2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService) | 979-8-3503-3379-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigDataService58306.2023.00037

APNet: Attention Mechanism with Point Sampling Loss Network for Remote Sensing Images Semantic Segmentation

Xuan Zhang
Beijing National Research Center for Information Science and Technology(BNRist)
Department of Automation
Tsinghua University
Beijing, China
zhangxua20@mails.tsinghua.edu.cn

Yushun Fan
Beijing National Research Center for Information Science and Technology(BNRist)
Department of Automation
Tsinghua University
Beijing, China
fanyus@tsinghua.edu.cn

Jia Zhang
Department of Computer Science
Southern Methodist University
Dallas, Texas, USA
jiazhang@smu.edu

***Abstract*—With the rapid development of remote sensing technology, high-resolution remote sensing images (HRRSIs) are becoming increasingly prevalent in various applications. HRRSIs contain a large amount of information on ground objects, exhibiting both diversity and complexity. The semantic segmentation of HRRSIs is a rapidly evolving field, emerging with the development of remote sensing technology. However, HRRSIs are typically captured from high altitudes, have a wide imaging range, and present a number of challenges, including foreground-background imbalance, multi-scale ground objects, large intra-class variance and small inter-class variance, all of which contribute to the difficulty of semantic segmentation. To address these challenges, this paper proposes an end-to-end semantic segmentation model named APNet, short for Attention Mechanism with Point Sampling Loss Network, which uses attention mechanism, multi-scale feature fusion, and a point sampling loss function. APNet aims to address the aforementioned challenges and improve the accuracy of semantic segmentation. We conducted comparative experiments with several classic methods on the LoveDA dataset. The results demonstrate the effectiveness of our proposed method in improving the results of semantic segmentation.**

***Index Terms*—Remote Sensing, Semantic Segmentation, Attention Mechanism, Point Sampling**

I. Introduction

Fueled by advances in computer vision, the semantic segmentation of high-resolution remote sensing images (HRRSIs) is developing in the direction of intelligence. As an application scenario of semantic segmentation, the purpose of semantic segmentation of HRRSIs is to assign each pixel in the image to a specific category. It is an important part of applications such as land planning, barren land identification, and landmark extraction.

Among various semantic segmentation methods, Convolutional Neural Networks (CNNs) have been widely used for various segmentation tasks [1] [2]. CNNs for image segmentation typically operate on regular grids: the input is a pixel-level regular grid, which is processed through multiple convolutions to obtain a regular grid containing feature information. Finally, various upsampling methods are used to obtain a regular grid containing labels. The "encoder-decoder" structure is a basic architecture of CNNs, where the encoder extracts features from the input image, and the decoder predicts the results based on the extracted features. However, convolution operations can reduce the resolution of the image, causing fine-grained details to be ignored, which poses a challenge to achieving fine-grained segmentation. Therefore, enriching the receptive field of the model and improving its feature extraction ability are crucial. In recent years, attention mechanism and multi-scale feature fusion have been commonly used in CNNs to achieve fine-grained semantic segmentation results with good performance.

Unlike general images, HRRSIs are usually captured from high altitudes, containing a large amount of object information, with significant scale variations of objects. Due to the large intra-class variance and small inter-class variance, the segmentation results have "salt and pepper" phenomenon and edge discontinuity problems. Meanwhile, the foreground pixels account for a much smaller proportion than the background, leading to poor segmentation quality of the foreground and an imbalanced foreground-background problem.To address these issues, advanced methods have been proposed. For example, multi-modal data like digital surface models (DSMs) can be used to increase the richness of feature information. Weighted focal loss [3] and dice loss [4] can be utilized to solve the class imbalance problem in HRRSIs. Incorporating object edge information [5] as prior knowledge in the model can also help tackle the edge discontinuity problem in segmentation results. These methods improve on a certain problem of HRRSIs, but it is difficult to achieve fine-grained segmentation of HRRSIs. In order to improve the feature extraction capability of the model and achieve fine-grained segmentation, we face three challenges.

