LSKSANet: A Novel Architecture for Remote Sensing Image Semantic Segmentation Leveraging Large Selective Kernel and Sparse Attention Mechanism

Miao Fu[1], *Feng Gao*[1], *Ruzhuang Hua*[1], *Yanhai Gan*[1], *Xiaowei Zhou*[1], *Yang Zhou*[2]

[1] School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[2] China Electronic Standardization Institute Huadong Branch, Suzhou 215124, China

ABSTRACT

In this paper, we proposed large selective kernel and sparse attention network (LSKSANet) for remote sensing image semantic segmentation. The LSKSANet is a lightweight network that effectively combines convolution with sparse attention mechanisms. Specifically, we design large selective kernel module to decomposing the large kernel into a series of depth-wise convolutions with progressively increasing dilation rates, thereby expanding the receptive field without significantly increasing the computational burden. In addition, we introduce the sparse attention to keep the most useful self-attention values for better feature aggregation. Experimental results on the Vaihingen and Postdam datasets demonstrate the superior performance of the proposed LSKSANet over state-of-the-art methods.

***Index Terms*—** Deep learning, hybrid attention, semantic segmentation, Transformer, urban planning.

1. INTRODUCTION

In remote sensing image analysis, semantic segmentation serves as a critical technology, offering valuable insights into the intricate nature of the Earth's surface. Semantic segmentation plays important roles in disaster assessment [1], urban planning [2], and environmental monitoring [3].

With the advancement of deep learning methods, significant progress has been made in the field of remote sensing image semantic segmentation. For instance, Xu et al. [3] proposed a lightweight Transformer model to accelerate the processing speed of remote sensing images and improve classification results. Wang et al. [4] introduced the Swin Transformer as the backbone and designed a unique DC-FAM decoder, more effectively extracting contextual information. Zhang et al. [5] developed a deep neural network that combines Transformer and CNN, demonstrating exceptional performance in remote sensing image segmentation tasks through an encoder-decoder structure and multi-scale contextual processing.

Although existing methods have achieved excellent segmentation performance, they still face challenges. In some cases, remote sensing image semantic segmentation is conducted in resource-constrained environments, such as unmanned aerial vehicles or low-power satellite systems. Existing methods can hardly work well in these resource-limited scenarios.

To solve the above mentioned problem, we propose **L**arge **S**elective **K**ernel and **S**parse **A**ttention **Net**work (**LSKSANet**) for remote sensing image semantic segmentation. The proposed LSKSANet is a lightweight network architecture that effectively combines the robust feature extraction capabilities of CNN with advanced attention mechanisms. Specifically, we design large selective kernel module to decomposing the large kernel into a series of depth-wise convolutions with progressively increasing dilation rates, thereby expanding the receptive field without significantly increasing the computational burden. In addition, we introduce the sparse attention to keep the most useful self-attention values for better feature aggregation. Experimental results on the Vaihingen and Postdam datasets demonstrate the superior performance of the proposed LSKSANet over state-of-the-art methods.

2. METHODOLOGY

As depicted in Fig. 1, LSKSANet uses the classical encoder-decoder architecture. The encoder is based on the well-known ResNet18, and it leverages the pre-trained model to efficiently extract the local features (corner or texture) within the complex remote sensing imagery. In the decoder, the Large Selective Kernel and Sparse Attention (LSKSA) block is designed to integrate large kernel convolution and sparse attention mechanisms. As shown in Fig. 1(d), the LSKSA block is comprised of the Top-k sparse attention and large selective kernel convolution. This module enhances the model's ability to recognize large-scale structures in remote sensing images and improves the efficiency of processing key features.

This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0117202 and in part by the Natural Science Foundation of Qingdao under Grant 23-2-1-222-ZYYD-JCH. (Corresponding author: Feng Gao, Email: gaofeng@ouc.edu.cn)

