Hindawi
Scientific Programming
Volume 2022, Article ID 1603273, 11 pages
https://doi.org/10.1155/2022/1603273

Research Article

U-Net: A Smart Application with Multidimensional Attention Network for Remote Sensing Images

Yao Wang, Jiayuan Kong, and Hesheng Zhang

School of Mining Engineering, Taiyuan University of Technology, Taiyuan 030024, China

Correspondence should be addressed to Hesheng Zhang; zhanghesheng@tyut.edu.cn

Received 7 January 2022; Revised 21 January 2022; Accepted 31 January 2022; Published 21 February 2022

Academic Editor: Muhammad Zakarya

Building segmentation is an important step in urban planning and development. In this work, we propose a new deep learning model, namely Multidimension Attention U-Net (MDAU-Net), to accurately segment building pixels and nonbuilding pixels in remote sensing images. Furthermore, we introduce a novel Multidimension Modified Efficient Channel Attention (MD-MECA) model to enhance the network discriminative ability through considering the interdependence between feature maps. Through deepening the U-Net model to a seven-story structure, the ability to identify the building is enhanced. We apply MD-MECA to the "skip connections" in traditional U-Net, instead of simply copying the feature mapping of the contraction path to the matching extension path, to optimize the feature transfer more efficiently. The obtained results show that our proposed MDAU-Net framework achieves the most advanced performance on publicly available building data sets (i.e. the precision over the Massachusetts buildings data set and WHU data set are 97.04% and 95.68%, respectively). Furthermore, we observed that the proposed framework outperforms several state-of-the-art approaches.

1. Introduction

With the rapid development of China's remote sensing satellite industry, building segmentation in remote sensing image is an important research field in the image interpretation problems. Extracting feature information from remote sensing images is related to urban planning and development. Therefore, timely updating of image information will have an impact on everything that depends on these systems [1]: for example, mapping, disaster analysis, and emergency response. For a long time, the acquisition of feature information in remote sensing images relies on the traditional manual visual interpretation method, which is time consuming and laborious, which restricts the development and application of high-resolution images. Therefore, the use of remote sensing images accurately, quickly, and automatically extraction of target features has attracted widespread attention of many researchers all over the world.

Many researchers have recommended a countless number of programmed building segmentation approaches for remote sensing images. However, the interdependence among the feature channels is often not discussed. Later, by presenting the attention segment, in the context of deep learning-based approaches, the features of dissimilar spaces and channels can be advanced to enrich the essential features and suppress the features that are not significant to the task. U-Net structure is widely used in the field of medical image segmentation. The channel attention mechanism can adjust the characteristic response value of each channel adaptively. The importance of different feature channels obtained by automatic learning is used to enhance the important features and suppress the features that are not important to the segmentation task. This method can be integrated into the U-Net model to improve its performance.

In this work, we introduce a multidimensional channel attention, which uses the average pool and the maximum pool features in multidimensional channel to further improve the performance of building segmentation in remote sensing images. In this study, we propose a new Multidimension Channel Attention Network model, called MDAU-Net, based on deep learning, which has achieved the latest performance in remote sensing image building

segmentation. We demonstrate through experimental results that the proposed MDAU-Net model achieves the best performance on two real data sets. Furthermore, the proposed model greatly reduces the network complexity in computer vision tasks (including image classification, object detection, and instance segmentation) while maintaining the performance. Specifically, we have the following contributions:

(1) Deepen the structure of U-Net, using 7-layer convolution and downsampling module for feature extraction and making full use of different levels of building feature details, so as to achieve the purpose of more fine segmentation of buildings

(2) We apply MD-MECA in order to "skip connections," so as to give weight to each feature map in the shrinking path in the feature transfer step, instead of copying them equally to the corresponding expansive path

(3) Batch normalization [2] is used after the feature stitching of the upsampling part, and Dropblock [3] is used after convolution to solve the over fitting problem in the network training process

(4) On the basis of the above work, we propose MDAU-Net and evaluate it on Massachusetts [4] and WHU data sets [5]

The rest of the paper is organized as follows. In Section 2, we offer an overview of the related work. Section 3 is about the proposed methodology. In Section 4, data sets and evaluation metrics are discussed. Moreover, experimental details are also presented. In Section 5, results are discussed. Moreover, various machine learning techniques are evaluated on the aforementioned data set to study validity of the proposed model. Finally, Section 6 concludes this paper and offers several directions for further research and investigation.

2. Related Work

In the past few decades, researchers have proposed a great number of automatic building segmentation methods for remote sensing images. For example, Zhong et al. [6] used k-means clustering, Kohonen et al. [7] used self-organizing mapping network, and Lin et al. [8] introduced the object-oriented Morphological Building Index (MBI). However, these methods are very dependent on the geometric texture of buildings and cannot adapt to the buildings under different conditions in the image. The sensitivity of their features and spectra is not enough to capture, similar objects are prone to mix and produce adhesion phenomenon, and the robustness is not enough good [9, 10].

Recently, deep learning-based methods have been used for automatic segmentation of buildings in remote sensing images and achieved excellent results. The deep learning method based on convolutional neural networks (CNN) proposed by Krizhevsky et al. [11] is usually used in various computer vision tasks. Long et al. [12] proposed full convolutional networks (FCN), which can classify images at the pixel level. Compared with the FCN structure, the SegNet structure proposed by Badrinalayanan et al. [13] transfers the maximum pooling index to the decoder, which improves the segmentation resolution and saves more storage space. Later, in order to improve the use of feature information in images, Ronneberger et al. [14] proposed the U-Net spanning connection structure, which has been widely used in image segmentation. This structure realizes the fusion of multiscale image information and improves the segmentation performance [15].

