An Improved Large Kernel-Based Remote Sensing Land Cover Segmentation Algorithm

Guohong Liu , Cong Liu , Xianyun Wu * , Yunsong Li , Xiao Zhang , Junjie Xu

Posted Date: 15 May 2024

doi: 10.20944/preprints202405.1035.v1

Keywords: Remote Sensing Images; Land Cover Segmentation; Dilated Convolution; Attention Mechanism; Large Kernel Convolution

Article

An Improved Large Kernel-Based Remote Sensing Land Cover Segmentation Algorithm

Guohong Liu [1], Cong Liu [2], Xianyun Wu [1,2,3,*], Yunsong Li [1], Xiao Zhang [3] and Junjie Xu [1]

[1] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China; ghliu_19990317@163.com (G.L.); ysli@xidian.edu.cn (Y.L); xjj2607035112@outlook.com(J.X.)
[2] Guangzhou institute of technology, Xidian University, Guangzhou 510555, China; 651031386@qq.com
[3] Hangzhou institute of technology, Xidian University, Hangzhou 311231, China; 1093834887@qq.com
* Correspondence: xywu@mail.xidian.edu.cn

Abstract: Land cover segmentation, a fundamental task within the domain of remote sensing, boasts a broad spectrum of application potential. In this article, we focus on the land cover segmentation tasks and complete the following research work: Firstly, to address the issues of uneven distribution of foreground and background and significant differences in target scales in remote sensing images, we propose a decoder called MDCFD based on multi-dilation rate convolution fusion. The decoder utilizes dilated convolution to expand the receptive field, enhancing the model's ability to capture global features and thus improving the model's ability to distinguish between foreground and background. Meanwhile, we design a multi-dilation rate convolution fusion module (MDCFM), which fuses the outputs of different dilation rate convolution layers. Secondly, aiming at the problems of diverse scenes, significant differences between categories, and many background interferences in remote sensing images, we propose a hybrid attention module called LKSHAM based on large kernel convolution. This module combines spatial attention and channel attention mechanisms and combines the two attention modules in series. By introducing large kernel convolution, the model's ability to extract contextual information is improved. At the same time, we adopt a strategy of decomposing large kernel convolution into multiple depthwise convolutions to reduce computational complexity. The improved network models designed by this paper can achieve an improvement of over 1.1% in the mIoU metric on segmentation tasks on the Potsdam and Vaihingen datasets.

Keywords: remote sensing images; land cover segmentation; dilated convolution; attention mechanism; large kernel convolution

1. Introduction

Remote sensing land cover segmentation is critical for analyzing remote sensing images, playing a key role in processing and utilizing remote sensory data. By employing image semantic segmentation algorithms, this technique assigns categories to each pixel in remote sensing data, identifying various landforms and extracting essential information. In military contexts, it provides crucial intelligence for tactical and strategic operations. Environmentally, it aids in quickly and accurately detecting ecological changes, while in urban development, it supports city planning and infrastructure enhancement. Geospatial clarity is also improved in geoscience, establishing an important base for earth studies. The wide-ranging utility of this technology underscores the need for precision in remote sensing segmentation methods.

Zheng [1] developed FarSeg, a foreground-aware relational network designed to address significant intra-class variance in background classes and the imbalance between foreground and background in remote sensing images. In 2021, Li [2] introduced Fctl, a geospatial segmentation approach based on location-aware contexts, which systematically crops and independently segments

 doi:10.20944/preprints202405.1035.v1

images, later merging these to form a cohesive output. Additionally, Ma [3] introduced FactSeg, utilizing the FA object representation to enhance detection and differentiation of smaller objects.

Recent studies on object detection highlight the critical role of foreground saliency in remote sensing image analysis. Various researchers, inspired by these findings, have been adapting Transformer architectures for remote object detection. Xu introduced the Efficient Transformer [4], based on the Swin Transformer, which exhibits improved computational efficiency and uses explicit and implicit edge-enhancement techniques for precise segmentation. Wang [5] presented a design using the Swin Transformer for context extraction from images and introduced a densely connected feature aggregation setup for resolution restoration and intricate segmentation. Xu proposed the RSSFormer [6], a remote sensing object segmentation framework with an adaptive Transformer conjunction module, attention layer, and a foreground prominence-based loss function, designed to reduce background noise and enhance foreground differentiation.

Land cover segmentation poses distinct challenges compared to standard semantic segmentation, which include: Significant variance in object sizes even within the same class, like large forests versus isolated trees. Complex background components in remote sensing images, making it difficult to classify certain elements into defined segments. Prevalence of background over foreground, which could bias the model to favor background segmentation during training, potentially affecting its optimization path.

In this paper, the prominent contributions are as follows:

- We propose a decoder called MDCFD based on multi-dilation-rate convolution fusion. A corresponding module integrates outputs from convolutions with different dilation rates, addressing dilated convolution's information loss and improving scale-variant target segmentation.
- We introduce a hybrid attention mechanism called LKSHAM, grounded on large kernel convolutions. In the spatial attention submodule, we embed a convolutional kernel selection strategy to accommodate varying segmentation scales. For channel attention, we consider a large kernel convolution-based attention to enhance the model's receptive scope, thereby improving its foreground-background distinction and suppressing unrelated background noise.

