DSCSNet: Dynamic Spatial Selective Model for Remote Sensing Image Semantic Segmentation

Yubo Zhang, Zhengpeng Li, Member, IANEG, Jun Hu, Bin Yang and Xiaolin Zhang *

Abstract—**In high-resolution complex remote sensing images, the task of semantic segmentation is crucial as it effectively extracts and classifies information from different regions of the image, thereby supporting remote sensing tasks. Fully convolutional networks (FCN) and Transformers dominate the mainstream model structures for local and global feature extraction in remote sensing images. However, FCN primarily rely on convolutional operations, excelling at local feature extraction but struggling to effectively capture global features. Transformers, on the other hand, focus more on global contextual features but are less effective in extracting fine local details, making it challenging to capture subtle nuances like convolutional networks. Inspired by the Swin Transformer, we pro- pose an innovative remote sensing semantic segmentation model, DSCSNet, which can effectively handle large-scale remote sensing images to simultaneously capture both local and global information. We improved the first layer of the Swin Transformer block and named it the contextual feature extraction swin-transformer block (CFESwin-Transformer block). In the decoder structure, we improved and designed two modules. The first is the dynamic space selection module (DSSM), used in the skip connection phase to fuse global and local feature maps at different scales. DSSM employs largescale convolutional kernels to effectively integrate multi-scale information, enhancing the accuracy of feature recognition in remote sensing images. Finally, we propose a global channel split block (GCSB) to enhance the correlation capabilities among multiple channels. We conducted extensive experiments and ablation studies on three challenging remote sensing semantic segmentation datasets: Vaihingen, LoveDA, and WHDLD. The results show that our model achieves superior segmentation metrics compared to previous methods, delivering outstanding performance.**

Index Terms—**Remote sensing images, semantic segmentation, Swin Transformer, global features, local features**

Manuscript received Oct 13, 2024; revised Mar 3, 2025.

The research work was supported by Fundamental Research Project of Higher Education Institutions by Liaoning Provincial Department of Education (LJ222410146057, LJ212410146040, LJ212410146006), and Graduate Education Reform and Scientific and Technological Innovation and Entrepreneurship Project of University of Science and Technology Liaoning (LKDYC202403).

Yubo Zhang and Zhengpeng Li contributed equally to this work.

Yubo Zhang is a postgraduate student of the University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 1169962657@qq.com).

Zhengpeng Li is a doctoral student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: lkdlzp0901@163.com).

Jun Hu is a professor of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 320083700074@ustl.edu.cn)

Bin Yang is an associate professor of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 320193700107@qq.com).

Xiaolin Zhang is an associate professor of University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: zhangxiaolin@ustl.edu.cn).

I. Introduction

With the continuous development of remote sensing and drone technology, it has become increasingly easy to acquire rich high-resolution remote sensing images. These images play a vital role in various fields, such as land cover change monitoring[1], urban planning[2, 3], disaster prediction[4, 5], and traffic management[6]. The primary goal is to accurately segment the images into independent regions of different semantic categories, making remote sensing data easier to analyze and understand[7]. Therefore, developing efficient and reliable methods is essential for extracting valuable information from remote sensing images.

With the rapid development of deep learning, FCN[8] replaced the fully connected layers[9] in traditional CNNs with convolutional layers, typically adopting an encoder-decoder framework featuring skip connections. This significantly improves the accuracy of semantic segmentation. Among existing models, U-Net[10] effectively retains more detail and semantic information by utilizing skip connections to transfer low-level features to higher levels, enhancing the capability for local feature extraction. This feature has led to the emergence of numerous U-Net-based variant models. Additionally, deep learning models based on Transformers[11, 12] have been widely applied to remote sensing semantic segmentation tasks[13]. The Transformer architecture is highly regarded for its efficient self-attention mechanism, which excels in capturing global contextual information and spatial dependencies. However, for high-resolution remote sensing image semantic segmentation tasks, single global feature extraction has not achieved the desired results, as it struggles to accurately capture fine details, affecting segmentation performance and potentially causing the loss of contextual information. Building on this, the innovation of the Swin Transformer[14] with its introduction of a sliding window mechanism, has successfully enhanced the ability to capture image features at different scales, achieving notable success in semantic segmentation.

Based on the aforementioned challenges, this paper proposes a semantic segmentation[15] model for remote sensing images, which combines a U-Net and Transformer-based architecture. To improve the accuracy of semantic segmentation, we adopted an encoder-decoder architecture, using the Swin Transformer as the primary encoder. In the first layer of the encoder, we introduced a new Swin Transformer block referred to as the CFESwin-Transformer block, which enables finer sliding windows. Additionally, we achieved feature integration and information transfer between the encoder and decoder through the use of skip

connections. In the decoder's upsampling stage, we designed the DSSM to adaptively process features at different scales, using our invented dynamic space selection mechanism to filter out the most suitable spatial features for use. Finally, in the output stage of the decoder, we incorporated our innovative GCSB to enhance the effectiveness of the output masks. Our main contributions are as follows:

1) During the downsampling process, the Swin Transformer suffers from a loss of small-scale feature map information. To address this issue, we propose an effective solution: the CFESwin-Transformer. This module improves the feature representation of each window by adaptively learning the content of each window and compressing finer window features. By implementing a channel attention mechanism, we achieve effective processing of local and global features, allowing each window's content to be more refined and enabling the network to focus more on key features.

2) We developed a DSSM that enhances feature representation by integrating global and local features and performing feature fusion. The DSSM performs adaptive weighted fusion of multi-scale and multi-source feature maps, thereby strengthening the feature representation capability. This module employs a dynamic space selection mechanism to extract and optimize the correlation between specific channels during the deep convolution process[16], efficiently captures semantic relationships across channels. Additionally, the DSSM can dynamically adjust features based on varying scenarios, specific task requirements, and inputs, providing a more adaptive feature representation to address a wide range of problems.

3) We designed a GCSB that captures multi-level features spanning local-global and channel-spatial domains, enabling fine-grained feature separation to enhance representation capability. This module utilizes the MFGU[17] mechanism to effectively reduce sensitivity to local noise, enhancing the robustness of the features. The efficient attention computation method allows the model to process high-resolution images while still maintaining a low computational cost.

4) We conducted extensive experimental comparisons on three challenging public datasets, and the results indicate that our method performs exceptionally well in processing high-resolution remote sensing images. The outstanding experimental results demonstrate the accuracy and superiority of the model.

II. Related Work

In remote sensing image processing, the semantic segmentation task involves pixel-level classification of images obtained from drone imagery to accurately identify and delineate different classes within the image. This includes assigning each pixel in the image to a predefined category, such as roads, vegetation, buildings, water bodies, etc., thereby presenting different semantic segmentation information to users.

