SwinAttention-HRNet: Optimizing Remote Sensing Image Recognition and Classification

Guinan Wu, Qinghong Wu*

***Abstract*—Remote sensing image recognition and classification are important tasks widely studied in fields such as agriculture, industry, mining, urban planning, and disaster monitoring, divided into two major aspects: recognition and classification. With the continuous development of deep learning models, remote sensing image recognition and classification tasks have also made constant progress in recent years. Traditional deep learning models, such as those represented by CNNs, although capable of mining relevant information within images, utilize less global and contextual information in remote sensing images, resulting in poor performance. While Transformer models and their variants based on attention mechanisms can utilize global information, the large redundant parameters lead to high computational overhead. To address these challenges, this paper proposes a novel SwinAttention-HRNet model, which includes three key optimization points: HRNet-SE, improved attention computation mechanism, and Patch Fusion. Specifically, HRNet-SE reduces the complexity of the model with minimal performance loss, the improved attention computation mechanism enables the model to utilize information from adjacent windows, enhancing classification capability, and Patch Fusion enhances the model's understanding of low-resolution images. Multiple experiments have demonstrated that SwinAttention-HRNet outperforms current mainstream models with a comprehensive performance improvement of 3.34%, providing better support for remote sensing image classification and recognition tasks.**

***Index Terms*—Remote Sensing, HRNet, Dual Stream Swin Transformer, Self-Attention**

I. Introduction

REMOTE sensing images are Earth surface image data obtained through remote sensing technology, typically collected by satellites, aircraft, or other sensors, and then converted into image form to display information such as surface features, topography, vegetation coverage, and land use. They have wide applications, including environmental monitoring, urban planning, agriculture, forestry, geological exploration, and disaster monitoring. By analyzing remote sensing images, changes in the Earth's surface, distribution of resources, and environmental conditions can be understood, providing important information support for scientific research and decision-making[1, 2].

However, remote sensing image recognition poses significant challenges due to data complexity, spectral information diversity, inconsistent spatial resolution, noise, and diversity of land cover types. Traditional classification methods include pixel-based classification, target recognition and supervised classification, and object-based classification. Although traditional methods are universal and operationally strong, emerging technologies such as deep learning gradually replace traditional methods due to their ability to handle complex data and improve classification accuracy[3].

Deep learning-based methods mainly involve extracting features using models such as Convolutional Neural Networks (CNNs)[4, 5] and Recurrent Neural Networks (RNNs)[6, 7] suitable for temporal features. However, these models need help with parameter tuning and better generalization. With the successful application of attention mechanisms and Transformer models based on them in image recognition, applying attention mechanisms to remote sensing image recognition becomes a natural choice. Therefore, this paper proposes a deep learning model focused on remote sensing images, called SwinAttention-HRNet, which innovates in the following aspects:

1) Improving HRNet to HRNet-SE as a Backbone to reduce model complexity and time overhead without a performance decrease exceeding 7%.
2) Improving the attention computation mechanism in Swin Transformer to utilize information from adjacent windows and enhance classification capability.
3) Proposing a new feature map fusion mechanism called Patch Fusion to enhance the model's understanding of details in low-resolution images.

Based on these innovations, this paper aims to provide a deep learning model with higher performance and lower computational overhead to better address remote sensing image recognition and classification tasks.

II. Related Work

A. HRNet

The High-Resolution Network(HRNet) is a deep convolutional neural network designed to address tasks such as image segmentation and pose estimation, drawing inspiration from high-resolution image processing[8]. Its structure is illustrated in Fig. 1. By integrating information from multiple resolutions, the network improves the effectiveness of image processing. Its core concept lies in constructing a network structure with multi-resolution feature maps, known as the "high-resolution" structure. The architecture of HRNet primarily comprises two modules: high-resolution representation learning and high-resolution fusion. The former aims to extract multi-scale feature representations by constructing a feature pyramid, while the latter combines feature maps of different resolutions to obtain superior feature representations. The experiments conducted in this study and those of other researchers indicate that HRNet, as the backbone network, exhibits a more robust feature representation capability than ResNet and EfficientNet[9]. The multi-branch convolutional neural network structure of HRNet facilitates

Manuscript received November 17, 2024; revised April 12, 2025.

Guinan Wu is a Postgraduate of School of Electronic Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China. (e-mail: 2608942251@qq.com).

Qinghong Wu* is a Professor of School of Electronic Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide e-mail: aswqh@163.com).