The rest of this paper is organized as follows. Section 2 introduces the principles and network structure of the fusion decoder based on multi-dilation rate convolution along with the hybrid attention module based on the large-kernel convolution selection. Section 3 describes the datasets and implementation details used. Section 4 provides discussions. Section 5 draws conclusions.

2. Methods

2.1. Multi-Dilation Rate Convolutional Fusion Module

To counter these challenges, we delineate the construction of a decoder tailored for the segmentation of objects within remote sensing imagery, dubbed the Multi-Dilation Rate Convolutional Fusion Decoder (MDCFD). Drawing inspiration from the decoders of Segformer [7] and SETR [8], alongside extant encoder-decoder frameworks, the conceived MDCFD embodies a comparatively simplistic structure of multiple Multi-Layer Perceptrons (MLPs). In a bid to augment the saliency of foreground features within the decoder's feature maps, this exposition integrates the Rssformer paradigm into the architecture of MDCFD through the adoption of dilated convolutions, with the intent to accentuate foreground details during the decoding phase. Dilated convolutions leverage enlarged kernels to span a wider sampling domain, thereby bridging distant pixels and infusing additional contextual data.

To adeptly tackle the issue of pronounced size variability among analogously classified objects within remote sensing imagery, we propose the Multi-Dilation Rate Convolutional Fusion Module (MDCFM). Said module employs dilated convolutions with minimal dilation rates for the delineation

 doi:10.20944/preprints202405.1035.v1

of fine-detail features inherent to smaller objects, while those with escalated dilation rates apprehend an extended feature spectrum. Subsequently, the MDCFM amalgamates the feature maps, processed via convolutional layers possessing divergent receptive fields across multiple convolutional trajectories, through an element-wise addition methodology. The structure of the MDCFM is as depicted in Figure 1.

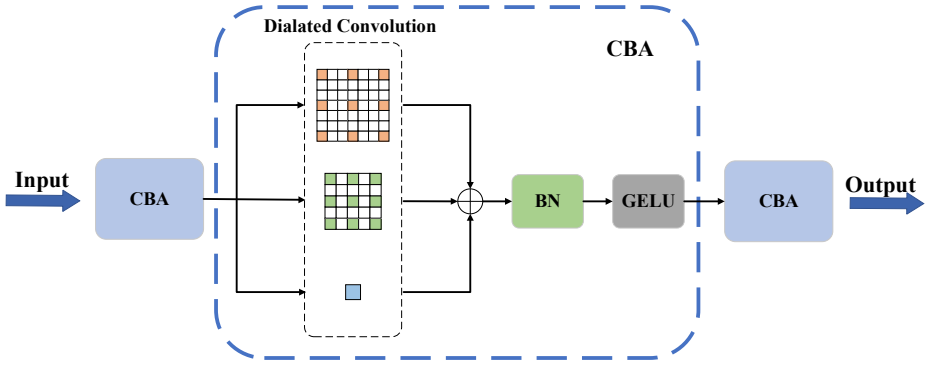

Figure 1. The structure of Multi-Dilation Rate Convolutional Fusion Module.

Input feature maps undergo sequential processing through three CBA (Convolution-BatchNorm-Activation) modules, each a composite of convolutional, batch normalization, and activation layers in a serial configuration. We incorporate the Gaussian Error Linear Unit (GELU) as the activation function within the CBA modules, with the computational progression of these modules delineated as

$$Y = GELU\left(BN\left(Conv(X)\right)\right) \tag{1}$$

where X is the input feature map and Y is the CBA output, with $BN(\cdot)$ indicating batch normalization. The CBA in the middle includes two dilated convolutions with dilations of 2 and 3, plus a convolution (dilation of 1), facilitating feature integration across varied receptive fields via element-wise addition. The module's computational progression is outlined as

$$Y_{Fusion} = Conv_{1\times1}^{R=1}(Y) + Conv_{3\times3}^{R=2}(Y) + Conv_{3\times3}^{R=3}(Y) \tag{2}$$

where Y_{Fusion} represents the output feature map produced by the fusion of multiple dilations, and $Conv_{3\times3}^{R=2}(\cdot)$ refers to the dilated convolution operation with a kernel size of 3 and a dilation rate of 2. This setup captures multilevel contextual information through different dilation rates, effectively generating enriched feature representation and enhancing the model's adaptability to complex, varied spatial structures in remote sensing images via element-wise addition operation. The construction of MDCFD is shown in Figure 2.

