Multi-Scale Residual Vision Transformer (MSRViT): A Robust Framework for Remote Sensing Image Segmentation

Guanlin Li, and Ji Zhao*

Abstract—**In recent years, Vision Transformers (ViT) have made significant progress in image classification tasks; however, they still face challenges in high-resolution remote sensing image segmentation and spatiotemporal video understanding tasks. To address the computational and memory bottlenecks caused by the complexity of the self-attention mechanism in Transformer models, this paper proposes improvements to the Multi-Scale Vision Transformer (MViT) architecture. First, a shift-invariant positional embedding strategy is introduced to overcome the limitations of absolute positional embeddings, enhancing positional information expression through a decomposed form of positional distance. Second, a residual pooling connection is proposed to address potential detail loss caused by pooling operations, preserving feature information before and after pooling to improve the model's performance in remote sensing image segmentation tasks. Additionally, the optimized MViT architecture is combined with the standard dense prediction framework, SegFormer, further enhancing segmentation performance in complex remote sensing images. Experimental results demonstrate that the proposed model, MSRViT, effectively enhances the model's ability to handle multi-scale and complex features while balancing computational efficiency and performance. These improvements provide a more powerful tool for applying Transformer models to remote sensing image segmentation tasks.**

Index Terms—**Image Segmentation, Deep Learning, Vision Transformer, Self-Attention**

I. Introduction

DESIGNING architectures for remote sensing image segmentation has long been a challenge. Classic architectures such as VGGNet and ResNet have been widely used for their simplicity and effectiveness. In recent years, Vision Transformers (ViT) [1–3] have demonstrated outstanding performance, rivaling Convolutional Neural Networks (CNNs)[4, 5] in many vision tasks. The ViT architecture has undergone various modifications to adapt to different visual tasks.

Although ViT has succeeded in image classification tasks, it still faces challenges in high-resolution remote sensing image segmentation and spatiotemporal video understanding tasks. The high resolution and complex scene structures of remote sensing images pose significant computational and memory challenges, as the complexity of the self-attention module in Transformer-based models increases quadratically with input resolution. Several strategies have been proposed to address these issues, with two popular approaches being.

One approach involves using window-based attention mechanisms, such as the Swin Transformer. The Swin Transformer applies self-attention hierarchically across different scales of the image, initially computing self-attention within local windows and progressively expanding the receptive field. This approach effectively reduces computational load while capturing global information and preserving local details[6–8]. The hierarchical design of the Swin Transformer allows it to be flexibly applied to various vision tasks, including remote sensing image classification, object detection, and semantic segmentation. However, while the Swin Transformer expands the receptive field through its hierarchical structure, its local window design is less precise in capturing global information than the global self-attention mechanism.

Another approach is the pooling attention mechanism, exemplified by the Multi-Scale Vision Transformer (MViT). MViT captures features at different scales through a multi-scale approach, allowing it to excel in processing remote-sensing images with complex structures. MViT effectively integrates information across scales, enhancing model performance, particularly in detailed segmentation and object detection tasks in high-resolution remote sensing images.

In this paper, the MViT model has been improved to enhance its performance in remote sensing image segmentation tasks through the following modifications:

1) A shift-invariant positional embedding strategy is proposed to overcome the limitations of absolute positional embeddings. This approach introduces positional information into Transformer blocks through a decomposed form of positional distance, enhancing the model's ability to perceive spatial positional information.

2) In pooling operations, downsampling is typically used to reduce computational load. However, such downsampling can lead to the loss of specific detailed information, which may affect model performance. To address this issue, a residual pooling connection is introduced. Adding a residual connection after the pooling operation combines the original input features with the pooled features, compensating for detail loss caused by the pooling stride in attention calculations.

3) The optimized MViT model is integrated with the standard dense prediction framework, SegFormer, further enhancing its performance in complex remote sensing image segmentation tasks. Experimental results demonstrate that the optimizations to the pooling attention mechanism and the improvements to the MViT architecture not only enhance the

Manuscript received Oct 24, 2024; revised May 4, 2025. The research work was supported by a scientific research project fund from the Liaoning Provincial Department of Education,and key project of Liaoning Provincial Department of Education(LJKZZ2022043)

Guanlin Li is a Postgraduate of University of Science and Technology Liaoning, Anshan, Liaoning, China. (e-mail: LGLin_2000@outlook.com).

Ji Zhao* is a Professor of University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide phone: +086-139-9808-6167; e-mail: zhaoji_1974@126.com).

model's ability to handle multi-scale and complex features but also significantly improve its performance in remote sensing image segmentation tasks when combined with a standard dense prediction framework. These enhancements provide a more powerful tool for Transformer models in remote sensing image segmentation, balancing computational efficiency and model performance.

II. Related Work

A. Residual Local Feature Network

In recent years, Transformer architectures have achieved significant success in Natural Language Processing (NLP) and have gradually expanded into computer vision, including remote sensing image segmentation tasks. Traditional CNNs have certain limitations in capturing global information and multi-scale features in visual tasks due to their local receptive fields and fixed convolutional kernels. Through their self-attention mechanism, transformers effectively model long-range dependencies and global features, thus attracting increasing interest from researchers in applying them to visual tasks[9, 10]. Vision Transformer (ViT) was one of the earliest attempts in this direction. Despite its high computational cost in high-resolution tasks, it laid the foundation for developing subsequent Transformer-based models. The Swin Transformer reduces computational load through a local window attention mechanism and performs well in handling multi-scale features. Models such as PVT and SegFormer combine the multi-scale feature representation capabilities of CNNs with the global modeling capabilities of Transformers, achieving significant progress in remote sensing image segmentation.