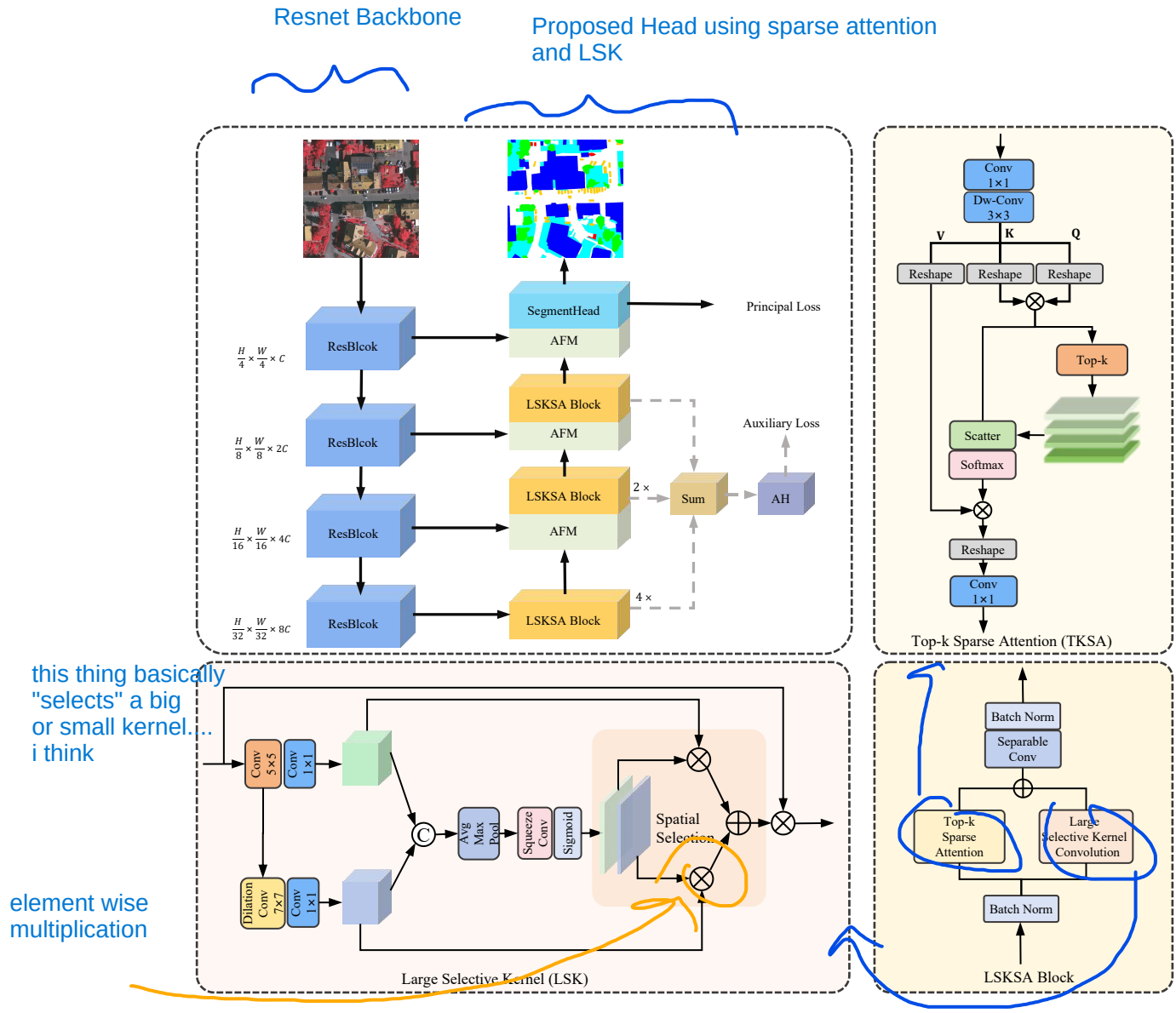

Fig. 1. LSKSANet Structure with LSK and TKSA Modules for Remote Sensing Image Semantic Segmentation

2.1. CNN-based Encoder

The proposed LSKSANet used the pre-trained ResNet18 as the encoder. The encoder consists of four stages, each of which systematically reduces the spatial dimensions of the input image. The weights of the encoder are initialized by the pre-trained weights from ImageNet. By combining the ResNet18 encoder into our network, detailed spatial information can be captured for accurate segmentation.

2.2. Large Selective Kernel and Sparse Attention Block-Based Decoder

The decoder of LSKSANet employs LSKSA blocks to enhance feature extraction and utilizes an Adaptive Fusion (AF) module to dynamically merge features from the encoder and the previous decoder layer. The details of the AF module is as follows:

$$FF = EF \cdot \alpha + DF \cdot (1 - \alpha) \quad (1)$$

where FF represents the fused features, EF denotes features from the encoder, DF denotes features from the decoder's LSKSA block, α is an learnable scalar.

The training process utilizes a cross-entropy loss function, supplemented with an auxiliary loss function to reinforce the training. The segment head concludes the decoder architecture, transforming the rich fused features into precise pixel-level classifications.

