DS-SwinUNet: Redesigning Skip Connection with Double Scale Attention for Land Cover Semantic Segmentation

Zirui Shen, Wanjie Liu, and Sheng Xu, *Member, IEEE*

***Abstract*—In recent years, the development of visual Transformer has gradually replaced convolutional neural networks in the visual domain with attention computation, causing pure Transformer networks to become a trend. Despite significant advancements in semantic segmentation models for remote sensing, a critical gap remains in effectively capturing both local and global contextual information. Existing models often excel in either fine-grained local detail or long-range dependencies, but not both. Our work addresses this research gap by proposing the DS-SwinUNet model integrating convolutional operations with Transformer-based attention mechanisms through the novel DS-Transformer block, which consists of a two-scale attention mechanism incorporating convolutional computation and a modified FFN, and the module is placed in the skip connection section with Swin-UNet as the backbone. Experiments demonstrate that the Transformer module proposed in this paper improves the mIoU by 2.73% and 0.41% over the original Swin-UNet when the WHDLD and OpenEarthMap dataset are used as the segmentation task. Code is available at: https://github.com/A1ray/DS-SwinUNet**

***Index Terms*—Deep learning, semantic segmentation, skip connection, Transformer.**

I. Introduction

REMOTE sensing images, as an important data source for obtaining the earth's surface coverage conditions, have a wide range of applications in the fields of urban planning, environmental monitoring, and resource management [1], [2]. Accurate semantic segmentation of these images is the key to understand the surface features, monitor environmental changes and make accurate decisions. However, traditional methods often face challenges such as complex feature categories and fuzzy boundaries when dealing with semantic segmentation of remotely sensed images, which constrain the accuracy of image interpretation and precise recognition of feature objects.

In recent years, the rise of deep learning techniques has provided new opportunities for solving remote sensing image semantic segmentation problems. Deep learning architectures such as convolutional neural network (CNN) and Transformer show great potential in image processing. CNN is capable of extracting image features through convolutional operations, but has limitations in handling global information and long-range dependencies [3]–[5]. In contrast, Transformer-like methods capture the correlation of different regions in an image with a self-attention mechanism, which is expected to better address the challenges of scale and boundary information in remote sensing images. Meanwhile, network structures such as U-Net fuse different layers of features through skip connection, as shown in Fig. 1, which improves the efficiency of utilizing details and contextual information [6].

However, the semantic segmentation of remote sensing images presents challenges for current deep learning methods. A key issue is the limited diversity and richness of available datasets, which constrains the improvement of model performance. Specifically, the lack of a large-scale and diverse land cover dataset hampers the model's generalization ability, especially in handling different feature objects and complex scenes. Remote sensing datasets such as WHDLD and OpenEarthMap present unique challenges in capturing both local and global contextual information. For instance, in the WHDLD dataset, the boundaries of water bodies are often blurred, making it difficult to distinguish between water and surrounding vegetation. Similarly, in the OpenEarthMap dataset, the varying scales of urban infrastructure such as buildings and roads introduce complexity in balancing fine-grained local details and broader spatial relationships. A critical gap persists in effectively capturing both local and global contextual information in semantic segmentation models designed for remote sensing, despite significant advancements. Existing models are adept at excelling in either fine-grained local detail or long-range dependencies, but not both simultaneously. CNN-based models like U-Net and SegNet excel at extracting local features but face challenges with capturing global context due to the inherent locality of convolution operations. On the other hand, Transformer-based models effectively capture global relationships but may lack the precision necessary for detailed segmentation of small or intricate structures. Swin-UNet, as a Transformer-based U-shaped network, has demonstrated the ability to capture long-range dependencies through attention mechanisms [7]. However, its emphasis on global features can result in the loss of fine-grained local details, which are essential for tasks such as land cover segmentation. In contrast, our proposed DS-SwinUNet enhances this architecture by introducing a Double-Scale Attention mechanism, which not only captures long-range dependencies through Transformer-based attention but also integrates convolutional operations to preserve local details more effectively. The research involves

This work was supported in part by the Scientific and Technological Innovation 2030-Major Projects under Grant 2023ZD0405605 and in part by the Practice Innovation Training Program Projects for Jiangsu College Students under Grant 202310298040Z, in part by the Practice Innovation Training Program Projects for Jiangsu College Students(Grant No. SJCX23 0320.)

Sheng Xu is the corresponding author.

comparing and analyzing the performance of different deep learning models in feature recognition and delving into the role of skip connections and other mechanisms in semantic segmentation tasks. Ultimately, the study seeks to provide new insights and technical support for enhancing semantic segmentation methods and advancing the field of remote sensing image processing. The contribution is as follows:

1) We propose a new two-scale attention mechanism, which preserves the linear relationship while learning the more complex non-linear one thus improving the precision of segmentation.
2) We have optimized the FFN in the traditional Transformer module to perform forward propagation, obtaining a better accuracy performance as well as a smoother loss drop during training, and reducing the training time.
3) We design a FPU(Final Processing Unit) at the end of the network model, which effectively mitigates the problem of possible information loss caused by decoder upsampling in the segmentation model and enhances the robustness of the model.

II. Related work

A. Semantic Segmentation

Semantic segmentation is a critical computer vision task with a wide range of applications in fields such as autonomous driving, robotics, and medical image analysis. Long et al. proposed the full convolutional neural network (FCN), an end-to-end neural network model for semantic segmentation [8]. The FCN model achieves state-of-the-art performance on semantic segmentation tasks by replacing the last convolutional layer of the convolutional neural network (CNN) with a transpositional convolutional layer, allowing for segmentation of the image into segmentation maps of arbitrary sizes. Chen et al. presented the Deep Hollow Convolutional Network (DeepLabV3), a deep neural network model for semantic segmentation [9]. The DeepLabV3 model employs hollow convolution to expand the receptive field, enabling the capture of a wider range of contextual information in an image [10]. He et al. developed the Mask R-CNN model, a deep neural network model for instance segmentation and semantic segmentation [11]–[13]. The addition of a segmentation branch to the Faster R-CNN model enables the Mask R-CNN model to perform target detection and semantic segmentation simultaneously. Wu et al. proposed the SETR model, an end-to-end Transformer model for semantic segmentation [14]. Applying the Transformer model to the semantic segmentation task and utilizing a new attention mechanism to aggregate features at different locations in the image has enabled the SETR to achieve a good performance. Cao et al. proposed Swin-UNet, the first U-Net shaped medical image segmentation network based on pure Transformer [7]. Labeled image chunks are fed into a Transformer-based U-shaped encoder-decoder architecture via skip connections for local global semantic feature learning.

B. Mutual attention

Mutual attention is a mechanism utilized in machine learning models, particularly in computer vision and natural language processing (NLP), to facilitate dynamic interaction among multiple input sources or different sections of the same input [15], [16]. This mechanism enables the model to prioritize the most pertinent aspects of the inputs, enhancing comprehension and information integration [17]. By dynamically adjusting its focus on varying sections of the input data based on their task-related relevance, mutual attention crucially aids tasks with components of fluctuating importance. Notably, in multi-modal learning scenarios involving various input sources such as text and image, mutual attention permits the model to interconnect information across these sources [18].

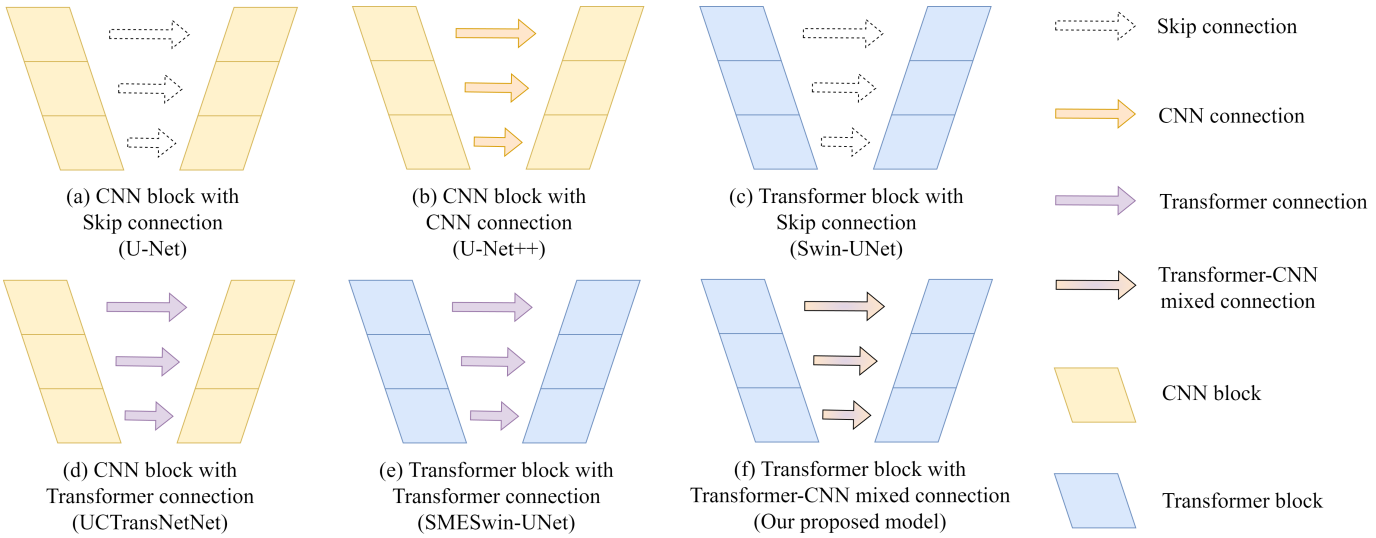


Fig. 1. The U-Net architecture and its variants. (a) uses skip connections to directly pass feature maps from encoder to decoder, aiding in spatial information recovery. (b) enhances U-Net by adding dense connections between convolutional layers, improving detail capture. (c) replaces CNN blocks with Transformer blocks while retaining skip connections for spatial information. (d) connects CNN blocks using Transformer layers to combine local and global feature extraction. (e) utilizes Transformer blocks exclusively, connected through additional Transformer layers for long-range dependencies. (f) combines Transformer blocks with both Transformer and CNN connections, leveraging strengths of both architectures.