B. Vision Transformer

ViT applies the Transformer architecture directly to image patches, achieving notable success in image classification tasks. The core idea of ViT is to divide an image into fixed-size patches and then treat these patches as a sequence input into the Transformer to capture global dependencies. Although ViT excels in image classification, its application to high-resolution images and spatiotemporal video understanding tasks remains challenging. Specifically, the complexity of the self-attention mechanism in ViT increases quadratically with the input image size, which limits its use

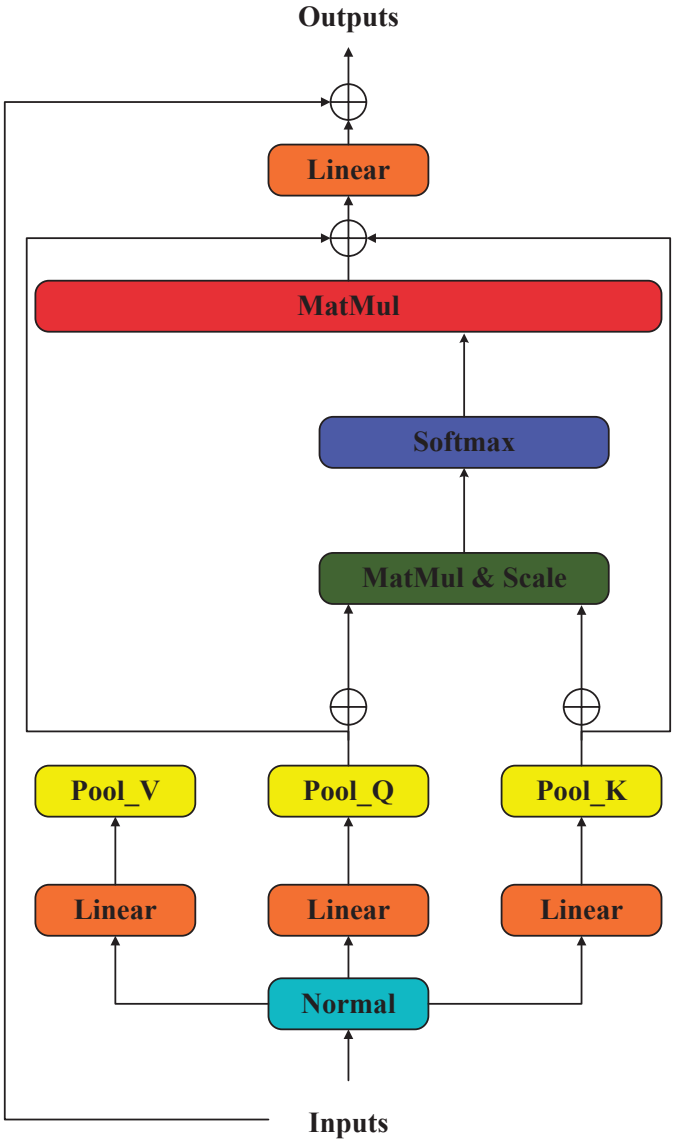

Fig. 1. Overall Structure of Self Attention

in tasks requiring the processing of high-resolution visual signals[11, 12].

C. Multi-Scale Vision Transformer (MViT)

The MViT was proposed to address the limitations of ViT in handling multi-scale information. MViT introduces a multi-scale mechanism to effectively capture features at different scales, accommodating complex structured visual tasks[13]. Unlike ViT, MViT gradually aggregates features within the model, reducing computational costs and enhancing its ability to process multi-scale information. Although MViT performs well across various visual tasks, it may still face challenges when dealing with high-resolution images rich in detail, which has spurred further research into its optimization[14–16].

D. SegFormer

SegFormer[17] is a Transformer architecture specifically designed for image segmentation tasks. It combines the multi-scale feature extraction capabilities of PVT with the global modeling abilities of Transformers. SegFormer employs an efficient self-attention mechanism and a simple decoder design to achieve efficient and accurate segmentation performance. SegFormer has demonstrated outstanding results in various semantic segmentation tasks and exhibits strong adaptability in complex scenarios, making it a powerful tool for remote sensing image segmentation tasks.

III. MSRViT

As shown in Fig. 1, in the standard multi-head self-attention mechanism, the model typically considers the interactions between each element and all other elements globally. While this approach excels at capturing complex dependencies, it leads to a sharp increase in computational cost when processing large-scale, high-resolution remote sensing images. To address this, pooling attention mechanisms have been introduced as an optimization strategy to reduce computational complexity while maintaining strong model performance. The overall structure of MSRViT is shown in Fig. 2. The research on MSRViT was conducted under this guiding principle, with further innovations and improvements specifically tailored for remote sensing image segmentation.

A. Shift-Invariant Positional Embedding

In the standard self-attention mechanism, positional embeddings are essential because the Transformer architecture cannot capture positional information[18]. Absolute positional encoding is typically used to explicitly indicate the position of each element[19, 20]. However, spatial transformations of remote sensing imagery (such as translation and rotation) can render absolute positional encoding suboptimal in remote sensing image segmentation tasks. To overcome the limitations of absolute positional embeddings, MSRViT introduces a shift-invariant positional embedding strategy. This approach incorporates positional information into Transformer blocks through a decomposed form of positional distance. Specifically, the model computes relative positional distances rather than absolute positions, integrating this information into the attention calculations. This enhances the model's robustness to positional changes in the input images, providing more remarkable translation invariance to the positions of different objects. Additionally, this relative positional encoding reduces the model's dependence on exact positions, allowing it to focus more on the relative spatial relationships between objects, thereby improving segmentation performance.

B. Residual Pooling Connection

Detailed information in high-resolution images is crucial in the remote sensing image segmentation task. The standard pooling operation reduces the computation by downsampling, but this often leads to the loss of detailed information, thus affecting the segmentation accuracy[21, 22]. To solve this problem, MSRViT proposed residual pooling connections. By adding residual connections after the pooling operation, the model combines the original input features with the pooled features to compensate for the detail loss caused by the pooling step in the attention calculation. The residual pooling connection retains the detailed information in the original input and effectively reduces the computational complexity through the pooling layer. This design ensures that the model does not sacrifice the ability to capture subtle features of objects in remote sensing images while maintaining efficient computation, thereby improving the overall performance in the remote sensing image segmentation task.