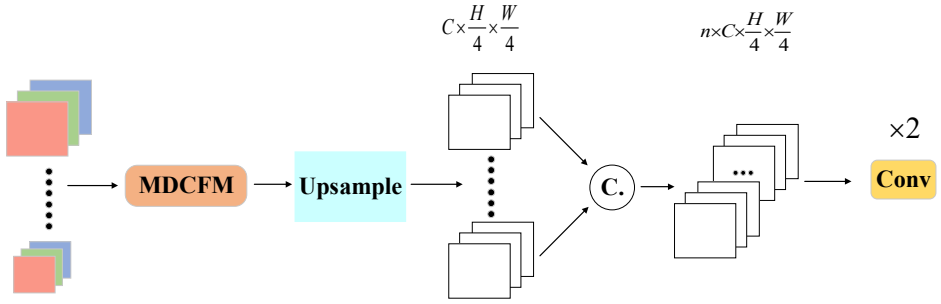

Figure 2. The configuration of Multi-Dilation Rate Convolutional Fusion Decoder.

 doi:10.20944/preprints202405.1035.v1

After MDCFM processing, the n-channel feature maps are subjected to bilinear upsampling within the upsampling module, ensuring spatial feature consistency. The interpolation increases the map resolution to a quarter of its original width and height, avoiding direct return to original resolution. This approach reduces computational and memory demands while maintaining the spatial detail needed for accurate semantic segmentation. Lastly, a channel concatenation followed by a convolutional layer with C output channels merges the feature maps to produce the decoder's output.

2.2. Large Kernel Selection Hybrid Attention Module

Inspired by the CBAM, we devise a Large Kernel Selection Hybrid Attention Module (LKSHAM) that adopts similar dual-submodule framework, as detailed in Figure 3. In this structure, the preliminary output X is passed through a channel attention module to generate an attention mask M_c (Equation (3)). This mask is then element-wise multiplied with X, resulting in a channel-enhanced feature map X_c (Equation (4)). The enhanced map X_c is then processed by the spatial attention module, generating a spatial attention mask M_S for boosting spatial focus (Equation (5)). The combination of the mask M_S and the enhanced feature map X_c returns OUT, an output denoting increased attention concentration as shown in Equation (6).

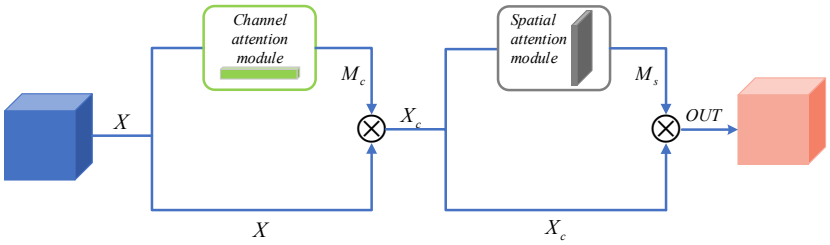

Figure 3. The configuration of Large Kernel Selection Hybrid Attention Module.

$$M_c = ChannelAttention(X) \tag{3}$$

$$X_c = M_c \cdot X \tag{4}$$

$$M_s = SpatialAttention(X_c) \tag{5}$$

$$OUT = M_s \cdot X_c \tag{6}$$

2.2.1. Large Kernel Selection Spatial Attention Module

Taking inspiration from SKNet and LSKNet [9], the present manuscript introduces a novel spatial attention module, denoted as the Large Kernel Selection Spatial Attention Module (LKSSAM), which integrates large kernel convolutions with a convolutional kernel selection mechanism.

Figure 4 delineates the procedural mechanics of the convolutional kernel selection mechanism. The output feature map X, emanating from the antecedent network layer characterized by a batch size B and a channel count N, is subjected to a triad of depthwise convolutions—operations $\widetilde{F_1}$, $\widetilde{F_2}$, and $\widetilde{F_3}$—originating from an expansive kernel convolution with a perceptive expanse of 17x17. Within this study, the dimensions of the kernel for operation $\widetilde{F_1}$ are stipulated as 3x3 with a corresponding dilation rate of 1, for $\widetilde{F_2}$ a kernel size of 5x5 with a dilation rate of 2, and for $\widetilde{F_3}$, a kernel size of 3x3 with a dilation rate of 3 is established. X traverses the aforementioned paths $\widetilde{F_1}$, $\widetilde{F_1} \rightarrow \widetilde{F_2}$, and $\widetilde{F_1} \rightarrow \widetilde{F_2} \rightarrow \widetilde{F_3}$ to procure respective outputs $\widetilde{O_1}$, $\widetilde{O_2}$, and $\widetilde{O_3}$, as expounded in the subsequent Equation.

 doi:10.20944/preprints202405.1035.v1

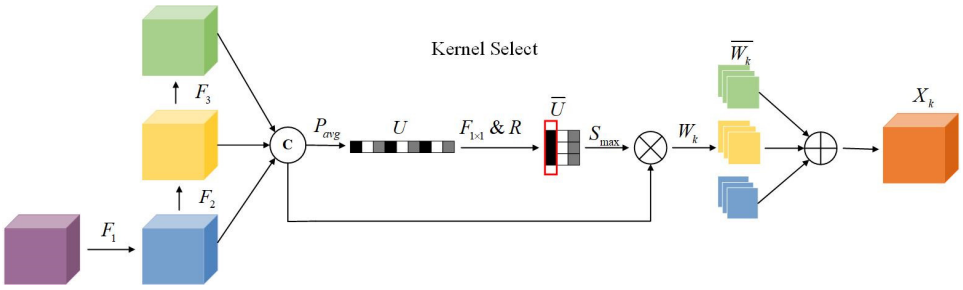

Figure 4. The configuration of Kernel Selection.

