Attention Guided Encoder-Decoder Network With Multi-Scale Context Aggregation for Land Cover Segmentation

SHUYANG WANG\textsuperscript{1}, XIAODONG MU\textsuperscript{1}, DONGFANG YANG\textsuperscript{2}, HAO HE\textsuperscript{3}, AND PENG ZHAO\textsuperscript{1}

\textsuperscript{1}Department of Information Engineering, Xi’an Hi-Tech Research Institute, Xi’an 710025, China
\textsuperscript{2}Department of Control Engineering, Xi’an Hi-Tech Research Institute, Xi’an 710025, China
\textsuperscript{3}Department of Measurement and Control, Qingzhou Hi-Tech Research Institute, Qingzhou 262500, China

Corresponding author: Dongfang Yang (yangdf301@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673017 and Grant 61403398.

\section*{ABSTRACT}
Land cover segmentation is an important and challenging task in the field of remote sensing. Even though convolutional neural networks (CNNs) provide great support for semantic segmentation, standard models are still difficult to capture global information and long-range dependencies in remote sensing images. To overcome these limitations, we proposed an attention guided encoder-decoder network with multi-scale context aggregation to achieve more accurate segmentation of land cover. Based on the structure of the encoder-decoder network, we introduce a multi-scale feature fusion module with two attention modules to the top of the encoder. The multi-scale feature fusion module is employed to aggregate multi-scale features and capture global correlations. The attention modules are used to exploit the long-range dependencies and the interdependence between channels from the perspective of space and channel respectively. The experimental results on the GF-2 images show that our proposed method achieves state-of-the-art performance, with an OA of 84.1\% and the mIoU of 62.3\%. Compared with the baseline network, our method improves the OA by 3.3\% and the mIoU by 4.4\%. The comparative experiments also demonstrate that the proposed approach can significantly improve the accuracy of land cover segmentation than other compared methods.

\section*{INDEX TERMS}
Remote sensing, semantic segmentation, encoder-decoder network, attention mechanism, multi-scale feature.

\section*{I. INTRODUCTION}
Automatic extraction of land cover information plays an important role in many applications, such as land use mapping, land resource management, urban planning and environmental monitoring [1]. With the rapid development of satellite and unmanned aerial vehicles (UAVs), a large number of high-resolution remote sensing images can be easily obtained from different sensors. Compared with the traditional methods such as field investigation, automatic land cover segmentation can reduce time-consuming and labor costs.

To achieve rapid and accurate land cover segmentation, people have tried many methods to deal with remote sensing data. The traditional land cover segmentation methods are mostly based on maximum likelihood classifier (MLC), clustering and logistic regression [2], [3]. Moreover, some more advanced methods such as K-nearest neighbor (KNN) [4], decision tree (DT) [5], random forest (RF) [6] and support vector machines (SVM) [7], spectral signal mixture analysis [8] and genetic programming [9], [10] are also used to address this issue. These methods perform well in some cases, but they usually only work in a small range of data, and can’t be validated on large datasets. It is because some parameters in these methods need to be tuned elaborately, and these parameters may vary in different images, which limits the generalization performance of these methods. In recent years, deep learning (DL) technology and convolutional neural networks have made great progress in the field of computer vision [11]–[13]. Therefore, many DL based semantic segmentation methods, such as U-net [14] and SegNet [15], are widely used in remote sensing segmentation [16]–[19]. Ghosh et al. [20] proposed a stacked U-net for ground material segmentation in remote sensing imagery. Seferbekov et al. [21] proposed a feature pyramid based fully convolutional network (FCN) for multi-category land cover segmentation.
However, the normal FCN based network is still limited by its scope of the receptive field, and can only perceive the context features within a certain range, so it can not perceive the long-range information in the whole image.