Although these deep learning-based methods have achieved significant results, the interdependence between the feature channels is often ignored. Later, by introducing the attention module, the features of different spaces and channels can be refined to enhance the important features and suppress the features that are not important to the task. At present, many researchers have applied attention module to the image extraction and have achieved good results [16, 17, 18].

3. Proposed Methodology

3.1. A.U-Net. U-Net structure was proposed by Ronneberger et al. [14] in May 2015 and was initially widely used in the field of medical image segmentation. As a very classic full convolutional network model, it is widely used in the field of remote sensing image segmentation at present. Its network structure is shown in Figure 1.

U-Net is divided into two parts, the left part is the feature extraction part, also known as the lower sampling part, the right part is the upper sampling part. In the feature extraction part, the deep semantic features of the image are extracted through convolution and pooling [19].

Each process includes the image is transformed into a matrix with the number of channels increased by 64 after two convolutions, and then the maximum pooling operation is carried out to reduce the length and width of the image to half of the original. In accordance with the same process, after four times of subsampling, the image becomes a $32 \times 32 \times 512$ matrix, and after two 3×3 convolution operations, the final feature map is obtained. For the upsampling part, the calculation starts from the information at the bottom of the network. After each 2×2 deconvolution, it is spliced with the downsampling feature map of the same layer, fused with the channel number corresponding to the feature extraction part at the same scale, and then the upsampling is completed once after two 3×3 convolution operations. Through the combination of feature images obtained by feature extraction, information is supplemented to optimize the segmentation results.

3.2. MECA and MD-MECA. Inspired by the recently proposed CAR-UNet (Channel Attention Residual U-Net) [20], Modified Efficient Channel Attention (MECA) module in CAR-UNet was improved. MECA module structure is shown in Figure 2, which greatly reduces the network complexity in image classification, object detection, and instance segmentation while maintaining performance. In

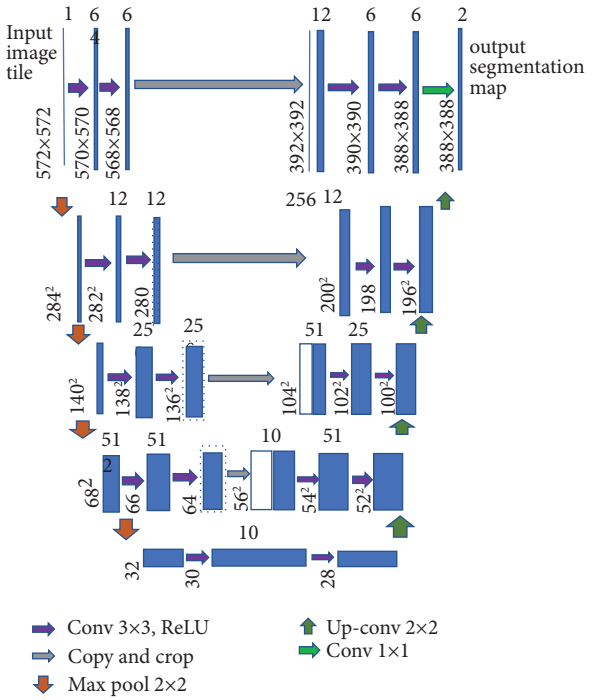

Figure 1: The classical U-Net network architecture.

this paper, the optimized module is named MD-MECA (Multidimension Modified Efficient Channel Attention). By adding MD-MECA to the encoding transfer features of U-Net architecture to the decoding module, compared with the original U-Net, the structure optimized the feature graph during the transmission of the feature graph, supervised the information of the feature graph of the encoding part in different ways, and then passed it to the decoding part for information supplement.

Channel attention (CA) was first used as a Squeeze-and-Excitation Networks for classification, which modeling by using the relationship between the channels [21]. It can adjust the characteristic response value of each channel adaptively. The importance of different feature channels obtained by automatic learning is used to enhance the important features and suppress the features that are not important to the segmentation task. And recent works showed that the channel attention mechanism has great potential in improving the performance of deep convolutional neural networks (CNN). Guo et al. [20] proposed a CAR-Unet (Channel Attention Residual U-net) network model for retinal vascular segmentation and achieved good results. The MECA module in CAR-Unet network model is based on an effective channel attention ECA module proposed by Wang et al. [22] and adopts both average pooling layer and maximum pooling layer to obtain more detailed channel attention, so as to better collect spatial information.

MECA module uses convolution to avoid dimensionality reduction in SE network blocks, thus greatly reducing the complexity of the model while maintaining superior performance. The MECA module is an embedding channel monitoring module, which extracts global features by using different global pooling calculations: average pooling can extract spatial information, while maximum pooling can obtain unique object characteristics, which can attract more detailed channel attention. Therefore, MECA module combines the global features extracted from the two to obtain a more refined channel monitoring weight and carries out channel attention monitoring based on C channel to obtain the weight parameters between different channels [23].

The feature graph in MECA module has different dimensions H, W, and C, which stand for the height, width, and the number of channels of the input feature F. Therefore, the MECA module can be strengthened by multiangle and all-aspect supervision, so that the feature graph can represent more detailed target information. Inspired by this, MECA (Modified Efficient Channel Attention) module in CAR-UNet was improved. Based on the dimension characteristics of the feature map, this paper optimized the MECA module and named the optimized module MD-MECA (Multidimension Modified Efficient Channel Attention Networks): on the basis of the C channel dimension, H and W channels are added simultaneously, and the same attention supervision module is designed, respectively, to obtain the supervision weight parameters of different dimensions. Then, MD-MECA module was added to the deeper U-Net network structure during the process of encoding, transmission, and decoding [24, 25]. Compared with the original U-Net, the structure optimized the feature graph during the transmission of the feature graph, supervised the feature graph of the coding part in different ways, and then transmitted to the decoding part for information supplement.