$$\widetilde{O_1} = \widetilde{F_1}(X), \widetilde{O_2} = \widetilde{F_2}(\widetilde{O_1}), \widetilde{O_3} = \widetilde{F_3}(\widetilde{O_2}) \tag{7}$$

After individual operations, $\tilde{O}_1$, $\tilde{O}_2$, and $\tilde{O}_3$ are combined through concatenation at the channel dimension (dim=1), forming O as shown in the Equation (8). O undergoes global average pooling P_{avg} to become $U \in \mathbb{R}^{B\times3N\times1\times1}$. U, after a 1x1 convolution maintaining channels at $3N$ and then expanding and reshaping, becomes the five-dimensional matrix $\overline{U} \in \mathbb{R}^{B\times3\times N\times1\times1}$, as described in the Equations (9) and (10). Applying Softmax S_{max} to $\overline{U}$'s next-to-last dimension yields the kernel selection weights matrix W_k, per Equation (11).

$$\tilde{O} = [\widetilde{O_1}; \widetilde{O_2}; \widetilde{O_3}]_{dim=1} \tag{8}$$

$$U = P_{avg}(\tilde{O}) \tag{9}$$

$$\overline{U} = reshape(F_{1\times1}(U)) \tag{10}$$

$$W_k = S_{max}^{dim=1}(\overline{U}) \tag{11}$$

$W_k \in \mathbb{R}^{B\times3\times N\times1\times1}$ contains $B \times N$ sets of convolutional kernel selection weights, where the three elements within the kernel selection weights correspond to the selection coefficients for feature maps with different receptive fields after operations $\widetilde{F_1}$, $\widetilde{F_2}$, and $\widetilde{F_3}$. After reshaping $\tilde{O}$ to the shape of $B \times 3 \times N \times H \times W$, multiplicative interaction with W_k yields $\overline{W_k} \in \mathbb{R}^{B\times3\times N\times H\times W}$, as depicted in Equation (12). Subsequently, $\overline{W_k}$ is divided along the penultimate dimension into three matrices, each with the shape of $B \times N \times H \times W$. By performing an element-wise addition operation on these three matrices, the feature map $X_k \in \mathbb{R}^{B\times N\times H\times W}$, post adaptive receptive field selection, is obtained. This process is detailed in Equation (13), where $chunk$ denotes the matrix partitioning operation.

$$\overline{W_k} = reshape(\tilde{O}) \times W_k \tag{12}$$

$$X_k = sum(\, chunk_{dim=1}^{3}(\overline{W_k})) \tag{13}$$

Figure 5 elucidates the comprehensive schema of the LKSSAM as proffered in this article. Emergent from the kernel selection module, the feature map X_k is subjected to both average pooling operation P_{avg} and maximal pooling operation P_{max}, culminating in the respective outputs $I_{avg} \in \mathbb{R}^{B\times1\times H\times W}$ and $I_{max} \in \mathbb{R}^{B\times1\times H\times W}$, explicated by Equation (14). The execution of a concatenative operation along the channel axis synthesizes I_{avg} and I_{max} into matrix $\tilde{I} \in \mathbb{R}^{B\times2\times H\times W}$, characterized in Equation (15). Subsequent to the processing of H through a convolutional maneuver I with an output channel quantum fixed at one and a 3x3 kernel dimension, succeeded by a Sigmoid function, a spatial attention mask matrix $M_s \in \mathbb{R}^{B\times1\times H\times W}$ with elemental values confined within the interval (0,1) comes to fruition, as delineated in Equation (16). The element-wise product of M_s with the antecedent input X begets the LKSSAM-augmented construct X_s, as depicted in Equation (17).

 doi:10.20944/preprints202405.1035.v1

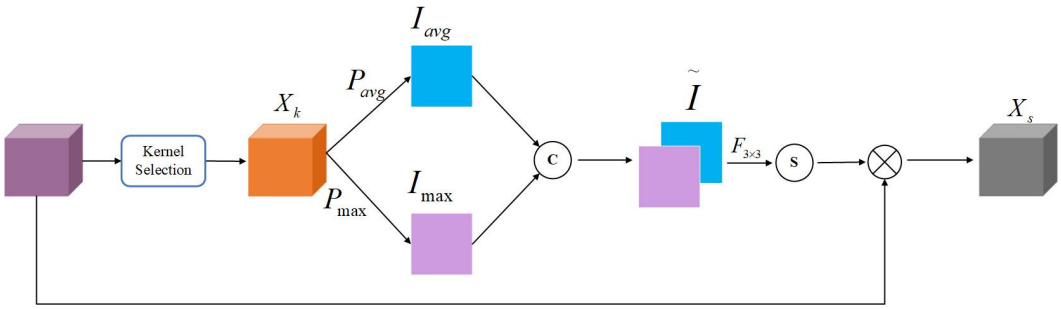

Figure 5. The configuration of Large Kernel Selection Spatial Attention Module.

$$I_{avg} = P_{avg}(X_k)$$
$$I_{max} = P_{max}(X_k) \tag{14}$$

$$\tilde{I} = [I_{avg}; I_{max}] \tag{15}$$

$$M_s = Sigmoid(F_{3\times3}(\tilde{I})) \tag{16}$$

$$X_s = M_s \cdot X \tag{17}$$

2.2.2. Large Kernel Channel Attention Module

We introduce the Large Kernel Channel Attention Module (LKCAM), predicated on substantial kernel convolutions. The integration of this module in tandem with the LKSSAM begets the LKSHAM, a hybrid attention assembly. This synthesized configuration is engineered to adeptly harness and synthesize informational vectors across both spatial and channel dimensions.