To enhance the context correlation between pixels, PSPNet [22] divides the feature mapping into multiple regions by pyramid pooling module, and the pixels in each region can be considered the global representation. Chen et al. [23] proposed a multi-scale atrous spatial pyramid pooling to aggregate context information. Vo and Lee [24] increased the accuracy of segmentation by using multi-scale images and multi-scale dilated convolutions. Li et al. [25] proposed a dilated-inception net to extract multi-scale features for semantic segmentation. Lan et al. [26] presented a global context based dilated CNN to capture and fuse the multi-scale features to achieve stronger feature representation. Chai et al. [27] used the distance map to learn spatial context information for high-resolution aerial image semantic segmentation. Mou et al. [28] proposed a relation augmented FCN, which introduced the spatial relation module and channel relation module to learn the relationship between any two pixels. Besides, attention mechanisms can effectively integrate local and global features to establish long-range context dependence, which is widely used in many tasks, such as machine translation, video classification, target detection and semantic segmentation. Wang et al. [29] proposed non-local operations as a spatial attention block for capturing long-range dependencies. PSANet [30] explains the attention mechanism from the perspective of information flow. It designs a point-wise spatial attention model, and infers the context dependence relationship between two points by their location relationship and semantic information. Li et al. [31] proposed a pyramid attention network for semantic segmentation, which obtains global context information through feature pyramid attention and global attention upsampling. Some studies focus on attention between channels. Hu et al. [32] presented SENet for image classification, with the Squeeze-and-Excitation block to model interdependencies between channels. DANet [33] introduces a dual attention module to enrich feature representation. These attention networks are integrated into the semantic segmentation network of natural scenes to improve accuracy.

However, targets in remote sensing images are special and complex. On the one hand, the ground objects in remote sensing images are usually scattered distributed, and there is often a long-distance between different objects. On the other hand, their semantic information often has multi-scale features, such as “river” and “lake” are on different scales. From a small-scale perspective, “river” and “lake” are both water regions, but the objects are different on large scales. Therefore, the context information of different scales is needed to identify these ground objects. Meanwhile, the features in high-resolution remote sensing images usually show strong intra-class heterogeneity and inter-class homogeneity, so it is necessary to capture global information and long-range context to enhance the feature representation for land cover segmentation [34].

Some studies have applied attention networks to the semantic segmentation of remote sensing images. Pan et al. [35] presented a generative adversarial network with attention mechanisms for building extraction. Ding et al. [36] designed a patch attention module and an attention embedding module, to enhance the embedding of context information and enrich the semantic information.

To better capture the global information and long-range dependencies for land cover segmentation, we proposed a semantic segmentation network that integrated attention mechanisms and multi-scale context aggregation for land cover segmentation. The main contributions of this paper are summarized as follows.

1. The multi-scale feature fusion module is integrated on top of the encoder network to capture multi-scale features and aggregate the global contextual information.
2. Two attention modules are designed to model the long-range contextual dependencies and channel dependencies.
3. A novel segmentation network is proposed for land cover segmentation by taking advantage of the modules above. The ablation experiments demonstrate the effectiveness of the proposed modules, and the comparative experiments on the GF-2 dataset show that the proposed approach achieves the state-of-the-art accuracy.

II. PROPOSED METHOD

A. OVERVIEW OF THE PROPOSED NETWORK

The scale of targets in high-resolution remote sensing images varies greatly. For small-scale targets, the network needs to have enough detail identification ability, and for large-scale targets, it also needs to be able to process global features. Also, different types of objects in remote sensing images may have similar spectral and texture features, and the same category of objects may have different features due to environmental factors. The traditional CNNs used for segmentation usually generate local feature representation from the local receptive field. Therefore, its weakness in representing long-range context features may lead to intra-class inconsistency and affect the segmentation accuracy.

To address this issue, we present a land cover segmentation network that combines attention mechanism and multi-scale feature fusion. Firstly, the multi-scale feature fusion (MFF) module is used to capture multi-scale features, then the integrated features are sent into a spatial attention module and a channel module simultaneously. The two attention modules can help to integrate the local features and global features, so that the long-range dependence between pixels can be established, and more accurate classification can be achieved. The architecture of the proposed network is shown in Fig 1.

For more detailed segmentation, the network is designed by following the architecture of the encoder-decoder network. The encoder-decoder networks have been successfully
applied to the semantic segmentation of remote sensing images [17], [20], [34]. Typically, the encoder-decoder networks contain an encoder part and a decoder part. The encoder part provides a mapping from the input image to the latent feature space to capture high semantic information. The decoder part can gradually recover the spatial information and provide a detailed segmentation map. Therefore, the encoder-decoder networks perform better in segmenting details.