The module structure of MD-MECA is shown in Figure 3: the supervision weight based on each dimension is extracted, respectively, with the supervision structure of MECA. Formally, input feature $F \in R^{H\times W\times C}$ through the channel-wise max pooling and average poling can generate $F_{mp} \in R^{1\times 1\times C}$ and $F_{ap} \in R^{1\times 1\times C}$, respectively, e.g., at the c^{th} channel:

$$F_{mp}^{c} = \text{Max}\left(F^{c}(i,j)\right), 0 < c < C, 0 < i < H, 0 < j < W,$$
$$F_{ap}^{c} = \frac{1}{H\times W}\sum_{u=1}^{W}\sum_{j=1}^{W} F^{c}(i,j), 0 < c < C, \tag{1}$$

where $\text{Max}(\cdot)$ represents the maximum value, and $P^{c}(\cdot)$ represents the pixel value at a specific position in channel c. The two calculated values are then transmitted to a 1D convolutional neural network with shared weights to generate a channel monitoring mechanism $M^{c} \in R^{1\times 1\times C}$ Then, the MECA module combines the eigenvectors of the convolutional layer output by channel addition, and the calculation is as follows:

$$M(F) = \sigma\left(\text{Conv1 } D\left(F_{ap}\right) + \text{Conv1 } D\left(F_{mp}\right)\right), \tag{2}$$

where Conv1D(•) represents the 1D convolutional layer and $\sigma(\bullet)$ denotes the Sigmoid function. Similarly, the monitoring weight based on H dimension and W dimension of feature map is obtained in the same calculation method, and then

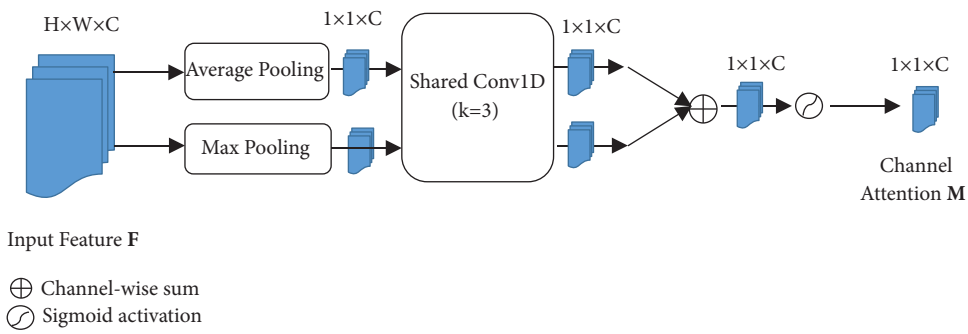

Figure 2: Diagram of the classical MECA technique.

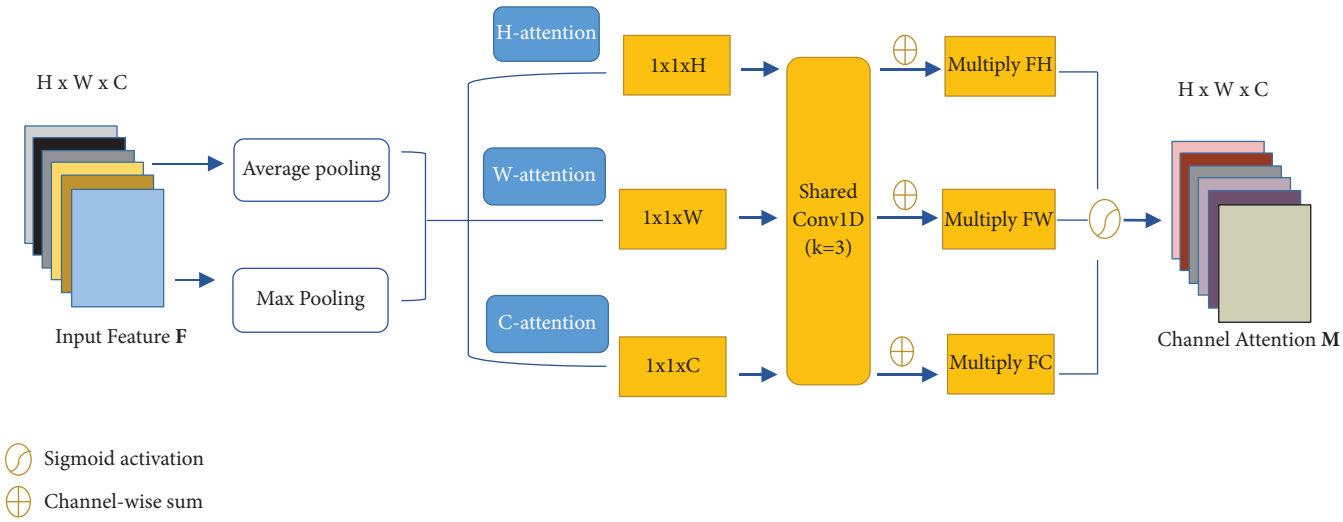

Figure 3: Structure of the proposed MD-MECA module.

the feature map is supervised in three dimensions, and then the required feature map is obtained by adding and combining them pixel by pixel.

Through the above method, the experiment has carried on the different dimension and the level optimization to the image feature image. By adding MD-MECA to the coding transfer features of U-Net architecture to the decoding module, the detailed texture features and semantic features transmitted by the coding part are refined. Therefore, the features obtained by the decoding part have better characterization ability to the target and can better segment the remote sensing image.

