Deep Attention and Multi-Scale Networks for Accurate Remote Sensing Image Segmentation

XINGQUN QI1,2, KAIQI LI1,2, PENGKUN LIU1,2, XIAOGUANG ZHOU3, AND MUYI SUN1,2
1School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
2Engineering Research Center of Information Network, Ministry of Education, Beijing 100876, China
3Minjiang University, Fuzhou 350108, China
Corresponding authors: Xiaoguang Zhou (zxg_bupt@126.com) and Muyi Sun (sunmuyi@bupt.edu.cn)

ABSTRACT
Remote sensing image segmentation is a challenging task in remote sensing image analysis. Remote sensing image segmentation has great significance in urban planning, crop planting, and other fields that need plentiful information about the land. Technically, this task suffers from the ultra-high resolution, large shooting angle, and feature complexity of the remote sensing images. To address these issues, we propose a deep learning-based network called ATD-LinkNet with several customized modules. Specifically, we propose a replaceable module named AT block using multi-scale convolution and attention mechanism as the building block in ATD-LinkNet. AT block fuses different scale features and effectively utilizes the abundant spatial and semantic information in remote sensing images. To refine the nonlinear boundaries of internal objects in remote sensing images, we adopt the dense upsampling convolution in the decoder part of ATD-LinkNet. Experimentally, we enforce sufficient comparative experiments on two public remote sensing datasets (Potsdam and DeepGlobe Road Extraction). The results show our ATD-LinkNet achieves better performance against most state-of-the-art networks. We obtain 89.0% for pixel-level accuracy in the Potsdam dataset and 62.68% for mean Intersection over Union in the DeepGlobe Road Extraction dataset.

INDEX TERMS
Remote sensing image, convolutional neural network, semantic segmentation, attention, multi-scale, dense upsampling convolution.

I. INTRODUCTION
Remote sensing image analysis is a hot research topic with academic and practical application values. Since remote sensing image includes rich geographic objects’ semantic information that can be applied to agriculture, environmental protection, geological exploration, and other aspects. Remote sensing image segmentation is one of the main directions for remote sensing image analysis. We can effectively extract the semantic information of various geographic objects contained in the image by segmenting the remote sensing image. This information contains features such as contours, edges, textures of ground objects which is very crucial in urban planning, crop planting, and even national defense military [1]. However, with the advancement of satellite and photography technology, the ultra-high resolution with a large shooting angle has become the main characteristic of remote sensing satellite images. This character has also become the main difficulty in remote sensing satellite image processing. Image segmentation is an imperative research field in image processing. In the traditional image segmentation methods, image information is mainly obtained by human eye recognition or some basic image feature processing methods such as scale-invariant feature transform(SIFT) [2] or histogram of oriented gradient(HOG) [3]. In our research, the remote sensing images with at least $1024 \times 1024$ pixels [4] or even $6000 \times 6000$ resolution (over 100MB) [5] are generally used in remote sensing image analysis. However, these traditional methods might not be adapted to ultra-high resolution because of the disadvantage of acquiring rich spatial and semantic information. In this situation, an ultra-high resolution image segmentation method with high precision and lightweight architecture is indispensable.

To our best knowledge, there are three types of methods for remote sensing image segmentation in previous research, the traditional image processing methods, machine learning methods for segmenting and semantic segmentation using deep learning methods.
In the field of traditional image processing methods, we find that the application of these methods can be used to perform image segmentation tasks for several simple situations. Pohl et al. [6] introduced a method of integrating and extracting digital remote sensing image features using image fusion. This method is compact and easy to operate. Saxton et al. [7] used a complex method by transforming, filtering, lattice averaging, and automatic pairing alignment to reconstruct the three-dimensional model for extracting feature information. Besides, this method should be implemented on an electron microscope. Franklin et al. [8] used semi-variograms to generate a variable window which could improve the texture analysis and edge detection results of the remote sensing images. However, these methods are limited by the fact that they can not be universally applied to various types of high-resolution remote sensing images. Meanwhile, their robustness and generalization are relatively poor.

In terms of the generalization and robustness of the various remote sensing image segmentation, several machine learning methods are proposed. Mitra et al. [9] proposed a technique using semi-supervised learning to train support vector machines (SVMs) for accurately classifying pixels of different semantics in remote sensing images. This method can be used on land classification, river distribution, and other geographical remote sensing images. Lary et al. [10] made a brief survey on machine learning with remote sensing images. He used the nonparametric regression and classification illustrative to build the model which validated some high resolution geosciences images. Pal and Mather [11] used SVM and maximum likelihood (ML) to perform segmentation tasks on multispectral and hyperspectral image datasets. Experiments show that SVM can be fully trained and can accurately segment multi-dimensional remote sensing images. However, in our study and experiments, these methods use machine learning methods also do not obtain high accuracy because of the inefficient use of spatial information.

