, vol. 23, no. 4, p. 435, Apr. 2021.

[26] Y. Li, H. Lu, Q. Liu, Y. Zhang, and X. Liu, "SSDBN: A single-side dual-branch network with encoder–decoder for building extraction," *Remote Sens.*, vol. 14, no. 3, p. 768, Feb. 2022.

[27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[28] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[29] Y. Zhang, W. Gong, J. Sun, and W. Li, "Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sens.*, vol. 11, no. 16, p. 1897, 2019.

[30] H. Wang, P. Cao, and J. Wang, "UCTransNet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2441–2449.

[31] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1857–1866.

[33] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention UNet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–12, Apr. 2020.

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

[35] H. Guo, Q. Shi, B. Du, L. Zhang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2020.

[36] J. Cai and Y. Chen, "MHA-Net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5807–5817, 2021.

[37] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Inf. Fusion*, vol. 91, pp. 376–387, Mar. 2023.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 4700–4708.

[39] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[40] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

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

[42] S. Liu, H. Ye, K. Jin, and H. Cheng, "CT-UNet: Context-transfer-UNet for building segmentation in remote sensing images," *Neural Process. Lett.*, vol. 53, no. 6, pp. 4257–4277, Dec. 2021.

[43] Y. Wang, J. Kong, and H. Zhang, "U-Net: A smart application with multidimensional attention network for remote sensing images," *Sci. Program.*, vol. 2022, pp. 1–11, Feb. 2022.

[44] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-Net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, Dec. 2022.

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

[46] Y. Wang, J. Kong, and H. Zhang, "U-Net: A smart application with multidimensional attention network for remote sensing images," *Sci. Program.*, vol. 2022, pp. 1–11, Feb. 2022.

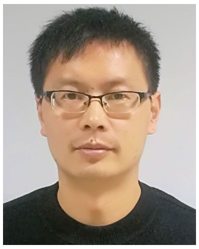
[47] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

HUANRAN YE was born in Henan, China, in 1996. He received the bachelor's degree from the School of Electrical Engineering, Zhejiang University of Science and Technology, in 2018, and the master's degree from the School of Computer Science and Technology, Zhejiang University of Technology, in 2021.

Since 2021, he has been working as an Assistant Teacher with the Yiwu Industrial & Commercial College. He has published articles at the *Neural Processing Letters*, Periodical, and IEEE International Conference on Pattern Recognition Conference. He has two patents for invention. His research interests include deep learning segmentation algorithms, building segmentation algorithms, and computer vision.

RUN ZHOU was born in Hunan, China, in 1982. He received the bachelor's degree from the Xi'an University of Electronic Science and Technology, in 2008, the master's degree from Xi'an Shiyou University, in 2011, and the Ph.D. degree from Northwestern Polytechnical University, in 2019. His current research interest includes key technologies of wireless communication networks.

JIANHAO WANG was born in Zhejiang, China, in 1995. He received the B.S., M.S., and Ph.D. degrees from the College of Electrical Engineering, Fuzhou University, Fuzhou, China, in 2017, 2019, and 2022, respectively. He is currently working as a Teacher with the Yiwu Industrial & Commercial College. His current research interests include current transformer, power conversion, high-frequency magnetics, and EMI of motor drive systems.



ZHILIANG HUANG (Member, IEEE) received the B.S. degree from the School of Materials Science and Engineering, Wuhan Institute of Technology, in 2004, the M.S. degree from the School of Mathematics-Physical and Information Engineering, Zhejiang Normal University, in 2009, and the Ph.D. degree from the School of Information Science and Engineering, Southeast University, in 2013. In 2015, he was a Visiting Researcher with Bilkent University, Ankara, Turkey. He is currently an Associate Professor with the School of Mathematics-Physical and Information Engineering, Zhejiang Normal University. His currently research interests include modern coding theory and signal processing for digital communications.