3.3. Batch Normalization and DropBlock. The Google team came up with a Batch Normalization method in 2015. In the deep network, if the network activation output is large, the gradient will be small and the learning rate will gradually slow down. In this way, the deeper the network structure is, the shallow layer of the gradient will be small and the learning rate will be slow. The higher the deep gradient, the faster the learning rate. For such a network structure, it loses its deep meaning. In order to solve this problem, the use of BN layer in the network can solve gradient disappearance and gradient explosion, improve the training speed and network convergence speed, and effectively prevent overfitting problems.

Overfitting is a serious problem in deep neural networks. The more complex the network structure is, the processing speed is slow, so it is difficult to deal with the overfitting of different complex neural networks in the test. Dropout proposed by Srivastava et al. [26] is also a technology to solve this problem. Its key idea is to temporarily discard neural units from the neural network in accordance with certain probability during training. Dropout has a significant effect in the full connection layer, but elements of adjacent positions in the feature graph of the convolutional layer share semantic information in space. Therefore, although a unit is discarded, its adjacent elements can still keep the semantic information of this location, and the information can still circulate in the convolutional neural network. Therefore, to solve this problem, this paper introduces a structural form of the dropout method, i.e., DropBlock. The DropBlock technology is a regularization technology used in convolutional neural network proposed by researchers of Google Brain in 2018. It can discard units in adjacent regions of feature graph at the same time to improve accuracy.

3.4. Network Architecture. The detailed architecture of Multidimension Attention U-Net (MDAU-Net) is displayed in Figure 4. Structure of MDAU-Net is derived from U-Net, the structure adopts multilevel subsampling structure module, which increases the model depth to increase the nonlinear mapping, thus enhancing the feature fitting ability. The experimental structure is designed with a 7-layer lower sampling module; each module consists of a 3×3 convolutional layer (ReLU) plus a 2×2 Maxpooling layer. For the input data with the size of 1024×1024, the image is extracted with the size of 8×8, and the size is 1/64 of the original image after 7 times of pooling at the coding end under the sampling structure. At the same time, 7 (seven) feature graph modules of different levels are obtained. In the upsampling part, seven (7) stages of upsampling calculation were also carried out. Furthermore, each layer was combined with the feature map of the downsampling coding part to supplement information and optimize the segmentation contour texture features. At the same time, MD-MECA module is added to the feature image transfer part of the coding part of each layer, and the feature image sent by the coding part is optimized through multidimensional supervised calculation to highlight the geometric features and suppress the background features.

For remote sensing image features, the shallow layer features have sufficient texture features, which is helpful to the contour restoration of segmentation algorithm. High-level semantic features help to distinguish target categories. Therefore, it is necessary to combine the features of the two to complete the feature information as much as possible. In the MDAU-Net network structure proposed in this paper, the multilevel coding and decoding structure is adopted, and the feature graphs of different levels in the coding part are made full use of to build a deeper network, and the features are extracted and combined in a fine way, so as to obtain better prediction effect.

4. Data sets and Evaluation Metrics

4.1. Data sets

4.1.1. Experimental Data. In order to avoid the impact of the uniqueness of the data set on the experimental results, the Massachusetts building data set and WHU data set were selected as the experimental data. Because the annotation accuracy and spatial resolution of different data sets are different, the conclusion is more convincing.

In this paper, 600 images in each data set were selected for training, 100 images for testing, and 100 images for verification. The input image size was 512 pixels ×512 pixels.

(1) Flip conversion: flip the image along the vertical or horizontal direction

(2) Random rotation transformation: randomly rotate the image by several angles

(3) Random clipping: local images at different positions can be obtained through random clipping of images

(4) Contrast transformation: the contrast transformation factor is randomly set for the image to adjust the image contrast

Among them, Figure 5(a) is the original image without processing; Figures 5(b)–5(d) is the clockwise rotation of 90°, 180°, and 270°, respectively; Figures 5(e) and 5(f) are the vertical and horizontal mirror flipping; Figure 5 5(g) is the contrast transformation; and Figure 5(h) is the random cropping. A total of 4000 images and labels were obtained after the geometric modification method data were expanded. Finally, in each data set, there are 1400 images as the training set, 400 images as the test set, and 200 images as the verification set.

4.1.2. Evaluation Metrics. Recall, Precision, F1-measure, and IoU were used to evaluate the experimental results. Recall rate is the ratio of correctly predicted positive samples to the total number of true positive samples. Accuracy refers to the ratio of correctly predicted samples to the total predicted samples, F1 value refers to the harmonic average of accuracy and recall rate, while IoU is the intersection of pixels labeled as building in the predicted results and ground truths, divided by the union of pixels labeled as building in the predicted results and ground truths. The calculation formula is as follows:

$$\begin{aligned} P_{rec} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ P_{pre} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ F1 &= 2 \times \frac{P_{pre} \times P_{rec}}{P_{pre} + P_{rec}}, \\ IoU &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \end{aligned} \tag{3}$$

where *TP* represents the correct number of pixels extracted, *FP* represents the number of pixels extracted with errors, and *FN* represents the number of missing pixels.

4.1.3. Implementation Details. In order to verify the feasibility of MDAU-Net proposed in this paper in remote sensing images and the superiority of the improved network MDAU-Net compared with U-Net network and CAR-UNet network, the same group of training samples and test samples were used for comparative experiments, and the experimental computer operating system was Windows. Based on the design of PyTorch deep learning framework of version 1.4.0, the CPU is configured as E2650, graphics card NVIDIA 1080Ti ×2, GPU is configured as GeForce GTX 1080, and video memory is 8G. Experimental parameters are shown in Table 1.