With the development of artificial intelligence, the application of deep learning has become more and more mature in the field of computer vision [12]. Using the relevant models and methods of deep learning to segment images and extract features of images has become the mainstream due to the higher accuracy [13]. The deep learning model can be applied to the image classification [14], object recognition [15] and semantic segmentation [16] of remote sensing images. We conduct a brief survey on the applications of deep learning with remote sensing images. Li et al. [17] has proposed a simple convolutional neural network (CNN) [18] with just 7 layers to achieve high accuracy on the oil palm tree detection in the remote sensing images. Kussul et al. [19] used a convolutional neural network with fully connected layers to perform image segmentation on crops. The network labels each pixel in the input remote sensing images to classify each crop. Fully connected network (FCN) [20] has improved the structure of CNN to gain better performance. Then, FCN became the main method in the semantic segmentation of remote sensing images. Based on the FCN, Fu et al. [21] added atrous-convolution and conditional random field (CRF) structure to segment remote sensing images containing ground building information. Liu et al. [22] compared the CNN structure of two different structures, FCN and Unet [24], and found that Unet showed better performance in the segmentation task of remote sensing image with ultra-high resolution. Haut [25] combined the attention mechanism with the residual cascade structure and performed a land cover segmentation experiment on high-resolution remote sensing images. This method can effectively combine the characteristics of high frequency and low frequency, and filter the useless low-frequency surface features which contributes to segment the object in the image. Although there are many kinds of research on deep learning methods with remote sensing images, these methods are not accurate enough and can not take advantage of the abundant context information effectively in remote sensing images.

Among the extensive research efforts on ultra-high resolution remote sensing image processing, there is limited attention on how to utilize the context information of the images. In this paper, we propose ATD-LinkNet, a deep convolutional neural network integrated with attention mechanism. ATD-LinkNet is based on D-LinkNet which was a convolutional network with outstanding image segmentation performance on high resolution images [26]. With the attention mechanism, the ATD-LinkNet can generate effective receptive fields of different sizes for different input scales. Our network has two branches with different convolution kernels which can yield different sizes of the receptive field. Besides, we also propose attention building blocks. Attention blocks can be easily used as a drop-in replacement for other stacking deep networks. Our proposed network can effectively combine contextual semantic information with a slightly increased in parameter and computational cost in high resolution images. In ultra-high resolution images, the effective use of the context information of the pixels around the segmented object is one of the key factors for the final image segmentation with high precision. The network demonstrates outstanding image segmentation performance on ultra-high-resolution remote sensing images. Meanwhile, we verify the performance of ATD-LinkNet on two ultra-high resolution remote sensing image datasets, Potsdam [5], [67] and DeepGlobe Road Extraction [4], [68].

During the process of road extraction for remote sensing images, the boundary information of the road is crucial for the final performance. In the previous researches, bilinear interpolation and transposed convolution are two commonly used techniques for generating the prediction of the deep network [24], [27]. However, bilinear interpolation and transposed convolution might not be customized in the road extracting process which including many nonlinear features. For the road extraction task, we utilize the dense upsampling convolution (DUC) [28] as the decoder part in ATD-LinkNet for coarse-to-fine prediction.
The main contributions of our study are listed as follows:

1) We propose ATD-LinkNet, a deep convolutional network with attention mechanism for ultra-high resolution remote sensing image segmentation.

2) We propose the AT building blocks which can be easily transplanted in stacking networks.

3) We utilize dense upsampling convolution as the decoder part in ATD-LinkNet to refine the boundary information in road extraction task.

4) We enforce sufficient comparative experiments on Potsdam and DeepGlobe Road Extraction datasets, and obtain state-of-the-art performance.

II. METHODS

A. SEMANTIC SEGMENTATION WITH EFFICIENCY AND QUALITY

In recent years, there is more and more attention to the combination of remote sensing images and semantic segmentation using deep learning methods. Efficiency and quality have become the main requirements for deep learning models. Attention mechanism and multi-scale are two techniques often used to improve the efficiency and quality of convolutional neural networks. Attention mechanism is first used in natural language processing (NLP) because of the outstanding ability to processing the characteristics of long sequences [29]. The use of attention mechanisms in the field of computer vision has become very popular in recent years [30], [31], [46]. The attention mechanism can effectively utilize the rich spatial and semantic information in the remote sensing images [32]. More close to our work, Hu et al. [31] proposed an efficient and quality attention block named Squeeze-and-Excitation (SE) module which can be easily replaced in most CNNs architecture. SE module uses the channel-wise trick to obtain global attention in the deeper-level feature map which make good use of the contextual information. In our AT building block, we not only use the channel-wise attention mechanism but also take advantage of the spatial information as we use multi-scale convolutional kernel. Multi-scale has been proven to be powerful in convolutional networks for classification and segmentation [33], [34]. Many high-performance CNNs use multi-scale methods [35]–[37] to obtain effective use of spatial information. In our model, we use two different-sized convolution kernels to obtain spatial information of different scales and then gain the final output through channel-wise attention mechanism. As we exploit both channel-wise and spatial attention mechanism in the network, our ATD-LinkNet has greatly improved on the two datasets Potsdam and DeepGlobe Road Extraction compared with D-LinkNet.

B. OVERALL ATD-LinkNet ARCHITECTURE

The depiction of our ATD-LinkNet architecture is in Fig. 1.