A. Application of Models Based on Local Feature Extraction in Computer Vision Tasks

Local feature extraction understands the content of images through information from local regions, discovering meaningful details such as edges and textures within small areas. AlexNet[18] pioneered an end-to-end FCN based on CNNs, which is applied to remote sensing semantic segmentation. FCN allows for end-to-end training from input images to output segmentation results, reducing the complexity of intermediate steps. By using convolutional operations, it preserves the spatial structure of the input image, avoiding spatial information loss caused by downsampling and enhancing semantic segmentation capabilities. However, FCNs can suffer from blurred edges and loss of detail due to overly simplistic upsampling operations, leading to inaccuracies in semantic segmentation. To achieve better results, several methods based on CNNs have been improved. DeeplabV3+[19] introduces dilated convolutions to expand the receptive field while enhancing the encoder-decoder structure, resulting in more accurate boundary preservation during detail recovery. HRNet[20] maintains high-resolution features and integrates multi-scale information, improving the handling of details and global context compared to FCNs, thus avoiding information loss issues in fully convolutional networks. PANet[21] strengthens feature fusion through the path aggregation module (PAM), enhancing the capability of semantic segmentation. OCRNet[22] enhances the contextual information of each pixel through a context representation module, improving the accuracy of semantic segmentation. PSPNet[23] effectively merges multi-scale contextual information by combining the powerful feature extraction capabilities of CNNs with a pyramid pooling module, thereby enhancing segmentation accuracy. GCNet[24] introduces a global context module to capture global information from images, improving segmentation performance through contextual modeling. DenseNet[25] significantly improves gradient propagation and feature reuse by using dense connections that allow each layer to receive feature maps from all preceding layers. MobileNetV3[26], as a lightweight convolutional network, combines depthwise separable convolutions with structural optimizations for lightweight design. Xception[27] replaces traditional convolutions with depthwise separable convolutions, demonstrating outstanding performance, especially in processing large-scale images. These CNN models optimize feature extraction and computational efficiency through various innovative approaches, showcasing the flexibility and strong adaptability of CNNs in image processing tasks.

B. Application of Models Based on Global Feature Extraction in Computer Vision Tasks

Global feature extraction analyzes the overall contextual information of images to capture long-range dependencies and global relationships. Unlike local feature extraction, global feature extraction covers a broader range, making it suitable for full-image patterns and overall structures. The Transformer architecture, initially applied mainly in natural language processing (NLP), has proven to be a groundbreaking neural network architecture and has successfully applied to the fields of computer vision and remote sensing image processing. The role of the Transformer is to capture the relationships between different parts of the input image, establishing dependencies

regardless of distance. The core component of the Transformer architecture is the attention mechanism, which allows the model to dynamically assign weights to each element in the input sequence, maintaining computational efficiency while capturing long-range dependencies. Transformers are crucial in various computer vision tasks such as image segmentation, image classification, object detection, video understanding, and time-series data processing. ViT is an innovative model that introduces the Transformer framework into image tasks by dividing images into fixed-size patches and using these patches as inputs, applying the self-attention mechanism of the Transformer to process the images. However, both approaches have high complexity when handling high-resolution images and require a large amount of training data to fully realize their potential. To address these complex issues, several new methods have been proposed in recent years. Data-efficient image transformer (DeiT)[28] introduces distillation techniques to reduce reliance on large-scale datasets, making it less demanding than ViT. Non-local Neural Networks[29] capture dependencies between all pixels in an image through a global attention mechanism. Attention U-Net[30] can dynamically adjust the weights of feature maps, allowing the model to focus more on salient regions and important information when processing complex scenes. Residual attention network (RAM)[31] dynamically adjusts weighted features by stacking residual and attention modules, allowing the model to focus on key content in complex tasks. Dual attention network (DANet)[32] employs a dual mechanism of channel attention and feature attention to concentrate on extracting multi-scale features. Efficient channel attention (ECA-Net)[33] improves upon SE-Net[34] by reducing computational overhead, making it suitable for lightweight networks.

C. Application of Combined Global-Local Feature Extraction in Computer Vision Tasks

In recent years, models that combine convolutional neural networks (CNNs) with Transformers have shown significant advantages. This combination leverages CNNs' strengths in extracting local features while using Transformers' attention mechanisms to handle global information. Global-local feature extraction integrates the benefits of both global and local feature extraction. Convolutional vision transformer (ConViT) [35] utilizes convolutions for local feature extraction and introduces self-attention mechanisms to enhance performance in visual tasks. Segmenter with transformer (SETR)[36] first extracts feature maps using CNNs for semantic segmentation tasks, then employs Transformers for global context modeling, improving segmentation accuracy. LeViT [37] is a CNN-Transformer hybrid model that enhances image classification accuracy while maintaining lower computational costs. Transformer in transformer (TNT) [38] effectively extracts local details and global semantics from images, primarily used for image classification and other visual tasks. Detection transformer (DETR)[39] features an innovative end-to-end detection mechanism for object detection tasks. The Swin Transformer introduces a local sliding window mechanism, integrating CNNs' local receptive fields into the Transformer architecture, allowing for more efficient processing of high-resolution images.

Thus, our encoder design is primarily based on the Swin Transformer, with innovations in the first layer of the Swin Transformer to better capture global-local features. We also implemented innovative modules and a Dynamic Space Selection Module in the decoder to handle upsampled feature information. Through this approach, we designed an efficient semantic segmentation network model for remote sensing images.

III. Research Method

In this section, we provide a detailed overview of the DSCSNet model architecture. Section 3A presents a comprehensive analysis of the model, focusing on the design principles and functional features of the CFESwin-Transformer block. Section 3B delves into the operating mechanisms of the GLB block and DSSM module, while Section 3C highlights the detailed design and implementation of the FHR and GCSB blocks within the decoder. These topics will be elaborated upon in the subsequent chapters.