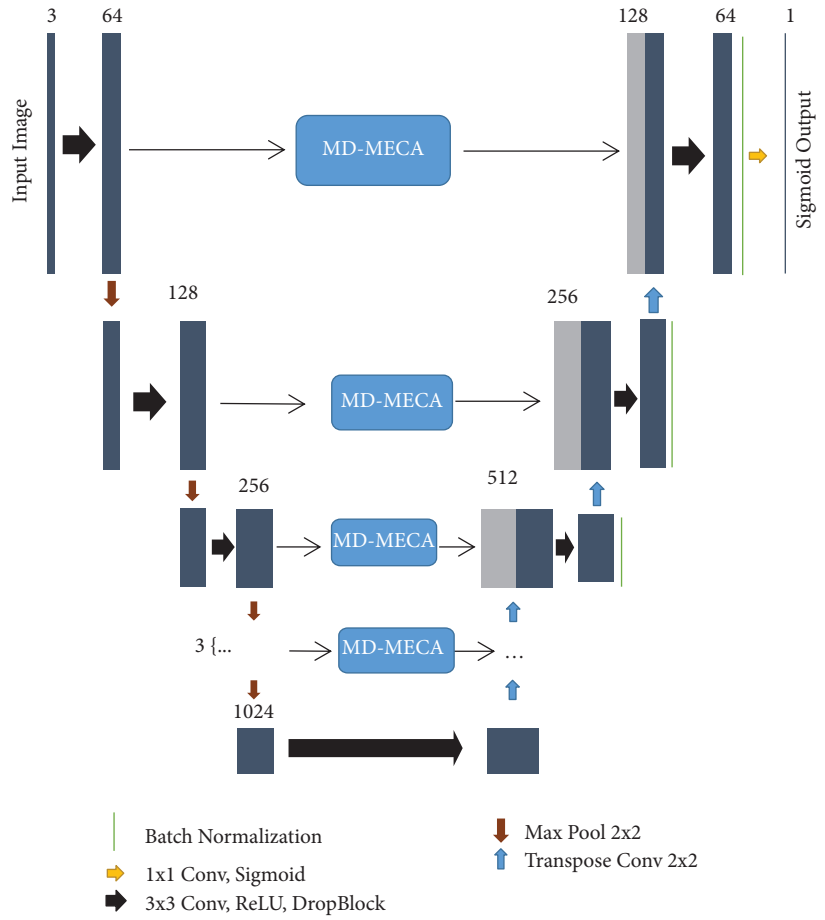

Figure 4: The MDAU-Net architecture (proposed model).

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5: Data expansion processing results (a) image, (b) rotate 90°, (c) rotate 180°, (d) rotate 270°, (e) flip vertical, (f) flip horizontal, (g) transform contrast, and (h) random cutting.

Table 1: Experimental parameters.

name of the parameter	Parameter values
Learning rate	1×10^{-3}
Optimizer	Adam
Loss function	Binary cross entropy
Batch size	2
Epochs	100
Dropout rate	0.15
Block size	7

5. Experiments and Results

5.1. Ablation Study

5.1.1. The Influence of Layer Number of Network Structure. Firstly, based on U-Net network model, the influence of model depth on experimental results is studied. The experiment was conducted on the Massachusetts building data set and the WHU data set, respectively. Tables 2 and 3 show the effect of different model depths on the experimental performance.

Table 2: Comparison of results of different deep network experiments on Massachusetts data set.

Methods	Layer number	Recall (%)	Precision (%)	F1-measure (%)
U-Net	4	86.36	91.94	89.06
U-Net5	5	89.92	92.54	91.21
U-Net6	6	92.41	94.87	93.62
U-Net7	7	94.38	96.62	95.49
U-Net8	8	92.53	94.33	93.42

Table 3: Comparison of results of different deep network experiments on WHU data set.

Methods	Layer number	Recall (%)	Precision (%)	F1-measure (%)
U-Net	4	84.62	90.64	87.53
U-Net5	5	88.74	91.32	90.01
U-Net6	6	91.28	93.63	92.44
U-Net7	7	95.14	95.81	95.47
U-Net8	8	92.46	94.19	93.32

As can be seen from the experimental results, with the increase of the number of network layers, each accuracy index will increase, but when the number of network layers reach 8, the influence of overfitting problem is very serious. By comparison, it can be found that U-Net7 has the best building extraction effect among the five network models. In different data sets, U-Net7 has the highest recall rate, accuracy and F1 values, which reach 96.62% and 95.81% respectively. Therefore, it can be concluded from the experimental results that the more layers of the network structure, the higher the accuracy of the experimental results. With the deepening of the network layer, although the receptive field will increase, the number of down sampling will increase, resulting in the loss of detail information. Meanwhile, the overfitting problem will increase with the deepening of the network, thus the experimental accuracy will be affected. Based on the above considerations, U-Net7 is selected as the best basic network model for the experiment.

5.1.2. The Impact of Batch Normalization and DropBlock. Based on the selected U-Net7 network model, the experiment explores the impact of Batch Normalization and DropBlock on the experimental results.

Tables 4 and 5 show the precision comparison of experimental results on Massachusetts data set and WHU data set after the introduction of BN layer and DropBlock, respectively, in U-Net7 network structure model. It can be seen that the selection of different data sets has a direct impact on the prediction results; however, under the same experimental conditions and data sets, the proposed U-Net7 network combined with BN and DropBlock at the same time, compared with the U-Net7 network combined with BN or DropBlock alone, the accuracy index has been improved to a certain extent, and the building extraction has reached high accuracy

Table 4: Massachusetts data set comparison of network models' experiment results.