We use 4 levels of downsampling blocks as the encoder parts of the overall network. And in this part, we use the AT building block to replace the Residual blocks which are used as feature extractors in the baseline network.
The encoder-decoder model is one of the most common structures in segmentation networks [38]. In the encoder part, the network must ensure the effective use of spatial information and semantic information. In D-LinkNet, residual block stacking in the encoder part as a feature extractor. The residual block is a highly modular structure. Similar to the AT building block we proposed, the residual block can be stacked with a large number in the convolutional neural network. In the encoder part of D-LinkNet, the residual block is stacked in four levels, and each level performs feature extraction on image semantic information of a fixed dimension, as shown in Fig. 1. Unlike the residual block, the AT building block we proposed uses the attention module and multi-scale module to extract the context semantic information more comprehensively. AT building block is a more excellent feature extractor designed for the residual block. The AT building block does not negatively affect the original D-LinkNet structure during the replacement of the residual block, which is also the motivation of our proposed AT building block. Considering the complexity and coherence of the object in the remote sensing image, ATD-LinkNet adds some dilated convolutional layers followed by D-LinkNet after the encoder part. Dilated convolution has been proven to be very effective in maintaining the spatial and semantic information of deep feature maps [40], [41]. We take two branches with different kernel sizes not only for reducing the computational cost but also for obtaining the different spatial information. In remote sensing images, the image contains many objects of different sizes because of the large resolution of the image. It is necessary to capture spatial and semantic information of different scales to segment these objects well. To make the two branches have the same size feature map as output, we also use different dilation rates for different kernel sizes. After the different spatial information obtainment from the multi-scale convolutional kernels of these two branches, we sum the two groups of feature maps and send them to the attention module.

C. MULTI-SCALE SPATIAL INFORMATION

We generate the multi-scale channel convolution by splitting a convolution layer into two branches with kernel sizes are 3 and 5 as illustrated in Fig. 2. At this point, we use the group convolution [43] to replace the traditional convolution operation. Group convolution is first used in AlexNet [44] because of the limitation of computational power. However, the method of dividing large convolution kernels into several branches and small convolution kernels has become more and more popular in lightweight models [43], [45], [47]. Group convolution is first used in AlexNet [44] because of the limitation of computational power. However, the method of dividing large convolution kernels into several branches and small convolution kernels has become more and more popular in lightweight models [43], [45], [47]. We take two branches with different kernel sizes not only for reducing the computational cost but also for obtaining the different spatial information. In remote sensing images, the image contains many objects of different sizes because of the large resolution of the image. It is necessary to capture spatial and semantic information of different scales to segment these objects well. To make the two branches have the same size feature map as output, we also use different dilation rates for different kernel sizes. After the different spatial information obtainment from the multi-scale convolutional kernels of these two branches, we sum the two groups of feature maps and send them to the attention module.

D. CHANNEL ATTENTION MODULE

As followed by multi-scale convolutional kernels, we produce the additive feature maps which contain semantic information from different kernel sizes. In the attention module, we obtain feature maps of different scales from the two branches of...
the multi-scale module. And when we gain reshaped feature maps, we must ensure that the two branch feature maps maintain a uniform dimension. It worth notice that these feature maps will be summed as pixels by pixels instead of combining. The reason for this is versatile. In this way, we can ensure it is convenient to obtain the weights of the two branches after the conditional softmax. Furthermore, it can also ensure the unity of the dimensions and reduce the computation cost. However, features in this form are not suitable for calculation and extraction. And for aggregating the spatial information, a global average pooling layer is adopted as illustrated in Fig.2 The purpose of the attention module is to gain semantic weights around the segmented object. In SE block [31], it only takes the single branch to calculate the global spatial and semantic information. We adopt the multi-scale convolutional kernels as we opt for a common convolutional layer with ReLU function [48] and Batch Normalization [49] to reduce the dimensionality of feature vectors after global average pooling. This can also reduce the computational cost and ensure that our attention model is lightweight. After that we conduct a conditional softmax function [23], [50], [51] to obtain the weight of the two branches. Notice that the conditional softmax function might be changed with the channel numbers in the multi-scale model. We generate the weight of the different branches in the form of the tensor. The core of our attention model is to get the different weights of the two branches with different kernel sizes. We multiply these weights with the different feature maps obtained in the multi-scale module. The final feature map will be the sum of the feature maps of the two branches.

E. COARSE-TO-FINE MODULE
Dense upsampling convolution (DUC) is used in the road extraction task of ATD-LinkNet as the decoder part restoring the resolution of feature map from $16 \times 16$ to $512 \times 512$ as illustrated in Fig.3 DUC is proven to be an effective alternative to bilinear interpolation and transposed convolution in semantic segmentation [27]. Especially in high resolution image segmentation tasks, DUC shows outstanding performance since it can alleviate the loss of information during the upsampling process [52]. Considering an input remote sensing image with resolution $H \times W \times C$, and the feature map $h \times w \times c$ after encoder part in ATD-LinkNet, where $c$ is the number of category in segmentation. We set the downsampling factor as $d$, so we have $h = H/d$ and $w = W/d$. The DUC operation applies the convolution between the input feature and output prediction map which generates the feature map with $h \times w \times (d \times d \times c)$. And then, the new feature map is reshaped in $H \times W \times C$ as the output prediction map. DUC can prevent uppooling operations like transposed convolution from inserting unrelated pixel values. Considering the characteristics of connectivity and narrowness in road extraction, we apply DUC to the decoder part of ATD-LinkNet. Furthermore, DUC can recoup the problem of gridding and can recompose the jagged boundaries caused by dilated convolution in road extraction. In the Potsdam dataset, ATD-LinkNet remains the same as D-LinkNet with the transposed convolution of the decoder part.