The proposed network consists of 3 parts, an encoder, a decoder and a middle bridge. The encoder is usually established from a normal CNN. We selected ResNet101 pre-trained on the ImageNet dataset as the backbone of the encoder. ResNet101 is a well-known convolutional neural network proposed by Kaiming He, which is widely used in image classification, and the details of ResNet101 refer to [37].

We designed the middle bridge by application of an MFF module and two attention modules. Concretely, the feature map produced by the encoder is first fed into the MFF module to capture multi-scale feature, and then the output is sent into the two parallel attention modules to model the long-range dependencies.

The decoder is used to gradually restore the condensed feature map to the size of the original image. The decoder used in our study is relatively simple, and its structure is illustrated in Fig.1. The decoder up-samples the feature map that output by the middle bridge by a factor of 4, and then concatenates with the feature maps that have the same resolution from lower layers of the encoder part to get detailed local features. Finally, after a few $3 \times 3$ convolutions, the feature map is up-sampled to the size of the original image by another bilinear up-sampling.

**B. MULTI-SCALE FEATURE FUSION MODULE**

The multi-scale feature fusion is achieved by atrous spatial pyramid pooling (ASPP). It is formulated by atrous convolution and spatial pyramid pooling [23].

Atrous convolution can control the receptive field of the convolutional layer to collect image features of different resolutions and capture multi-scale context information. Atrous convolution is realized by filling zeros between adjacent weights of the ordinary convolution kernel. The distance between the adjacent weights is called the dilation rate. Forwarding the input image $X$ by a filter $w$, its output $y$ is

$$y[i] = \sum_k x[i + r \cdot k] w[k]$$

where $i$ is the index of a pixel, $r$ is the dilation rate of the atrous convolution, and $k$ is the index of the elements in the filter $w$. So that the normal convolution is a special kind of atrous convolution, and its dilation rate is 1. Variety dilation rates can be used to adjust the range of receptive fields and capture the features of different scales.

Since atrous convolution is obtained by filling zeros between adjacent weights of a convolutional kernel, the number of parameters is not increased, but the receptive field becomes larger, so more global features can be extracted.

When the dilation rate of atrous convolution is $r$ and the size of convolution kernel is $k$, the size of receptive field $F$ is calculated as follows:

$$F = (r - 1) \times (k - 1) + k$$

By utilizing several parallel atrous convolutions, the pyramid model can capture the multi-scale features. In detail, a smaller dilation rate means a smaller receptive field, through which local details can be effectively learned. In the same way, a larger dilation rate means a larger receptive field, through which global context features can be aggregated.

We designed the multi-scale feature fusion module according to the ASPP of DeepLab v3+ [38], as shown in Fig.2. The multi-scale feature fusion module is a parallel structure composed of several multiple branches. These branches process the input feature map simultaneously and produce an output that incorporates multi-scale information. The multi-scale feature fusion module uses atrous convolution with
different dilation rates in each branch. Besides, there is a global average pooling branch to merge the global context information. Then, the bilinear interpolation is applied to the feature maps generated by the above branches, and their sizes are unified to the size of the input feature map. Finally, these feature maps are merged and then a $1 \times 1$ convolution is used to make the output result consistent with the channel number of the original feature map.

C. ATTENTION MODULE

After multi-scale feature fusion, the feature map is sent into the attention module to establish a more extensive and richer context representation. In essence, the attention module calculates the correlation through matrix transposition and multiplication operation in mathematics, which increases the feature weight with strong dependence, and improves the utilization of effective information.

The attention module is composed of two parallel attention modules, namely spatial attention module and channel attention module, so that the semantic association of features is modeled from the perspective of space and channel. The spatial attention module gathers features by weighted summation of all location features, and captures the spatial dependence between any two positions in the feature map. No matter how far the spatial location is, similar features will obtain a closer distribution distance in the feature space. Channel feature maps can also be regarded as specific responses to different classes, and different channels are related to different categories of semantic information. Therefore, the feature representation to enhance specific semantics can be realized by improving the dependency relationship between channel mappings through the channel attention module.