A. Network Model and CFESwin Transformer Block

The overall architecture is shown in Fig. 1 This model enhances and optimizes both the encoder and decoder, effectively combining them through skip connections, demonstrating exceptional performance in remote sensing semantic segmentation tasks. In the encoder part, we improved the structure of the original Swin Transformer block's first layer, creating a new Swin Transformer module called the CFESwin-Transformer module. The traditional Swin Transformer block primarily consists of several basic elements: Layer normalization (LN), Multi-Layer perceptron (MLP), window multi-Head self-attention (W-MSA), and Cross-window self-attention (SW-MSA)[14]. However, the window boundaries may create discontinuities, leading to inconsistent contour lines, especially for important features located at the window edges. Another issue is that SW-MSA is highly sensitive to window size and sliding stride, as different tasks may require different window configurations. If not set correctly, this can hinder optimal performance and even degrade the model's effectiveness. The CFESwin-Transformer module incorporates global pooling, normalization layers, and ReLU functions to help the model finely capture global context information：

$$Q_C = F_{GAP} = \frac{1}{\mathrm{H}\times\mathrm{W}}\sum_{\mathrm{i}=1}^{\mathrm{H}}\sum_{\mathrm{j}=1}^{\mathrm{W}} P_{\mathrm{c}}(\mathrm{i},\mathrm{j}) \tag{1}$$

Here, H and W denote the height and width, P_{c} represents the two-dimensional matrix generated under feature details, F_{GAP} is the process of global average pooling, Q_C is the matrix after global average pooling.

$$T_{\mathrm{c}} = F_{FRFS}(Q_{\mathrm{c}}, D) = \sigma(D_2\delta(D_1 Q_{\mathrm{c}})) \tag{2}$$

$D_1 Q_{\mathrm{c}}$ represents the first fully connected operation, $\sigma(\bullet)$ is represented by the Sigmoid function, $\delta(\bullet)$ represents the ReLU function, D , D_1 , D_2 is a different downsampling layer, F_{FRFS} is the name of the function processed by the

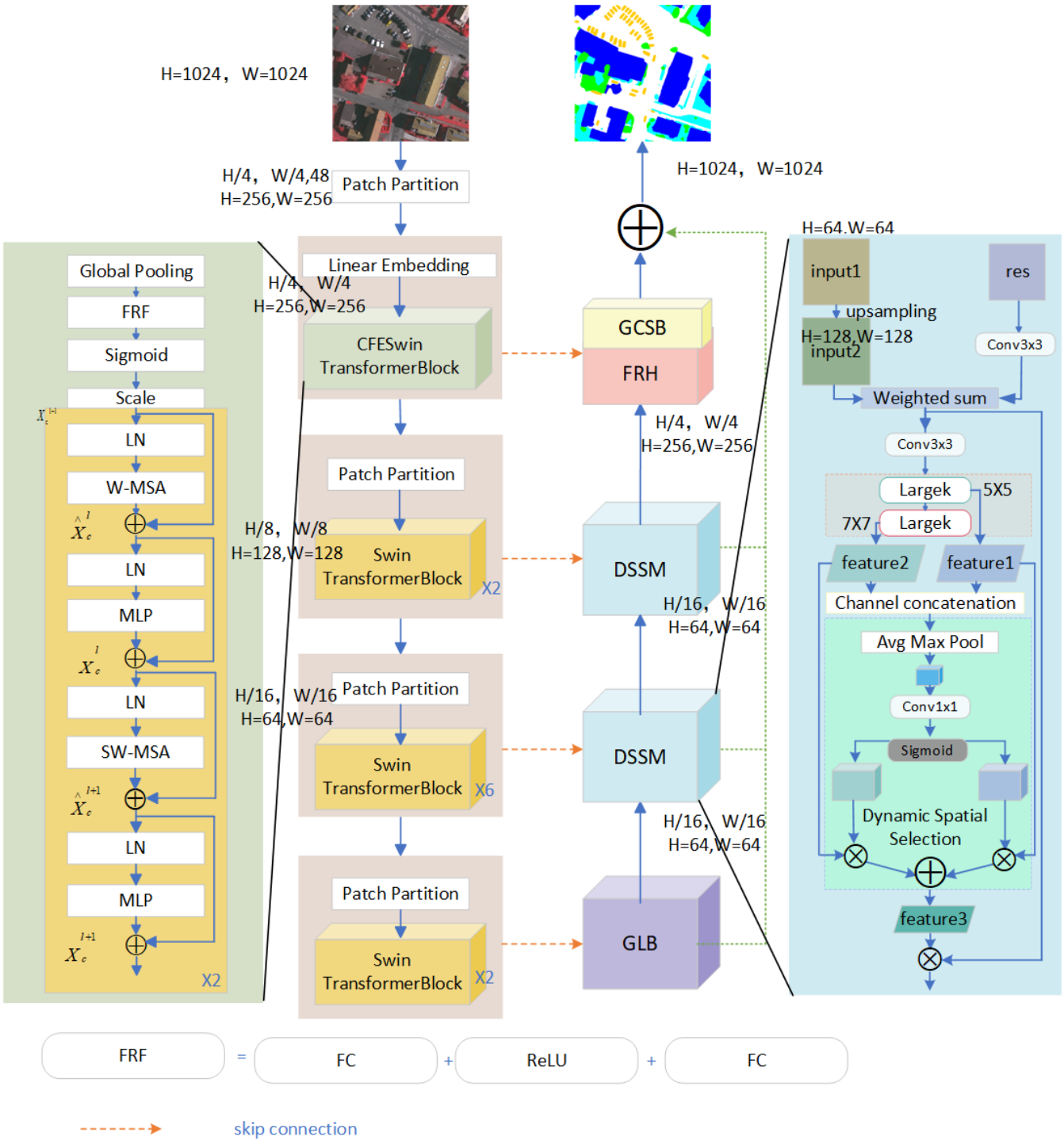

Fig. 1. DSCSNet Overall Architecture

normalization layer, ReLU function, and Sigmoid function, T_c is the weight for the corresponding channel:

$$X_c = F_{scale}(P_c, T_c) \tag{3}$$

In Equation (3), the normalized weights T_c after processing are assigned to P_c. Theoretically, this module reduces the high computational cost of the Swin Transformer and improves clarity in windowed images by decreasing the computational complexity of the first layer. The encoder incorporating the CFESwin-Transformer module adopts a feature construction strategy from local to global, further enhancing the multi-scale feature extraction capability and effectively improving feature fusion, making it suitable for various remote sensing semantic segmentation tasks. The calculation formula is as follow:

$$\begin{aligned}
\hat{X}_c^l &= W - MSA(\ln(\hat{X}_c^{l-1})) + \hat{X}_c^{l-1} \\
X_c^l &= MLP(\ln(\hat{X}_c^l)) + \hat{X}_c^l \\
\hat{X}_c^{l+1} &= SW - MSA(\ln(X_c^l)) + X_c^l \\
X_c^{l+1} &= MLP(\ln(\hat{X}_c^{l+1})) + \hat{X}_c^{l+1}
\end{aligned} \tag{4}$$

Here, $\hat{X}_c^l$ represents the output after W-MSA. X_c^l and X_c^{l+1} represent the outputs after passing through the MLP, $\hat{X}_c^{l+1}$ is the result after passing through the SW-MSA, however, since the W-MSA operates within non-

overlapping windows, the new decoder consists of three modules: We improved the global local transformer (GLTB)[40], and renamed it as the global local block (GLB), the DSSM, and GCSB. These modules work together to effectively process multi-scale features while preserving details, significantly enhancing the overall performance of the model.