F. IMPLEMENTATION DETAILS
In the training process, in order to prevent the model from overfitting, we use a series of data augmentation methods, such as random rotation images, flip, rescale, shift, etc. We use binary cross-entropy (BCE) loss function and Adam [53] as our training optimization strategy. BCELoss is a loss function for the two-class problem and is well utilized on the DeepGlobe Road Extraction dataset. However, when training the Potsdam dataset, we must convert the data into a one-hot form for training firstly. The initial learning rate of the network is set to $2 \times 10^{-4}$. As the network continues to optimize, the learning rate is also updated. The batch size of all the experiments we conducted is set to 4.

G. EVALUATION METRICS
Evaluation Metrics is the primary way we visualize the results of our experiments. Both Potsdam and DeepGlobe Road Extraction datasets have their official indicators. For Potsdam...
dataset, the official evaluation indicator is overall accuracy (OA). Kappa, F1 score, Model size, and other indicators are calculated by ourselves in order to fully reflect the excellent performance of our proposed model. For DeepGlobe Road Extraction dataset, the official evaluation indicator is mIOU, and this benchmark doesn’t provide the label of its testing set. So we have to submit our prediction results to the official website to get the score. Besides, in order to ensure the consistency of the experiments we conducted on the two datasets, we also use mIOU in the Potsdam dataset as an evaluation metric. The main calculation formulas are as follows:

\[
\text{accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(y_i = \hat{y}_i) \tag{1}
\]

\[
\text{mIOU} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{2}
\]

where \(\hat{y}_i\) means the predicted values of \(i-th\) samples and \(y_i\) represents the true value in formula 1 which is the main evaluation metric in Potsdam benchmark. And in formula 2, True Positive (TP) means the total number of positive pixels which are correctly predicted; False Positive (FP) means the total number of negative pixels which are incorrectly predicted; False Negative (FN) means total number of positive pixels which are incorrectly predicted. \(i-th\) means the index of predicted images.

III. RESULTS

We evaluate the ATD-LinkNet on two remote sensing image benchmark: Potsdam dataset and DeepGlobe Road dataset. We mainly perform the experimental comparison between D-LinkNet50 and ATD-LinkNet50. ATD-LinkNet obtains the best performance on both benchmarks. The ablation study is also presented carefully on the road extraction task. And in the Potsdam and DeepGlobe Road Extraction, all the comparisons above on these two benchmarks are based on using the same training and testing sets. But every researcher can process the datasets by themselves. The official websites doesn’t make any restrictions on the use of the datasets.

A. DATASETS AND PRE-PROCESSING

We mainly use two image datasets to verify and compare the optimization model. The two datasets are the Potsdam Remote Sensing Image Dataset and the DeepGlobe Road Extraction Challenge Dataset, both of which are already open-sourced.

Potsdam is an open-source remote sensing image dataset provided by ISPRS (International Society for Photogrammetry and Remote Sensing). The entire dataset consists of a total of 38 images. These 38 images are with a spatial resolution of 6000 × 6000. The ISPRS official website divides the dataset, of which 24 images are used as training sets, and 14 images are used as testing sets. The ground truth of the testing set is officially given by ISPRS. Because of the limited computing resources, we crop all the images in the training set and testing set to 512 × 512 size patches of Potsdam dataset. Partial pixels overlap between patches. We crop an image in the original dataset into 144 patches according to 12 rows and 12 columns. In the patches from rows 1 to 11 and columns 1 to 11, there is no overlapping of pixels. In the 11th to 12th rows and 11th to 12th columns of patches, there is an overlap of 144 × 512 pixels between patches. This will not affect our training process. But in the testing process, we need to stitch these patches into the original size. We take patches of size 72 × 512 from the 11th row (column) and the 12th row (column) according to a 50% ratio for splicing. Finally, we obtain 3,456 training images and 2,016 testing patches with a spatial resolution of 512 × 512. Potsdam is a multispectral dataset that contains three types of images. They are RGB, RGBIR, RG-IR, and IR refers to the infrared channel. Due to the limitation of computational resources, we have to use three-channel images as experimental inputs. Combining the previous research on the Potsdam dataset, we find that though some researchers used the RG-IR band as input [60], most researchers still choose the RGB band [55], [56], [65], [66]. And we use RGB band images for experiments. The specific comparison is shown in Fig.4

The DeepGlobe Road Extraction Challenge is a binary image semantic segmentation task set up by IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) in 2018. This task is mainly to segments and labels some of the main roads in the satellite image, as shown in Fig.5 The entire dataset contains 6,226 images of the training set, all of which contain ground truth. At present, the official version of the challenge only provides 1,243 images for testing. The images of these testing do not provide ground truth, but the researchers can submit the results of the test to the challenge website and test it. The image spatial resolutions of the training set and the testing set are both 1024×1024. Similar to the processing method of the Potsdam dataset, we also crop the training and testing images of the DeepGlobe Road Extraction dataset into patches of size 512 × 512. Accordingly, we finally obtain 24,904 training patches and 4,972 testing patches. There are no pixels that overlap between patches. And finally, we use stitching to restore the test results to 1024 × 1024 for evaluation.