Network structure	Recall (%)	Precision (%)	F1-measure (%)
U-Net7+BN	95.36	96.94	96.14
U-Net7+DropBlock	94.42	95.82	95.11
U-Net7+BN + DropBlock	96.68	97.04	96.86

Table 5: WHU data set Comparison of network models' experiment results.

Network structure	Recall (%)	Precision (%)	F1-measure (%)
U-Net7+BN	94.21	94.83	94.52
U-Net7+DropBlock	94.26	95.12	94.69
U-Net7+BN + DropBlock	94.63	95.68	95.15

requirements. The accuracy rates are 97.04% and 95.68%, respectively, indicating that BN and DropBlock can effectively solve the gradient disappearance and gradient explosion, reduce the overfitting problem, and improve the accuracy of building identification. The feasibility and potential of this method in remote sensing image target extraction are proved.

5.2. Comparison to State-of-the-Art Methods. Based on the experimental data set, the classical U-Net network, CAR-UNet network, and MDAU-Net network were trained, respectively, and the experimental results were compared in detail after testing.

The three index values of the three methods in the two test data sets are shown in Table 6. The statistical data in Table 6 show that in the two tests, the three index values of the method in this paper are better than the corresponding index values of U-Net and CAR-UNet. Taking the Massachusetts data set as an example, the accuracy rates of the method, U-Net, and CAR-UNet are 97.04%, 94.82%, and 92.34%; recall are 87.68%, 82.42%, and 82.36%; IoU are 78.35%, 72.31%, and 70.96%, respectively. It can be seen from Table 6 that the selection of data sets has a direct impact on the predicted results. However, under the same experimental conditions and data sets, the accuracy indexes of the MDAU-Net network proposed in this paper are improved to some extent compared with the U-Net network and CAR-UNet network and meet the high accuracy requirements for target extraction.

Part of the visualization results are shown in Figures 6 and 7. The visual effect shows that although there are a few extraction errors, the overall building extraction effect is good. It can be seen from the results that the overall effect of MDAU-Net extraction of the target is better than that of U-Net and CAR-UNet network structures, and the result image extracted by MDAU-Net is closer to the label map. As can be seen from the figure, the extraction of irregular buildings in the image is incomplete, and the overall extraction effect is not good. Adhesion phenomenon exists in small -scale buildings; In addition, there are a few extraction

Table 6: Quantitative evaluation results of building detection for different methods.

Data sets	Test area	IoU			Recall			Precision		
		U-Net	CAR-UNet	MDAU-Net	U-Net	CAR-UNet	MDAU-Net	U-Net	CAR-UNet	MDAU-Net
Massachusetts building data set	1	69.74	68.96	76.95	83.46	80.23	86.46	88.48	91.34	96.32
	2	66.98	69.73	77.82	80.27	81.32	85.24	94.24	93.21	95.11
	3	73.26	74.22	80.93	81.64	78.43	88.72	93.31	90.23	96.68
	Mean	**70.96**	**72.31**	**78.35**	**82.36**	**82.42**	**87.68**	**92.34**	**94.82**	**97.04**
WHU data set	1	69.42	69.83	76.42	83.24	81.47	82.03	90.18	90.21	93.02
	2	68.33	65.24	73.92	80.66	82.56	87.49	89.32	91.37	92.83
	3	74.26	73.82	78.46	81.24	80.33	89.32	86.87	93.06	94.37
	Mean	**71.53**	**72.41**	**77.97**	**82.21**	**83.26**	**90.63**	**88.94**	**93.12**	**95.68**

Bold shows the values of IOU, recall, and precision on the whole dataset.

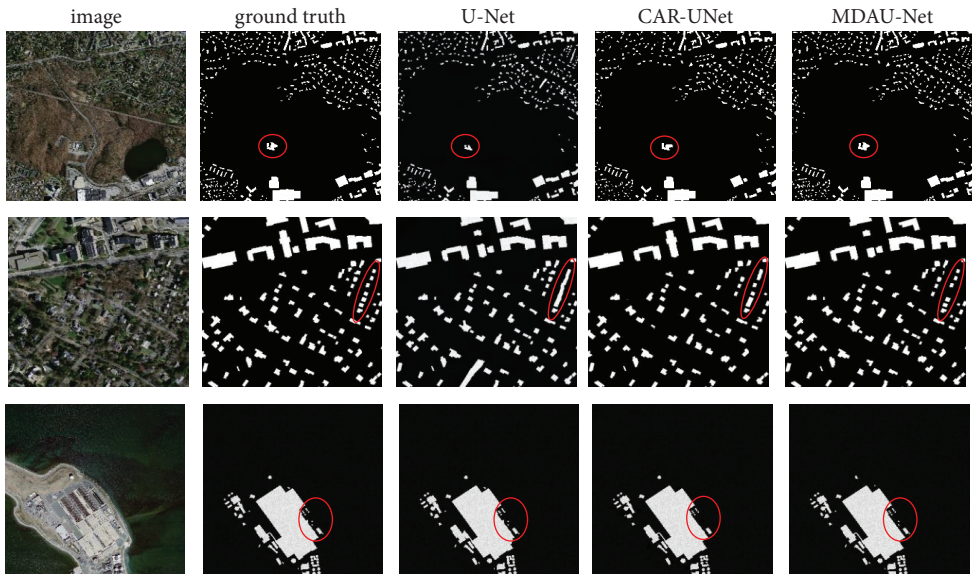

Figure 6: Experimental comparison of extraction results from the Massachusetts data set.