B. Backbone GLB and Dynamic Space Selection Module

We utilized the GLB, which comprises various operations, including global-local attention The mechanism of global context is crucial for complex scenes in remote sensing images, the focus of local information lies in preserving diverse spatial content. Therefore, the global-local attention mechanism employs parallel branches to extract global and local contexts separately. The global branch typically captures global contextual information by segmenting the image through windowing operations. The local branch utilizes two parallel convolutional layers of different sizes to obtain local contextual information, with kernel sizes of 1 and 3, respectively followed by two batch normalization operations[40]. Finally, the global and local contexts are summed and fused together. To achieve dynamic spatial selection for long-range contextual feature extraction, we first perform a weighted summation of the upsampled image and the residual image generated from the skip connection to produce the feature tensor F, and then combine the result with the features from a large convolutional kernel：

$$DR_{i-1} < DR \le RF_{i-1}, \quad DR_1 = 1, \quad k_{i-1} \le k_i \tag{5}$$

$$RF_i = DR_i(k_i - 1) + RF_{i-1}, RF_1 = k_1 \tag{6}$$

As can be seen, the definitions of the kernel size k, dilation rate DR, and receptive field RF for the i-th depth convolution are as follows. After integrating the detailed information from the image and the residual image, we selected two convolution kernels of sizes 5×5 and 7×7. These kernels are fused with the features obtained from the weighted summation to establish a dynamic spatial selection mechanism. This allows us to dynamically select feature maps from large convolution kernels of multiple scales based on different contextual needs, thereby better extracting features of diverse levels. This mechanism strengthens the perceptual ability of the spatial context area, enabling the network to adaptively choose the most relevant features across multiple scales:

$$\hat{f} = \{\hat{f}_1, \ldots, \hat{f}_i\} \tag{7}$$

Next, we use max pooling and average pooling to select spatial relationships：

$$DSSF = F_{MAP}(\hat{f}) \tag{8}$$

Here, F_{MAP} represents the dynamic spatial masks obtained from max pooling and average pooling, while $DSSF$ is the feature descriptor generated after max pooling and average pooling. Next, the Sigmoid function is applied:

$$SSF = \sigma(DSSF) \tag{9}$$

The features generated from the convolution are dynamically processed through a corresponding spatial selection mechanism, and convolution operations are continued to obtain the attention features SS：

$$SS = F_{conv1\times1}(\sum_{i=1}^{N} SSF, \hat{f}) \tag{10}$$

Finally, the input features F are element-wise multiplied with SS to obtain the result：

$$E = F \bullet SS \tag{11}$$

C. FRH and Global Channel Split Block

UNetFormer introduces the feature refinement head (FRH) Fig. 2 to reduce the semantic gap between different features, thereby further improving accuracy. First, two features are weighted and summed to fully leverage precise semantic segmentation and spatial details. Provide the synthesized features as input to the FRH. Next, two paths are constructed to enhance feature representation in both the channel and spatial dimensions. In the channel path, global average pooling is used to generate a channel attention map, followed by reduction and expansion operations. Two 1×1 convolution layers are used to first reduce the channel dimensions to a quarter of their original size, then expand them back to their original size. In the spatial path, depthwise convolution is used to generate the spatial attention map. The attention features from both paths are then fused through summation.

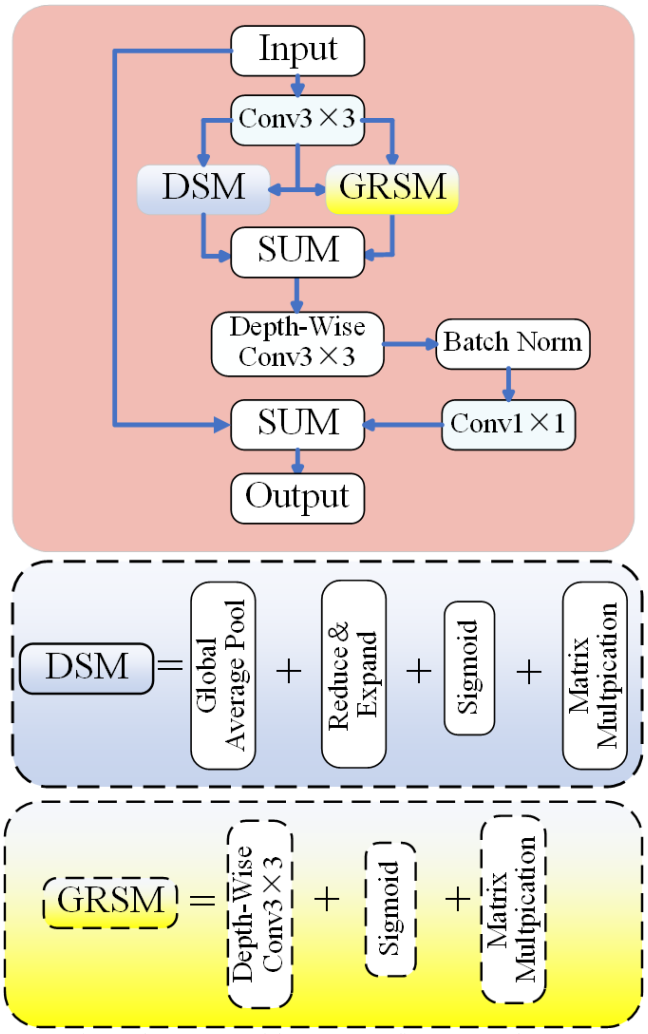

Fig. 2. The structure of the Feature Refinement Head (FRH).

At the same time, we designed a novel GCSB block, emphasizing a lightweight design concept. Its purpose is to