**FIGURE 5.** The instances of DeepGlobe Road Extraction dataset. Row (a) represents the original images; row (b) represents the ground truth.
TABLE 1. The results on Potsdam dataset with model D-LinkNet50 and ATD-LinkNet50.

| Model                  | mIOU(%) | Kappa(%) | OA(%) | F1 score | Model Size(kB) | Time(hours) |
|------------------------|--------|---------|-------|----------|----------------|-------------|
| SVL [57]               | -      | -       | 77.8  | -        | -              | -           |
| UZ [58]                | -      | -       | 85.8  | -        | -              | -           |
| PTHRCRF [55]           | -      | -       | 88.4  | -        | -              | -           |
| RFT [59]               | -      | -       | 88.6  | -        | -              | -           |
| UFMG [60]              | -      | -       | 87.9  | -        | -              | -           |
| KLab [69]              | -      | -       | 86.7  | -        | -              | -           |
| DownPCN [65]           | -      | -       | 84.1  | -        | -              | -           |
| D-LinkNet50 (baseline) | 65.12  | 79.22   | 84.2  | 0.84     | 850,487        | 46.16       |
| ATD-LinkNet50 (ours)   | 74.31  | 85.60   | 89.0  | 0.90     | 858,171        | 48.71       |

B. RESULTS ON POTSDAM DATASET

Our experiments on the Postdam dataset compare the performance with the previous researches with D-LinkNet50 and ATD-LinkNet50. The evaluation indicators mainly include overall accuracy, mIOU, Kappa, F1 score and Model Size, as well as the Training Time of the model. On the Potsdam dataset, there are six types of image segmentation tasks.

In our study, we conduct several experiments on the Postdam dataset. We use D-LinkNet50 and ATD-LinkNet50 to perform and average the testing results after training to obtain the final model performance. We also make comparisons with previous studies. The comparisons of these models are listed in Table 1. It’s worth noting that the official evaluation metric of Potsdam is Overall Accuracy (OA). Other indicators, such as Kappa, F1 score, mIOU, etc., could not be obtained from the work of other researchers, so we use “−” in Table 1 to indicate. As we can see that ATD-LinkNet achieves the highest accuracy in these models. ATD-LinkNet50 is nearly 5% higher than D-LinkNet50 on overall accuracy, 6% higher than D-LinkNet on F1 score and even 9% higher than D-LinkNet on mIOU. Samples of the test results of the two models for the Potsdam dataset are shown in Fig.6 We implement the state-of-the-art performance on the Potsdam dataset.

Meanwhile, we also conduct comparative experiments of the images with different channels in the Potsdam dataset as shown in Table 2. We use two types of RG-IR and RGB images for training and testing in the same computing environment. The results show that different channel images only have a small effect on the final prediction. This is negligible in a practical application environment with a larger number of images and more complicated image distribution. More importantly, it also shows that our network can fully extract and use the semantic information in the image. Our network is robust against different types of images.

C. RESULTS ON DeepGlobe ROAD EXTRACTION DATASET

Similar to the experiments on Potsdam, we also use the D-LinkNet50 and ATD-LinkNet50 models to test the DeepGlobe Road Extraction dataset. Unlike the previous experiments, the DeepGlobe Road Extraction does not provide ground truth for the testing set. All of the results of our experiments are evaluated on the DeepGlobe Road Extraction Challenge website. This benchmark mainly uses mIOU to measure the testing results. The image spatial resolution of the DeepGlobe Road Extraction Challenge is 1024 × 1024. We crop origin image to 512 × 512 for training and testing as patch-wise and then restore them to 1024 × 1024 for evaluation.

In our study, our experiment results have been successful in the ATD-LinkNet50 model. As can be seen from Table 3, the ATD-LinkNet50 is 1.5% higher than the D-LinkNet50 in accuracy. To show the outstanding performance of our proposed ATD-LinkNet network structure, we also compare it with related models in the workshop of IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW) 2018 in DeepGlobe Road Extraction Challenge. To our best knowledge, ATD-LinkNet has achieved state-of-the-art results. Testing samples for the baseline and ATD-LinkNet models of the DeepGlobe Road
TABLE 2. The prediction score for each category on the Potsdam dataset based on RG-IR and RGB channel.

| Category | Indicator | Impervious Surface | Building | Low Vegetation | Tree | Car | Clutter/Background | Overall |
|----------|-----------|--------------------|----------|----------------|------|-----|--------------------|---------|
| RG-IR    | Accuracy(%)| 91.9               | 94.8     | 86.7           | 82.6 | 94.0| 42.1               | 88.4    |
| RGB      | Accuracy(%)| 92.0               | 95.3     | 88.8           | 83.3 | 93.5| 52.4               | 89.0    |
| RG-IR    | F1 score(%)| 91.1               | 95.1     | 83.9           | 84.1 | 93.9| 44.0               | 89.6    |
| RGB      | F1 score(%)| 91.6               | 95.3     | 84.0           | 85.0 | 94.2| 49.9               | 90.0    |