979-8-3503-3379-4/23/$31.00 ©2023 IEEE
DOI 10.1109/BigDataService58306.2023.00037

- **Foreground-background imbalance**. In HRRSIs, some objects occupy a small area relative to the background, such as buildings in large forests or water in large agricultural fields. This foreground-background imbalance issue results in low segmentation accuracy of small objects, leading to missed detection or false detection issues in downstream tasks.
- **Multi-scale**. HRRSIs are usually observed from a high-altitude perspective, and the objects on the ground typically exhibit large-scale variations. On the one hand, different spatial resolutions can affect the features of the objects. On the other hand, different object categories also have significant differences in scale [6].
- **Large intra-class variance and small inter-class variance**. The texture and structure of objects in HRRSIs vary greatly, resulting in significant differences in the features of the same type of object on the same image. At the same time, the objects in HRRSIs also have some similar features, and the differences in appearance between different objects are not significant. These two issues lead to the common "salt and pepper" phenomenon (where some pixels of a certain type of object are labeled as other categories) and non-smooth edges of objects in existing segmentation methods.

In our work, we propose a semantic segmentation network called APNet (Attention Mechanism with Point Sampling Loss Network) for HRRSIs. APNet utilizes an attention mechanism, multi-scale feature fusion, and a point-sampling loss function to enhance semantic segmentation accuracy.

To address the foreground-background imbalance issue, we modeled the network in both channel and spatial dimensions, which enables the model to focus on important features and suppress unimportant ones, thereby improving the recognition of small target objects. To tackle the multi-scale problem, we utilized atrous convolutions with multiple rates and two global pooling methods to enrich the model's receptive field, capturing information of different object scales. To resolve the problem of large intra-class variance and small inter-class variance in HRRSIs, we proposed a point sampling loss function to supervise the difficult-to-classify points in the output, thereby improving the segmentation accuracy of the edge parts and mitigating the "salt and pepper" phenomenon.

The contributions of this study are as follows:

(1) We propose an attention block to address the small object problem and foreground-background imbalance problem in HRRSIs. By modeling in both channel and spatial dimensions, The model can pay more attention to the features of objects that are easily overlooked or less noticeable.

(2) We propose a point sampling loss function to tackle the issue of large intra-class variance and small inter-class variance in HRRSIs. By calculating the uncertainty of the output results and obtaining a set of difficult-to-classify points, the points are further modified to make the object edges smoother.

(3)Experiments on LoveDA dataset [7] demonstrate competitive results by learning representations for HRRSIs via APNet and show remarkable performance improvements on small objects and edge segmentation.

II. Related Work

Existing semantic segmentation methods are usually divided into two categories: traditional machine learning methods and deep learning methods.

A. Traditional methods

Traditional methods are generally designed for medium to low resolution images. Before the emergence of deep learning, traditional image segmentation methods relied heavily on expertise in digital image processing, topology, mathematics, and related fields. These methods can be broadly classified into three categories: threshold-based segmentation [8] [9], region-based segmentation [10], and edge detection-based segmentation [11]. Image segmentation is currently defined as the process of dividing an image into several non-overlapping regions based on its grayscale, color, texture, shape, and other features. The goal is to ensure consistency or similarity of these features within the same region and to show significant differences between different regions. However, it is important to note that segmented images lack semantic information. Traditional machine learning methods require manually designed features, which can be limited when dealing with high-resolution and multi-dimensional images.