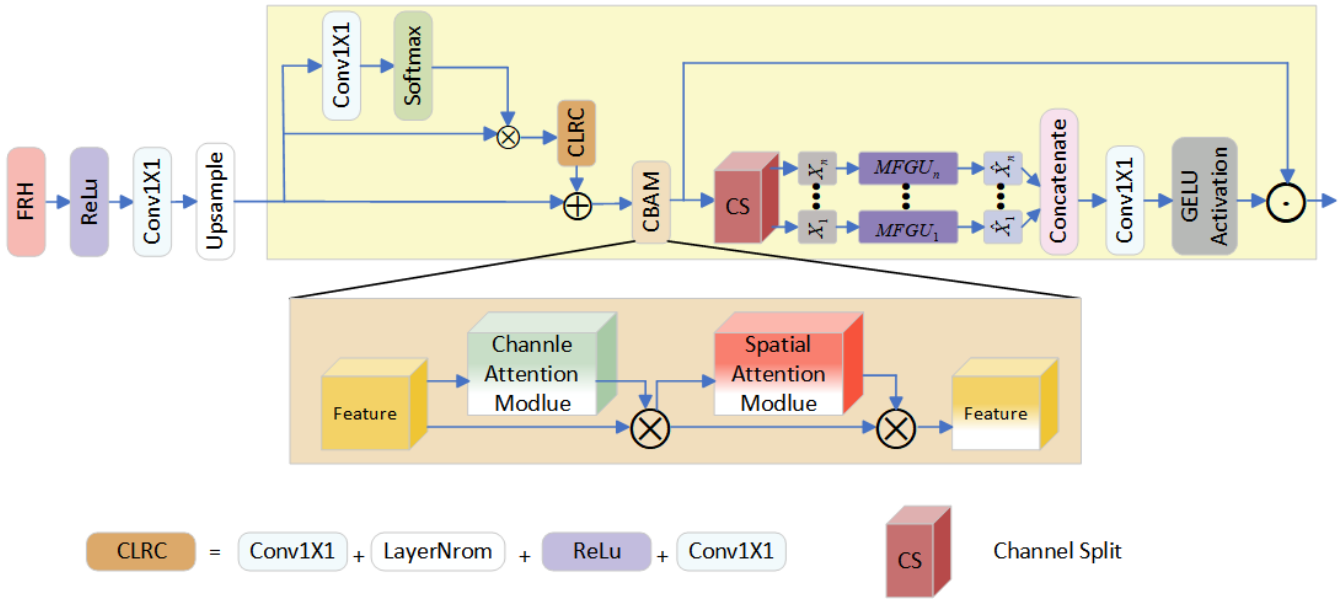

Fig. 3. Refinement of the GCSB module.

reduce the number of parameters and complexity while achieving efficient feature learning and effectively constructing global contextual information. Connected to the FRH module, the GCSB block integrates the advantages of non-local blocks and SE blocks in lightweight computation, effectively capturing long-range dependencies and contextual information between channels. It can also adjust the weights of different channels to enhance key features. The GCSB block performs particularly well when handling details in large-scale remote sensing images. The GCSB module incorporates the concept of the Multi-Scale Feature Generation Unit (MFGU)[17], further enhancing the expressive power of convolutions. Through the multi-scale feature generation unit, this module can capture both local and global features within different receptive fields, strengthening the model's ability to process complex images. This module not only maintains a lightweight design but also significantly improves the model's performance in remote sensing semantic segmentation tasks. The structure of this module is shown in Fig 3. The global attention pooling CLRC mechanism utilizes convolution and ReLU functions to capture long-range image features. It reduces redundancy in the global context features and minimizes errors through bottleneck transformation:

$$z=\sum_{j=1}^{N}\frac{e^{V_k x_j}}{\sum_{m=1}^{N} e^{V_k x_m}}x_j \tag{12}$$

$$Y_i = x + V_{w2}\delta(LN(V_{w2}z_i)) \tag{13}$$

In this process, V_k represents the weight matrix, N is the total number of positions in the feature map, and $e^{V_k x_m}$ calculates the attention weight for each position. In the GCSB, we incorporated the capabilities of the CBAM module. The features processed by CLRC are passed to the CBAM module, where the Channel Attention Module keeps the spatial dimensions of the features unchanged while compressing the channel dimensions to emphasize the target location information[41]. The features are first processed through average pooling and max pooling, then passed through the MLP, which reduces the number of channels to 1/R (where R is the reduction rate) and then expands it back to the original number of channels[41]. Finally, the output is generated through the Sigmoid function:

$$M_C(F)=\sigma(MLP(F_{AP}(F))+MLP(F_{MP}(F))) \tag{14}$$

The output feature is then passed into the Spatial Attention Module. First, max pooling and average pooling are applied to obtain a $1\times H\times W$ feature map. This is followed by a 7×7 convolution, and finally, the Sigmoid function is applied, resulting in a feature map of size $C\times H\times W$:

$$M_S(F)=\sigma(F_{conv7\times7}(F_{AP}(F);F_{MP}(F))) \tag{15}$$

After passing through the CBAM module, a channel splitting operation is performed, dividing the feature map into n smaller features. The channel splitting method is represented as CS:

$$CS(X)=[X_1,\ldots,X_n] \tag{16}$$

Input them into the MFGU, where X_1 undergoes 3×3 depthwise convolution, while the remaining i-1 parts perform pooling operations:

$$\hat{X}_1 = F^{DW}_{conv3\times3}(X_1) \tag{17}$$

$$\hat{X}_n = F_{Pool}(F^{DW}_{conv3\times3}(X_n)), 1<n\le i \tag{18}$$

Then, we concatenate these features and combine them using a 1×1 convolution:

$$X = F_{conv1\times1}(F_{concat}([\hat{X}_1,\ldots,\hat{X}_n]) \tag{19}$$

Finally, we apply the GELU non-linear function for normalization on the features, estimate the attention map, and then adaptively adjust the output $\bar{X}$ through element-wise multiplication based on the estimated attention:

$$\bar{X} = F_{GELU}(\hat{X})\odot X \tag{20}$$

IV. EXPERIMENT

This section is organized into three parts. The first part introduces the dataset employed in the experiments, the second outlines the baseline models used for comparison, and the third details the evaluation metrics.

A. Datasets

The Vaihingen dataset consists of 33 images, each with an average size of 2500×2064 pixels. Each orthophoto includes three bands: near-infrared, red, and green, along with a normalized digital surface model (DSM) and a normalized digital surface model (NDSM). The dataset includes labels such as buildings, trees, low vegetation, vehicles, various water surfaces, and background clutter. In the experiments, we selected images with IDs: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 35, and 38 for testing, with image ID: 30 used for validation, while the remaining 15 images were used for training[42].

The WHDLD dataset consists of high-resolution images captured by drones, comprising a total of 4,940 images with extremely high spatial resolution. 80% of the images were randomly selected as the training set, 10% as the validation set, and the remaining 10% as the test set. The dataset includes six label categories: buildings, roads, bare soil, sidewalks, vegetation, and water bodies.