TABLE 3. The results on DeepGlobe Road Extraction dataset with model D-LinkNet50, ATD-LinkNet50 and previous researchers.

| Model          | mIOU(%) |
|----------------|---------|
| GPS-CNN [61]  | 44.02   |
| EosResUNet [62]| 55.96   |
| StackedUNet [63]| 60.60  |
| NU-Net [70]   | 57.80   |
| IBN [71]      | 59.65   |
| ResIncepSkipNet [64] | 61.30 |
| D-LinkNet50   | 61.12   |
| ATD-LinkNet50 (ours) | **62.68** |

TABLE 4. The ablation study about DUC module and attention module on road extraction task.

| Model                                      | mIOU(%) |
|--------------------------------------------|---------|
| D-LinkNet50 (baseline)                     | 61.12   |
| D-LinkNet50 + DUC                         | 61.13   |
| D-LinkNet50 + attention module             | 62.37   |
| D-LinkNet50 + attention module + DUC      | **62.68** |

TABLE 5. The ablation study about multi-scale convolution kernels on road extraction task.

| Model                                      | mIOU(%) |
|--------------------------------------------|---------|
| D-LinkNet50 + DUC                         | 61.13   |
| D-LinkNet50 + DUC + 5 × 5 kernel           | 62.07   |
| D-LinkNet50 + DUC + 3 × 3 kernel           | 62.16   |
| D-LinkNet50 + DUC + multi-scale kernel     | **62.68** |

IV. DISCUSSION

Our research is based on the combination of relevant models of computer vision in deep learning and high resolution remote sensing images analysis. We use D-LinkNet as the baseline network model, which performed satisfactorily in DeepGlobe Road Extraction Challenge. We propose a novel replacement structure that could be easily utilized in the baseline network named AT blocks. Inspired by the SE blocks [31], our proposed AT blocks could be integrated on most deep convolutional networks. We name our proposed model as ATD-LinkNet. Meanwhile, we experimentally verify our network on two remote sensing datasets which are Potsdam and DeepGlobe Road Extraction.
ATD-LinkNet utilizes the attention mechanism to efficiently extract and gain global context semantic information in high-resolution images. And our proposed network demonstrates excellent performance in both the Potsdam and DeepGlobe Road Extraction dataset. ATD-LinkNet increases accuracy by nearly 5% on Potsdam based on D-LinkNet. Besides, considering the connectivity and narrowness of the road extraction task, we adopt the dense upsampling convolution as the decoder part in ATD-LinkNet to refine the road boundaries. With the attention mechanism and dense upsampling convolution module, ATD-LinkNet increases more than 1.5% on the road extraction task. We also conduct careful ablation study with attention model and dense upsampling convolution module based on D-LinkNet50.

Our novel method is validated and improved on both datasets, compared to the previous method. Although these improvement reactions are not so obvious on indicators, which also has certain research and practical value. The semantic segmentation task needs to label each pixel of the input image, which is very difficult. For datasets with high-spatial-resolution such as Potsdam and DeepGlobe Road Extraction, a marginal improvement on OA and mIOU is also of research significance. Besides, on the DeepGlobe Road Extraction dataset, although there is an only marginal improvement on the indicator, the segmentation effect of the image is greatly improved, just as shown in Fig.7 and Fig.8.

We conduct the detailed statistics and classification on the prediction results of the Potsdam dataset, and calculate the prediction results of accuracy for each category, as shown in Table. 7. In the 6 categories, the scores obtained by ATD-LinkNet50 are better than D-LinkNet50. Among them, we find that ATD-LinkNet scores more than 90% in the three categories of Impervious Surface, Building and Car. This may because our proposed AT Block could make more
effective use of semantic information in images. We find that in the three types of predictions of Low Vegetation, Tree, and Clutter / Background, although ATD-LinkNet50 has higher scores than D-LinkNet50, it doesn’t exceed 90%. For the two categories of Low Vegetation and Tree, their semantic information is very similar, and segmenting these two categories is a challenging task. In the last category, both D-LinkNet and ATD-LinkNet score lower. This may be related to our experimental methods and experimental computing resources. Clutter/ Background doesn’t have obvious semantic features, and during our experiments, we use patches as training and testing samples. This might cause the semantic information of Clutter/ Background to be incoherent between the patches, which harm the segmentation of the global context semantic information extracted by the model.

In the experiments of the Potsdam dataset, we add time complexity as one of our evaluation indicators. Time complexity can reflect the excellent performance of our experimental methods. Under the same calculation conditions, our method only slightly increases the inference time than the baseline model, but our model achieves a much better result. In practical applications, the inference time used in the training process can often be shortened a lot by better equipment. In this case, the improvement in accuracy is often the most important. Our method has a higher value in practical application scenarios. Meanwhile, we also calculate the FLOPS (Floating-point Operations Per Second) and Parameters of D-LinkNet50 and ATD-LinkNet50 as shown in Table. 8. Compared with D-LinkNet50, ATD-LinkNet50 only slightly increases FLOPS and Parameters in the same order of...
magnitude, but its accuracy improves significantly. This reflects the superiority and applicability of our method.