C. Integration of Multi-Scale Feature Extraction with Dense Prediction Framework

The MViT has demonstrated excellent performance in computer vision tasks due to its capability in multi-scale feature extraction. However, for remote sensing image segmentation tasks, MSRViT further optimizes the MViT structure and combines it with the standard dense prediction framework—SegFormer—significantly enhancing the segmentation performance of the model. This integration leverages the expressive power of multi-scale features, improving the model's performance in handling complex object segmentation tasks, particularly in high-resolution remote sensing imagery. By combining multi-scale feature extraction with dense prediction, MSRViT exhibits outstanding performance in fine-grained object segmentation, object detection, and instance segmentation tasks.

D. Overall Process

Firstly, input normalization is performed as shown in Equation (1), where X represents the input features, and X denotes the normalized features. Next, a linear transformation is applied to the normalized features to generate the query(Q), key(K), and value(V) vectors, as specified in Equation (2). Here, W_Q, W_K, and W_V are the linear transformation matrices for the query, key, and value, respectively. Subsequently, the query, key, and value vectors are pooled to reduce computational complexity, as depicted in Equation (3).

$$\hat{X} = \mathrm{Norm}(X) \tag{1}$$

$$Q = W_Q\hat{X}, \quad K = W_K\hat{X}, \quad V = W_V\hat{X} \tag{2}$$

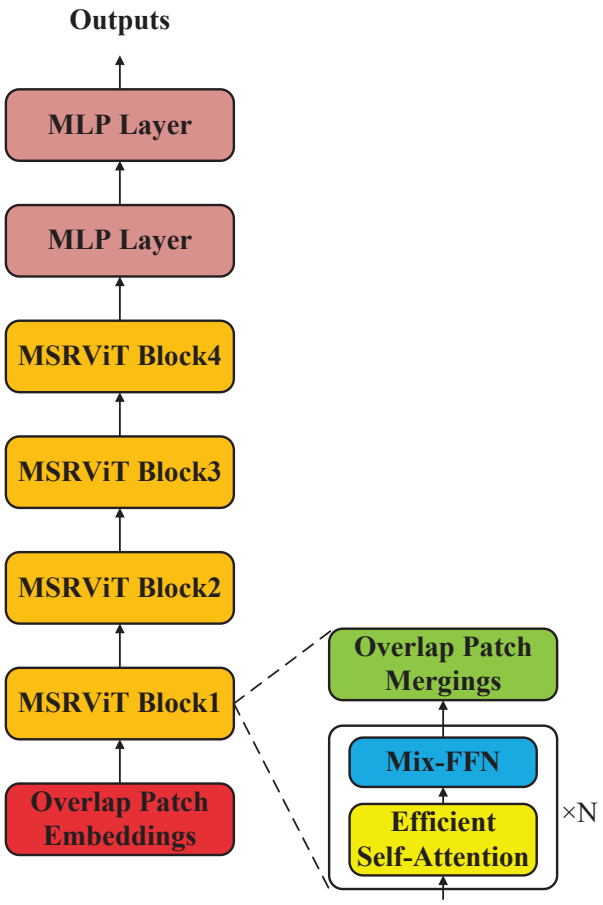

Fig. 2. Multiscale Vision Transformer Overall Structure

$$Q_{\text{pool}} = \text{Pool}(Q), \quad K_{\text{pool}} = \text{Pool}(K), \quad V_{\text{pool}} = \text{Pool}(V) \tag{3}$$

Following this, relative positional embeddings are computed and incorporated into the attention scores, as detailed in Equation (4), where $E(i,j)$ and $F(i,j)$ represent the computed relative positional embeddings. The attention weights are then calculated using the Softmax function, and the value vectors are weighted and summed based on these attention weights to obtain the self-attention output, as shown in Equation (5) The next step involves combining the input features before pooling with the self-attention output after pooling to form a residual connection, as illustrated in Equation (6). Finally, the output of the residual connection is subjected to a linear transformation to yield the final output, as indicated in Equation (7), where W_O is the linear transformation matrix for the output.

$$A_{ij} = \frac{Q_{\text{pool},i} K_{\text{pool},j}^{\top}}{\sqrt{d}} + E(i,j) + F(i,j) \tag{4}$$

$$\alpha_{ij} = \text{Softmax}(A_{ij}) \tag{5}$$

$$Z = \sum_{j} \alpha_{ij} V_{\text{pool},j} \tag{6}$$

$$\text{Output} = W_O Z_{\text{res}} \tag{7}$$

This process integrates the standard self-attention mechanism with pooling operations, residual connections, and relative positional embeddings, enhancing the model's positional information perception while preserving detailed information and reducing computational complexity.

IV. Experimental Setting

A. Datasets

1) INRIA Aerial Image Labeling Dataset (IAIL): The IAIL is a high-resolution aerial imagery dataset primarily designed for semantic segmentation tasks related to buildings. This dataset focuses on aerial imagery from large urban areas, covering multiple cities in Europe and the United States, encompassing diverse architectural styles and building densities. The dataset consists of aerial images with a spatial resolution of 0.3 meters and includes 180 large-scale images, each with a resolution of 5120×5120 pixels. The dataset spans approximately 360 square kilometers and includes cities such as Austin, Detroit, and Vienna. It provides pixel-wise semantic segmentation labels, categorizing each pixel as "building" or "non-building," making it particularly well-suited for binary building segmentation tasks. Although INRIA focuses specifically on buildings, the variation in building shapes and sizes introduces additional complexity to the segmentation task.

2) SpaceNet Dataset: The SpaceNet Dataset is an open-source dataset developed collaboratively by CosmiQ Works, Maxar, and NVIDIA, aiming to advance satellite image analysis techniques. The building extraction task within SpaceNet requires delineating building boundaries from high-resolution satellite imagery. SpaceNet 2, in particular, is dedicated to building segmentation tasks. The imagery in this dataset is sourced from commercial satellites such as WorldView-3, with resolutions ranging from 0.3 to 1 meter. Covering a total area exceeding 1000 square kilometers, the dataset includes

satellite images from major cities worldwide, including Atlanta, Chicago, Las Vegas, and Kansas City. The dataset provides precise polygon annotations (vector labels) for each building, which is particularly useful for building outline extraction. The diversity in building density and structure, especially in densely populated city centers, increases the difficulty of accurate building boundary extraction.