1) SPATIAL ATTENTION MODULE

It is important to obtain distinctive feature representation in remote sensing image segmentation. To build a more abundant context model on local features, the spatial attention module is introduced to model the location relationship between different pixels, so that a long-distance context information dependence can be established, and the representation ability of features is improved. For each location feature, the weighted summation of all positions is used to update the feature, and the weight can be determined according to the similarity. In this case, no matter how far the distance in the spatial dimension is, any two pixels with similar characteristics can be extracted. The spatial attention module is shown in Fig.3.

Concretely, given a feature map $X \in \mathbb{R}^{C \times H \times W}$, where $C, H, W$ represent the channel, height and width of the feature map. Firstly, the feature map $X$ is sent into the convolutional layer with three different convolution kernels, and three new feature maps $A, B$ and $D$ are generated, which have $\{A, B, D\} \in \mathbb{R}^{C \times H \times W}$. Reshape them into $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of the pixels. Then the transposed feature map $A^T$ is multiplied with feature map $B$, and calculate the result with softmax function to have the spatial attention feature map $S \in \mathbb{R}^{N \times N}$.

$$s_{ji} = \frac{\exp(A_j \cdot B_j)}{\sum_{i=1}^{N} \exp(A_j \cdot B_j)} \tag{3}$$

where $s_{ji}$ represents the relationship between $i_{th}$ pixel and $j_{th}$ pixel. The more similar the two features are, the more likely they are to belong to the same category.

At the same time, we send $X$ into another convolutional layer to generate a new feature map $D \in \mathbb{R}^{C \times H \times W}$ and reshape it to $\mathbb{R}^{C \times N}$. Then, the feature map $D \in \mathbb{R}^{C \times N}$ is multiplied with the transposed feature map $S^T$, and the calculated result is reshaped into $\mathbb{R}^{C \times H \times W}$. Finally, the output $E \in \mathbb{R}^{C \times H \times W}$ is calculated as follows:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ij}D_i) + X_j \tag{4}$$

where $\alpha$ is the spatial attention parameter, which is initialized to 0, and can be updated through the training process.

From equation (4), it can be learned that the result feature map $E$ is the weighted summation of all the spatial features and the original features in each position. Therefore, it reflects both global context features and selective statistical features derived from spatial attention feature maps. Similar semantic features will complement each other, thus, it can improve the semantic consistency of intra-class.

2) CHANNEL ATTENTION MODULE

The channel attention module simulates the inter-dependence between channels and highlights the feature mapping of interdependence, which improves the feature representation of specific semantics. The design of the channel attention module is composed of two parallel attention modules, namely channel attention module and feature attention module, so that the semantic association of features is modeled from the perspective of channel and feature. The channel attention module simulates the inter-dependence, which improves the feature representation and the original features in each position. Therefore, the feature representation to enhance specific semantics can be realized by improving the dependency relationship between channel mappings through the channel attention module.

FIGURE 2. Detailed design of the multi-scale feature fusion module.

FIGURE 3. The details of the spatial attention module.
module is similar to the spatial attention module, but it is in the channel dimension. The structure of the channel attention module is depicted in Fig.4.

Different from the spatial attention module, the channel attention module directly calculates the channel attention map \( M \in \mathbb{R}^{C \times C} \) through \( \lambda \in \mathbb{R}^{C \times H \times W} \), without sending it through a convolutional layer. Firstly, we reshape the given feature map \( \lambda \in \mathbb{R}^{C \times H \times W} \) into \( X \in \mathbb{R}^{C \times N} \), and then multiply the reshaped result \( X \) by its transposition matrix. Then, a softmax layer is applied to obtain the channel attention map, \( M \in \mathbb{R}^{C \times C} \).

\[
m_{ij} = \frac{\exp(X_i \cdot X_j)}{\sum_{i=1}^{C} \exp(X_i \cdot X_j)}
\]

where \( m_{ij} \) represents the impact of the \( i \)th channel on the \( j \)th channel.

Similarly, the channel attention map \( M \) is multiplied by \( X \), and the result is reshaped as \( \mathbb{R}^{C \times H \times W} \). Multiply the result by the coefficient \( \beta \), and sum with original feature map \( \lambda \) to get the final output \( \mathbf{F} \in \mathbb{R}^{C \times H \times W} \).

\[
F_j = \beta \sum_{i=1}^{N} (m_{ij}X_i) + X_j
\]

where, \( \beta \) is the channel attention parameter, which is initialized as 0.