Although our approached methods obtain satisfactory predictions, there are still some limitations. Firstly, our method optimize the encoder portion of the baseline network model and change the way of the original convolution operation. These optimizations will result in inadequate utilization of the relevant parameters in the pre-trained model. This might harm the final experimental results. Our experiments are based on the environment in which the pre-trained model is not loaded. Secondly, due to our limited computational budget, we conduct all the experiments with 512 × 512 spatial resolution in the road extraction task which may cause the network cannot fully capture the overall context and semantic information in the remote sensing images. The advantages of our attention model can not be fully utilized.

V. CONCLUSION
In conclusion, we have proposed a deep network with attention mechanism and multi-scale convolutions for the segmentation of the remote sensing images which have ultra-high resolution. We named our proposed network as ATD-LinkNet. Meanwhile, we proposed the replacement building blocks named AT blocks. Extensive comparison experiments are implemented on two benchmarks named Potsdam and DeepGlobe Road Extraction. Furthermore, we adopted dense upsampling convolution to refine the boundaries in the road extraction task.

In future work, we will further optimize ATD-LinkNet. And we will validate the performance of ATD-LinkNet with more remote sensing image datasets. Meanwhile, we will use some large datasets to pre-train ATD-LinkNet to get the pre-trained model. Besides, our proposed replacement structure can be easily utilized in other networks, we will conduct more experiments with various baseline networks which can reflect the excellent performance of our proposed method more strongly.

ACKNOWLEDGMENT
(Xingjun Qi and Kaizi Li contributed equally to this work.)

REFERENCES
[1] S. Dawn, V. Saxena, and B. Sharma, “Remote sensing image registration techniques: A survey,” in Proc. Int. Conf. Image Signal Process. (ICISP), Trois-Rivières, QC, Canada, Jun./Jul. 2010, pp. 103–112.
[2] W. Cheung and G. Hamarneh, “n-SIFT: n-dimensional scale invariant feature transform,” IEEE Trans. Image Process., vol. 18, no. 9, pp. 2012–2021, Sep. 2009, doi: 10.1109/TIP.2009.2024578.
[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), San Diego, CA, USA, Jun. 2005, pp. 886–893.
[4] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “DeepGlobe 2018: A challenge to parse the earth through satellite images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 1712–1720.
[5] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, “THE ISPRS benchmark on urban object classification and 3D building reconstruction,” ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci., vol. 1, no. 1, pp. 293–298, Jul. 2012, doi: 10.5194/isprsannals-I-3-293-2012.
[6] C. Pohl and J. L. Van Genderen, “Review article multisensor image fusion in remote sensing: Concepts, methods and applications,” Int. J. Remote Sens., vol. 19, no. 5, pp. 823–854, Jan. 1998, doi: 10.1080/014311698215748.
[7] W. O. Saxton, T. J. Pitt, and M. Horner, “Digital image processing: The semiper system,” Ultramicroscopy, vol. 4, no. 3, pp. 343–353, 1979, doi: 10.1016/0304-3991(79)90044-3.
[8] S. E. Franklin, M. A. Walder, and M. B. Lavigne, “Automated derivation of geographic window sizes for use in remote sensing digital image texture analysis,” Comput. Geosci., vol. 22, no. 6, pp. 665–673, 1996, doi: 10.1016/0098-3004(96)00009-X.
[9] P. Mitra, B. U. Shankar, and S. K. Pal, “Segmentation of multispectral remote sensing images using active support vector machines,” Pattern Recognit. Lett., vol. 25, no. 9, pp. 1067–1074, Jul. 2004, doi: 10.1016/j.patrec.2004.03.004.
[10] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, “Machine learning in geosciences and remote sensing,” Geosci. Frontiers, vol. 7, no. 1, pp. 3–10, Jan. 2016, doi: 10.1016/j.gsf.2015.07.003.
[11] M. Pal and P. M. Mather, “Support vector machines for classification in remote sensing,” Int. J. Remote Sens., vol. 26, no. 5, pp. 1007–1011, Mar. 2005, doi: 10.1080/014311605123314083.
[12] I. Arel, D. C. Rose, and T. P. Karnowski, “Deep machine learning—A new frontier in artificial intelligence research,” IEEE Comput. Intell. Mag., vol. 5, no. 4, pp. 13–18, Nov. 2010, doi: 10.1109/MCI.2010.938364.
[13] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” IEEE Geosci. Remote Sens. Mag., vol. 4, no. 2, pp. 22–40, Jun. 2016, doi: 10.1109/MGRS.2016.2540798.
[14] E. Maggiori, Y. Tarabalka, G. Charpillet, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 2, pp. 645–657, Feb. 2017, doi: 10.1109/TGRS.2016.2612821.
[15] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: 10.1109/TGRS.2016.2601622.
[16] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, “Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2016, pp. 1–9.
[17] H. Santoso, H. Tani, and X. Wang, “A simple method for detection and counting of oil palm trees using high-resolution multispectral satellite imagery,” Int. J. Remote Sens., vol. 37, no. 21, pp. 5122–5134, Nov. 2016, doi: 10.1080/0143116161265527.
[18] X. Qi, “The understanding of convolutional neuron network family,” DEStech Trans. Comput. Sci. Eng., Dec. 2017, doi: 10.12783/dctese/cii2017/17254.
[19] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep learning classification of land cover and crop types using remote sensing data,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 5, pp. 778–782, May 2017, doi: 10.1109/LGRS.2017.2681128.
[20] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.
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.