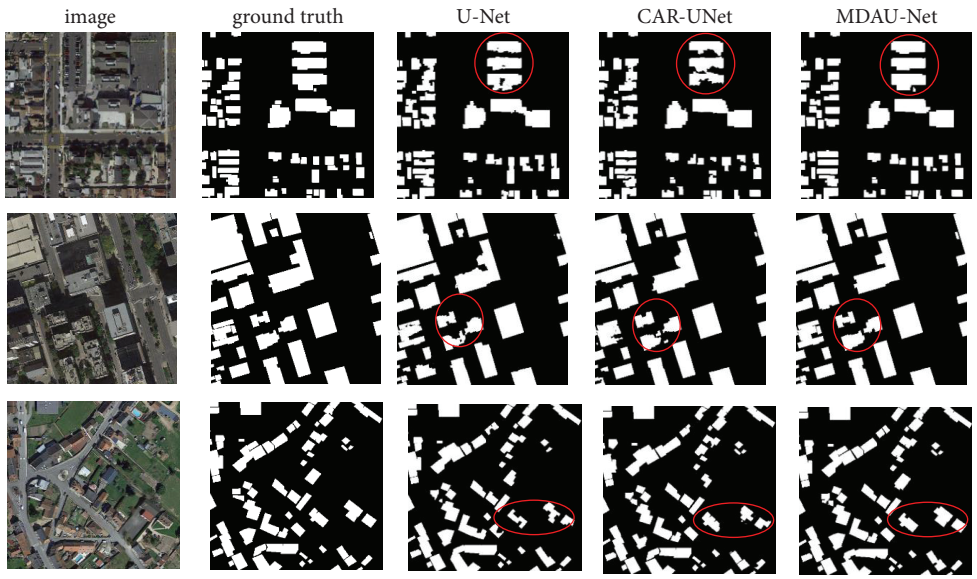

Figure 7: Experimental comparison of extraction results from the WHU data set.

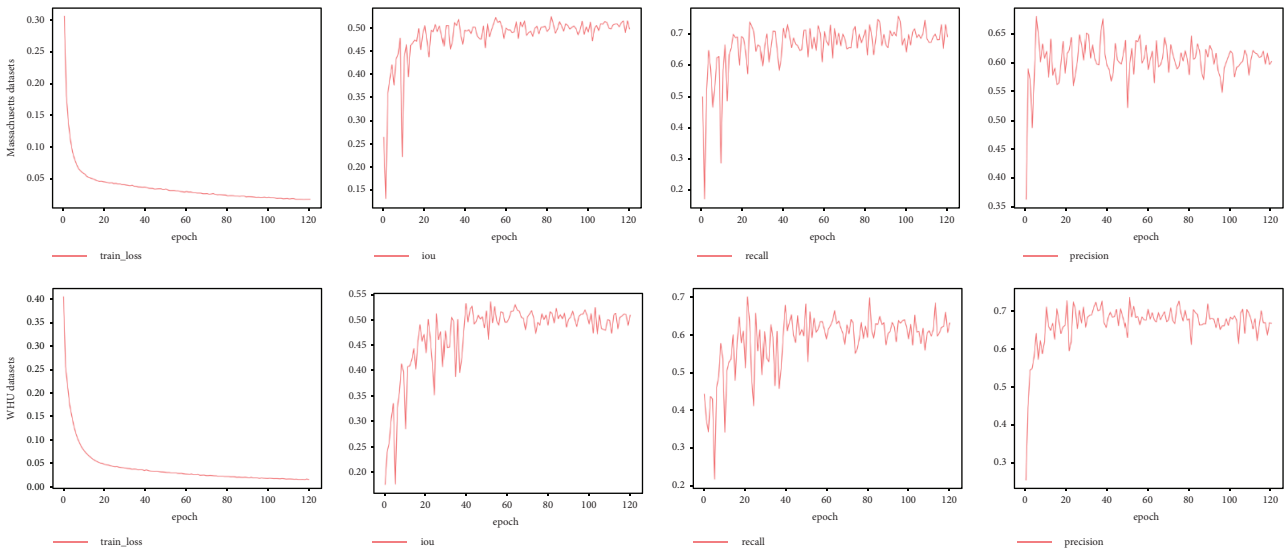

Figure 8: Accuracy of the proposed MDAU-Net model using evaluation graph on both Massachusetts and WHU data sets.

errors in this method, mainly in that it is easy to mistake the bright hardened ground and bare ground as buildings in the test image; and the cement hardened ground is connected with the building, so it is difficult to eliminate this false extraction through postprocessing. However, in the case of irregular building boundaries and a large number of isolated points, the MDAU-Net network structure proposed in this paper can effectively extract targets comprehensively, and the extraction effect is better in dense buildings or shaded areas, especially for some small buildings or irregular buildings with good recognition effect. It can be seen that the improved network model is more complete in extracting the target details, segmenting the target edges accurately, and has a better recognition effect on some subtle feature details.

Figure 8 shows the change curves of train loss, IoU, recall, and precision of this method in the training process of Massachusetts data set and WHU data set. The horizontal axis represents the number of training steps, as shown in Figure 8. After 100 steps of training, the change rate of loss value gradually decreased and later became stable. It shows that the network type has good stability. In Figure 8, IoU, recall, and precision of two different data sets all increased step by step with the increase of training steps. The precision reached 97.04% and 95.68% in the end.

In other experiments, Softmax, Sigmoid, and ReLU were the three types of activation functions we employed [9, 27, 28, 29]. Over the acquired findings, we noticed significant variations that may be attributed to the data sets. Because the Massachusetts data set is so big, there were more variances than in the WHU data set. This indicates that the proper activation function should be employed based on the data set and image attributes. Furthermore, each activation function generates distinct and significantly different results when paired with a certain training model. When compared to the standard U-Net and MECA methods, the MDAU-Net model has a tiny overlap. Smaller overlaps guarantee that the model delivers outcomes that are more similar to one another, i.e., stray less from one another.