3) DeepGlobe Building Extraction Dataset (DBE): The DBE was introduced as part of the 2018 CVPR challenge to promote advancements in remote sensing image understanding. This dataset is designed explicitly for building extraction from high-resolution satellite imagery collected from various global regions. The dataset comprises satellite images with a resolution of approximately 0.5 meters, featuring thousands of images, each with a resolution of 650×650 pixels. These images encompass diverse geographical regions with varying levels of urbanization. Each image is accompanied by a binary segmentation mask, where building pixels are labeled as "1" and non-building pixels as "0," making it highly suitable for supervised semantic segmentation tasks. The dataset exhibits substantial variability in building size, shape, and spatial distribution, with certain regions featuring sparse and irregularly shaped buildings, posing significant challenges for model generalization.

4) WHU Building Dataset (WHUB): The WHUB, released by the School of Remote Sensing and Information Engineering at Wuhan University, is curated explicitly for building segmentation tasks. This dataset focuses on urban buildings and provides high-resolution aerial imagery from multiple cities in China. The aerial images have an exceptionally high resolution of 0.075 meters, enabling precise identification of building boundaries and structures. The dataset comprises 30,000 aerial images, covering a total urban area of 450 square kilometers and featuring diverse urban environments with varying building densities and structures. Each building in the images is annotated with detailed polygonal labels, making the dataset particularly suitable for building outline extraction and change detection. The high resolution of the WHU dataset ensures that intricate building details are preserved, imposing stringent requirements on segmentation algorithms' precision and detail extraction capabilities.

These datasets involve geospatial data with well-defined annotations, making them highly suitable for automated analysis of remote sensing imagery. They have extensive applications in various fields and are either open-access or freely available, making them valuable resources for remote sensing image classification and building segmentation research.

B. Evaluation Metrics

In this paper, the performance of the proposed MSRViT model is compared against baseline models using commonly adopted evaluation metrics in remote sensing image segmentation. These metrics include Mean Intersection over Union (mIoU), F1 Score, Dice Coefficient, Pixel Accuracy, and Boundary IoU. Prior to introducing these evaluation metrics, several fundamental concepts must be clarified.

True Positive (TP): A correctly predicted positive sample, where both the model and the ground truth classify the instance as positive.

False Positive (FP): A misclassified sample where the model predicts a positive instance, but the ground truth is negative.

True Negative (TN): A correctly predicted negative sample, where both the model and the ground truth classify the instance as unfavorable.

False Negative (FN): A misclassified sample where the model predicts a negative instance, but the ground truth is positive.

Precision: Precision measures the proportion of correctly predicted positive samples among all samples classified as positive. It evaluates the accuracy of optimistic predictions and helps reduce false positives. The formula for precision is given in Equation (8).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall: Recall (also known as sensitivity) quantifies the proportion of actual positive samples that the model has correctly identified. It focuses on the model's ability to recognize positive samples and helps reduce false negatives. The formula for precision is given in Equation (9).

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Intersection over Union (IoU): IoU evaluates the ratio of the overlapping area between the predicted segmentation and the ground truth to their total combined area. IoU is one of the most widely used metrics in image segmentation, particularly in building segmentation tasks, as it accurately quantifies the degree of overlap between predicted and actual building regions. The IoU calculation formula is provided in Equation (10).

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

Given these fundamental concepts, the following evaluation metrics are introduced:

1) Mean Intersection over Union (mIoU): mIoU is the mean IoU computed across different classes (e.g., buildings and background). In building segmentation tasks, mIoU is typically used for binary classification (building vs. non-building) and assesses the overall model performance by averaging the IoU values of both classes. Since mIoU considers building and background segmentation quality, it provides a more comprehensive evaluation. The formula for mIoU is given in Equation (3), with a value range from 0 to 1. A higher mIoU, closer to 1, indicates better segmentation performance.

2) F1 Score: The F1 Score is the harmonic mean of precision and recall, designed to balance both metrics. It is particularly suitable for imbalanced data scenarios, such as building segmentation, where building pixels are often significantly fewer than background pixels. The F1 Score provides a more reliable evaluation in such cases, addressing false positives and negatives. The formula for F1 Score is presented in Equation (4), with a value range from 0 to 1. A higher F1 Score signifies a better balance between precision and recall.

3) Dice Coefficient: The Dice Coefficient, similar to IoU, emphasizes the extent of overlap between the predicted and ground truth regions. Although commonly used in medical

TABLE I
MSRViT Performance Under Different Configurations

Model	Channels	Blocks	Heads	Param	FLOPs	mIOU
MSRViT-T	[96,192,384,768]	[1,2,5,2]	[1,2,4,8]	58.4	85.4	36.4
MSRViT-S	[96,192,384,768]	[1,2,11,2]	[1,2,4,8]	61.3	88.5	38.4
MSRViT-B	[96,192,384,768]	[2,3,16,3]	[1,2,4,8]	63.4	92.4	40.1
MSRViT-L	[144,288,576,1152]	[2,6,36,4]	[2,4,8,16]	66.9	98.9	41.2
MSRViT-H	[192,384,768,1536]	[4,8,60,8]	[2,4,8,16]	72.5	105.4	42.4

image segmentation, it is also effective in building segmentation tasks, particularly when evaluating models focusing on overlapping regions. The Dice Coefficient is more sensitive to small target regions, making it well-suited for tasks such as building extraction. The formula for the Dice Coefficient is provided in Equation (5), with a value range from 0 to 1. Like the F1 Score, a higher Dice Coefficient indicates better segmentation performance.

4) Pixel Accuracy: Pixel Accuracy represents the proportion of correctly classified pixels out of the total number of pixels in the image. While this metric is intuitive and straightforward, it may be influenced by class imbalance in building segmentation tasks—particularly when the number of non-building pixels significantly exceeds the number of building pixels. This can result in an overestimated performance of the model while failing to capture its true segmentation capability. Pixel Accuracy is more reliable in scenarios where buildings occupy a substantial portion of the image. The formula for Pixel Accuracy is provided in Equation (6), with a value range from 0 to 1. A higher value indicates greater classification accuracy at the pixel level.