TABLE 1
THE RESULTS OF MODULE ABLATION ON THE VAIHINGEN DATASET

CFES	DSSM	GCSB	OA(%)	MF1 (%)	mIoU(%)
√			93.74	91.60	84.85
	√		93.69	91.54	84.76
		√	93.71	91.19	84.15
√	√		93.62	91.42	84.56
	√	√	93.73	81.46	84.65
√		√	93.70	91.32	84.41
√	√	√	**93.76**	**91.67**	**84.98**

TABLE 2
THE RESULTS OF MODULE ABLATION ON THE WHDLD DATASET

CFES	DSSM	GCSB	OA(%)	MF1 (%)	mIoU(%)
√			87.82	72.70	58.46
	√		87.41	72.96	58.72
		√	87.68	72.54	58.31
√	√		87.81	72.79	58.55
	√	√	88.06	73.34	59.14
√		√	87.91	72.83	58.62
√	√	√	**88.14**	**73.62**	**59.50**

TABLE 3
THE RESULTS OF MODULE ABLATION ON THE LOVEDA DATASET

CFES	DSSM	GCSB	mIoU(%)
			53.03
√			53.44
	√		53.19
		√	53.19
√	√		53.41
	√	√	53.20
√		√	53.27
√	√	√	**53.58**

B. Training Setup

In this experiment, we used a computer equipped with an NVIDIA RTX 3090 GPU and implemented the experiments using the PyTorch framework. The Adam optimizer was employed to optimize the model, with a learning rate of insert learning rate. We set the batch size to 8 and the number of epochs to 100. The evaluation metrics for the experimental results included overall accuracy (OA), Mean F1 Score, and mean intersection over union (mIoU). We primarily utilized a combination of the cross-entropy loss function and the Dice loss function for supervised learning.

The LoveDA dataset contains 5,987 high-resolution remote sensing images, each with a resolution of 0.3 meters and a size of 1024×1024 pixels. The purpose of this dataset is to address the semantic segmentation problem of remote sensing images in different scenes (urban and rural). There are seven label categories, including buildings, roads, water bodies, wasteland, forests, farmland, and background. We divided the dataset as follows: the training set contains 2,522 images, the validation set includes 1,669 images, and the test set consists of 1,796 images. Due to the dataset's complex backgrounds, multi-scale objects, and uneven class distribution, it presents significant challenges for research.

The loss function formula is as follows:

$$L = 0.5L_C + 0.5L_D \tag{21}$$

The formula for the evaluation complexity is as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{22}$$

The formula for overall accuracy (OA) is as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

The formula for the average F1 score is as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{24}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

TABLE 4
ABLATION STUDY RESULTS OF THE CFESWIN TRANSFORMER BLOCK ACROSS DIFFERENT ENCODER LAYERS ON THE VAIHINGEN DATASET

1	2	3	4	OA(%)	MF1 (%)	mIoU(%)
	√			93.61	91.52	84.73
		√		**93.81**	91.62	84.88
			√	93.62	91.42	84.57
√				93.76	**91.67**	**84.98**

TABLE 5
ABLATION STUDY RESULTS OF DSSM ON DIFFERENT DECODER LAYERS OF THE VAIHINGEN DATASET

1	2	OA(%)	MF1 (%)	mIoU(%)
√		93.80	91.55	84.77
	√	93.61	91.62	84.29
√	√	**93.76**	**91.67**	**84.98**

C. Ablation study

To validate the effectiveness of each innovative module in our network model, we conducted ablation studies on the Vaihingen and WHDLD and LoveDA datasets, as shown in Tables 1, 2and 3. Our main innovative modules include the CFESwin Transformer Block, the DSSM, and the GCSB.

TABLE 6
MODEL COMPARISON ON THE VAIHINGEN DATASET.

Method	IoU					mF1(%)	mIoU(%)	mOA(%)
	IOU_ImSurf	IOU_Building	IOU_LowVeg	IOU_Tree	IOU_Car			
ABCNet[42] *	93.29	91.18	73.41	76.00	72.36	89.38	81.25	92.75
BAnet[43] *	92.96	91.25	74.01	76.39	81.16	90.61	83.15	92.82
CMTFNet[6] σ	86.09	94.09	66.64	82.19	81.60	89.91	82.12	------
CSTUet[8] σ	75.16	83.00	61.45	73.72	60.88	82.85	70.75	------
DCSwin[44] *	94.38	93.63	74.61	76.43	81.50	91.15	84.11	93.53
Eight-Directional[19] σ	82.83	89.46	67.30	77.77	58.86	------	75.24	------
FTransUet[45] *	93.34	**97.37**	**81.63**	**91.35**	**88.44**	90.42	82.94	92.12
FTUNetfomer[40] *	94.31	93.31	74.64	77.25	82.33	91.31	84.37	93.53
GE-swin[11] σ	87.09	92.01	73.38	80.64	76.64	90.85	81.97	------
Hybrid[18] σ	85.35	90.96	71.57	80.88	79.01	85.93	76.37	90.29
MBT-UNet[17]	84.25	88.14	69.62	79	64.34	86.73	77.07	------
RS3Mamba[6] σ	86.62	93.83	67.84	83.66	81.97	90.34	82.78	------
Segformer[20] σ	------	------	------	------	------	89.22	80.08	90.66
TransUet[6] σ	83.10	89.25	65.32	82.70	70.45	87.46	78.16	------
UNetFormer[40] *	94.35	93.18	74.67	77.10	79.99	91.00	83.56	93.52
Ours	**94.40**	94.20	75.26	77.48	83.50	**91.67**	**84.98**	**93.76**

TABLE 7
MODEL COMPARISON ON THE WHDLD DATASET.

Method	IoU						mF1(%)	mIoU(%)	mOA(%)
	IOU_rode	IOU_Building	IOU_Pave	IOU_Vege	IOU_bare	IOU_water			
ABCNet[42]*	43.56	38.23	28.79	77.26	31.83	87.90	59.23	43.93	81.23
BAnet[43]*	53.52	55.22	39.30	83.35	39.59	93.26	68.99	54.19	86.55
DCSwin[44]*	56.52	55.84	37.27	83.06	42.18	93.22	69.65	54.97	86.50
Deeplabv3+[19]σ	47.02	56.87	31.19	85.45	36.41	83.85	70.13	56.79	86.46
FactSeg[45]σ	48.35	55.31	29.45	85.27	34.21	84.43	69.42	56.17	69.42
FTUNetfomer[40]*	61.84	58.94	42.66	84.71	45.46	94.16	72.92	58.72	87.66
HRNetV2[45]σ	52.41	56.11	31.74	85.65	31.58	86.63	70.33	57.35	86.73
TransUnet[46]σ	52.39	51.62	37.64	79.61	37.69	92.56	------	58.58	------
U-net[10]σ	43.14	55.32	28.16	84.68	34.85	83.19	68.25	54.89	85.82
Unetformer[40]*	59.15	56.78	38.85	83.40	36.08	93.85	69.34	54.85	86.95
Ours	**62.15**	**59.22**	**44.63**	**84.99**	**46.51**	**94.29**	**73.62**	**59.50**	**88.14**

TABLE 8
MODEL COMPARISON ON THE LOVEDA DATASET.