The architectural delineation of the LKCAM formulated within this paper is characterized in Figure 6. As the output feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, bearing a batch size B and channel enumeration C from the neural network's antecedent stratum, enters the module, it is initially engaged by a 1x1 convolutional exertion $F_{1\times1}$ outputting a channel number commensurate with C/R. To curtail the parameterization scale of the model, a channel reduction factor denoted as R is invoked, effectively attenuating the channel count of the input feature map to a fraction C/R, as elucidated in Equation (18).

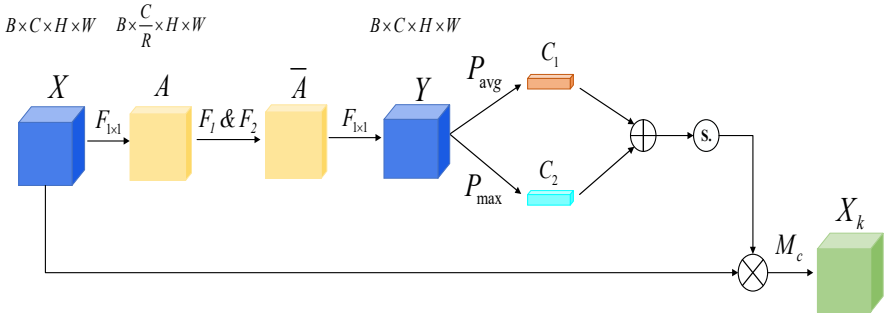

Figure 6. The configuration of Large Kernel Channel Attention Module.

$$A = F_{1\times1}(X) \tag{28}$$

Subsequent to operation $F_{1\times1}$, the convolutional output $A \in \mathbb{R}^{B\times C/R\times H\times W}$ is processed through two subsequent depthwise convolutional operations, designated as $\widetilde{F_1}$ and $\widetilde{F_2}$, encapsulated within Equation (19). Within the scope of this research, the kernel dimension for operation $\widetilde{F_1}$ is configured

 doi:10.20944/preprints202405.1035.v1

to a scale that corresponds to a dilation rate of 1, while operation $\widetilde{F_2}$ utilizes a kernel size of 7x7, paired with a dilation rate of 4. The sequential application of operations $\widetilde{F_1}$ and $\widetilde{F_2}$ is tantamount to a singular large kernel convolution with a receptive field of 29x29. The feature map $\overline{A}$, refined by the large kernel convolution, progresses through a 1x1 convolutional layer with an output channel tally equating to C. This procedure culminates in the reinstatement of the channel magnitude for the resultant feature map $Y \in \mathbb{R}^{B\times C\times H\times W}$ to C, as elucidated in Equation (20).

$$\overline{A} = \widetilde{F_1}(\widetilde{F_2}(A)) \tag{19}$$

$$Y = F_{1\times 1}(\overline{A}) \tag{20}$$

Y is concurrently processed by a global max pooling layer and a global average pooling layer, with their respective outputs being matrices $C_1 \in \mathbb{R}^{B\times C\times 1\times 1}$ and $C_2 \in \mathbb{R}^{B\times C\times 1\times 1}$, as presented in Equation (21). After an element-wise addition of C_1 and C_2, the outcome enters a Sigmoid activation layer. The output from the Sigmoid activation layer constitutes the channel attention mask M_c, this process being referenced in Equation (22). The elemental values within M_c span between 0 and 1, representing the weight values assigned by the channel attention module to $B \times C$ channels. By element-wise multiplication of the output from the previous layer of the neural network with M_c, the LKCAM-enhanced output X_c is obtained, as depicted in Equation (23). This design enables the model to discern the significance of different channels, which aids in emphasizing channel features that contribute significantly to the task at hand, while suppressing irrelevant channel information.

$$\begin{aligned} C_1 &= P_{avg}(Y) \\ C_2 &= P_{max}(Y) \end{aligned} \tag{21}$$

$$M_c = Sigmoid(C_1 + C_2) \tag{22}$$

$$X_c = M_c \cdot X \tag{23}$$

The construction of the model is shown in Figure 7.

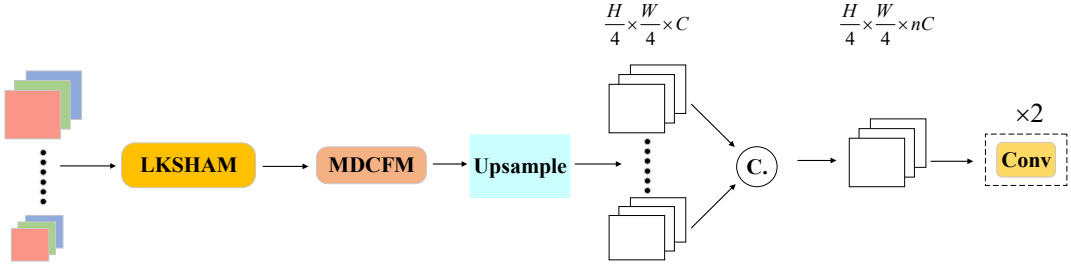

Figure 7. The structural diagram of the model improved by MDCFD and LKSHAM.