5) Floating Point Operations (FLOPs): FLOPs measure the total number of floating-point operations required for a given computational task, indicating model complexity. A lower FLOP count generally indicates a more computationally efficient model.

6) Frames Per Second (FPS): FPS is a performance metric commonly used in deep learning and computer vision tasks. It represents the number of frames the model can process per second. FPS is crucial for evaluating model efficiency and real-time performance, particularly in applications such as video processing, real-time detection, and segmentation tasks. A higher FPS indicates a faster model processing speed.

Among the five primary evaluation metrics—mIoU, F1 Score, Dice Coefficient, and Pixel Accuracy—all have a value range of 0 to 1, with higher values indicating better segmentation performance. In contrast, FLOPs should be minimized to improve computational efficiency. Under comparable accuracy conditions, a higher FPS is preferred for enhanced real-time processing. To enhance clarity, mIoU, F1 Score, Dice Coefficient, and Pixel Accuracy are presented as percentage values in this paper.

C. Baseline Models

For remote sensing image recognition tasks, various semantic segmentation models exhibit distinct advantages in handling large-scale images, complex backgrounds, and fine-grained details. The following provides a brief introduction to the baseline models considered in this paper:

1) Fully Convolutional Networks (FCN): FCN is the first end-to-end convolutional neural network for semantic segmentation. By replacing fully connected layers with convolutional layers, FCN enables the network to process input images of arbitrary size while producing segmentation outputs of the exact dimensions. Although it establishes the foundation for semantic segmentation, its performance preserving fine-grained details and object boundaries remains limited.

2) Pyramid Scene Parsing Network (PSPNet): PSPNet introduces the Pyramid Pooling Module, which facilitates multi-scale feature fusion to capture global and local contextual information. This model is particularly effective in processing large-scale objects with rich semantic information, making it well-suited for scene-parsing tasks that require a strong understanding of global semantics.

3) High-Resolution Network (HRNet): HRNet excels in preserving high-resolution feature representations. By maintaining parallel multi-resolution feature streams and continuously fusing them, HRNet enhances segmentation accuracy for fine-grained details. Retaining high-resolution features throughout the network makes it particularly effective in handling complex scenes that demand precise segmentation results.

4) DeepLab V3+: DeepLab V3+ incorporates Atrous Convolution and the Atrous Spatial Pyramid Pooling (ASPP) module while introducing an encoder-decoder structure to refine boundary details. This model demonstrates outstanding performance in multi-scale contextual feature extraction and edge refinement, making it highly suitable for complex semantic segmentation tasks.

5) SegFormer: SegFormer is a lightweight segmentation model based on the Transformer architecture. It utilizes a Mix Transformer (MiT) encoder for long-range dependency modeling and employs a simple MLP decoder to enhance computational efficiency. Striking a balance between performance and efficiency, SegFormer is particularly well-suited for edge-devices deployment.

Each model has its strengths, excelling in different segmentation scenarios and demonstrating superior performance in various remote sensing applications.

D. Data Cleaning and Experimental Setup

The experiments were conducted on a system equipped with an Intel(R) Xeon(R) Bronze 3104 CPU @ 1.70GHz processor, 128GB of RAM, and two NVIDIA GeForce GTX TITAN XP GPUs, running on Ubuntu 22.04 operating system. The experiments were based on the PaddleRS 1.0 deep learning framework built on PaddlePaddle 2.4. During

training, the experiment employed a fixed interval learning rate scheduler with proportional decay and explored various decay strategies, incorporating a warmup phase to improve model performance. The learning rate was set to 0.0004, with a batch size of 32, and the AdamW optimizer was used. Additionally, the training process utilized the Momentum optimizer, combined with linear learning rate decay and Exponential Moving Average (EMA) methods. One hundred epochs were conducted to train the model fully, enhancing its performance and generalization capability.

V. Experimental and Analysis

A. Model Comparison

To comprehensively evaluate the performance of the proposed MSRViT model, Experiment 1 conducted comparative analyses across multiple remote sensing image segmentation datasets, including IAIL, SpaceNet, DBE, and WHUB. Several mainstream baseline models—such as FCN, PSPNet, HRNet, DeepLabV3+, and SegFormer—were included for benchmarking. The results are summarized in Table II, where the best-performing results for each metric are bolded, and the second-best results are underlined, allowing readers to identify the relative advantages easily. To provide a more intuitive representation of MSRViT's relative superiority, Figure 3 visualizes the normalized performance of all baseline models across all datasets (excluding FLOPs), using MSRViT as the reference baseline (normalized to 100). This comparative visualization clearly illustrates the performance differences among models across various evaluation dimensions. Such analysis not only enhances the interpretability of model performance but also offers valuable insights for future model design and selection.

The experimental results demonstrate that MSRViT consistently achieves the best or second-best performance across multiple datasets and evaluation metrics. Precisely, on the IAIL dataset, MSRViT attains the highest scores in mIoU, F1 Score, and Pixel Accuracy, surpassing the second-best model by 7.31%, 10.08%, and 9.82%, respectively. It also achieves the second-best performance in Dice Coefficient and FLOPs, with a marginal gap of 2.81% and 2.59% , respectively, compared to the best-performing model. However, regarding FPS, MSRViT lags behind the top-performing model by 9.66% . On the SpaceNet Dataset, MSRViT attains the highest scores in mIoU, F1 Score, and Pixel Accuracy, surpassing the second-best model by 1.85%, 14.29%, and 9.82%, respectively. It also achieves the second-best performance in Dice Coefficient and FLOPs, with a marginal gap of 2.80% and 2.91% , respectively, compared to the best-performing model. However, regarding FPS, MSRViT lags behind the top-performing model by 10.83% . On the DeepGlobe Building Extraction Dataset, MSRViT attains the highest scores in mIoU, F1 Score, and Pixel Accuracy, surpassing the second-best model by 2.32%, 14.28%, and 9.82%, respectively. It also achieves the second-best performance in Dice Coefficient and FLOPs, with a marginal gap of 3.62% and 2.59% , respectively, compared to the best-performing model. However, regarding FPS, MSRViT lags behind the top-performing model by 4.49% . On the WHU Building Dataset, MSRViT attains the highest scores in F1 Score, Dice Coefficient, and Pixel Accuracy, surpassing the second-best model by 9.09%, 3.60%, and 9.19%, respectively. Additionally, it achieves the second-best performance in mIoU and FLOPs, with a marginal gap of 3.63% and 2.59% , respectively, compared to the best-performing model. However, regarding FPS, MSRViT lags behind the top-performing model by 10.38% .

mIoU is the most commonly used evaluation metric for segmentation tasks, representing the average IoU across all classes. The superior mIoU score of MSRViT indicates its effectiveness in distinguishing different categories with high precision while maintaining a low sensitivity to misclassified regions. In remote sensing imagery, a higher mIoU implies improved differentiation of land cover types, such as buildings, roads, and vegetation.