Therefore, it can combine the weighted features of all channels into the original features, and highlight the class related feature mapping.

After the parallel spatial attention module and the channel attention module, the feature maps obtained from the two attention modules are summed in pixel-wise, so that the result is the output of the attention module.

### D. LOSS FUNCTION

Besides the traditional cross-entropy loss function \( L_{\text{seg}} \), we use the auxiliary supervision \( L_{\text{segA}} \) on the feature map behind the attention module to get more semantic distinguishing features. \( L_{\text{segA}} \) is used to evaluate the segmentation results of the feature map obtained after the attention module, and the cross-entropy loss is also employed as the segmentation loss.

The final loss function can be expressed as follows:

\[
L = \lambda_1 L_{\text{seg}} + \lambda_2 L_{\text{segA}}
\]

where \( \lambda_1 \) and \( \lambda_2 \) are two parameters to balance each term in the loss function.

### III. EXPERIMENTAL RESULTS AND DISCUSSION

#### A. THE DATASET AND PREPROCESSING

In this paper, we selected images from the GF-2 dataset as experimental data [39]. The dataset contains 20 images and the corresponding annotations. Each image is in \( 7200 \times 6800 \). The images not only provide RGB channels, but also provide the near-infrared channel. The corresponding spectral range is red \((0.63-0.69 \mu m)\), green \((0.52-0.59 \mu m)\), blue \((0.45-0.52 \mu m)\) and the near-infrared \((0.77-0.89 \mu m)\), respectively, with a resolution of 4 m/pixel.

These satellite images are collected from different cities in China, from December 5, 2014 to October 13, 2016. It provides 15 categories of land cover annotations, including paddy field, irrigated land, dry cropland, garden plot, arbor woodland, shrub land, natural grassland, artificial grassland, industrial land, urban residential, rural residential, traffic land, river, lake, and pond. The other uncovered areas are labeled as background. Some 4-band images, with their corresponding RGB images and annotations of the GF-2 dataset are shown in Fig.5.

The experiment is conducted in the framework of Pytorch, using one GPU of NVIDIA GeForce GTX 1080Ti (11G) to accelerate. Limited to the memory of GPU, the training data is randomly cropped into \( 512 \times 512 \). To improve the stability of the network training process, the input image is normalized to \([0, 1]\). We augmented the training data by horizontal flip, vertical flip, random rotation, and random scaling.

Although it is possible to obtain better scores and results by using the pre-processing enhancement strategy, in the comparative experiments, we do not use any extra pre-processing strategy in any method, which is fair to all methods [40]–[42].

The Adam optimizer [43] is used to optimize the network, the learning rate is set to \( \alpha = 10^{-3} \), and the batch size is set to 6.

#### B. EVALUATION METHOD

To evaluate the result of the segmentation network, we used the confusion matrix, overall accuracy (OA), and mean intersection over union (mIoU) as the metrics for experiments.

The confusion matrix is a standard method to evaluate the classification performance of an algorithm. It is represented by a matrix of \( n \times n \), where \( n \) is the number of categories. For this experiment, \( n = 16 \). The confusion matrix is defined as \( x_{ij} = \{x_{ij} \mid i, j = 1, 2, \cdots , 16\} \), where \( x_{ij} \) indicates that one pixel is predicted as category \( i \) and the ground truth is category \( j \).

According to the confusion matrix, the number of true positive (TP), false positive (FP), and false negative (FN) of any class can be obtained as follows.

\[
TP_i = x_{ii}
\]

\[
FP_i = \sum_{j=1}^{16} x_{ij} - x_{ii}
\]

\[
FN_i = \sum_{j=1}^{16} x_{ji} - x_{ii}
\]
Overall accuracy refers to the percentage of correctly classified pixels in all pixels, as defined in equation (11).

\[
OA = \frac{\sum TP_i}{N}
\]  

where \(N\) is the number of total pixels.