B. Deep learning methods

In recent years, deep learning methods [12] [13] [14] have become one of the primary techniques for solving image segmentation problems. This is due to the availability of various fundamental algorithm frameworks, such as the classic models Fully Convolutional Networks (FCN) [15] and U-Net [16]. Based on these frameworks, many methods have been proposed to further improve the performance of semantic segmentation. SENet [17] improves model performance by modeling channels, while Dual Attention Network (DANet) [18] and Convolutional Block Attention Module (CBAM) [19] propose spatial and channel attention modules respectively to model dependencies in the spatial and channel dimensions. Object Context Network (OCNet) [20], inspired by the self-attention mechanism, represents semantic features by calculating the weighted similarity between pixels. MaskFormer [21] uses Transformer [22] structure to view semantic segmentation as mask-level classification rather than pixel-level, greatly improving accuracy but also resulting in significant computational overhead. In addition, atrous convolution [13] [14] and multi-level pooling [23] are also commonly used to enrich model receptive fields and improve multi-scale feature capture ability. Feature Pyramid Networks (FPN) [24] construct a bottom-up and top-down structure to fuse high-level features with low-level features, producing more expressive fused features. Methods that were originally developed for object detection have shown promising results in semantic segmentation. For instance, Mask R-CNN [25] uses Region Proposal Network (RPN) [26] to generate a set of candidate

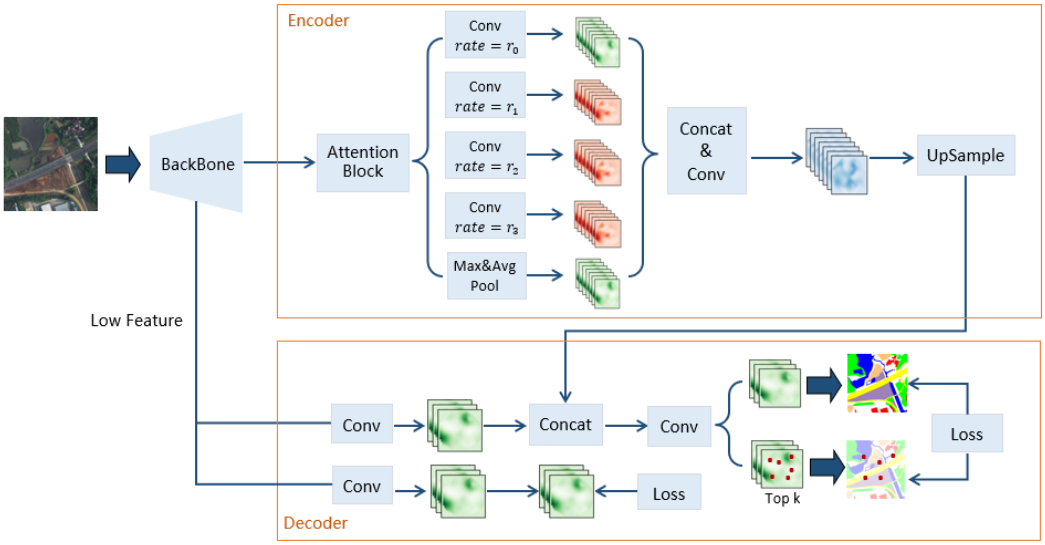


Fig. 1. Model structure

regions, and solves the problem of region mismatch caused by pooling operations through the ROIAlign layer. Mask Scoring R-CNN [27] further enhances the segmentation accuracy by adding a MaskIoU Head to evaluate the segmentation quality, building on the Mask R-CNN framework. However, most of these methods are proposed based on general images and may not perform optimally for HRRSIs.

Most existing segmentation methods for HRRSIs are based on generic semantic segmentation methods. For instance, [28] proposed a combined network based on ResNet and U-Net to improve the segmentation accuracy of HRRSIs. [29] introduced multispectral information into deep convolutional neural networks (DCNNs) to improve HRRSIs segmentation results. [30] built a convolutional neural network specifically for the problem of small object recognition in urban areas. In addition, many improved methods have been proposed focus on extracting specific features, such as roads and buildings. For example, [31] used dynamic labels to train CNNs to improve the accuracy of CNNs decision functions for road extraction. The above method usually only improves a specific problem in HRRSIs and it is difficult to comprehensively improve the fine segmentation results of HRRSIs.