F1 Score, as the harmonic mean of Precision and Recall, is particularly useful in scenarios where data distributions are imbalanced. The high F1 Score achieved by MSRViT suggests a balanced capability in minimizing false positives and negatives, ensuring robust segmentation of target regions while reducing misclassification errors. In remote sensing applications, the model effectively detects most targets (high recall) while maintaining low false alarms (high precision), enhancing segmentation reliability.

Pixel Accuracy measures the proportion of correctly classified pixels across the entire image. The high Pixel Accuracy achieved by MSRViT indicates a strong overall classification capability, effectively segmenting most land cover regions. However, in remote sensing segmentation, Pixel Accuracy alone may not be sufficient, as it tends to favor dominant classes while being less sensitive to small-scale objects. Therefore, it is often combined with mIoU and F1 Score to provide a more comprehensive assessment of model performance.

MSRViT exhibits slightly lower Dice Coefficient values compared to SegFormer across multiple datasets. This is primarily due to SegFormer's lightweight MLP decoder, which enhances segmentation efficiency and preserves boundary sharpness. Unlike complex decoders introducing excessive smoothing, SegFormer's design maintains finer details, improving Dice Coefficient scores. Additionally, the lightweight nature of MLP contributes to reduced FLOPs, explaining why MSRViT has a more considerable computational cost.

Due to the multi-scale feature integration and dense prediction strategy employed in MSRViT, its FPS is approximately 3–4 frames lower than the best-performing model. However, considering its superior mIoU, F1 Score, and Pixel Accuracy, this trade-off is deemed acceptable in achieving a balance between accuracy and computational efficiency.

MSRViT performs excellently in comparative experiments, demonstrating strong semantic segmentation capabilities and computational efficiency advantages. It achieves the best or near-best results in key metrics such as mIOU, F1 Score, Dice Coefficient, and Pixel Accuracy across multiple datasets, ensuring higher segmentation precision and more accurate category differentiation. While maintaining high performance, MSRViT optimizes computational cost compared to traditional methods, allowing it to sustain a high inference speed while preserving accuracy, making it suitable for real-world applications. MSRViT exhibits strong generalization ability, consistently adapting to different data distributions and ensuring stable segmentation performance,