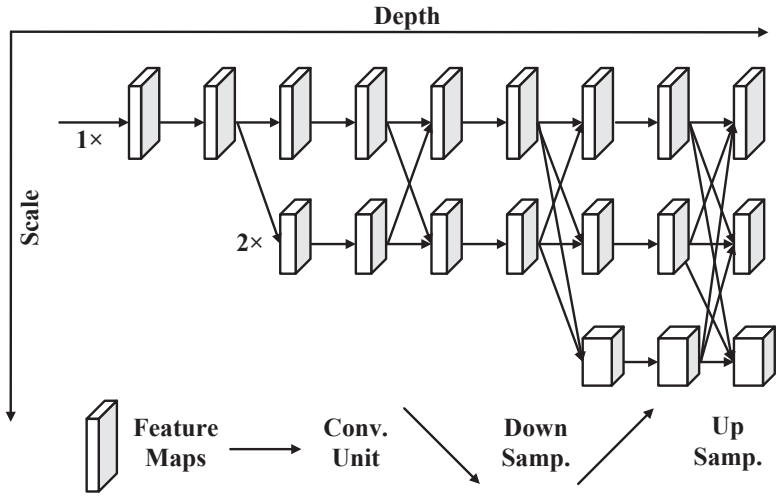

Fig. 1. The Structure of HRNet

information exchange between branches, thereby enhancing the fusion of features at different levels and consequently improving feature representation. In contrast, ResNet and EfficientNet, based on single-branch network structures, struggle to integrate components at different levels effectively and exhibit limitations when dealing with high-resolution images. Hence, selecting HRNet as this study's backbone network is paramount.

B. Attention Mechanism

The attention mechanism is a widely employed technique in artificial intelligence that mimics human behavior in information processing. This mechanism enables models to focus on specific parts of input data, thereby enhancing the performance of models in handling complex tasks. Within the attention mechanism, models can dynamically allocate attention based on the importance of input data, implying that during the processing of input sequences, models can, akin to humans, assign varying degrees of importance to different parts based on different time steps or positions, rather than uniformly averaging processing[10].

This mechanism encodes input sentences or images through encoding operations. It uses learnable weight matrices $W_Q \in R^{D\times d_k}$, $W_K \in R^{L\times d_k}$ and $W_V \in R^{L\times d_k}$ to transform each input into three matrices: Query, Key, and Value.

1) Query is used to compute attention weight vectors, where each query vector at each position undergoes a dot product operation with all positions in the sequence to calculate its correlation with other positions.
2) Key is used to compute attention weight vectors, where each key vector at each position undergoes a dot product operation with all positions in the sequence to calculate its correlation with other positions.
3) Value is used to compute weighted sum vectors, where value vectors are weighted based on the correlation weights of queries and keys to generate the final output representation.

Subsequently, attention weights for each position regarding queries are obtained by computing the dot product between query vectors and key vectors to measure the correlation between queries and keys and transforming them into a probability distribution via the softmax function. Finally, the final output representation is derived by multiplying and summing the value vectors with attention weights. This computation process can be represented by the Equation (1):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In remote sensing images, objects may appear at varying scales, positions, and lighting conditions. The attention mechanism assists models in focusing attention on the most important regions of images, thereby improving the accuracy of object detection and recognition[11–13]. By dynamically adjusting the attention weights of different regions, models can better handle challenging situations, such as occluded targets or blurred images.

C. Dual-Stream Swin Transformer

The Swin Transformer is a deep learning model based on the Transformer architecture, specifically designed for image classification tasks. Its structural diagram is depicted in Fig. 2. Proposed by Microsoft Research Asia, it was initially introduced in early 2021 to effectively handle large-scale information within images, thereby achieving leading performance in large-scale image classification tasks[14]. The Swin Transformer employs a series of optimization strategies to optimize image processing tasks. Firstly, it introduces a hierarchical attention mechanism, dividing the image into different blocks and performing self-attention operations at various levels to effectively capture global and local information, thus enhancing image classification accuracy. Secondly, to handle large-scale images, it utilizes a Window Attention mechanism, segmenting input features into different windows and performing self-attention operations within each window to improve the efficiency of processing large-scale images. Despite its powerful capabilities, the Swin Transformer is designed with efficient computational and parameter control mechanisms. It is characterized by low computational complexity and parameter volume, thereby demonstrating outstanding performance in large-scale image classification tasks and possessing better scalability[15, 16].