III. Methodology

A. Model structure

The proposed model is composed of three parts: the backbone network, the encoder, and the decoder, as shown in Figure 1. The input image is first processed by the backbone network to extract features. The encoder further strengthens the feature extraction, including attention block and multi-scale feature extraction. In the decoder, the output of the encoder is concatenated and convolved with the low-level feature maps from the backbone network to obtain the prediction results. The loss is calculated based on the prediction and the ground truth labels, as well as the point sampling loss. In addition, to ensure the accuracy of the low-level feature map from the backbone network, we also supervise it by loss function.

B. Backbone for preliminary feature extraction

In this study, we used ResNet101 as the backbone [32] for preliminary feature extraction. ResNet101 is one of the more popular semantic segmentation networks, which effectively alleviates the network degradation problem caused by network depth by using multiple residual modules, and has achieved good results in many semantic segmentation tasks. In our research, ResNet101 initially extracts image features, which have a total of 4 stages of output. The low-level features contain more image detail information, while the high-level features provide a more comprehensive understanding of image features. Therefore, we use the outputs of the first and fourth stages for subsequent processes.

C. Encoder with attention and multi-scale feature fusion

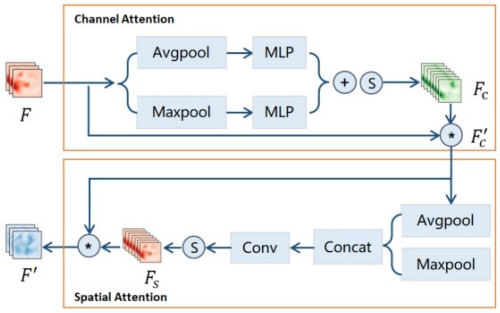


Fig. 2. Attention block structure

Attention block: We propose a method based on attention mechanism to solve the problem of small objects and

foreground-background imbalance in HRRSIs. The feature map $F \in \mathbb{R}^{C\times H\times W}$ (where C, H, and W denote the number of channels, the height, and the width of the feature map, respectively) from the fourth stage of the backbone network is used as input, and feature weighting is performed in the channel and spatial dimensions through the attention block. The structure of the attention block is shown in Figure 2. After multiple convolutions, the channels of the feature map often express different features, and these features have different effects on the segmentation result. The channel attention mechanism captures the dependency relationship between any two channels, assigns weights between channels, and thereby improves the model's ability to represent target features. Specifically, we first summarize the spatial features of F through global max pooling and global average pooling to generate two feature descriptors ($C \times 1 \times 1$). We send them to an MLP network to learn the parameters. Then, we element-wise add the output features, obtain the attention weights of each channel F_c after applying the sigmoid function, and multiply them with F channel by channel to obtain $F_c^{'}$. The calculation of the channel attention mechanism is shown in Formula (1):

$$F_c = sigmoid(MLP(Avgpool(F)) + MLP(Maxpool(F))) \tag{1}$$

The purpose of spatial attention mechanism is to capture the spatial dependency between any two positions. We first generate two feature descriptions for each spatial position through global average pooling and global max pooling. After concatenating the two feature descriptions, a 5x5 convolution operation is performed to obtain the spatial attention weight. The resulting spatial attention feature map F_s, obtained after passing through the sigmoid function, is then element-wise multiplied with the input $F_c^{'}$ to restore the original size. The calculation of the spatial attention mechanism is shown in Formula (2).

$$F_s = sigmoid(f([Avgpool(F_c^{'}) : Maxpool(F_c^{'})])) \tag{2}$$

Multi-scale feature extraction: Inspired by DeepLabv3 [14], this study uses atrous convolution with different dilation rates and two types of global pooling to enlarge the receptive field and tackle the multi-scale problem in HRRSIs. In our approach, the feature map F' output by the attention block is fed into four atrous convolutions with dilation rates of r_0, r_1, r_2, and r_3 to generate four feature maps of different scales. Meanwhile, to capture global information, we use global average pooling and global max pooling to obtain two different global vectors, which are added element-wise and then processed by a 1×1 convolution and upsampling to restore the size. Finally, we concatenate and convolve the outputs from the five sources and upsample the result to obtain the output of the encoder F_{en}.