Method	IoU							mIoU(%)
	Background	Building	Road	Water	Barren	Forest	Agriculture	
ABCNet[42]*	44.28	56.76	54.00	79.31	18.87	45.67	59.44	50.76
BANet[43]*	44.08	53.86	52.61	77.22	16.86	47.50	60.09	50.32
CMTFNet[47]σ	38.98	58.96	50.50	54.27	30.72	37.41	22.56	46.68
DCSwin[44]*	41.72	58.91	57.33	79.69	22.68	47.04	56.18	51.94
FTUNetfomer[40]*	45.04	60.60	57.62	81.52	18.99	46.76	60.71	53.03
FTUNetfomer+SAM[7]σ	39.27	62.91	**59.67**	63.00	28.57	36.33	28.34	50.68
MBT-Uet[48]σ	51.92	54.33	47.15	57.68	26.16	41.06	43.50	45.97
MSCAT-Uet[49]σ	**53.77**	**63.25**	56.70	71.26	31.48	41.32	57.26	53.48
RS3Mamba[47]σ	39.72	58.75	57.92	61.00	**37.27**	39.36	33.98	50.93
Unetformer[40]*	44.57	58.84	55.16	80.02	19.74	45.70	**61.51**	52.21
Ours	46.34	60.70	57.25	**81.57**	19.16	**47.87**	62.18	**53.58**

The first row of both tables represents our baseline model, the second row uses only the CFESwin Transformer Block, the third row employs the DSSM module, and the fourth row utilizes the GCSB. Rows five to seven present the experimental results of pairwise combinations of the three modules. As shown in Table 1, there is still room for improvement in the standalone GCSB module; the model combining all three modules achieved a 0.83% improvement over the model with only the GCSB, with slight enhancements compared to other combination models. In Table 2, the performance of different layer module combinations on the WHDLD dataset is somewhat inferior to the baseline, but the combination of all three modules shows better results, with a 0.88% improvement in mIoU.

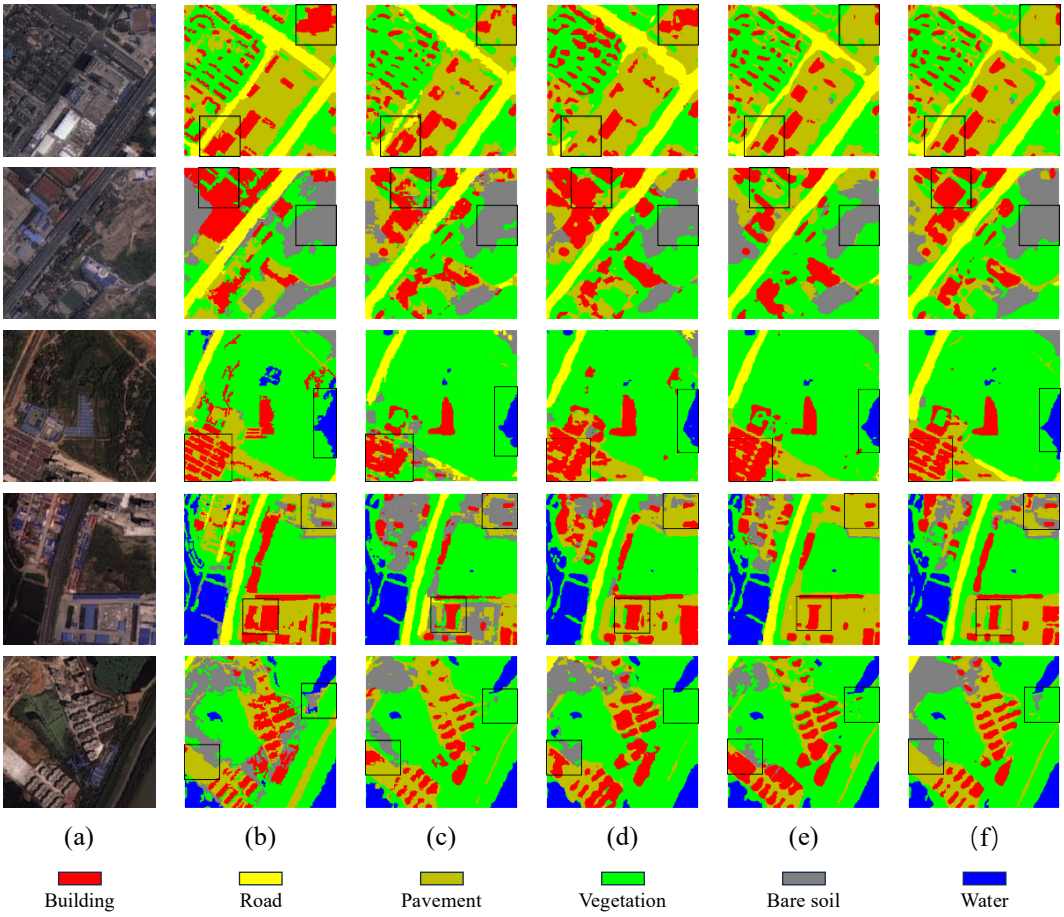

Fig.4. Output mask comparison on the WHDLD dataset: (a) original image, (b) ground truth label, (c) BANet, (d) DCSwin, (e) UNetFormer, （f) our own model. We present five sets of visualization results for each model.

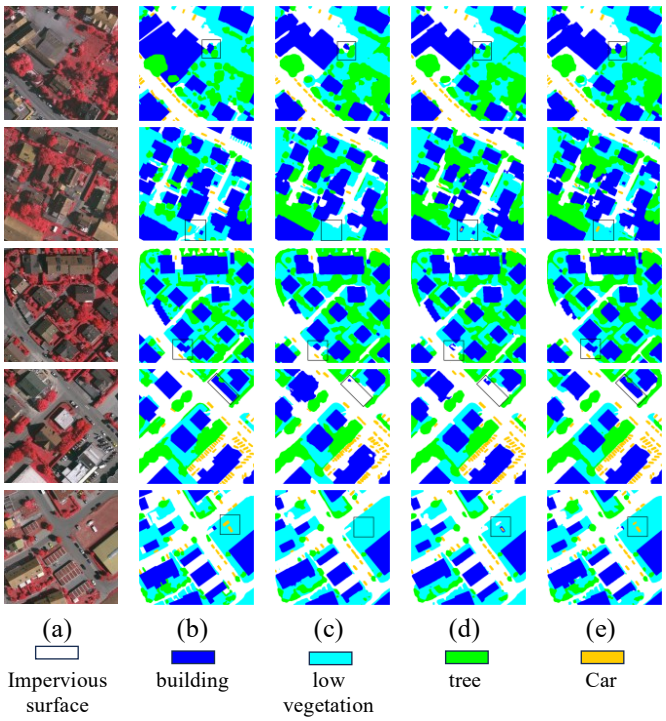

Fig. 5. Output mask comparison on the ISPRS Vaihingen dataset: (a) original image, (b) ground truth label, (c) ABCNet, (d) UNetFormer, (e) our own model. We present one experimental result for each model.