Intersection on union (IoU), also known as the Jaccard index, is defined as follows:

\[
IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}
\]  

And mIoU is calculated as follows:

\[
mIoU = \frac{\sum IoU_i}{n}
\]  

C. EXPERIMENTAL RESULTS AND ANALYSIS

We randomly selected 5 out of 20 images in the GF-2 dataset as test data, and the remaining 15 images as training data.

Firstly, we need to determine the parameters of the loss function in equation (7). Usually, \(\lambda_1\) and \(\lambda_2\) are set empirically. We set \(\lambda_1 = 1\), then chose different \(\lambda_2\) to carry out the contrast experiment. The results are shown in Table 1. Therefore, we set \(\lambda_1 = 1, \lambda_2 = 0.4\) in the following experiments.

We further compared the experimental results with several state-of-the-art semantic segmentation models including U-net [14], SegNet [15], Deeplab V3+ [38] and DANet [32]. The results are shown in Table 2.

Table 2 shows the OA and mIoU of each method on the GF-2 dataset. The results show that the OA and mIoU of the proposed method are 84.1% and 62.3% respectively, which are better than other methods obviously. In general, in the control group, the segmentation metrics of U-net and SegNet are worse than other methods, while DANet and Deeplab V3+ are better. Compared with U-net and SegNet, the Deeplab V3+ can utilize multi-scale features to aggregate context information, and DANet can establish long-range context correlations by using attention modules. It reflects that multi-scale features and attention mechanisms are instrumental in land cover segmentation.

Figure 6 show some examples of the land cover segmentation results on the GF-2 dataset. It can be observed that our method can recognize confusing areas and capture the contours of different land cover categories better, so it makes fewer mistakes.
To further validate the effect of multi-scale feature fusion and attention module in land cover segmentation, we carried out the ablation study. The backbones of the following groups of experiments are the same. Firstly, the multi-scale feature fusion module is added to the network. Further, the multi-scale feature fusion module and attention modules are added to the network simultaneously. The comparison results are shown in Table 3.

Compared with baseline, the overall accuracy and mIoU are improved by 1.5% and 2.1% respectively after adding the MFF module. It shows that the MFF module can effectively capture targets of different scales and improve the segmentation accuracy of the network. After adding the attention modules, the overall accuracy and mIoU are further improved by 1.8% and 2.3%. Overall, compared with baseline, the OA and mIoU of our proposed method are improved by 3.3% and 4.4% respectively. It shows that multi-scale fusion and attention mechanisms improve the segmentation results of remote sensing images. Therefore, the effectiveness of the MFF module and attention modules are validated.

Table 3 shows the classification accuracy of each category to evaluate the segmentation performance of the model for different categories of land covers. The results show that for different categories, the multi-scale feature fusion module and attention module can capture the global context, enhance the feature representation, and suppress the local information interference, which can improve the classification ability of each category.

Figure 7 shows the visualization results of several comparative experiments. The rivers in Fig. 7 (a) and (b) have a similar appearance and texture features with lakes and ponds, so it is difficult to make accurate classification by local context information. By adding the multi-scale feature fusion module and attention module, the segmentation network has a better ability to capture context relations and reduce error classification. The road width in Figure 7 (c) and (d) is...
thin, belonging to the small-scale land cover, but its length appears large-scale features. The segmentation results of the traffic land category by the baseline network divided the road target into discrete sections. By using the multi-scale fusion strategy, more continuous road segmentation results can be obtained, which effectively improves the detection quality of

**FIGURE 7.** Visual comparisons between the baseline method and the proposed method using test samples from the GF-2 dataset. Each image was cropped into $1024 \times 1024$. For each row in (a)–(g), it contains an RGB image, two result maps and its corresponding label.
this kind of target. Moreover, as shown in Fig. 7 (e), (f), (g), compared with the baseline network, the proposed method can effectively extract all kinds of land cover information, and has fewer outliers and noise.

The confusion matrix of segmentation results on the GF-2 dataset is shown in Figure 8. From the confusion matrix, garden plot, arbor woodland, and shrub land have the worst segmentation accuracy. The pond and lake are all closed.
winters of different sizes, which are easy to be confused, so the segmentation accuracy is also low. The features of the farmland and built-up area are obvious, and the segmentation accuracy is better. Because the annotations of land cover segmentation can’t be very accurate, there are a lot of background areas that can’t be labeled, which makes all categories more likely to be classified into “other” category when they can’t be accurately distinguished, which also reduces the overall segmentation accuracy.