Mutual attention mechanisms have been employed in image classification to capture dependencies between various parts of an image, thereby enhancing the accuracy of classification. Vision Transformers (ViTs), as introduced by Dosovitskiy et al., utilize self-attention mechanisms across image patches to more effectively capture the global context compared to traditional CNNs [19]. Through the attention mechanism, the model can determine the significance of different image regions, consequently boosting classification performance. In the context of image segmentation, mutual attention mechanisms play a crucial role in accurately delineating objects within an image by incorporating contextual information from diverse regions of the image. An illustrative instance is the Dynamic Dual Attention Network (DDAN) presented by Chen et al., where both spatial and channel attention mechanisms are leveraged to improve segmentation [20]. While spatial attention focuses on pertinent image regions, channel attention captures dependencies among different feature maps, thereby enhancing segmentation accuracy. Liu et al. proposed S^2MA, by combining self-attention and mutual attention, long-range contextual dependencies are enhanced, allowing for more precise learning and propagation of contexts with the integration of multi-modal information [21].

C. Transformers and CNNs for remote sensing applications

Recent advancements in remote sensing applications have significantly benefited from the integration of Transformer models and Convolutional Neural Networks (CNNs). Transformer models such as Swin Transformer proposed by Liu et al., have shown remarkable capabilities in handling the vast and complex data typical in remote sensing, providing enhanced accuracy and efficiency in various tasks such as land cover classification, object detection, and change detection [22]. The Swin Transformer incorporates a hierarchical feature extraction process that significantly improves the efficiency and accuracy of remote sensing image analysis. Its ability to handle different scales of features makes it particularly suitable for high-resolution satellite imagery, where objects can vary greatly in size. On the other hand, CNNs have long been the cornerstone of remote sensing image analysis due to their proficiency in capturing local spatial features through convolutional operations [23]–[25]. Current state-of-the-art CNN models such as ConvNeXt provide comparable performance to Transformer while maintaining a clean design, and in some cases are more efficient [26].

III. Method

A. Model Architecture

The model proposed in this paper enhances the Swin-UNet model, which is depicted in Fig. 2. The green and orange rectangles in the figure represent the encoder and decoder in Swin-UNet respectively. Specifically, both the encoders and decoders consist of 4 layers of Swin Blocks, with 2, 2, 6 and 2 blocks in each layer respectively [27], [28]. It is important to note that the skip connection between encoders and decoders in each layer only performs the concat operation [29]. To better preserve the linear feature relationship between the encoders and decoders in the same layer, this paper introduces a new Double Scale Attention(DS-Attention) mechanism designed to achieve this purpose. DS-Attention combines the features of two neighboring layers of encoders of different dimensions, and the attention computation is done separately by these two layers of features, so that the same input learns its own semantic information at different scales, and provides richer and more accurate semantic information for the subsequent splicing with the output of the decoder. We designed the final processing unit at the output position of the network to minimize the loss of image position information resulting from up-sampling. In addition, we implemented a module with depth-separable convolution to reduce the number of parameters added to the network, thereby lowering the computation load. This design choice maintains the expressive power of the convolutional layer while minimizing the overall parameter count.

The DS-Transformer block integrates the strengths of CNNs and Transformers to enhance multi-scale feature extraction for image segmentation, as depicted in Fig. 1. The block starts by normalizing the input feature maps z^l and z^{l+1} using Layer Normalization (LN) [30]. These normalized feature maps are then fed into the DS-Attention module, which combines convolutional operations for local feature extraction with Transformer attention for capturing global dependencies. The output from the DS-Attention module is normalized again through another LN layer and passed through a Multi-Layer Perceptron (MLP) for further feature transformation [31], [32]. Finally, a residual connection adds the MLP output back to the original input, enhancing the feature representation. This mixed connection of convolutional and Transformer operations enables the DS-Transformer block to effectively capture both local and global features, improving segmentation performance.

B. DS-Attention

Traditional skip connections in U-shaped networks help fuse low-level and high-level features by directly passing feature maps from the encoder to the decoder. However, these methods often result in the loss of fine details or fail to accurately capture long-range dependencies in complex scenes, particularly when there are significant scale variations or ambiguous boundaries, such as in the case of water bodies or bare soil in remote sensing images. While skip connections assist in retaining spatial information, they do not explicitly account for the varying scales of features within the same image. This inadequacy motivates the development of our DS-Attention mechanism, which integrates multi-scale feature extraction within the skip connections to better preserve both local and global information, ensuring a more robust segmentation [33], [34]. The cross-attention mechanism facilitates the comparison and fusion of two sequences from different sources, enabling the model to establish associations between them and derive comprehensive insights [35], [36]. Specifically, the approach discussed in the paper involves using the output of the current encoder and the next encoder layer as inputs to DS-Attention, facilitating the relationship between the two sequences and

enabling the extraction of information from one sequence to enhance understanding of the other, as shown in Fig. 3 [37]. The methodology involves utilizing the output of the lth layer of the Swin encoder block as the Query, and the output of the $l+1$th layer of the Swin encoder block as the Key and Value for attention calculation. Subsequently, the $l+1$th layer is used as the Query, and the lth layer as the Key & Value, with the resultant attention computation being concatenated with the earlier result. The inclusion of multi-scale features ensures that both fine and coarse details are preserved, leading to more accurate segmentation outcomes. This process allows the two sequences to effectively learn and comprehend each other's sequential information, resulting in the acquisition of features with enriched semantic content. Traditional self-attention mechanisms, such as those used in ViT, operate on a single scale, which may not be sufficient for tasks requiring multi-scale feature aggregation while DS-Attention enhances this by incorporating multiple scales [19].

In contrast to traditional attention modules, where the input undergoes linear layer processing to compute correlation between all features, the proposed method replaces the linear layer with a convolutional layer. This adaptation localizes correlation computation, thereby reducing computational complexity. Moreover, the translation invariance property of the convolutional layer ensures model robustness to image translation.

Specifically, given an input z^l and its next layer of features z^{l+1}, it is first transformed into the corresponding sequence Query, Key and Value with the following formula:

$$\tilde{Q} = z^l W_{\tilde{Q}} + B_{\tilde{Q}} \qquad \hat{Q} = z^{l+1} W_{\hat{Q}} + B_{\hat{Q}} \tag{1}$$

$$\tilde{K} = z^{l+1} W_{\tilde{K}} + B_{\tilde{K}} \qquad \hat{K} = z^{l} W_{\hat{K}} + B_{\hat{K}} \tag{2}$$

$$\tilde{V} = z^{l+1} W_{\tilde{V}} + B_{\tilde{V}} \qquad \hat{V} = z^{l} W_{\hat{V}} + B_{\hat{V}} \tag{3}$$

where $\tilde{Q}(\hat{Q})$, $\tilde{K}(\hat{K})$ and $\tilde{V}(\hat{V})$ is transformed from the input and into a learnable linear mapping matrix, $W_{\tilde{Q}}(W_{\hat{Q}})$, $W_{\tilde{K}}(W_{\hat{K}})$ and $W_{\tilde{V}}(W_{\hat{V}})$ represents the weight matrix. $B_{\tilde{Q}}(B_{\hat{Q}})$, $B_{\tilde{K}}(B_{\hat{K}})$ and $B_{\tilde{V}}(B_{\hat{V}})$ represents the bias of the