2.3. Large Selective Convolution

In traditional convolutional neural networks, large convolutional kernels typically lead to increased computational costs. We design the Large Selective Kernel (LSK) module to address this issue by decomposing the large kernel into a series of depth convolutions with progressively increasing dilation rates, thereby expanding the receptive field without significantly increasing the computational burden.

As shown in Fig. 1(c), the LSK module processes the input feature map x through the following steps: Firstly, the module applies 5×5 and 7×7 depth-wise convolutional kernels to capture features at different scales. Then, a 1×1 convolution is used for feature mixing. The obtained features are concatenated to U. Next, to selectively learn important features, the LSK module employs an adaptive selective mechanism. First, average pooling P_{avg} and max pooling P_{max} are used to extract two distinct spatial descriptors as:

$$U_{\text{avg}} = P_{\text{avg}}(U), \quad U_{\text{max}} = P_{\text{max}}(U) \tag{2}$$

Then, both spatial descriptors are concatenated, and then transformed into N spatial attention maps by a convolutional layer $F_{2\to N}$. After that, the Sigmoid activation σ is employed to generate the spatial selective masks SM. This process can be represented as:

$$SM = \sigma(F_{2\to N}(\text{Concat}(U_{\text{avg}}, U_{\text{max}}))) \tag{3}$$

Finally, SM is used to weight the original feature maps by element-wise multiplication, enhancing the important parts of the input feature. Details of the computation can be found in Fig. 1.

2.4. Sparse Attention

In contrast to the traditional densely connected self-attention paradigm that computes an attention map across all query-key pairs, our work integrates the sparse attention mechanism. It significantly reduces the computational burden of attention while preserving salient semantic information.

As depicted in Fig. 1(b), self-attention is applied along the channel dimension, followed by calculating the similarity between all reshaped query and key pixel pairs. Subsequently, a top-k strategy is employed to selectively discard elements with lower attention weights in the transition attention matrix M, of dimension $R^{\hat{C}\times\hat{C}}$.

The parameter k, which varies with the number of channels, dynamically controls the sparsity level. Specifically, by using various top-k proportions, different amounts of information are retained for each attention head. The most important attention scores are identified using diverse top-k strategies (implemented from mask_1 to mask_4). For each row of the matrix M, only the top-k values are normalized for the softmax, thus focusing on the most critical features. Elements falling below these top-k scores are addressed using a scatter function, which sets the corresponding probabilities to 0. This dynamic selection transitions attention from dense to sparse, encapsulated by the formula:

$$\text{SparseAtt}(Q, K, V) = \text{softmax}(\text{T}_k(QK^T) > \lambda)V, \tag{4}$$

where $\text{T}_k(\cdot)$ denotes the learnable top-k selection operator, defined as:

$$[\text{T}_k(S)]_{ij} = \begin{cases} S_{ij} & \text{if } S_{ij} \in \text{top-}k \text{ of row } j, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Finally, outputs from multiple heads are concatenated and linearly projected to obtain the final output.

3. EXPERIMENTAL RESULTS AND ANALYSIS

To validate the effectiveness of our proposed network in semantic segmentation of remote sensing images, experiments were conducted on the Vaihingen and Potsdam datasets, each comprising the same six categories: impervious surfaces, buildings, low vegetation, trees, cars, and background clutter.