3. Results

3.1. Datasets and Data Pre-Processing

In order to ascertain the efficacy of the MDCFD decoder delineated, empirical assessments were conducted on the decoder's ability to augment segmentation precision within models, utilizing the ISPRS Potsdam and ISPRS Vaihingen datasets. This was followed by a thorough examination and dialectic of the obtained outcomes.

(1) *Potsdam Dataset:* The Potsdam dataset, recognized for advancing object segmentation and scene analysis, includes various image formats such as RGB, IRRG, and DSM. Our study selectively employed the RGB format for evaluation. Categorically, it covers six types of land cover: roads, buildings, low vegetation, trees, cars, and miscellaneous areas (commonly termed as 'clutter').

 doi:10.20944/preprints202405.1035.v1

(2) *Vaihingen Dataset:* This public dataset is key for advancing remote sensing object segmentation research and applications. It mainly includes 33 'TOP' high-resolution aerial images, with a typical resolution of 2494*2064. The Vaihingen dataset additionally provides detailed elevation data through Digital Surface Models (DSM) and Normalized Digital Surface Models (NDSM), with precise ground sampling accuracy up to 9 centimeters. Its object categories align with those in the Potsdam dataset. We exclusively utilized the 'TOP' images, without DSM and NDSM data.

(3) *Data Pre-Processing:* The high-resolution images from the Potsdam and Vaihingen datasets necessitate preprocessing through cropping to reduce memory load, hence images were resized to 512*512. We analyzed six land cover categories from Potsdam and five corresponding categories from Vaihingen, excluding clutter, to evaluate the segmentation capability of our model.

3.2. Implementation Details

3.2.1. Hardware Environment

The investigative methodology employed herein pivots around the following hardware orchestration: an Intel Core i9-9900K CPU indexed at a base frequency of 3.6GHz and structured with an aggregate of 16 threads; an NVIDIA GeForce GTX 2080Ti GPU; alongside a computational platform appointed for experimental conduction, provisioned with a memory allocation of 64 GB and storage capability of 4 TB.

3.2.2. Software Environment

The computational experiments outlined in this article employed Ubuntu 20.04 as the operating system, alongside CUDA 11.3 for the parallel computing framework. To facilitate the management of virtual environments crucial for the training and inference of models, this research makes use of the Anaconda virtual environment management system. Detailed configurations are delineated in the table below:

Table 1. The list of hardware and software environments relied on by the experiment.

Hardware/Software	Parameter/Version
CPU	Intel Core i9-9900K
GPU	GeForce GTX 2080Ti
Memory Allocation	64GB
Storage Capability	4TB
Operation System	Ubuntu 20.04
Python	3.8.2
CUDA	11.3
PyTorch	1.12.1
mmcv	2.0.0
mmsegementation	1.2.2
numpy	1.24.4
opencv	4.9.0

3.2.3. Hyperparameter Settings

During the training regimen, the Adam optimizer is selected, featuring the beta coefficients of the Adam algorithm designated as 0.9 for B_1 and 0.999 for B_2. The instructional rate for model training is established at 6e-5, while the regularization term for weight decay is anchored at 0.01. The learning rate experiences adaptive alterations in line with the Poly learning rate policy, where the Poly power is defined as 1. We stop after 160000 epochs.

 doi:10.20944/preprints202405.1035.v1

3.3. Ablation Study

To establish the MDCFD decoder and hybrid attention module's accuracy, universality, and effectiveness in remote sensing segmentation, we merged them with top-performing models for comparative evaluation against original models. The hybrid attention module was also added to MDCFD-enhanced models to determine its performance impact. In our ablation studies, we first paired the MDCFD and LKSHAM with the Segformer's MiT encoder, choosing the parameter-rich MiT-B5 encoder. Secondly, we combined them with HRNetV2 [10], using the robust HRNetV2-W48 model. The outcomes are presented in Table 2.

Table 2. The results of ablation experiments for combinations of MDCFD and LKSHAM on the Vaihingen dataset and Potsdam dataset.

Dataset	**Model**	**MIoU(%)**		**mF1(%)**	
		MiT	HRNetV2	MiT	HRNetV2
Vaihingen	Baseline	80.41	79.11	88.93	88.17
	Baseline + MDCFD	81.56	79.49	89.68	88.41
	Baseline + MDCFD +LKSHAM	82.14	80.27	90.06	88.89
Potsdam	Baseline	78.10	77.42	86.39	85.96
	Baseline + MDCFD	79.13	78.19	87.21	86.46
	Baseline + MDCFD +LKSHAM	79.63	78.80	87.56	87.07

Figure 7 graphically shows the improved segmentation accuracy of the model due to the decoder, using a set of image comparisons. The first column displays the input images for the neural network, the second column presents the labels (ground truth), as manually marked by the data compilers. The third column illustrates the predictions from the baseline model used as the control, and the fourth column shows the predicted results from the enhanced model per the architectural advancements discussed in this paper.