D. Decoder for refined segmentation

The encoder extracts features from the image, while the decoder uses the feature map to make segmentation predictions. Generally, low-level features contain more image detail information that may have been lost in higher-level features. In the decoder, the output from the first stage of the backbone network is first reduced in channel dimension via 1x1 convolution and then concatenated with F_{en} to supplement the missing details. Subsequently, a 1x1 convolution is used to map the concatenated feature map to K channels (K is the number of classes), representing the predicted probability for each class and used for calculating the loss function.

Point sampling loss: In semantic segmentation tasks, the accuracy of object edges is often low [33]. To address the problem of large intra-class variance and small inter-class variance in HRRSIs and achieve fine-grained segmentation results, especially for edge segmentation, we propose a point sampling loss method. Specifically, we calculate the uncertainty of each pixel in the prediction results to obtain a set $N \in \mathbb{R}^{K\times k}$ containing k difficult-to-classify points. Then, we perform upsampling on the output results to match the size of the label. We obtain the corresponding label point set $N_l \in \mathbb{R}^{K\times k}$ from the label using the point set N and finally compute the loss between N and N_l. This approach improves the prediction accuracy while reducing computational costs. In the output results, the numerical values represent the probability of the pixel belonging to a certain category. The uncertainty calculation method first selects the two largest predicted values k_1 and k_2 for each class at each pixel in the output result and calculates the difference between k_1 and k_2 for each pixel. The smaller the difference, the greater the uncertainty of the model for that pixel. The calculation of uncertainty is shown in Formula (3):

$$uncertainty = |(P_{i,j}^{c_{max1}}) - (P_{i,j}^{c_{max2}})| \tag{3}$$

Therefore, the loss function in APNet consists of three parts. They are the loss calculation of the output result and the label L_{output}, the loss calculation of the difficult-to-classify point set and the corresponding label point set L_{point}, and the loss calculation of the low-level features of the backbone network and the label $L_{backbone}$. The calculation of the APNet loss function is shown in Formula (4):

$$L = L_{output} + L_{point} + L_{backbone} \tag{4}$$

IV. EXPERIMNTS

A. Training dataset and evaluation metrics

The dataset we used is LoveDA [7]. LoveDA has a spatial resolution of 0.3 meters and is divided into seven land cover classes (including the background class): background, building, road, water, barren, forest, and agriculture. Each

TABLE I
COMPARISON OF EXPERIMENTAL RESULTS ON TEST DATASET OF LOVEDA

Models	Mean (MIoU)	building	road	water	barren	forest	agriculture	background
FPN [24]	48.83	53.43	53.17	75.69	14.89	43.69	57.38	43.58
FCN-8s [15]	49.34	54.61	53.43	77.09	14.67	44.15	57.48	43.98
PSPNet [23]	49.99	55.36	51.55	77.88	14.18	46.13	60.86	43.99
OCRNet [34]	48.85	55.04	53.41	76.91	16.26	44.03	55.05	41.27
UperNet [35]	47.56	51.46	51.84	74.51	11.06	43.82	58.72	41.51
DeepLabV3+ [36]	49.68	54.54	55.31	77.01	13.94	45.32	58.34	43.29
ours	**52.79**	**59.08**	**58.75**	**80.13**	**18.4**	**46.33**	**61.31**	**45.52**