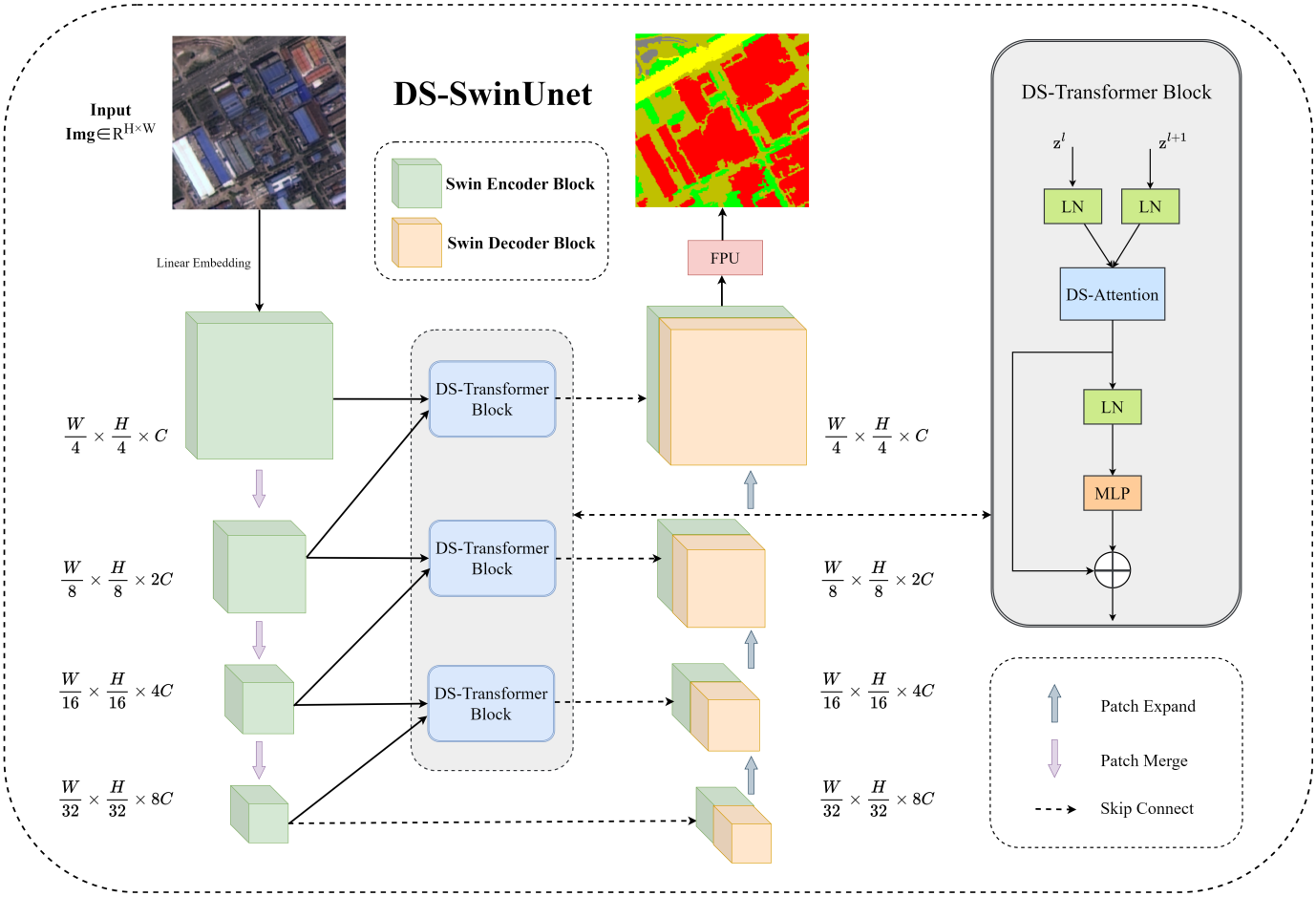


Fig. 2. The proposed model architecture. Our model comprises an encoder, bottleneck, and decoder. Initially, the input image is transformed into feature maps of size $\frac{W}{4} \times \frac{H}{4} \times C$ through linear embedding. The encoder utilizes Swin Encoder Block to process these feature maps, gradually reducing the spatial dimensions to $\frac{W}{32} \times \frac{H}{32} \times 8C$. Subsequently, the bottleneck stage employs DS-Transformer to ensure continuous connections with the decoder. Within the decoder, Swin Decoder Block reverses the process, increasing the dimensions back to $\frac{W}{4} \times \frac{H}{4} \times C$. Ultimately, the Final Processing Unit (FPU) merges the final feature maps to generate the segmentation map. Additionally, the diagram illustrates various connection types: Patch Expand, Patch Merge, and Skip Connection, elucidating the feature transfer mechanisms operating within the model.

linear mapping [38]. Then, the attention scores of the two are calculated separately with the following formula:

$$\tilde{S} = \text{softmax}(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_k}})\tilde{V} \tag{4a}$$

$$\hat{S} = \text{softmax}(\frac{\hat{Q}\hat{K}^T}{\sqrt{d_k}})\hat{V} \tag{4b}$$

$$S = \tilde{S} \oplus \hat{S} \tag{4c}$$

Where d_k represents the dimension of the input, $\oplus$ represents the final score is spliced from the results of two attention calculations in the channel dimension. Since the length of V is half of the input dimension, the output dimension is the same size as the input, which facilitates the encoder part of the channel splicing and reduces the computational complexity while effectively avoiding overfitting.

C. SwiGLU for FFN

The FFN, or feed-forward neural network layer, is an integral component of the Transformer block, tasked with reshaping the inputs of the Transformer block to create new representations.

As a critical element within the Transformer model, the FFN serves to enhance the model's representation capability and overall performance. Typically, the FFN's structure involves two fully-connected layers, linked by an activation function placed between them, such as GeLU. While GeLU is effective, SwiGLU's combination of Swish and GLU offers a superior balance between non-linearity and model capacity. Traditional activation functions like ReLU and its variants (e.g., Leaky ReLU) are simpler but often insufficient for capturing the complex patterns required for high-performance segmentation tasks. SwiGLU provides a more sophisticated alternative [39]. SwiGLU, which combines the Swish activation function with the Gated Linear Unit (GLU), provides enhanced flexibility by allowing the model to adaptively control the information flow through gating mechanisms. This property is particularly beneficial in deep networks, as it helps mitigate the vanishing gradient problem commonly associated with deeper architectures. SwiGLU introduces a more complex non-linear relationship compared to GeLU, enabling the model to capture more intricate patterns in the data. By incorporating gating mechanisms, SwiGLU increases the representational capacity of the FFN layer, leading to improved performance. The formula for SwiGLU is depicted as follows:

$$\text{Swish}_\beta(x) = x\sigma(\beta x) \tag{5}$$

$$\begin{aligned}\text{SwiGLU}(x, W, V, b, c, \beta) =& \text{Swish}_\beta(xW + b) \\ & \otimes (xV + c)\end{aligned} \tag{6}$$

$$\begin{aligned}\text{FFN}_{\text{SwiGLU}}(x, W, V, W_2) =& (\text{Swish}_1(xW) \otimes xV)W_2 \\ =& (xW\sigma(xW) \otimes xV)W_2\end{aligned} \tag{7}$$

where β is a hyperparameter, taken as 1 in this experiment. σ denotes the sigmoid activation function, W, W_2 and V represents the 3-layer weight matrix in the modified FFN (the original FFN has only 2 layers), b and c represents the bias of the weight matrix [40]. The modified FFN usually does a scaling of the size of the hidden layers due to the introduction of more weight matrices, thus ensuring that the overall number of parameters remains the same.

D. Final Processing Unit

In this section, a final processing unit (FPU) is developed with the aim of preserving the semantic feature information of the input data and minimizing the loss of accuracy, as depicted in Fig. 4, the FPU consists of several stages designed to achieve these objectives [41]. FPU plays a crucial role in maintaining the integrity of semantic feature information during upsampling, where traditional methods often suffer from information loss due to the spatial resolution increase.

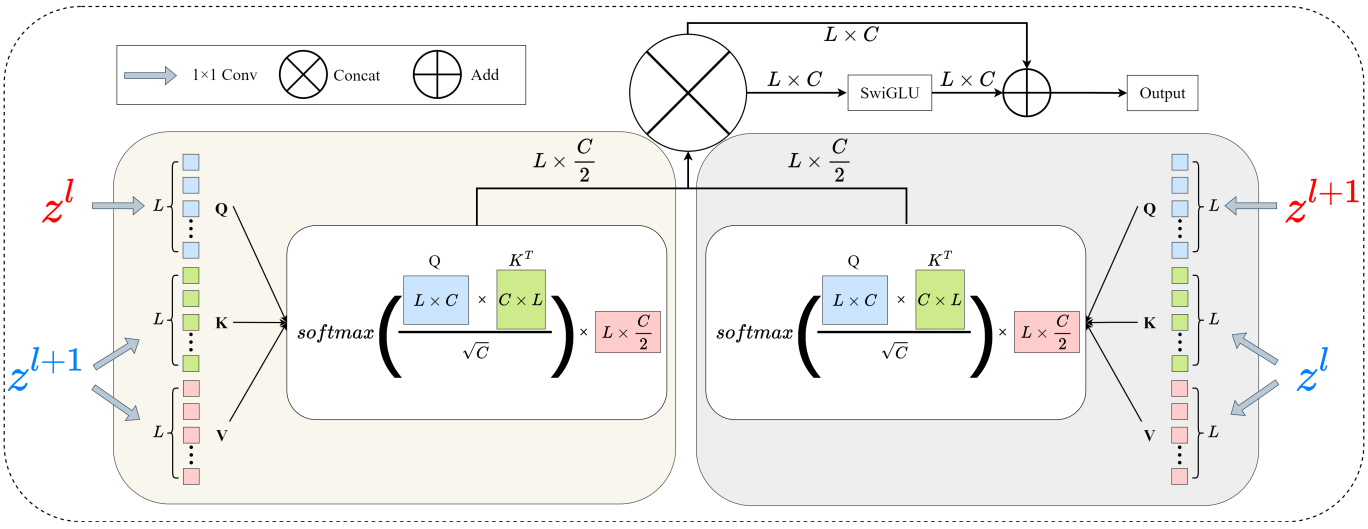


Fig. 3. Structure of DS-Attention. DS-Attention starts with convolutional layers to capture local features, followed by a Transformer encoder to integrate global context. The figure highlights the mixed connection approach, where convolutional operations are seamlessly combined with Transformer-based attention.