Figure 9 shows the segmentation result from a larger range of scenes. Our proposed approach can interpret and analyze the satellite image from the spatial distribution of different land cover categories well.

IV. CONCLUSION

This paper presents a method of land cover segmentation based on attention mechanism and multi-scale feature fusion. To improve the accuracy of land cover segmentation, a multi-scale fusion module and two attention modules are introduced. The multi-scale fusion module can extract the multi-scale features of the image, expand the receptive field of the convolutional neural network, and fuse the multi-scale context information to capture the global information. The attention module is composed of a spatial attention module and a channel attention module, which can adaptively capture the long-range context information and the interdependence between channels. The experimental results on the GF-2 images show that the proposed method can achieve better performance of land cover segmentation compared with the other methods, and the effectiveness of multi-scale fusion module and attention module is validated by an ablation study.

ACKNOWLEDGMENT

The authors would like to thank the State Key Laboratory LJESMARS of Wuhan University for providing the GF-2 dataset used in the experiments.

REFERENCES

[1] B. Cui, X. Chen, and Y. Lu, “Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection,” IEEE Access, vol. 8, pp. 116744–116755, 2020.

[2] P. K. Srivastava, D. Han, M. A. Rico-Ramirez, M. Bray, and T. Islam, “Selection of classification techniques for land use/land cover change investigation,” Adv. Space Res., vol. 50, no. 9, pp. 1250–1265, Nov. 2012.

[3] A. R. Cabral, S. Silva, P. C. Silva, L. Vanneschi, and M. J. Vasconcelos, “Burned area estimations derived from Landsat ETM+ and OLI data: Comparing genetic programming with maximum likelihood and classification and regression trees,” ISPRS J. Photogramm. Remote Sens., vol. 142, pp. 94–105, Aug. 2018.

[4] L. Samaniego, A. Bardossy, and K. Schulz, “Supervised classification of remotely sensed imagery using a modified k-NN technique,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 7, pp. 2112–2125, Jul. 2008.

[5] D. Phiri, M. Simwanda, V. Nyirenda, Y. Murayama, and M. Ranagalage, “Decision tree algorithms for developing rulesets for object-based land cover classification,” ISPRS Int. J. Geo-Inf., vol. 9, no. 5, p. 329, May 2020.

[6] M. Pal, “Random forest classifier for remote sensing classification,” Int. J. Remote Sens., vol. 26, no. 1, pp. 217–222, 2005.

[7] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” IEEE Trans. Geosci. Remote Sens., vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[8] L. Stancic, K. Ostir, and Ž. Kokalj, “Fluvial gravel bar mapping with spectral signal mixture analysis,” Eur. J. Remote Sens., pp. 1–16, Aug. 2020, doi: 10.1080/22797254.2020.1811776.

[9] C. Puente, G. Olague, S. V. Smith, S. H. Bullock, A. Hinojosa-Corona, and M. A. González-Botello, “A genetic programming approach to estimate vegetation cover in the context of soil erosion assessment,” Photogramm. Eng. Remote Sens., vol. 77, no. 4, pp. 363–376, Apr. 2011.

[10] C. Puente, G. Olague, M. Trabucchi, P. Arjona-Villicaña, and C. Souubrieville-Montalvo, “Synthesis of vegetation indices using genetic programming for soil erosion estimation,” Remote Sens., vol. 11, no. 2, p. 156, Jan. 2019.

[11] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.

[12] K. Simonany and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., San Diego, CA, USA, May 2015, pp. 1–14.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Med. Image Comput. Comput.-Assist. Intervent., Munich, Germany, 2015, pp. 234–241.

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

[16] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[17] C. Ma, Y. Xu, and D. Sha, “Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing,” IEEE Access, vol. 7, pp. 121685–121694, 2019.

[18] X. Liu, L. Jiao, X. Tang, Q. Sun, and D. Zhang, “Polarimetric convolutional network for PolSAR image classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 5, pp. 3040–3054, May 2019.