In Table, the combination of the three modules is more effective. Overall, the model structure we created is effective for remote sensing semantic segmentation tasks.

Table 4 demonstrates the effectiveness of our innovative CFESwin Transformer Block at different layers. Although the OA value in the third layer is slightly higher than that in the first layer, overall, the first layer still shows the best performance, and Table 5 demonstrates the performance advantage of the two the DSSM groups on the decoder.

D. Experimental Results and Analysis

The symbol * indicates the experimental results from our own tested and verified model, while the symbol σ refers to results adapted from other papers. In Table 6, our model shows significant improvements compared to the baseline model FTUNetformer on the Vaihingen dataset, with mIoU, mF1, and mOA increased by 0.61%, 0.36%, and 0.23%respectively. In terms of detailed performance, our model outperforms all comparison models in the ImSurf category, while it slightly lags behind FTransUet in the Building, LowVeg, Tree, and Car categories. However, overall metrics show that mIoU, mF1, and mOA exceed the FTransUet model by 1.04%, 1.25%, and 1.64%, respectively. The experimental results indicate that our model demonstrates significant improvements on the Vaihingen dataset, primarily due to its efficient global-local feature extraction capabilities.

In the experimental comparison on the WHDLD dataset, our model demonstrates significant improvements over the other models listed in the table, with mIoU, mF1, and mOA being higher than the baseline model by 0.78%,0.70%, and 0.48%, respectively. In terms of fine-grained category performance, our model outperforms all comparison models in the categories of ImSurf, Building, LowVeg,Tree, and

Car. The experimental results indicate that our model achieves the best semantic segmentation performance on the WHDLD dataset as shown in the Table6.In Table 7 presents the model comparisons on the LoveDA dataset, where our model achieved the highest values for the Water and Forest categories. However, it performed 7.43% and 2.55% lower than the MSCAN-Uet model for the Background and Building categories, indicating a notable capability in object recognition within complex natural environments. The model's performance on the Road category was also 2.42% lower than FTUNetformer+SAM. Our results for the Barren and Agriculture categories were not satisfactory, and we aim to improve in these areas in future research. Overall, the mIoU value reached the highest level in the LoveDA dataset, exceeding the baseline by 0.55%.

V. CONCLUSION

In this paper, we constructed an encoder with a CFESwin Transformer as the first layer, forming a global-local feature fusion model, DSCSNet, with a decoder composed of three main modules: GLB, DSSM, and GCSB. The model we created not only achieves more refined feature extraction in local areas but also enhances overall performance in global context understanding. This model demonstrates better expressiveness and accuracy when handling high-resolution complex models. The CFESwin Transformer, as an improved Swin Transformer block, strengthens the window features, while DSSM uses a dynamic spatial separation mechanism and large convolution kernels to flexibly extract features at different scales. GCSB captures more complex remote sensing images through channel separation, thereby enhancing segmentation effects. In future work, we will create more diverse types of Swin Transformer blocks and innovate in computational speed.

REFERENCES

[1] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation," *IEEE Transactions on Image Processing,* vol. 32, pp. 1052-1064, 2023.

[2] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 58, no. 8, pp. 5367-5376, 2020.

[3] J. Shen, C. Li, and J. Chen, "Strong Convergence Theorems of Generalized Viscosity Implicit Rules for Fixed Points of Total Asymptotically Nonexpansive Mappings in Hilbert Spaces," *IAENG International Journal of Applied Mathematics,* vol. 54, no. 10, 2024.

[4] L. Sahar, S. Muthukumar, and S. P. French, "Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 48, no. 9, pp. 3511-3520, 2010.

[5] D. Huang, Y. Tang, and R. Qin, "An evaluation of PlanetScope images for 3D reconstruction and change detection–experimental validations with case studies," *GIScience & Remote Sensing,* vol. 59, no. 1, pp. 744-761, 2022.

[6] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, "Road segmentation for remote sensing images using adversarial spatial pyramid networks," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 59, no. 6, pp. 4673-4688, 2020.

[7] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Transactions on Geoscience and Remote Sensing,* 2024.

[8] 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.

[9] Z. Zhao, Z. Li, J. Miao, K. Wu, and J. Wu, "Global Context-Enhanced Network for Pixel-Level Change Detection in Remote Sensing Images," *IAENG International Journal of Computer Science,* vol. 51, no. 8, pp. 1060-1070, 2024.

[10] 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*, 2015: Springer, pp. 234-241.

[11] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems,* 2017.

[12] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array,* vol. 14, p. 100157, 2022.

[13] 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.

[14] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.

[15] 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.

[16] Z. Liu *et al.*, "Proceedings of the IEEE/CVF international conference on computer vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012-10022.

[17] L. Sun, J. Dong, J. Tang, and J. Pan, "Spatially-adaptive feature modulation for efficient image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13190-13199.

[18] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian conference on computer vision*, 2016: Springer, pp. 180-196.

[19] 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.

[20] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 43, no. 10, pp. 3349-3364, 2020.

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759-8768.

[22] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 2020: Springer, pp. 173-190.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.

[24] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0-0.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[26] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314-1324.

[27] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, 2021: PMLR, pp. 10347-10357.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794-7803.

[30] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999,* 2018.

[31] F. Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156-3164.

[32] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146-3154.
[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534-11542.
[34] 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.
[35] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *International conference on machine learning*, 2021: PMLR, pp. 2286-2296.
[36] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881-6890.
[37] B. Graham *et al.*, "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12259-12269.
[38] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in neural information processing systems,* vol. 34, pp. 15908-15919, 2021.
[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020: Springer, pp. 213-229.
[40] L. Wang *et al.*, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 190, pp. 196-214, 2022.
[41] 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.
[42] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 181, pp. 84-98, 2021.
[43] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sensing,* vol. 13, no. 16, p. 3065, 2021.
[44] 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 Geoscience and Remote Sensing Letters,* vol. 19, pp. 1-5, 2022.
[45] W. Zhu and Y. Jia, "High-resolution remote sensing image land cover classification based on EAR-HRNetV2," in *Journal of Physics: Conference Series*, 2023, vol. 2593, no. 1: IOP Publishing, p. 012002.
[46] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306,* 2021.
[47] X. Ma, X. Zhang, and M.-O. Pun, "RS 3 Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation," *IEEE Geoscience and Remote Sensing Letters,* 2024.
[48] B. Liu, B. Li, V. Sreeram, and S. Li, "MBT-UNet: Multi-Branch Transform Combined with UNet for Semantic Segmentation of Remote Sensing Images," *Remote Sensing,* vol. 16, no. 15, p. 2776, 2024.
[49] T. Wang *et al.*, "MCAT-UNet: Convolutional and Cross-shaped Window Attention Enhanced UNet for Efficient High-resolution Remote Sensing Image Segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2024.