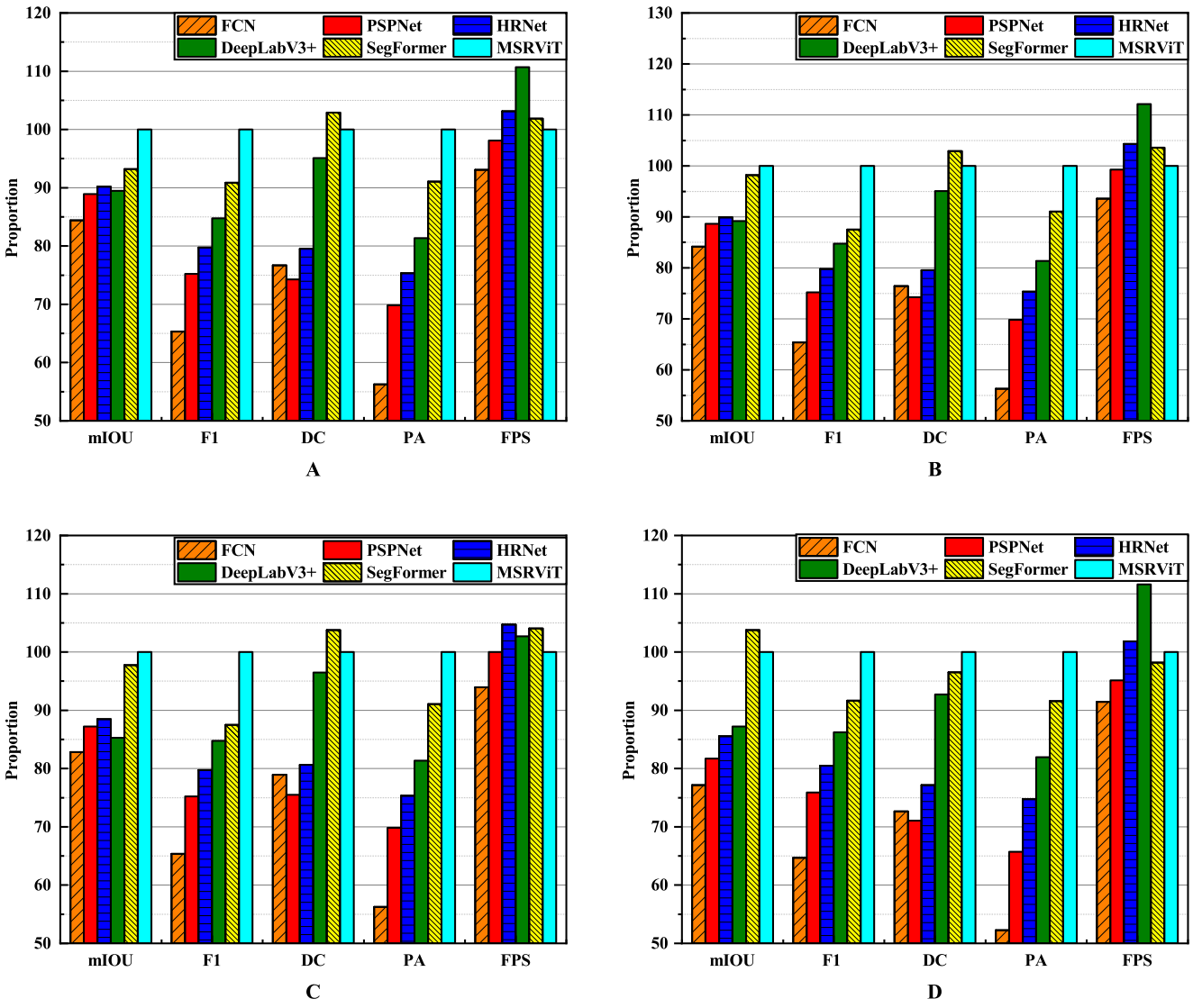

Fig. 3. Proportional Performance Comparison of Baseline Models

particularly on complex datasets. Compared to other models, it balances accuracy and speed well, making it well-suited for various high-demand semantic segmentation tasks.

B. Impact of Positional Encoding

Following Experiment 1, which demonstrated the superior performance of MSRViT in image segmentation tasks, Experiment 2 further investigates the influence of different positional encoding strategies on model performance. Specifically, this paper evaluates MSRViT on the WHUB dataset using four distinct positional encoding strategies: No Positional Encoding (No pos), Absolute Positional Encoding (Abs. pos), Joint Relative Positional Encoding (Joint rel. pos), and Decomposed Relative Positional Encoding (Decomposed rel. pos). The analysis focuses on three key metrics: mIoU, FLOPs, and FPS.The comparative results are summarized in Table III

The experimental results reveal that Absolute Positional Encoding and Joint Relative Positional Encoding enhance mIoU compared to No Positional Encoding, with Joint Relative Positional Encoding achieving the best performance, reaching 41.2%. This improvement can be attributed to several factors:

1) More Effective Spatial Relationship Modeling

Compared to No Positional Encoding and Absolute Positional Encoding, Relative Positional Encoding captures the relative positional information between pixels or features, enabling the model to better learn local and global spatial relationships. In image segmentation tasks, the relative positioning of pixels is crucial, as adjacent pixels are often highly correlated. Joint Relative Positional Encoding provides a more precise modeling capability, enhancing segmentation accuracy.

2) Improved Local and Global Feature

Joint Relative Positional Encoding integrates horizontal, vertical, and channel-wise relative positional information, allowing the model to perceive features at different scales. This approach enhances the model's ability to capture local structures, such as edges and textures, and strengthens global information integration, leading to improved segmentation precision.

3) Enhanced Generalization Capability

A key advantage of Relative Positional Encoding is its ability to adapt to images of varying sizes and structures without being constrained by fixed input dimensions. This means that MSRViT, when utilizing Joint Relative Positional

TABLE II
COMPARISON OF DIFFERENT POSITION ENCODINGS IN MSRVIT

Datasets	Metrics	Model					
		FCN	PSPNet	HRNet	DeepLab V3+	SegFormer	MSRViT
IAIL	mIOU	34.82	36.68	37.21	36.91	38.44	**41.25**
	F1 Score	56.31	64.84	68.75	73.06	78.31	**86.20**
	Dice Coefficient	59.15	57.28	61.35	73.33	**79.35**	77.12
	Pixel Accuracy	47.38	58.80	63.44	68.50	76.67	**84.20**
	FLOPs	280.20	264.10	223.80	201.60	**96.40**	98.90
	FPS	14.80	15.60	16.40	**17.60**	16.20	15.90
SpaceNet	mIOU	31.00	32.65	33.12	32.85	36.16	**36.83**
	F1 Score	50.15	57.71	61.19	65.02	67.13	**76.72**
	Dice Coefficient	52.46	50.98	54.60	65.26	**70.62**	68.64
	Pixel Accuracy	42.17	52.33	56.46	60.97	68.24	**74.94**
	FLOPs	237.50	235.50	199.10	179.40	**85.80**	88.30
	FPS	13.10	13.90	14.60	**15.70**	14.50	14.00
DBE	mIOU	33.17	34.93	35.44	34.15	39.14	**40.05**
	F1 Score	53.66	61.75	65.47	69.58	71.83	**82.09**
	Dice Coefficient	57.12	54.62	58.35	69.83	**75.08**	72.36
	Pixel Accuracy	45.12	56.00	60.41	65.23	73.01	**80.18**
	FLOPs	266.83	251.50	213.12	191.98	**91.80**	94.18
	FPS	14.00	14.90	**15.60**	15.30	15.50	14.90
WHUB	mIOU	29.90	31.65	33.15	33.78	**40.20**	38.74
	F1 Score	52.21	61.27	64.97	69.60	74.00	**80.73**
	Dice Coefficient	54.30	53.13	57.67	69.30	72.15	**74.75**
	Pixel Accuracy	41.90	52.70	59.95	65.73	73.45	**80.20**
	FLOPs	261.90	249.57	211.49	190.51	**91.10**	93.46
	FPS	15.00	15.60	16.70	**18.30**	16.10	16.40

Encoding, can better generalize across different image resolutions and spatial configurations, resulting in more robust performance on the WHUB dataset.

4) Balanced Computational Complexity and Performance While Decomposed Relative Positional Encoding reduces the number of parameters and computational complexity, the experimental results indicate that this simplification leads to a drop in mIoU. This decline is likely due to the loss of global relative positional information caused by decomposition. In contrast, Joint Relative Positional Encoding strikes an optimal balance between computational cost and performance improvement, ensuring a significant boost in accuracy without incurring excessive computational overhead.

5) Better Adaptation to the Transformer-Based Self-Attention Mechanism The core of Transformer-based architectures lies in the self-attention mechanism. The incorporation of Relative Positional Encoding enhances attention efficiency by avoiding reliance solely on absolute position indexing. Joint Relative Positional Encoding allows the model to dynamically focus on critical regions within the image rather than being constrained by fixed coordinates, thereby facilitating more precise object segmentation.