[19] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, “Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 10, pp. 5385–5399, Oct. 2017.

[20] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Schellappa, “Stacked U-Nets for ground material segmentation in remote sensing imagery,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, Jun. 2018, pp. 257–261.

[21] S. Feferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, “Feature pyramid network for multi-class land segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, Jun. 2018, pp. 272–275.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 6230–6239.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587. [Online]. Available: http://arxiv.org/abs/1706.05587

[24] D. M. Vo and S.-W. Lee, “Semantic image segmentation using fully convolutional neural networks with multi-scale image and multi-scale dilated convolutions,” Multimedia Tools Appl., vol. 77, no. 14, pp. 18689–18707, Jul. 2018.

[25] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, “Dilated-inception net: Multi-scale feature aggregation for cardiac right ventricle segmentation,” IEEE Trans. Biomed. Eng., vol. 66, no. 12, pp. 3499–3508, Dec. 2019.

[26] M. Lan, Y. Zhang, L. Zhang, and B. Du, “Global context based automatic road segmentation via dilated convolutional neural network,” Inf. Sci., vol. 535, pp. 156–171, Oct. 2020.

[27] D. Chai, S. Newsam, and J. Huang, “Aerial image semantic segmentation using DCNN predicted distance maps,” ISPRS J. Photogramm. Remote Sens., vol. 142, pp. 18689–18707, Jul. 2018.
[30] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, “PSANet: Point-wise spatial attention network for scene parsing,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 270–286.

[31] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in Proc. Brit. Mach. Vis. Conf., Newcastle, U.K., 2018, p. 285.

[32] J. Hu, L. Shen, and G. Sun, “Squeeze- and-excitation networks,” in Proc. IEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[34] C. Liu, D. Zeng, H. Wu, Y. Wang, S. Jia, and L. Xin, “Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network,” Remote Sens., vol. 12, no. 2, p. 311, Jan. 2020.

[35] X. Fan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren, “Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms,” Remote Sens., vol. 11, no. 8, p. 917, Apr. 2019.

[36] L. Ding, H. Tang, and L. Bruzzone, “LA-Net: Local attention embedding to improve the semantic segmentation of remote sensing images,” IEEE Trans. Geosci. Remote Sens., early access, May 27, 2020, doi: 10.1109/TGRS.2020.2994150.

[37] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2016, pp. 770–778.

[38] 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), Munich, Germany, 2018, pp. 833–851.

[39] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, “Learning transferable deep models for land-use classification with high-resolution remote sensing images,” 2018, arXiv:1807.05713. [Online]. Available: http://arxiv.org/abs/1807.05713

[40] L. Teng, F. Xue, and Q. Bai, “Remote sensing image enhancement via edge-preserving multiscale retinex,” IEEE Photon. J., vol. 11, no. 2, pp. 1–10, Apr. 2019.

[41] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, “Edge-enhanced GAN for remote sensing image superresolution,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[42] J. Liu, C. Zhou, P. Chen, and C. Kang, “An efficient contrast enhancement method for remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 10, pp. 1715–1719, Oct. 2017.

[43] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3rd Int. Conf. Learn. Represent., San Diego, CA, USA, 2015, pp. 1–15.

SHUYANG WANG received the B.S. and M.S. degrees in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Xi’an Hi-Tech Research Institute. Her current research interests include remote sensing image processing and computer vision.

XIAODONG MU is currently a Professor with the Xi’an Hi-Tech Research Institute, China. His research interests include intelligent information processing, data mining, and computer vision.

DONGFANG YANG received the B.S., M.S., and Ph.D. degrees in 2006, 2009, and 2017, respectively.

He is currently a Professor with the Xi’an Hi-Tech Research Institute. His research interests include computer vision, intelligent image processing, and navigation technologies.

HAO HE received the B.S., M.S., and Ph.D. degrees in 2013, 2016, and 2020, respectively.

He is currently a Lecturer with the Qingzhou Hi-Tech Research Institute. His research interests include road extraction, computer vision, and navigation technologies.

PENG ZHAO is currently a Professor with the Xi’an Hi-Tech Research Institute, China. His research interests include intelligent information processing, data mining, and communication technology.