To assess the detection performance of the two structures suggested in this article ('Ours'), we have compared the improved Segformer and HRNetV2 models based on our proposal with conventional land cover segmentation algorithms. The results are illustrated in Tables 3

Table 3. Quantitative comparison results on the Vaihingen dataset and Potsdam dataset.

Dataset	Model	Imp. Surf.	Building	Low veg.	Tree	Car	mF1	mIoU
Vaihingen	DABNet [11]	87.8	88.8	74.3	84.9	60.2	79.2	70.2
	ERFNet	88.5	90.2	76.4	85.8	53.6	78.9	69.1
	PSPNet	89.0	93.2	81.5	87.7	43.9	79.0	68.6
	FANet [12]	90.7	93.8	82.6	88.6	71.6	85.4	75.6
	ABCNet [13]	92.7	95.2	**84.5**	89.7	85.3	89.5	81.3
	BoTNet [14]	89.9	92.1	81.8	88.7	71.3	84.8	74.3
	BANet [15]	92.2	95.2	83.8	**89.9**	86.8	89.6	81.4
	Segmenter [16]	89.8	93.0	81.2	88.9	67.6	84.1	73.6
	Deeplabv3+	90.1	93.2	82.1	88.0	84.1	87.5	78.0
	Segformer	92.0	95.5	83.3	89.2	84.6	88.9	80.4
	HRNetV2	91.0	94.4	82.8	88.8	83.8	88.2	79.1
	MiT&Ours	**92.8**	**95.7**	85.1	89.6	**87.0**	**90.1**	**82.1**
	HRNetV2&Ours	91.8	95.1	84.0	89.0	84.5	88.9	80.3
Potsdam	DeeplabV3+	92.6	96.4	86.3	87.8	95.4	85.6	77.1
	DANet [17]	88.5	92.7	78.8	85.7	73.7	77.1	65.3
	CCNet	88.3	92.5	78.8	85.7	73.9	75.9	64.3
	EMANet [18]	88.2	92.7	78.0	85.7	72.7	77.7	65.6
	Segformer	92.9	96.4	86.9	88.1	95.2	86.4	78.1

 doi:10.20944/preprints202405.1035.v1

PFNet [19]	91.5	95.9	85.4	86.3	91.1	84.8	58.6
HRNetV2	92.7	96.4	87.1	88.2	94.4	86.0	77.4
MiT&Ours	**93.3**	96.8	**87.9**	**89.3**	**96.2**	**87.6**	**79.6**
HRNetV2&Ours	93.7	**96.9**	87.6	88.8	96.2	87.1	78.8

Study results confirm the effectiveness of the two designs introduced in this paper for improving remote sensing-based land cover segmentation. Implemented designs in models from our experiments showed marked improvements in mIoU and mF1 accuracy metrics. Specifically, models using these designs in the Potsdam dataset improved mIoU by up to 1.5% and mF1 by up to 1.2%. All categories saw F1 score increases, especially 'Car' by 1.8%. Vaihingen dataset models noted mIoU rises of up to 1.7%, mF1 by up to 1.2%, with 'car' category F1 jumping 2.4%. These restructured models outperformed traditional segmentation algorithms in precision.

Figure 8 visually displays how the two architectural configurations discussed in this paper enhance model segmentation precision, as illustrated by four image set examples.

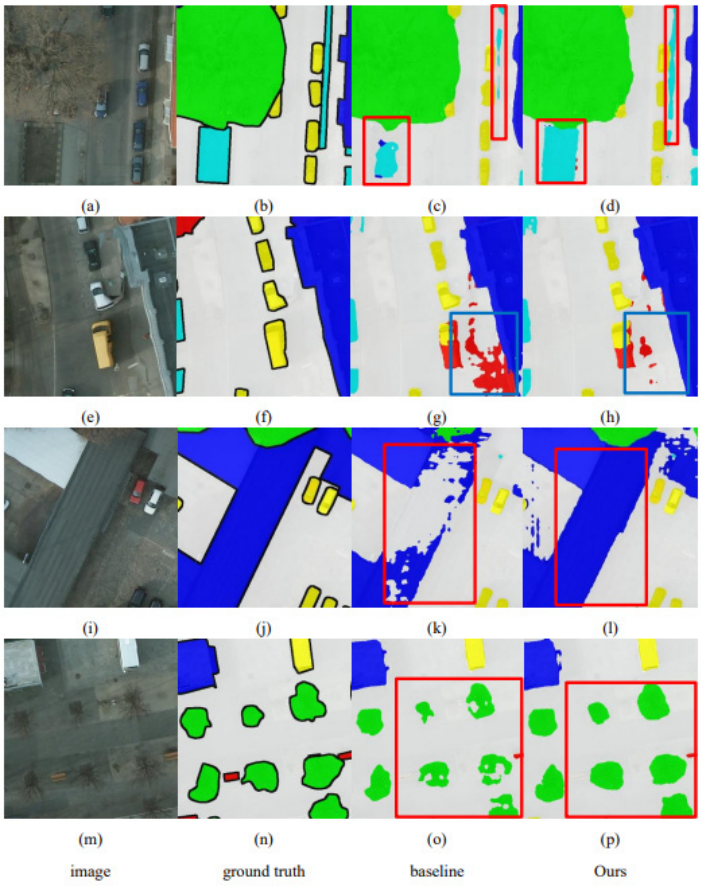

Figure 8. Comparative Visualization of Model Segmentation Efficacy Pre- and Post-Enhancement.