The superior performance of Joint Relative Positional Encoding on the WHUB dataset can be attributed to its advantages in spatial relationship modeling, local-global feature integration, generalization enhancement, computational efficiency, and adaptability to Transformer-based architectures. MSRViT achieves higher segmentation accuracy by effectively utilizing positional information while maintaining reasonable computational costs. These factors collectively contribute to the significant improvement in the mIoU metric when employing Joint Relative Positional Encoding.

TABLE III
THE IMPACT OF DIFFERENT RESIDUAL POOLING STRATEGIES OF MSRVIT ON MODEL PERFORMANCE

Positional embeddings	mIOU	FLOPs	FPS
No pos	40.80	**96.60**	16.40
Abs. pos	40.90	97.00	16.50
Joint rel.pos	**41.20**	107.20	16.50
Decomposed rel. pos	41.00	97.60	**16.70**

C. Impact of Different Pooling Strategies

Building on the findings of Experiments 1 and 2, this paper further examines the effect of different pooling strategies on MSRViT. Specifically, the experiments conducted on the WHUB dataset evaluate the impact of the following pooling configurations: Without Residual Pooling (w/o), X Residual, X Residual + Full Q/K Residual, X Residual + Q/K Pooling and X Residual + Full Q/K Pooling + Q/K Residual.

Table IV presents the results, indicating that incorporating X Residual increases mIoU from 39.5% to 40.2%, suggesting

TABLE IV
COMPARISON OF MSRVIT WITH OTHER SEGMENTATION MODELS

Residual Pooling	mIOU	FLOPs	FPS
w/o	39.50	**86.60**	15.10
X_Residual	40.20	92.50	15.30
X_Residual + full Q/K Residual	40.70	95.60	15.30
X_Residual + Q/K Pooling	41.00	96.90	15.50
X_Residual+full Q/K Pooling + Q/K Residual	**41.20**	98.90	**15.70**

that this strategy enhances the model's feature extraction capability. The introduction of Q/K Residual further improves segmentation boundary precision, as residual connections provide an additional information flow, enabling the network to learn spatial relationships between different regions more effectively. Q/K Residual allows the model to better focus on critical regions while reducing interference from irrelevant areas.

With the incorporation of Q/K Pooling, the spatial dimension is compressed, allowing the model to capture better local structural information, which is particularly beneficial for refining target edges and intricate details in image segmentation. Additionally, reducing the Q/K dimension in attention computation lowers computational complexity, thereby improving inference speed while preventing feature representations from becoming overly dispersed. The pooling operation enables Q/K embeddings to integrate a broader contextual understanding, making it easier for the model to focus on distant pixels and enhancing overall segmentation performance. As a result, more continues to improve progressively.

X Residual facilitates efficient feature propagation across layers, mitigating information loss. Q/K Residual enhances positional information learning within the attention mechanism, making the self-attention distribution more structured and meaningful. Q/K Pooling reduces computational costs while improving local feature aggregation, making the model more efficient and enhancing global contextual awareness.

Combining these three strategies yields the most significant performance improvement, demonstrating their complementary nature within Transformer-based architectures and their efficacy in improving segmentation accuracy. Notably, the final configuration, which integrates residual and pooling strategies, achieves a mIoU of 41.2%, matching the best-performing positional encoding strategy.

VI. Conclusions

This paper addresses the challenges of high-resolution remote sensing image segmentation by proposing improvements based on the MViT. Traditional Transformer architectures face significant computational and memory challenges when processing high-resolution images. To tackle these issues, this work introduces shift-invariant positional embedding and residual pooling connections, which effectively enhance the model's multiscale feature extraction and detail preservation capacity.

The shift-invariant positional embedding more accurately captures spatial information by decomposing positional distances, improving the model's understanding of complex remote-sensing image structures. Meanwhile, the residual pooling connection maintains computational efficiency while preserving as much detail as possible that could be lost during pooling operations. The proposed architecture demonstrates significant advantages in complex remote sensing image segmentation tasks by integrating the optimized MViT with standard dense prediction frameworks such as U-Net and FPN.

Experimental results show that the proposed methods effectively address the computational complexity issues inherent in Transformer architectures for high-resolution image segmentation, achieving a balanced trade-off between performance and efficiency. This work provides valuable insights for applying Transformer models in remote sensing image analysis and lays a foundation for developing more efficient and accurate segmentation models.

References

[1] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[2] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[3] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 042–12 051.

[4] H. Ajmal, S. Rehman, U. Farooq, Q. U. Ain, F. Riaz, and A. Hassan, "Convolutional neural network based image segmentation: a review," *Pattern Recognition and Tracking XXIX*, vol. 10649, pp. 191–203, 2018.

[5] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International Symposium on Computer-based Medical Systems (CBMS)*. IEEE, 2020, pp. 558–564.

[6] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[7] 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.

[8] C. Wei, S. Ren, K. Guo, H. Hu, and J. Liang, "High-resolution swin transformer for automatic medical image segmentation," *Sensors*, vol. 23, no. 7, p. 3420, 2023.

[9] Y. Gao, X. Liu, J. Li, Z. Fang, X. Jiang, and K. M. S. Huq, "Lft-net: Local feature transformer network for point clouds analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 2158–2168, 2022.

[10] K. Zeng, Q. Ma, J. Wu, S. Xiang, T. Shen, and L. Zhang, "Nlfftnet: A non-local feature fusion transformer network for multi-scale object detection," *Neurocomputing*, vol. 493, pp. 15–27, 2022.

[11] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 589–598.

[12] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[13] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2998–3008.

[14] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 094–12 103.

[15] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[16] 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.

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[18] F. Wang, J. Li, Q. Yuan, and L. Zhang, "Local–global feature-aware transformer based residual network for hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[19] F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, F. Chen, and L. Fu, "Residual local feature network for efficient super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 766–776.

[20] Y. Gao, X. Liu, J. Li, Z. Fang, X. Jiang, and K. M. S. Huq, "Lft-net: Local feature transformer network for point clouds analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 2158–2168, 2022.

[21] M. Ding, A. Qu, H. Zhong, Z. Lai, S. Xiao, and P. He, "An enhanced vision transformer with wavelet position embedding for histopathological image classification," *Pattern Recognition*, vol. 140, p. 109532, 2023.

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