In contrast to conventional bilinear or deconvolution-based upsampling, FPU leverages depthwise separable convolutions to minimize the number of parameters while preserving fine-grained details. This operation allows the model to focus on critical semantic features during the upsampling process, reducing the potential for blurring or misclassification of smaller objects. First, the input data undergoes a Depth Separable Convolution(DWConv), which is a specialized convolution operation [42]. This operation applies a separate convolution kernel to each input channel, thereby reducing computational complexity and the number of parameters in the convolution layer while maintaining its expressive power. Subsequently, the output of the depth convolution is convolved using a 1x1 convolution kernel, resulting in the generation of the final output feature map. Unlike the final patch expand of Swin-UNet, the FPU does not solely enlarge the resolution of the data before outputting the final result; it focuses on preserving the semantic feature information of the input data throughout the process of enlargement. This distinctive approach aims to restore the output to the same size as the input data, while also ensuring the preservation of important details. The equation for FPU is as follows:

$$\begin{aligned} y =& \text{DWConv2D}(x, k_{DW}) \\ &+\text{PWConv2D}(\text{DWConv2D}(x, k_{DW}), k_{PW}) \end{aligned} \tag{8}$$

where x is the input feature map, k_{DW} and k_{PW} is the number of deep convolutional kernels and dot convolutional kernels, both set to 6 in this experiment, while the size of the deep convolutional kernel is 3 and the size of the dot convolutional kernel is 1. y is the output feature map. The semantic richness of the data is positively correlated with the length of the data, where longer sequences tend to contain richer semantic information. As such, in both PW convolutions, the number of channels is scaled up by a factor of 2 in order to capture more information about the data before applying finer processing. Subsequently, the scaled-up channels are then scaled back to their original size to enable finer processing of this enriched information. This approach enables the model to better capture the semantic nuances within the data, particularly in longer sequences.

E. Loss Function

Both the WHDLD and Xuanwu Lake datasets exhibit class imbalance, with certain classes (e.g., water bodies and vehicles) having significantly fewer examples compared to more prevalent classes like vegetation or buildings. To mitigate the impact of class imbalance, we employed a combination of Cross-Entropy Loss and Dice Loss with the following formula:

$$\begin{aligned} \text{Loss} =& 0.4 * \text{CrossEntropyLoss} \\ &+ 0.6 * \text{DiceLoss} \end{aligned} \tag{9}$$

$$\text{CrossEntropyLoss} = -\sum_{i=1}^{N} y_i \log(p_{\hat{y}_i}) \tag{10}$$

$$\text{DiceLoss} = 1 - \frac{2\,|X \cap Y| + \text{Smooth}}{|X| + |Y| + \text{Smooth}} \tag{11}$$

where y_i and $\hat{y}_i$ denote the true label and predicted label respectively, $p_{\hat{y}_i}$ denotes the probability value of correct prediction. $|X|$ and $|Y|$ represents the mask values of the true and predicted targets respectively, $|X \cap Y|$ represents the intersection of the two, and Smooth as a hyperparameter is set to 1 in this experiment. We set weights for the two loss functions, a weight of 0.4 was set for CELoss, while DiceLoss accounted for 0.6. This loss function ensured that the model paid more attention to underrepresented classes during training, reducing the likelihood of bias towards the dominant classes.

IV. Evaluation Indicators and Datasets

A. Evaluation Indicators

We use five evaluation metrics for evaluating the performance of the model, which are OA, AA, Kappa, F1, and mIoU, each formula will be explained below. OA, AA and F1 are the metrics to measure the performance of the classification model. OA calculates the correct prediction rate for all samples and AA calculates the average of the accuracy rates for all categories. Given the TP(True Positive), TN(True Negative), FP(False Positive) and FN(False Negative) for all categories, metrics for OA, AA, and F1 scores can be derived with the following formulas:

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{12}$$

$$AA = \frac{\sum_{i=1}^{N} Accuracy_i}{N} \tag{13}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{14}$$

where N is the number of categories and i is the current category. The mIoU is one of the most commonly used evaluation metrics for segmentation models, which calculates the average of the ratio of the intersection and concatenation of all categories with the following formula:

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP}{FN + FP + TP} \tag{15}$$

Kappa coefficient is a statistic that measures the consistency of a classification model. It can be used to compare the performance of different classification models, and it is more robust to sample imbalance, while other metrics such as accuracy are susceptible to sample imbalance. The formula is as follows:

$$K = \frac{p_0 - p_{\text{exp}}}{1 - p_{\text{exp}}} \tag{16}$$

where p_0 represents the proportion of the number of samples correctly predicted by the classification model to the number of all samples, p_{exp} represents the probability that a sample will be correctly classified without taking into account the predictions of the classification model.

B. Wuhan Dense Labeling Dataset

The WHDLD dataset, derived from a large RS image of the Wuhan urban area, comprises 4940 RGB images, each manually labeled with six classes, namely building, road, pavement, vegetation, bare soil, and water [43]. The images have a spatial size of 256 × 256 and a resolution of 2m. In Fig. 5(a), sample images along with their pixelwise labeling results are depicted, providing a visual representation of the dataset.

C. OpenEarthMap Dataset

OpenEarthMap presents a major advance over existing data with respect to geographic diversity and annotation quality [44]. It consists of 2.2 million segments of 5000 aerial and satellite images covering 97 regions from 44 countries across 6 continents, as shown in Fig. 5(c), with manually annotated 8-class land cover labels at a 0.25–0.5m ground sampling distance. Semantic segmentation models trained on the OpenEarthMap generalize worldwide and can be used as off-the-shelf models in a variety of applications.

D. Xuanwu Lake Dataset

The Xuanwu Lake Dataset is a private dataset produced by our research group, as shown in Fig. 5(b), and we plan to make it public in the future. It was collected by a drone from Xuanwu Lake and its surrounding area in Nanjing, Jiangsu Province, China. The dataset has a total of 2869 256 × 256 images in seven categories, which is one more category of vehicle than WHDLD. Due to the complexity of real-world environments, the images captured by our drones could not be more accurately segmented into vegetation and soil categories, and the categorization of soils and roads was problematic in certain shaded environments. Since our dataset still suffers from a small amount of labeling, it is only used as an evaluation of the effect of generalization in subsequent experiments.

V. Experimental results and analysis

A. Experimental Setup

In the training process of DS-SwinUNet, we used the following hyperparameters:

- Learning rate: An initial learning rate of 1e-4 with a cosine annealing schedule to gradually reduce the learning rate as the training progressed.
- Batch size: 16.
- Optimizer: Adam optimizer with default parameters β_1 = 0.9, β_2 = 0.999, and an epsilon value of 1e-8 to ensure stable convergence.
- Weight decay: 5e-5.
- Epochs: 150 epochs, and early stopping was applied if the validation loss did not improve for 10 consecutive epochs.

To enhance the model's generalization ability, various preprocessing and data augmentation techniques were employed on the training datasets. Initially, each image was standardized by subtracting the dataset's mean and dividing by its standard deviation, ensuring consistent input values. To further increase the diversity of the training data and mitigate overfitting, several augmentation methods were applied. These included random rotations, horizontal and vertical flipping, random cropping and resizing, as well as color jittering.

B. Performance Comparison of Different Models on WHDLD

We compared the proposed model's overall segmentation performance with that of six state-of-the-art methods. The experimental results, detailed in Table I, indicated that our model slightly outperformed the other methods across various metrics, each evaluation metric is expressed as a percentage, with higher numbers indicating better performance. Meanwhile, we also compared the loss convergence of our model with various types of U-shaped networks during training, the comparison is shown in Fig. 8(a). Our model exhibits more stable convergence and smoother curves compared to Swin-UNet. This observation underscores the effectiveness of our network optimization strategies. This success was the result of a series of optimizations and enhancements made to the original model, which ultimately led to a 2.726% increase in the mIoU. Fig. 8(b) shows the number of parameters and flops of various types of segmentation networks compared to ours, where all the remaining ones are CNN-based networks except Swin and ours. It is obvious and well recognized that the number of parameters of the Transformer-based networks is much larger than that of CNN, but the flops are smaller than that of CNN. Compared with Swin, our network has an increase in both the number of parameters and the FLOPs, but this small increase is acceptable in exchange for an increase in accuracy. The DS-Attention mechanism enhances the model's ability to capture both fine local details and global contextual information, which is particularly beneficial in classes such as water and bare soil. For instance, in the water class, the ability to attend to broad, uniform areas while maintaining sharp boundaries at the water's edge contributes to higher accuracy. Similarly, in the bare soil class, DS-Attention's focus on intricate textures and edges helps differentiate between soil and similar surrounding categories. This dual-scale attention allows the model to outperform Swin-UNet, which may struggle with balancing fine detail preservation and long-range dependencies in these classes.

Class imbalance can have a noticeable effect on certain evaluation metrics, particularly F1-score and mIoU, where underrepresented classes might receive lower scores due to fewer correctly predicted examples. However, our model demonstrated relatively stable performance even in categories with fewer examples, such as water bodies and bare soil, thanks to the loss function. As shown in Table I, the F1-score and mIoU for these classes, although slightly lower than for majority classes, were still competitive, indicating the model's ability to generalize well across imbalanced datasets.