Figure 8c,d with red borders highlight the model's advanced land cover classification accuracy after adopting the architectural upgrades discussed. While Figure 8c shows the baseline model's misclassified pixels over a large area, Figure 8d demonstrates significant improvement, indicating precise pixel classification due to the new designs. This suggests the proposed architecture enhances land cover differentiation. The azure borders in Figure 8g depict many background misclassifications, greatly reduced in Figure 8h by the modified model, showing the designed attention modules reduce background noise. Figure 8o,p,k,l represent small and large land cover features, respectively, illustrating the model's improved segmentation across different scales, highlighting the system's adaptability to process varied target sizes efficiently.

 doi:10.20944/preprints202405.1035.v1

4. Discussion

We introduced a multi-dilation rate convolution fusion decoder and a large-core convolution-based hybrid attention module LKSHAM, aiming to boost remote sensing image segmentation accuracy. These designs aim to broaden the model's receptive field, enhancing global feature perception and foreground-background distinction. Notably, our approach considerably enhanced mIoU and mF1 segmentation metrics. On the Potsdam dataset, our model showed achieved mIoU improvements of 1.5% and 1.4% and mF1 gains of 1.2% and 1.1%. The Vaihingen dataset results corroborated this, with mIoU rises of 1.7% and 1.2%, and mF1 improvements of 1.2% and 0.8%. These achievements were made possible by our convolution kernel redesign, which not only minimized computational demands but also elevated model segmentation efficiency across varying scales through tailored convolution kernel strategies.

5. Conclusions

Our study advances remote sensing image segmentation by integrating novel structural enhancements into models, particularly a multi-dilation rate convolution fusion decoder (MDCFD) and a large-kernel based hybrid attention module (LKSHAM). These innovations have proven to enhance accuracy in differentiating complex landscape features. By addressing common challenges such as uneven foreground-background distribution and varying target scales, our work significantly improves segmentation capabilities across varied scenarios and scales depicted in remote sensing imagery. The consistent improvement in performance metrics on the Potsdam and Vaihingen datasets validates the effectiveness of our approach.

Author Contributions: Conceptualization, G.L. and X.W.; methodology, G.L. and C.L.; software, C.L. and X.Z.; validation, Y.L.; formal analysis, X.Z. and X.W.; writing—original draft preparation, J.X. and X.W.; visualization, J.X.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation (2013M540735); by the National Nature Science Foundation of China under Grant 61901388, 61301291, 61701360; by the 111 Project under Grant B08038; by the Shaanxi Provincial Science and Technology Innovation Team; by the Fundamental Research Funds for the Central Universities; by the Youth Innovation Team of Shaanxi Universities.

Data Availability Statement: Data is available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020; pp. 4096–4105.
2. Liu, W.; Li, Q.; Lin, X.; Yang, W.; He, S.; Yu, Y. Ultra-High Resolution Image Segmentation via Locality-Aware Context Fusion and Alternating Local Enhancement. *arXiv preprint arXiv:2109.02580* **2021**.
3. Ma, A.; Wang, J.; Zhong, Y.; Zheng, Z. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–16.
4. Xu, Z.; Zhang, W.; Zhang, T.; Yang, Z.; Li, J. Efficient Transformer for Remote Sensing Image Segmentation. *Remote Sensing* **2021**, *13*, 3585.
5. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5.
6. Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; Zhang, X. Rssformer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation. *IEEE Transactions on Image Processing* **2023**, *32*, 1052–1064.
7. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in neural information processing systems* **2021**, *34*, 12077–12090.

 doi:10.20944/preprints202405.1035.v1

8. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 6881–6890.
9. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; pp. 16794–16805.
10. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3349–3364.
11. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-Wise Asymmetric Bottleneck for Real-Time Semantic Segmentation. *arXiv preprint arXiv:1907.11357* **2019**.
12. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.; Saenko, K.; Sclaroff, S. Real-Time Semantic Segmentation with Fast Attention. *IEEE Robotics and Automation Letters* **2020**, *6*, 263–270.
13. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery. *ISPRS journal of photogrammetry and remote sensing* **2021**, *181*, 84–98.
14. Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 16519–16529.
15. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sensing* **2021**, *13*, 3065.
16. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 7262–7272.
17. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019; pp. 3146–3154.
18. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-Maximization Attention Networks for Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2019; pp. 9167–9176.
19. Li, X.; He, H.; Li, X.; Li, D.; Cheng, G.; Shi, J.; Weng, L.; Tong, Y.; Lin, Z. Pointflow: Flowing Semantics through Points for Aerial Image Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021; pp. 4217–4226.