TABLE II
INFLUENCE OF ATTENTION BLOCK AND POINT SAMPLING LOSS

Models	Mean (MIoU)	building	road	water	barren	forest	agriculture	background
ours	**52.79**	**59.08**	**58.75**	**80.13**	**18.4**	**46.33**	**61.31**	**45.52**
-POI	50.35	56.3	56.44	76.9	15.38	45.44	57.93	44.04
-ATT	51.34	57.49	57.11	78.0	14.96	46.71	61.32	43.77
-N	49.9	54.96	54.48	77.83	13.05	46.41	60.18	42.39

image has a size of 1024×1024, and there are 2522, 1669, and 1796 images in the training, validation, and test sets, respectively. Compared with existing datasets, LoveDA has three main features: **1) Multi-scale objects.** Land cover of the same class exhibits different scale features in different scenes, which increases the multi-scale variation characteristics of the dataset. **2) Complex background.** The diverse background is an inevitable problem in HRRSIs segmentation tasks, which is particularly evident in the LoveDA dataset. The high resolution and diverse scenes bring richer detail information and larger intra-class differences to the background, posing higher demands on the models. **3) Class distribution inconsistency.** Urban and rural areas have different feature distributions. Urban areas with high population density contain more man-made buildings and roads, while rural areas contain more natural landscapes, such as forests and water. LoveDA focuses on the differences between urban and rural areas, bringing special challenges to the models. It is worth noting that LoveDA has class imbalance, with the barren class being relatively underrepresented compared to other classes.

Like previous work, we use mean inter-section over union (MIoU) as our evaluation criterion.

B. Implementation details

To improve the generalization ability of the model, we used three data augmentation techniques: random cropping, image flipping, and image distortion. The size of the random cropping was 512×512, and during cropping, the number of pixels for each class did not exceed 0.75 times the total number of pixels in the image. The probability of image flipping and distortion was set at 50%.

We used popular models as baselines, including FPN, FCN-8s, PSPNet, OCRNet, UperNet, and DeepLabV3+. They all used ResNet-101 as the backbone network. The dilation rates of APNet are 1, 12, 24, and 36, and the number of sampling points is 2048. we set the learning rate of APNet on the LoveDA dataset to 1e-2. All models adopt Adam as the optimizer and cross-entropy as the loss function. And the training epochs are set to 40000. Considering limited computing resources, the batch size is set to 8. We conduct all our experiments in the Tensorflow platform with NVIDIA 3090Ti GPU and the model training took 21 hours.

C. Experiments results and analysis

1) Experiments results: Table I shows the experiment results of our model on the LoveDA dataset. The results show that we achieve good results on LoveDA dataset. Compared with classical methods, the model using attention mechanism and point sampling loss performs better in the semantic segmentation task of HRRSIs. Compared with FPN which uses feature pyramids, APNet has an 8.1% improvement in the MIoU metric. Compared with OCRNet which uses self-attention mechanism, APNet has an 8.07% improvement in the MIoU metric. Compared with PSPNet which uses multi-level pooling, APNet has a 5.6% improvement in the MIoU metric. Additionally, we found that APNet also has a good effect on solving the problem of class imbalance, such as increasing the IoU value of the barren class to 18.4.

Figure 3 shows the visualization results of APNet and the best-performing PSPNet in the baseline models. As can be seen, APNet can effectively improve the accuracy of small object segmentation and greatly alleviate the problem of non-smooth edges of objects.

2) Further analysis: We use the model without attention block and point sampling loss for further analysis of the model to show the effect of attention mechanism and point sampling loss. In addition, we also discuss the impact of the position of the attention block and the number of sampling points on the performance of the model. We also discuss the inference speed and parameter count of APNet. The following experiments are conducted on the LoveDA dataset.