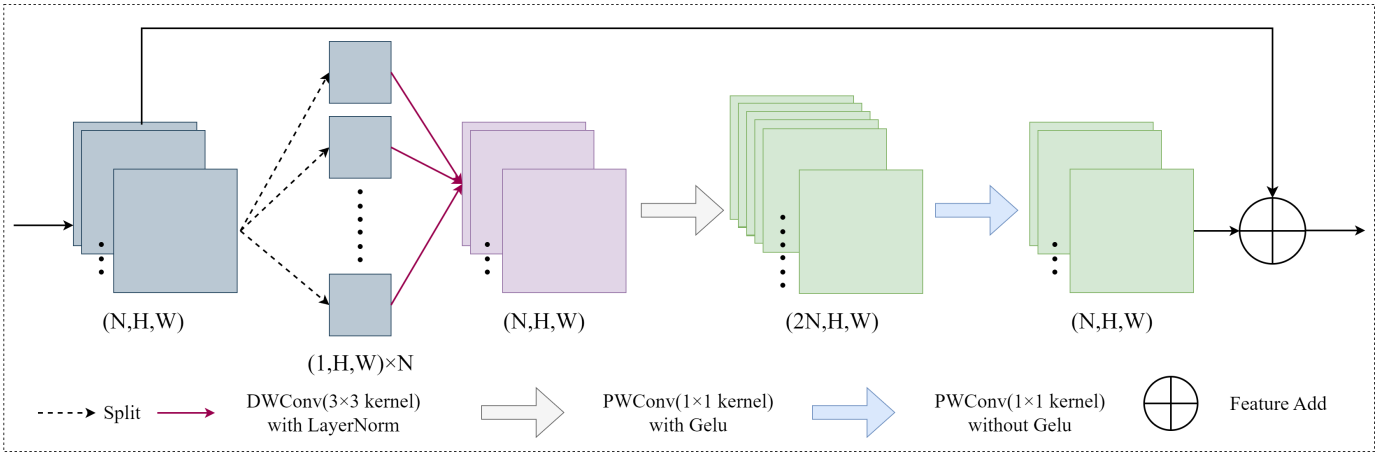


Fig. 4. Structure of FPU. The input feature maps, denoted as (N,H,W), where N represents the number of feature maps, H the height, and W the width, undergo a splitting operation. The split feature maps are processed by depthwise convolution (DWConv) layers with a 3×3 kernel and LayerNorm. These processed features are then concatenated to form feature maps of size (1,H,W)×N. Subsequently, pointwise convolutions (PWConv) with a 1×1 kernel and Gelu activation function are applied, followed by another PWConv without Gelu activation. The resulting feature maps are of size (2N,H,W). Finally, a feature addition operation integrates these feature maps back into the original dimensions (N,H,W), completing the FPU processing pipeline.

In Fig. 6, we present the segmentation results obtained from the comparable U-Net family models. Notably, the black boxes in these figures delineate areas where our model demonstrates superior performance compared to the other methods. The box demonstrates our model's ability to effectively segment various classes of small objects among a wide range of similar objects. Our findings indicate that DS-SwinUNet yields segmentation results that closely align with the ground truth, surpassing those produced by the baseline model. Specifically, our proposed method effectively identifies the correct salient regions while mitigating the occurrence of false positive lesions, resulting in the creation of coherent boundaries.

C. Performance Comparison of Different Models on OpenEarthMap

To comprehensively evaluate the performance of our proposed DS-SwinUNet model, we compared it with several state-of-the-art semantic segmentation models using the OpenEarthMap dataset. The quantitative results are presented in Table III, while Fig. 7 offers a visual comparison of the segmentation outputs. Table III demonstrates that our DS-SwinUNet model achieves superior performance across multiple key metrics, including Intersection over Union (IoU) for various land cover classes and mean Intersection over Union (mIoU). Specifically, our model achieves an impressive mIoU of 67.49%, improving a little bit over the original SwinUnet in all classes of IoU, mIoU improves by 0.41%, and comparing to the other models in the table it is also at the top of the list. Our DS-SwinUNet model excels in multiple classes, particularly in challenging ones like "Bareland" and "Developed," where it significantly reduces misclassification and captures finer details compared to other models. The integration of convolutional operations and Transformer-based attention in the DS-Attention block is a pivotal factor in our model's performance. The convolutional layers effectively capture local spatial details, while the Transformer layers provide a global context, leading to more precise and coherent segmentations. This hybrid approach allows our model to leverage the strengths of both techniques, resulting in enhanced feature extraction and robust performance across diverse land cover types.

D. Pure CNN Networks with DS-Transformer

We conducted experiments involving the integration of the DS-Transformer module into three distinct CNN networks [45]–[47]. Specifically, we introduced the module in the cross-layer feature reconstruction section of each network. The resulting experimental outcomes are presented in Table II, pure CNN means the original model and with DST means the concat opreation of the original model with DS-Transformer. Overall, the evaluation metrics for each model exhibited a slight improvement, with the exception of the mIoU metric of PSPNet, which instead displayed a minor decrease. We attribute this finding to the distinction between the local attention mechanism of CNN and the global attention mechanism of the Transformer. Consequently, it is plausible that PSPNet may have learned an incorrect local feature, leading to our proposed module unintentionally amplifying this erroneous feature.

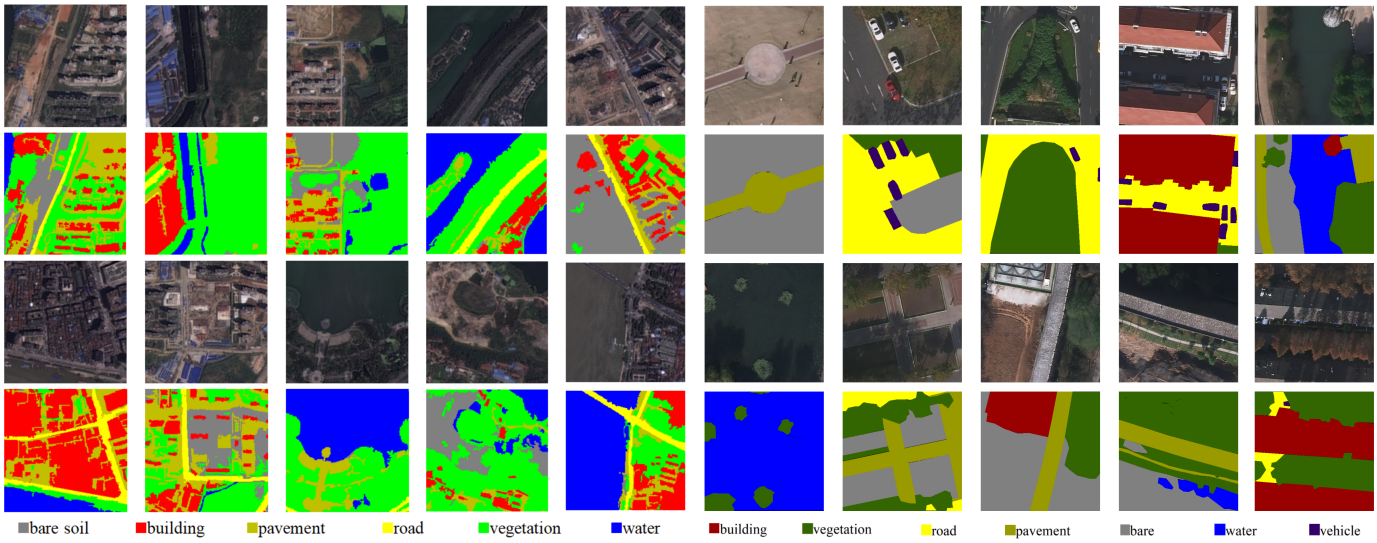
E. Ablation Study

Table IV displays the results of individual experiments conducted for each proposed module, providing evidence of the effectiveness of our designed DS-Attention. We reevaluate the role of feedforward neural networks in the Transformer block by incorporating the latest activation functions, introducing additional weight matrices, and scaling the size of the hidden layer in order to maintain a constant overall number of parameters. This modification enables DS-Attention to capture more intricate relationships and enhance the expressive capacity of this attention mechanism. Incorporating the strengths of both the Swish and GLU activation functions, SwiGLU offers several advantages. The Swish function is known for its ability

to generate smoother gradients compared to ReLU, thereby addressing issues related to gradient vanishing and explosion. By leveraging the combined form of SwiGLU, models are empowered to capture more intricate non-linear relationships. This increased expressiveness results in the model being able to more effectively fit the training data, thus leading to improved performance across a range of tasks. To verify the effectiveness of SwiGLU, it was imperative to conduct experiments in conjunction with DS-Attention. In order to compare and demonstrate the effectiveness of SwiGLU, traditional Gelu and SwiGLU were included as components of DS-Attention. Empirical results from our experiments indicated that using SwiGLU led to a more stable training process and improved performance, as evidenced by a smoother loss curve and better convergence compared to models using GeLU.