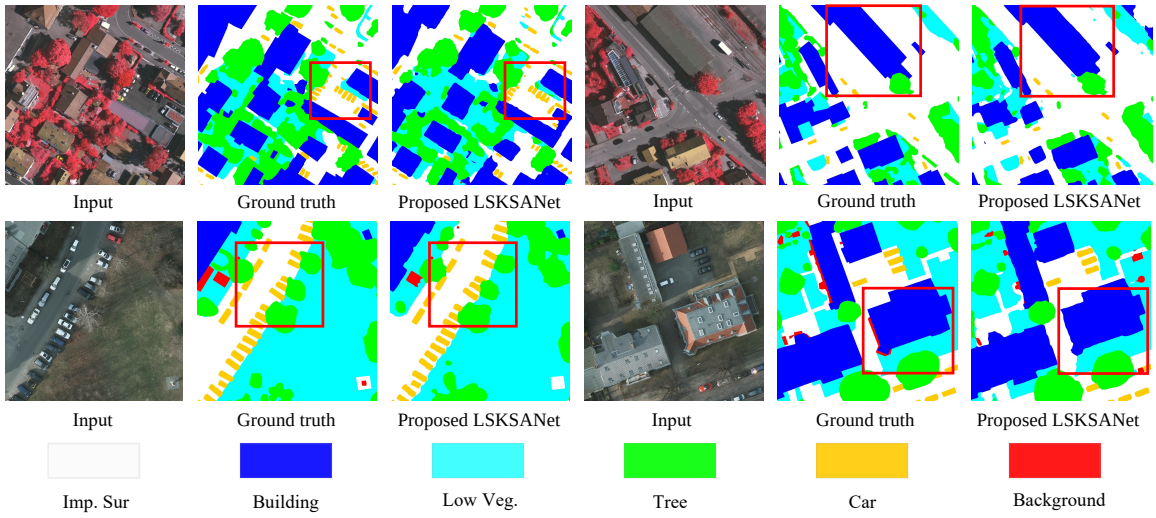

Fig. 2. Performance of LSKSANet on Vaihingen Dataset (first row) and Potsdam Dataset (second row).

Table 1. Comparison of Different Methods on Vaihingen

Method	Backbone	Imp.surf	Building	Lowveg.	Tree	Car	Params(M)	mIoU	OA	F1
SwiftNet	ResNet18	92.3	94.7	84.1	89.2	81.0	11.8	79.7	90.3	88.2
UNetFormer	ResNet18	92.7	95.3	84.9	90.1	88.5	11.8	82.2	90.8	90.2
ABCNet	ResNet18	92.6	95.1	84.5	89.8	85.3	13.4	81.4	90.6	89.5
DCSwin	Swin-S	93.6	96.2	84.6	90.0	87.6	66.9	83.0	91.6	90.4
Proposed LSKSANet	ResNet18	94.3	95.7	85.2	90.2	87.9	12.0	84.0	92.2	90.7

Table 2. Comparison of Different Methods Potsdam

Method	Params(M)	mIoU	OA	F1
SwiftNet	11.8	83.8	91.0	89.3
UNetFormer	11.8	86.8	92.6	91.3
ABCNet	13.4	86.4	92.6	91.2
DCSwin	66.9	87.1	93.0	91.6
Proposed LSKSANet	12.0	86.9	92.8	91.6

Comparative experiments were performed against four models: DCswin [4], UNetFormer [6], SwiftNet [7], and ABCNet [8], with the main metrics being mIoU, OA, and F1 scores. According to the results in Tables 1 and 2, our model exhibited superior performance on the Vaihingen test set with a lower parameter count, achieving an mIoU of 84.0%. It was also competitive on the Potsdam test set with an mIoU of 86.9%, performing on par with or slightly below the DCwin model, striking a good balance between parameters and segmentation performance.

Fig. 2 illustrates the visualization results of our model on both datasets. The first and fourth columns show the input images, the second and fifth columns represent ground truth, and the third and sixth columns display our segmentation results, showcasing LSKSANet's precision in segmenting smaller objects locally while maintaining excellent performance in edge detailing and segmentation of large objects.

4. CONCLUSION

In this paper, we propose LSKSANet for remote sensing image segmentation. The network can effectively capture extensive contextual information and focus on key features within images. LSK module decomposes the large kernel into a series of depth-wise convolutions, thereby expanding the receptive field without significantly increasing the computational burden. In addition, the sparse attention is introduced to keep the most useful self-attention values for better feature aggregation. Experimental evaluations on two datasets demonstrate the effectiveness of the proposed LSKSANet.

5. REFERENCES

[1] T. Chowdhury and M. Rahnemoonfar, "Attention-based semantic segmentation on UAV dataset for natural disaster damage assessment," In *IGARSS* 2021, pp. 2325-2328.

[2] Q. Li and Q. Zhao, "Weakly-supervised semantic segmentation of airborne LiDAR point clouds in Hong Kong urban areas," In *JURSE* 2023, pp. 1-4.

[3] K. Gao, A. Yu, X. You, et al. "Integrating multiple sources knowledge for class asymmetry domain adaptation segmentation of remote sensing images," *IEEE Trans. on Geosci. Remote Sens.*, vol. 62, pp. 1-18, 2024.

[4] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng and S. Fang, "A Novel Transformer-based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1-5, 2022.

[5] C. Zhang, W. Jiang, Y. Zhang, et al. "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. on Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.

[6] L. Wang, R. Li, C. Zhang, et al. "UNetFormer: A UNet-like Transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022.

[7] M. Oršić and S. Šegvić, "Efficient semantic segmentation with pyramidal fusion," *Pattern Recognition*, vol. 110, 2021.

[8] R. Li, S. Zheng, C. Zhang, et al. "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery," *ISPRS J Photogramm. Remote Sens.*, vol. 181, pp. 84–98, 2021.