a) Ablation study: After we added attention block and point sampling loss to the model, we tested the gains they can bring, and the experiment results are shown in Table II. APNet-ATT is the APNet model without point sampling loss, APNet-POI is the APNet model without attention block,

and APNet-N is the model without both point sampling loss and attention block. By observing the results, We can find that the attention block and point sampling loss both have a positive impact on the model, with progressive growth in MIoU. The attention block models the spatial and channel dimensions to make the model focus on more important features. From the results, it can be seen that compared to APNet-N, APNet-ATT with the addition of the attention block effectively improves the classification accuracy of various land objects, especially in the recognition of small objects such as buildings and roads, APNet-ATT performs surprisingly well, indicating that the attention mechanism is important for improving the recognition of small targets. The model APNet-POI with the addition of point sampling loss also has better segmentation performance on small targets such as roads and buildings compared to APNet-N. In addition, it increases the MIoU value of the barren class, indicating that point sampling loss also has a certain effect on dealing with class imbalance problems in the dataset.

b) Sensitivity Analysis: We conducted a sensitivity analysis to investigate the impact of attention block position on model performance based on APNet-ATT, and the impact of the number of sampling points in point sampling loss on model performance based on APNet-POI.

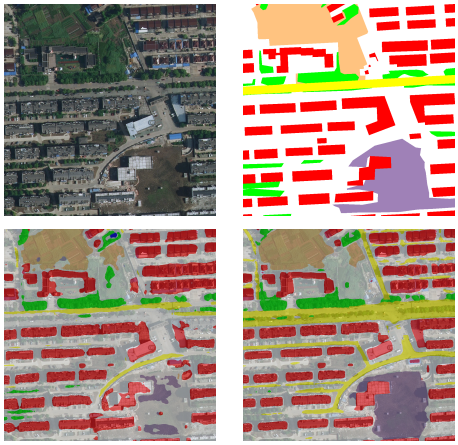
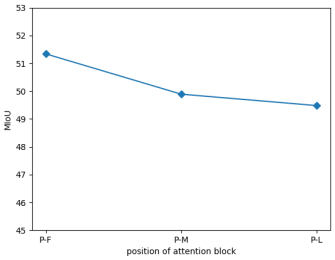


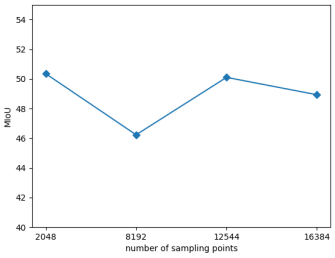
Fig. 3. Visualization results of the LoveDA dataset. From left to right, from top to bottom: original image, ground truth, PSPNet output results, and ours.

The position of attention block: The attention mechanism plays an important role in improving model performance, and its position in the model can also affect the segmentation results. In APNet, we place the attention block after the backbone network and before the multi-scale feature fusion, with the aim of enabling the subsequent processes of the model to operate on the feature map after attention weighting, so as to fully utilize the weighted information. In the experimental section, we discuss the impact of the attention block's position in the model on the performance, based on APNet-ATT. The results, as shown in Figure 4(a), indicate that the closer the attention block is to the front of the encoder, the better the model performance. The attention block is located respectively before multi-scale feature fusion after backbone (P-F), before convolution after concatenating multi-scale features (P-M), and before upsampling after multi-scale feature fusion (P-L). The experimental results show that the MIoU is 3.8% higher at location P-F than at location P-L.

The number of sampling points: The number of sampling points k in point sampling loss has an impact on the model's performance in our experiments. Therefore, we tested the performance of the model with different values of k based on APNet-POI. We set k to 2048, 8192, 12544, 16384, and the results are shown in Figure 4(b). According to the experimental results, a larger k value does not necessarily lead to better model performance. The optimal number of sampling points may be related to the data structure. In the dataset used in this study, good results were obtained when k was set to 2048 and 12544.