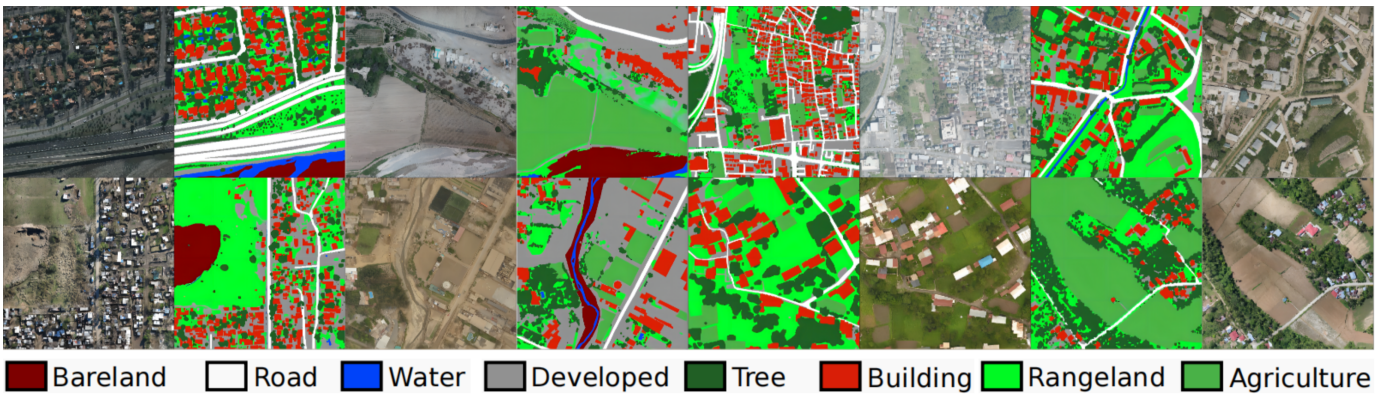
F. Performance of Generalization on Xuanwu Lake Dataset

In our private dataset, we relabeled some images to align the number and types of categories with those in WHDLD, and ensured that their masks correspond one-to-one with the masks used for training in WHDLD. The same class (e.g., vegetation) can appear very differently in various parts of the dataset due to differences in seasonal conditions, lighting, and vegetation density. This variability makes it difficult for models to generalize and accurately segment the entire class. As depicted in Fig. 9, both Swin-UNet and our enhanced model were tested for generalization in an untrained dataset. Remarkably, our proposed model demonstrated superior performance compared to SwinUNet in accurately segmenting target edges. The dataset features many complex and irregular boundaries, especially between natural and man-made structures. Accurately capturing these boundaries is crucial for precise segmentation, but it is challenging for models that rely solely on local context. The presence of mixed scenes with both urban infrastructure and natural elements requires the model to be versatile and capable of distinguishing between different types of features accurately.



(a) WHDLD

(b) Xuanwu Lake Dataset



(c) OpenEarthMap Dataset

Fig. 5. Example images and labels of WHDLD, Xuanwu Lake Dataset and OpenEarthMap Dataset.

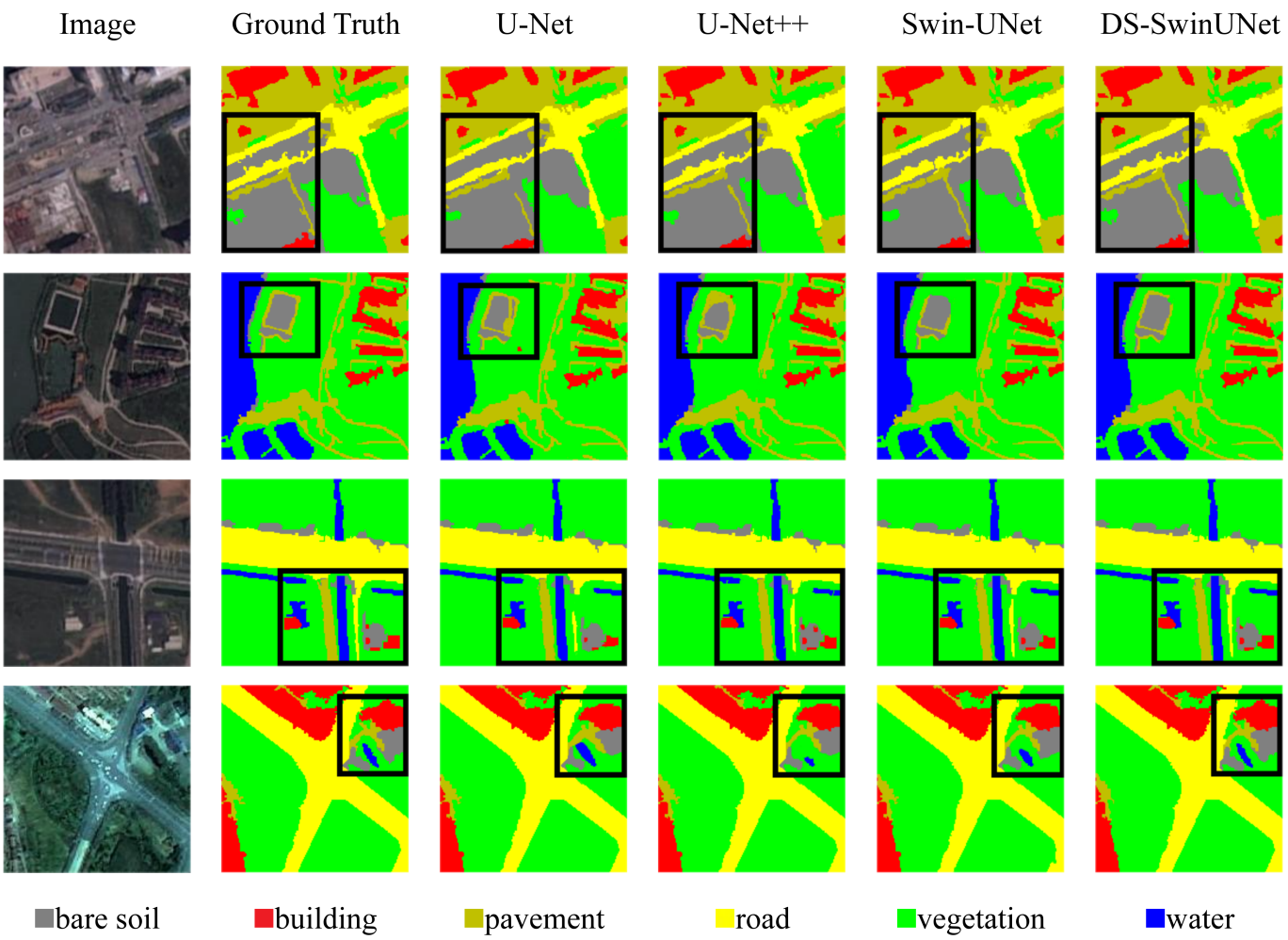


Fig. 6. Visualization of various UNet-like networks on WHDLD. The visualisation in the black box highlights that the output of our model is much closer to the segmentation similarity of Ground Truth than to the other models.

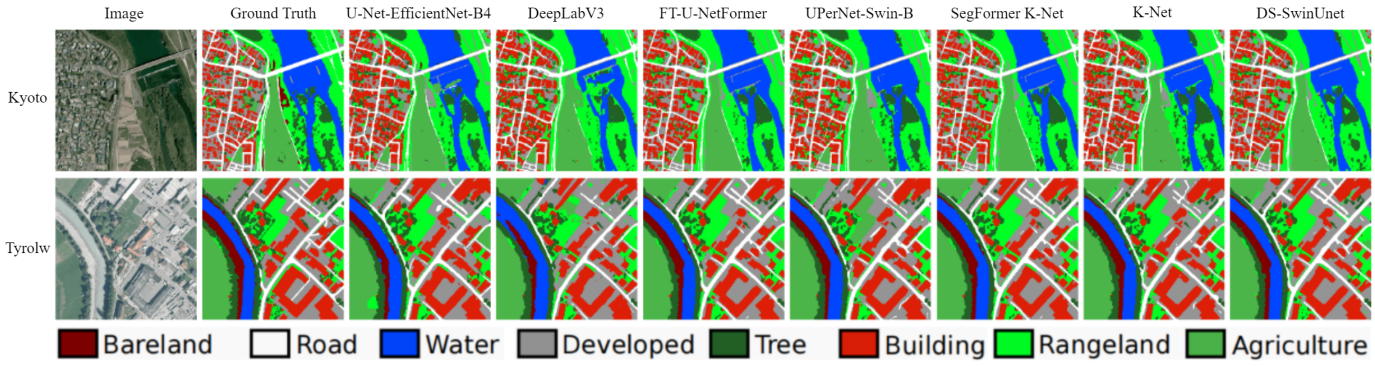
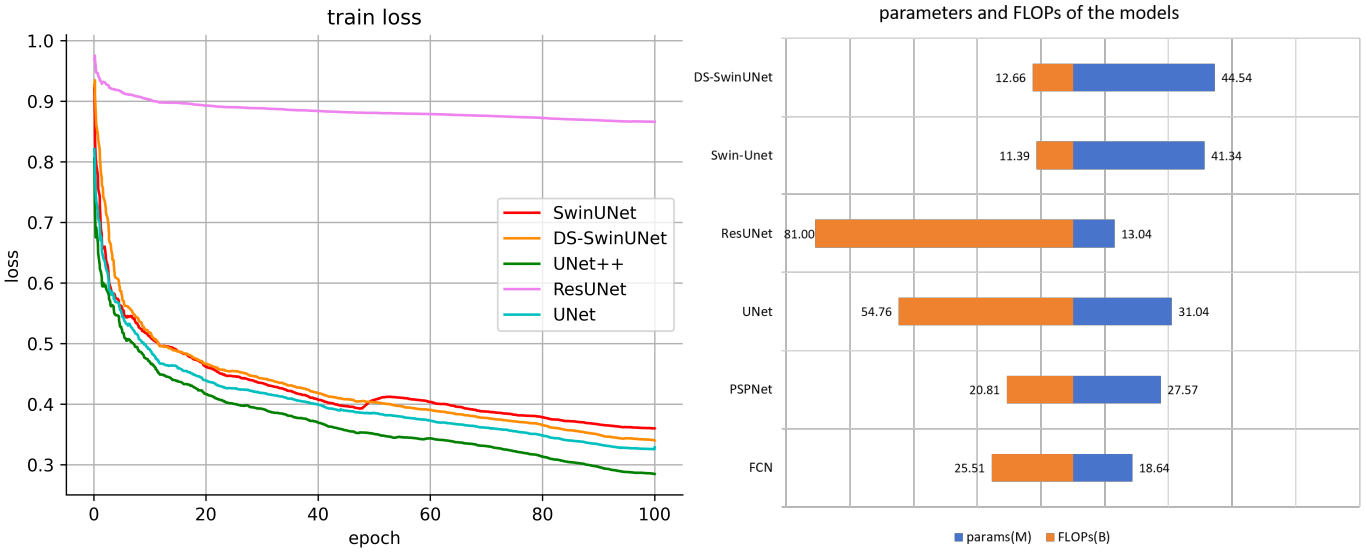