6. Conclusions and Future Work

In this paper, we presented a Multidimension Attention Network model for building segmentation in remote sensing images. The method considers the relationship between feature channels and introduces a new channel attention mechanism to enhance the network discrimination ability. Specifically, we increase multiple dimensions through the recently proposed modified efficient channel attention (MECA). Then, we apply MD-MECA to "skip connections," assigning weights to the element map from the shrink path rather than equally copying to the corresponding expansive path. The DropBlock is added after the convolutional layer and BN is added in the decoding path, which improves the accuracy of the segmentation algorithm in remote sensing images and effectively solves the problems of missed detection, wrong detection, and irregular edge in the segmentation and extraction of buildings in remote sensing images.

It is compared with the classical U-Net network structure and car U-Net network structure. The results show that the recall, accuracy, F1 value, and IoU of this method have been significantly improved, and the final extraction effect can detect and classify buildings more accurately. Our experiments show that the algorithm achieves the highest performance of building segmentation on two data sets, i.e., Massachusetts and WHU. However, we observed that the activation function used in this method cannot activate all neurons and the improvement accuracy is limited. Therefore, the optimization of the model structure and the search for the optimal activation function are further research work [30].

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Shanxi Province Key R&D Plan (International Science and technology cooperation) (project 201903D421089) and Shanxi Postgraduate Education Innovation Project, Research on discontinuous deformation law of mining surface based on particle flow theory (2019SY126).

References

[1] C. Liu, X. Huang, Z. Zhu, H. Chen, X. Tang, and J. Gong, "Automatic extraction of built-up area from ZY3 multi-view satellite imagery: analysis of 45 global cities," *Remote Sensing of Environment*, vol. 226, no. 226, pp. 51–73, 2019.

[2] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, https://arxiv.org/abs/1502.03167.

[3] G. Ghiasi, T. Lin, and Q. Le. DropBlock, "A regularization method for convolutional networks," 2018, https://arxiv.org/abs/1810.12890.

[4] V. Mnih, *Machine Learning for Aerial Image Labeling*, pp. 84–88, University of Toronto, Toronto, Canada, 2013, PhD_Thesis.

[5] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[6] Y. Zhong and L. Zhang, "Initialization methods for remote sensing image clustering using K-means algorithm," *Systems Engineering and Electronics*, vol. 32, no. 9, pp. 2009–2014, 2010.

[7] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[8] X. Lin and J. Zhang, "Object-based morphological building index for building extraction from high Resolution.Remote sensing imagery," *Acta Geodaetica et Cartographica Sinica*, vol. 46, no. 6, pp. 724–733, 2017.

[9] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-attentioned object detection in remote sensing imagery," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3886–3890, Taipei, Taiwan, September 2019.

[10] D. Yu, R. Zhang, and S. Qin, "Cascade saliency attention network for object detection in remote sensing images," in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 217–223, Milan, Italy, January 2021.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science*, in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Munich, Germany, October 2015.

[15] S. T. Seydi, M. Hasanlou, and M. Amani, "A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets," *Remote Sensing*, vol. 12, no. 12, p. 2010, 2020.

[16] M. Chen, J. Wu, L. Liu et al., "DR-net: an improved network for building extraction from high resolution remote sensing image," *Remote Sensing*, vol. 13, no. 2, p. 294, 2021.

[17] A. Abdollahi, B. Pradhan, and A. M. Alamri, "An ensemble architecture of deep convolutional segnet and unet networks for building semantic segmentation from high-resolution aerial images," *Geocarto International*, vol. 18, no. 3, 2020.

[18] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data," 2019, https://arxiv.org/abs/1904.00592.

[19] B. Wang, L. Wang, J. Chen, Z. Xu, T. Lukasiewicz, and Z. Fu, "W-net: dual supervised medical image segmentation model with multi-dimensional attention and cascade multi-scale convolution," 2020, https://arxiv.org/abs/2012.03674.

[20] C. Guo, M. Szemenyei, Y. Yi, Y. Hu, W. Wang, and W. Zhou, "channel attention residual U-net for retinal vessel segmentation," 2020, https://arxiv.org/abs/2004.03702.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, June. 2020.

[23] X. Qi, K. Li, P. Liu, X. Zhou, and M. Sun, "Deep attention and multi-scale networks for accurate remote sensing image segmentation," *IEEE Access*, vol. 8, Article ID 146627, 2020.

[24] X. Lai, W. Yang, and R. Li, "DBT masses automatic segmentation using U-net neural networks," *Computational and mathematical methods in medicine*, vol. 2020, Article ID 7156165, 10 pages, 2020.

[25] J. Kong, Y. Gao, Y. Zhang, H. Lei, Y. Wang, and H. Zhang, "Improved attention mechanism and residual network for remote sensing image scene classification," *IEEE Access*, vol. 9, no. 2021, Article ID 134800.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] A. Ali, Y. Zhu, and M. Zakarya, "Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks," *Information Sciences*, vol. 577, pp. 852–870, 2021.

[28] X. Tang, H. Zhang, J. Ma, X. Zhang, and L. Jiao, "Supervised adaptive-RPN network for object detection in remote sensing images," in *Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2647–2650, Waikoloa, HI, USA, 26 September.-2 October 2020.

[29] L. Hou, J. Xue, K. Lu, L. Hao, and M. M. Rahman, "A single-stage multi-class object detection method for remote sensing images," in *Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, Sydney, Australia, December 2019.

[30] S. Dong and Z. Chen, "Block multi-dimensional attention for road segmentation in remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2021.