(a) The the influence of attention block position



(b) The the influence of sampling points number

Fig. 4. The the influence of attention block position and sampling points number

TABLE III
INFERENCE SPEED AND PARAMETER COMPARISON

Models	Time(s)	Parameter(Mb)
FPN	6	47.49
FCN-8s	5	66.12
PSPNet	6	65.6
OCRNet	7	55.51
UperNet	7	83.04
DeepLabV3+	6	62.19
ours	6	61.14

c) Inference speed: In order to compare the performance of our proposed model with the baseline model, we analyzed the number of parameters and the time required for predicting a single image. The experimental results are presented in Table

III. From the results, it can be seen that APNet achieves faster inference speeds while also having a smaller number of parameters. This indicates that APNet not only delivers better prediction results but also operates more efficiently.

V. Conclusion

In this study, we propose a semantic segmentation method for high-resolution remote sensing images (HRRSIs). Our approach employs an attention block, multi-scale feature fusion, and point sampling loss to tackle the challenges of foreground-background imbalance, multi-scale, and large intra-class variance and small inter-class variance. The attention block models relationships between any two positions in both channel and spatial dimensions, capturing correlations between different features and enhancing the feature extraction capability of the model. The attention block is effective in extracting small objects. The point sampling loss function calculates the uncertainty values of the model's final stage prediction results, obtains the set of difficult-to-classify points, and calculates the loss function corresponding to the point set of the label, thereby improving the segmentation results of difficult-to-classify points, and addressing the problem of "salt and pepper" phenomenon and non-smooth edges of segmented objects. In addition, our model also uses atrous convolution and global pooling to obtain multi-scale information, and combines high-level and low-level features to refine the segmentation results. To verify the performance of the model, we conduct experiments on the public LoveDA dataset to compare with other models, and perform ablation learning and sensitivity analysis for each innovative structure and parameter in our model, demonstrating the effectiveness of our model. In future research, we will attempt to utilize self-attention to further enhance the model's ability to capture contextual information. Moreover, the multimodal data of HRRSIs, such as geographic coordinates, has the potential to improve the semantic segmentation results.

References

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] R. Dong, X. Pan, and F. Li, "Denseu-net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65 347–65 356, 2019.

[4] Q. Zhu, Y. Zheng, Y. Jiang, and J. Yang, "Efficient multi-class semantic segmentation of high resolution aerial imagery with dilated linknet," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 1065–1068.

[5] C. He, S. Li, D. Xiong, P. Fang, and M. Liao, "Remote sensing image semantic segmentation based on edge information guidance," *Remote Sensing*, vol. 12, no. 9, p. 1501, 2020.

[6] Z. D. A, H. S. A, S. Z. A, J. Z. B, L. L. A, and H. Z. A, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, 2018.

[7] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[8] K. Bhargavi and S. Jyothi, "A survey on threshold based segmentation technique in image processing," *International Journal of Innovative Research and Development*, vol. 3, no. 12, pp. 234–239, 2014.

[9] S. Zhu, X. Xia, Q. Zhang, and K. Belloulata, "An image segmentation algorithm in image processing based on threshold segmentation," in *2007 Third International IEEE Conference on Signal-image Technologies and Internet-based System*. IEEE, 2007, pp. 673–678.

[10] Z. Wang, J. R. Jensen, and J. Im, "An automatic region-based image segmentation algorithm for remote sensing applications," *Environmental Modelling & Software*, vol. 25, no. 10, pp. 1149–1165, 2010.

[11] B. Alhadidi, M. H. Zu'bi, and H. N. Suleiman, "Mammogram breast cancer image detection using image processing functions," *Information Technology Journal*, vol. 6, no. 2, pp. 217–221, 2007.

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

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

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

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[18] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[20] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[21] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *IEEE Computer Society*, 2016.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[27] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 6409–6418.

[28] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[29] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,"

ISPRS Journal of Photogrammetry and Remote Sensing, vol. 145, pp. 60–77, 2018.

[30] 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 *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[31] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "Roadtracer: Automatic extraction of road networks from aerial images," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 4720–4728.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 3193–3202.

[34] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019.

[35] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.