Fig. 7. Visual comparison of land cover mapping results of some of the baseline models presented in Table III.

adds to the complexity as the model must adapt to different contexts within the same image. Notably, out model excelled particularly in delineating the edges of regular-shaped targets like circular plazas encompassing sidewalks, square buildings, and road edges. In the second row of Fig. 9, our model accurately segments out six boats in the lake, and although it was not labeled with that species during training, it was also able to separate out tiny non-water objects in the water of just a few pixels, which can indicate that our model performs well in segmenting small objects as well.



(a) (b)

Fig. 8. Training loss curves and the number of parameters and FLOPs of the models.

VI. Discussion

In urban planning, this method could assist in accurately identifying and segmenting buildings, roads, and green spaces, providing valuable insights for infrastructure development. In agriculture, the model could be applied to monitor crop health and detect land use changes, which are critical for precision farming. Environmental monitoring is another potential domain, where the model could be used to track deforestation, water bodies, and natural resource depletion with high accuracy. The model's ability to generalize across diverse datasets suggests that it could be adapted to various tasks where both fine details and broader context are essential for decision-making. The key factor contributing to this performance improvement is the integration of a two-scale attention mechanism within the skip connections. This mechanism includes mutual attention mechanisms working at varying scales, enabling the model to better capture both fine and coarse features. By adopting this dual-scale approach, the model can effectively utilize detailed local information and broader contextual cues simultaneously, resulting in more accurate segmentation results.

While the DS-SwinUNet model demonstrates strong overall performance across various land cover types, there are specific conditions where it struggles. For instance, in shadowed areas, the model often has difficulty accurately segmenting features due to reduced contrast and lack of distinct visual cues. This is particularly evident in urban environments where buildings create significant shadow effects, leading to misclassification of adjacent features like roads and vegetation. Additionally, in regions with highly mixed land cover, such as transitional areas between urban and rural environments, the model may face challenges in distinguishing between similar classes (e.g., bare soil versus constructed surfaces). In these cases, the overlapping characteristics of different land cover types can confuse the model, resulting in lower accuracy for specific classes. These limitations highlight the need for further refinement and possibly additional training data focused on these challenging conditions to enhance the model's robustness.

VII. Conclusion

In this paper, we present DS-SwinUNet, a novel semantic segmentation network developed specifically for the land cover remote sensing dataset. DS-SwinUNet focuses on optimizing the hopping connections within the U-shaped network to address the challenge of effectively integrating low-dimensional features from the encoder part with high-dimensional features from the decoder part. Our evaluation of DS-SwinUNet was conducted using the publicly available land cover remote sensing dataset WHDLD. The experimental findings demonstrate that DS-SwinUNet achieves superior segmentation accuracy on this dataset.

Our model introduces substantial improvements over existing models by effectively combining local and global feature extraction capabilities. The increase in segmentation accuracy demonstrates the effectiveness of the DS-Attention. While the DS-Attention adds some complexity, the significant performance improvements justify this addition. Future research could focus on optimizing the computational efficiency of the DS-Attention, potentially by reducing parameters and FLOPs. Extending the model to other computer vision tasks, such as object detection, and evaluating its performance on larger and more diverse datasets would be valuable. The balanced trade-off between computational efficiency and accuracy makes our model a valuable contribution, especially for applications requiring high accuracy.

TABLE I
EVALUATION METRICS OF DIFFERENT MODELS ON WHDLD

Model	OA(%)↑	AA(%)↑	K(%)↑	mIoU(%)↑	F1 score(%)↑					
					Building	Road	Pavement	Vegetation	Bare Soil	Water
SegNet [48]	80.229	63.787	71.403	52.940	63.253	54.649	51.466	86.473	47.682	95.649
U-Net [6]	81.830	67.724	74.422	55.706	70.752	58.668	52.609	89.185	48.097	97.089
Tiramisu [49]	82.188	65.712	74.903	58.167	68.918	**70.047**	53.576	88.206	50.313	96.598
FGC [50]	82.975	68.855	75.927	57.368	72.642	57.931	53.842	89.651	50.282	97.294
MSFCN [51]	84.168	72.081	77.558	60.366	74.499	68.797	55.176	90.024	52.178	**97.511**
EfficientNet-B4+FCN [52]	82.019	68.675	75.019	56.895	72.714	63.226	54.515	89.341	50.398	96.843
UNetFormer [53]	83.633	71.514	77.118	58.453	73.957	65.590	54.211	89.717	51.027	96.483
Swin-UNet [7]	82.290	69.388	76.529	57.944	72.508	64.104	53.759	89.809	50.255	96.915
Ours	**84.446**	**72.629**	**77.934**	**60.670**	**75.486**	69.471	**55.314**	**90.228**	**52.790**	97.415

TABLE II
THE PERFORMANCE OF DS-TRANSFORMER IN DIFFERENT MODELS

Method	OA(%)↑		AA(%)↑		mIoU(%)↑	
	pure CNN	with DST	pure CNN	with DST	pure CNN	with DST
U-Net [6]	81.830	**82.037**	67.724	**67.940**	55.706	**56.034**
U-Net++L^3 [54]	82.040	**82.159**	68.117	**68.291**	56.610	**56.933**
PSPNet [55]	82.227	**82.304**	68.227	**68.454**	**56.785**	56.763

TABLE III
EVALUATION METRICS OF DIFFERENT MODELS ON OPENEARTHMAP

Model	Backbone	IoU(%)↑								mIoU (%)↑
		Bareland	Rangeland	Developed	Road	Tree	Water	Agriculture	Building	
U-Net	VGG-11	40.69	56.76	53.99	62.16	72.44	82.81	73.14	77.77	64.97
U-Net	ResNet-34	40.35	57.75	54.92	62.87	72.65	82.24	74.06	78.58	65.43
U-Net	EfficientNet-B4	50.63	58.17	56.27	64.83	73.20	86.02	76.28	80.20	68.20
U-NetFormer	ResNeXt101	46.09	60.67	**58.12**	65.07	**73.77**	86.34	76.98	79.96	68.37
FT-U-NetFormer	Swin-B	**50.19**	**60.84**	57.58	**65.85**	73.33	**87.44**	**77.50**	**80.29**	**69.13**
DeepLabV3	ResNet-50	39.11	56.16	52.28	60.57	71.25	79.32	70.75	75.83	63.16
HRNet	W48	39.71	55.50	53.49	59.22	71.10	79.03	71.38	75.12	63.07
UPerNet	ViT	34.39	54.45	50.64	54.57	69.73	79.24	66.22	74.92	60.52
UPerNet	Swin-B	44.52	58.98	54.78	63.43	72.20	83.71	72.97	78.11	66.09
SegFormer	MiT-B5	36.84	57.94	53.53	63.60	70.51	80.11	72.21	77.35	64.01
SETR PUP	ViT-L	45.35	55.72	51.31	55.47	67.63	73.12	67.14	75.48	61.40
UPerNet	Twins	37.29	57.62	53.83	60.23	72.32	81.93	71.71	77.49	64.05
UperNet	ConvNeXt	40.61	54.94	51.76	58.47	70.44	75.95	68.94	74.30	61.93
K-Net	Swin-B	44.02	57.81	54.85	62.91	71.76	85.18	73.41	78.91	66.11
Swin-UNet	-	48.13	58.59	56.91	64.11	72.17	84.83	73.24	78.67	67.08
Ours	Swin-UNet	48.45	58.94	57.16	64.19	72.35	84.06	76.25	78.57	67.49

REFERENCES

[1] L. Di and E. Yu, "Remote sensing," in *Remote Sensing Big Data*. Cham: Springer International Publishing, 2023, pp. 17–43.