A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representations for efficient semantic segmentation,” in Proc. IEEE Vis. Commun. Image Process. (VCIP), St. Petersburg, FL, USA, Dec. 2017, pp. 10–13.

X. Zhang, Z. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6848–6856.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, F. Pereira and C. J. C. Burges, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1251–1258.

S. Wang, Y. Hu, and N. Liu, “Signal separation of phase-sensitive optical time-domain reflectometry considering thermo-mechanical coupling and 3D data matching,” Traitement du Signal, vol. 36, no. 1, pp. 65–77, Apr. 2019, doi: 10.18280/ts.360109.

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, pp. 7794–7803. [Online]. Available: http://arxiv.org/abs/1704.04861.

V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. IEEE Int. Conf. Mach. Learn. (ICML), Haifa, Israel, Feb. 2010, pp. 807–814.

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems, C. J. C. Burges and L. Bottou, Eds. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

X. Qi, T. Wang, and J. Liu, “Comparison of support vector machine and softmax classifiers in computer vision,” in Proc. 2nd Int. Conf. Mech. Control Comput. Eng. (ICMCE), Harbin, China, Dec. 2017, pp. 151–155.

B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8697–8710.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), Vancouver, BC, Canada, Dec. 2017, pp. 8026–8037.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems, C. J. C. Burges and L. Bottou, Eds. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

Y. Liu and G. E. Hinton, “Akari: A deep neural network for 3D scene understanding,” in Proc. Int. Conf. Comput. Vis. (ICCV), St. Petersburg, FL, USA, Dec. 2017, pp. 1429–1438.

M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, pp. 7794–7803. [Online]. Available: http://arxiv.org/abs/1704.04861.
[62] O. Filin, A. Zapara, and S. Panchenko, “Road detection with EOSResUNet and post vectorizing algorithm,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 211–215.

[63] T. Sun, Z. Chen, W. Yang, and Y. Wang, “Stacked U-Nets with multi-output for road extraction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 202–206.

[64] J. Doshi, “Residual inception skip network for binary segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 216–219.

[65] J. Sherrah, “Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,” 2016, arXiv:1606.02585. [Online]. Available: http://arxiv.org/abs/1606.02585.

[66] N. Audebert, B. Le Saux, and S. Lefevre, “Joint learning from earth observation and OpenStreetMap data to get faster better semantic maps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 67–75.

[67] 2D Semantic Labeling Contest—Potsdam. Accessed: Mar. 15, 2020. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[68] DeepGlobe Road Extraction Challenge. Accessed: Mar. 15, 2020. [Online]. Available: https://competitions.codalab.org/competitions/18467

[69] R. Kemker, C. Salvaggio, and C. Kanan, “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 60–77, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.04.014.

[70] M. Samy, K. Amer, K. Eissa, M. Shaker, and M. EIHelw, “NU-Net: Deep residual wide field of view convolutional neural network for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 267–271.

[71] Y. Li, B. Peng, L. He, K. Fan, Z. Li, and L. Tong, “Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks,” Sensors, vol. 19, no. 19, p. 4115, Sep. 2019, doi: 10.3390/s19194115.

XINGQUN QI is currently pursuing the master’s degree in control science and engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include deep learning, computer vision, and so on.

KAIQI LI is currently pursuing the master’s degree in logistics engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include machine learning, computer vision, neural networks, and so on.

PENGKUN LIU received the B.S. degree from the Logistics Engineering College, Shanghai Maritime University, Shanghai, China, in 2019. He is currently pursuing the master’s degree in control science and engineering with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include pattern recognition, deep learning, and computer vision.

XIAOGUANG ZHOU received the M.S. degree from the Department of Precision Instrument, Tsinghua University, in 1984, and the Ph.D. degree in engineering from the Tokyo University of Agriculture and Technology, Japan. He was a Visitor Professor with the Tokyo University of Agriculture and Technology, from 2001 to 2002, and a JSPS Researcher with The University of Tokyo, from 2013 to 2014. He is currently the Min Du Scholars Professor. He is the Vice Director of the Engineering Research Center of Information Networks, Ministry of Education, and the Ph.D. Supervisor of the Beijing University of Post and Telecommunications. He is the author of over 10 books, over 100 articles, and over 16 inventions. His research interests include control theory and its application in engineering, Internet of Things and automated logistics systems, and mechatronics technology. He is a Permanent Member of the Chinese Association of Automation/Manufacturing Technology Committee and the China Institute of Communications/Equipment Manufacturing Technical Committee.

MUYI SUN received the B.S. degree from the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree in control science and engineering. His research interests include pattern recognition, pathological image analysis, and computer vision.