[2] C. Cao and N. S.-N. Lam, "Understanding the scale and resolution effects in remote sensing and gis," in *Scale in remote sensing and GIS*. New York: Routledge, 2023, pp. 57–72.

[3] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106669, 2023.

[4] M. M. Taye, "Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions," *Computation*, vol. 11, no. 3, p. 52, 2023.

[5] M. Krichen, "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, p. 151, 2023.

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

[7] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[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] 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.

[10] C. Chen, C. Wang, B. Liu, C. He, L. Cong, and S. Wan, "Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[11] P. K. Balasubramanian, W.-C. Lai, G. H. Seng, and J. Selvaraj, "Apest-net with mask r-cnn for liver tumor segmentation and classification," *Cancers*, vol. 15, no. 2, p. 330, 2023.

[12] S. Fang, B. Zhang, and J. Hu, "Improved mask r-cnn multi-target detection and segmentation for autonomous driving in complex scenes," *Sensors*, vol. 23, no. 8, p. 3853, 2023.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[14] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *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.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation

TABLE IV
ABLATION EXPERIMENTS. EACH COLUMN REPRESENTS WHETHER OR NOT THE COMPONENTS WERE USED IN ONE EXPERIMENT IN ADDITION TO THE BASELINE MODEL SWIN-UNET.

	Choice						
Baseline(Swin-UNet)	✓	✓	✓	✓	✓	✓	✓
DS-Attention		✓	✓	✓			
Gelu for FFN			✓				
SwiGLU for FFN				✓			
FPU					✓		
FPU without DWConv						✓	
FPU without PWConv							✓
mIoU(%)↑	57.944	58.695	59.831	60.294	58.570	57.951	58.030

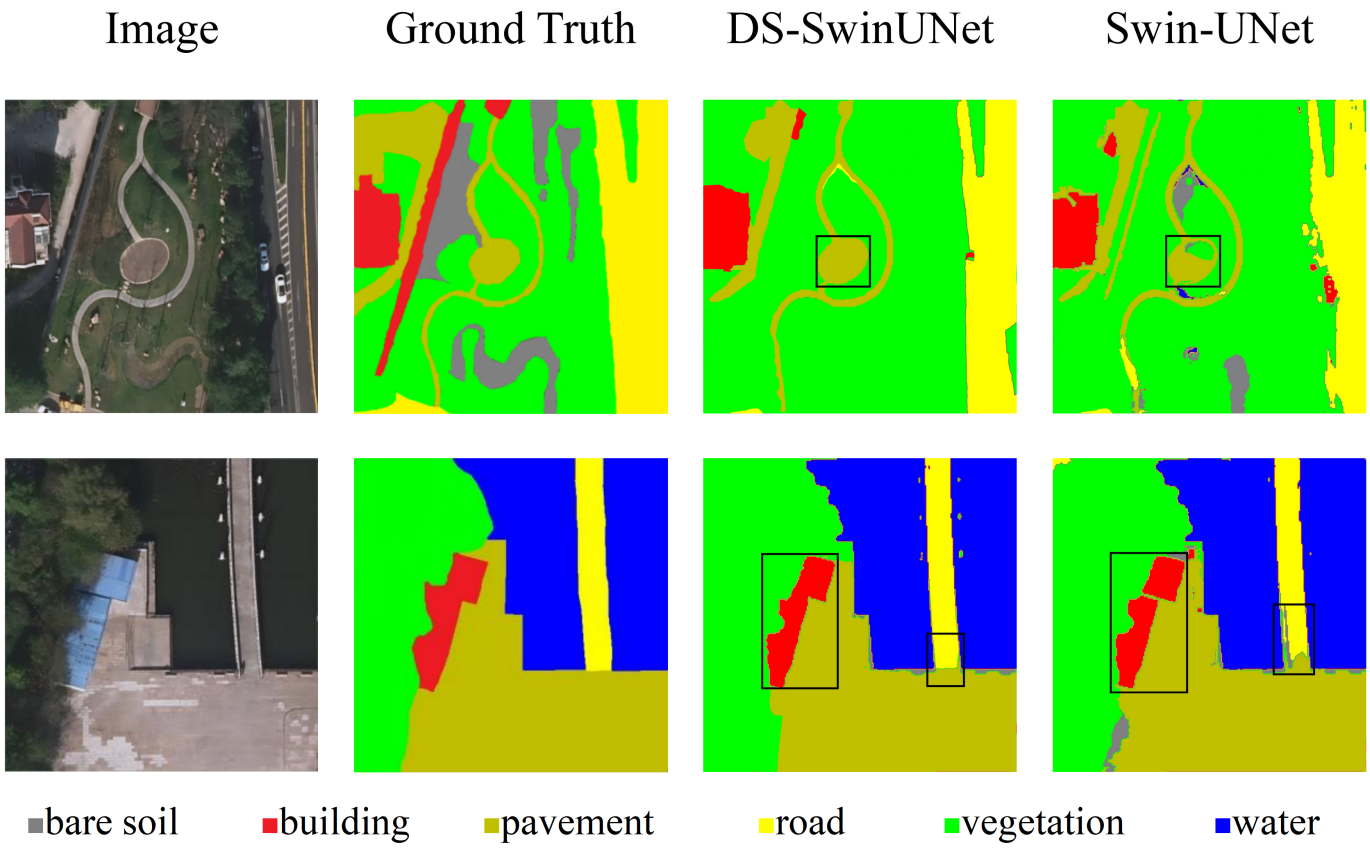


Fig. 9. Generalization ability test on untrained Xuanwu Lake dataset.

with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[17] C. Liu, H. Ding, Y. Zhang, and X. Jiang, "Multi-modal mutual attention and iterative interaction for referring image segmentation," *IEEE Transactions on Image Processing*, 2023.

[18] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] F. Li, H. Bai, and Y. Zhao, "Learning a deep dual attention network for video super-resolution," *IEEE transactions on image processing*, vol. 29, pp. 4474–4488, 2020.

[21] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 756–13 765.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[23] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[24] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, "An improved swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sensing*, vol. 13, no. 23, p. 4779, 2021.

[25] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[27] Y. Li, W. Cai, Y. Gao, C. Li, and X. Hu, "More than encoder: Introducing transformer decoder to upsample," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1597–1602.

[28] C. Jin, R. Tanno, T. Mertzanidou, E. Panagiotaki, and D. C. Alexander, "Learning to downsample for segmentation of ultra-high resolution

images," *arXiv preprint arXiv:2109.11071*, 2021.
[29] J. Díaz García, P. Brunet Crosa, I. Navazo Álvaro, and P. P. Vázquez Alcocer, "Downsampling methods for medical datasets," in *Proceedings of the International conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2017 and Big Data Analytics, Data Mining and Computational Intelligence 2017: Lisbon, Portugal, July 21-23, 2017*. IADIS Press, 2017, pp. 12–20.
[30] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[31] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.
[32] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 809–821, 2015.
[33] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.
[34] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proceedings of the IEEE/CVF International Conference on Computer vision*, 2021, pp. 16 269–16 279.
[35] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 267–276.
[36] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.
[37] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," *arXiv preprint arXiv:1906.04284*, 2019.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[39] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, and G. Yang, "Nonpeaked discriminant analysis for data representation," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 12, pp. 3818–3832, 2019.
[40] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.
[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
[42] 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.
[43] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020.
[44] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6254–6264.
[45] Z. Li, G. Chen, and T. Zhang, "A cnn-transformer hybrid approach for crop classification using multitemporal multisensor images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 847–858, 2020.
[46] Y. Liu, G. Sun, Y. Qiu, L. Zhang, A. Chhatkuli, and L. Van Gool, "Transformer in convolutional neural networks," *arXiv preprint arXiv:2106.03180*, vol. 3, 2021.
[47] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.
[48] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[49] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.
[50] S. Ji, Z. Zhang, C. Zhang, S. Wei, M. Lu, and Y. Duan, "Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images," *International Journal of Remote Sensing*, vol. 41, no. 8, pp. 3162–3174, 2020.
[51] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-spatial information science*, vol. 25, no. 2, pp. 278–294, 2022.
[52] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
[53] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "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.
[54] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.
[55] 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.

Zirui Shen received his B.S.degree in software engineering in 2021 from Shanghai Polytechnic University, Shanghai, China. Now, he is studying for the master degree at the College of Information Technology in Nanjing Forestry University.

His research interests include Semantic Segmentation and 3D Object Detection.

Wanjie Liu received the B.S. degree in computer science and technology from Henan Institute of Science and Technology, Xinxiang, China, in 2021. He is currently working toward the M.D. degree in computer technology from Nanjing Forestry University, Nanjing, China.

His research interests include computer vision, remote sensing, pattern recognition and sensor fusion.

Sheng Xu (IEEE member) received the B.Eng. degree in computer science and technology from Nanjing Forestry University, Nanjing, China, in 2010, and the Ph.D. degree in digital image systems from the University of Calgary, Calgary, AB, Canada, in 2018. In 2018, he joined the College of Information Science and Technology, Nanjing Forestry University, where he is currently an Associate Professor. His research interests include mobile mapping, vegetation mapping, and computer vision.