In the traditional Swin Transformer, input images are divided into different blocks, and self-attention mechanisms are applied to these blocks to capture global and local

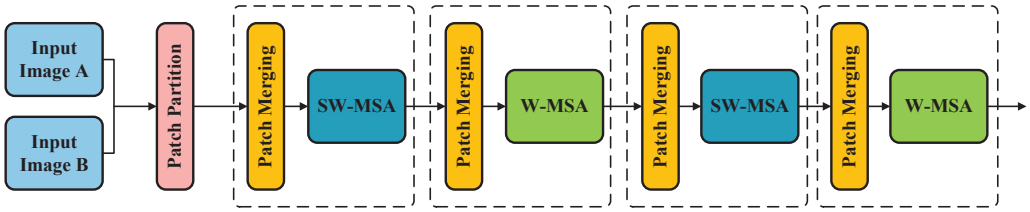

Fig. 2. The Structure of Swin Transformer

information. In the Dual-Stream Swin Transformer, however, two parallel streams are introduced to process image information of different resolutions[17]. One stream handles the original input image, while the other processes a low-resolution version. Through attention mechanisms or feature fusion modules across streams, features of different scales and resolutions are interactively learned and fused, better-utilizing information from different streams to enhance image understanding accuracy. This approach aims to fully capture semantic information within images and improve the perception of multiscale features[18].

Compared to the single-stream Swin Transformer, the Dual-Stream Swin Transformer typically achieves better performance in image understanding tasks due to its introduction of a dual-stream structure, enabling more careful consideration and utilization of multiscale information within images, thereby enhancing task accuracy and generalization capability[19]. However, the Dual-Stream Swin Transformer still inherits some deficiencies of Transformer models, such as high computational costs and memory consumption[20].

III. SwinAttention-HRNet

A. Overall Process

This paper proposes a model specifically designed for remote sensing image classification tasks, named SwinAttention-HRNet (SA-HRNet), based on the Dual-Stream Swin Transformer and incorporating improved attention mechanisms and the HRNet model. Its overall structure is illustrated in Fig.3. The mathematical representation of the model as a whole is described by Equations (2) to (7).

$$\hat{z}^{l} = PatchMerging\left(z^{l-1}\right) \tag{2}$$

$$\hat{z}'^{l} = Patchmerging\left(z'^{l-1}\right) \tag{3}$$

$$z^{l} = SWBlock(\hat{z}^{l} \oplus \hat{z}'^{l}) \tag{4}$$

$$\widehat{z}'^{l+1} = WBlock(\hat{z}^{l} \oplus z'^{l-1}) \tag{5}$$

$$\hat{z}^{l+1} = WBlock(z^{l} \oplus z^{l-1}) \tag{6}$$

$$z^{l+1} = ABS(\hat{z}^{l+1} - \widehat{z}'^{l+1}) \tag{7}$$

In the above Equations, z^{l-1} and z'^{l-1} denote the feature maps produced by HRNet-SE, with $\oplus$ denoting the concatenation operation. The resulting feature map, represented as z^{l+1}, signifies the output of the improved Dual-Stream Swin-Transformer.

B. HRNet-SE

Due to HRNet's necessity to concurrently handle multiple-resolution feature maps, its computational resource consumption is relatively high compared to some lightweight network architectures. This could potentially lead to limitations in resource-constrained scenarios. This paper proposes improvements to HRNet to reduce its computational load.

The original HRNet extensively employs 1×1 convolution operations to reduce feature map dimensions and enhance network expressiveness. However, the widespread use of 1×1 convolutions in HRNet increases computational costs, especially when these operations are applied to feature maps of multiple resolutions, necessitating computations at each resolution and increased computational overhead. Additionally, although 1×1 convolution operations have relatively fewer parameters, they incur a certain computational burden during feature fusion and channel adjustment, particularly in deep networks like HRNet, potentially escalating computational resource consumption, especially during training[21].

To mitigate this, the paper introduces Res-Conv as an alternative to the 1×1 convolution operation, drawing inspiration from residual computation to further enhance network computational efficiency. The Res-Conv module consists of two branches: one branch conducts 1×1 convolution, 3×3 depthwise convolution, and another 1×1 convolution on input features, concatenates the output with input features, performs shuffle operation, and obtains the final output features. Through computation, it is observed that when the channel numbers of input and output feature maps are $\mathbb{C}$, the time complexity of 1×1 convolution is $\mathcal{O}(\mathbb{C}^2)$. In contrast, the time complexity of 3×3 depthwise convolution is $\mathcal{O}(9\mathbb{C})$. Therefore, when $\mathbb{C}$>5, the computational load of 1×1 convolution exceeds that of 3×3 depthwise convolution, such operations can effectively reduce HRNet's computational load. By substituting Res-Conv for the 1×1 convolution in HRNet, HRNet-SE is obtained, whose overall network structure remains consistent with HRNet, as illustrated in Fig. 1.

C. SW-Block and W-Block

The traditional Transformer relies on Multi-head Self-Attention (MSA) for computation, resulting in exceedingly high computational complexity. In contrast, the Swin Transformer introduces Window-based Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA). The original Swin-Block comprises these two attention mechanisms. This paper enhances the algorithm by decomposing the Swin-Block containing SW-MSA and W-MSA. According to the design, they are respectively named SW-Block and W-Block. The structures of SW-Block and W-Block are illustrated in Equations(8) and (9). The calculation

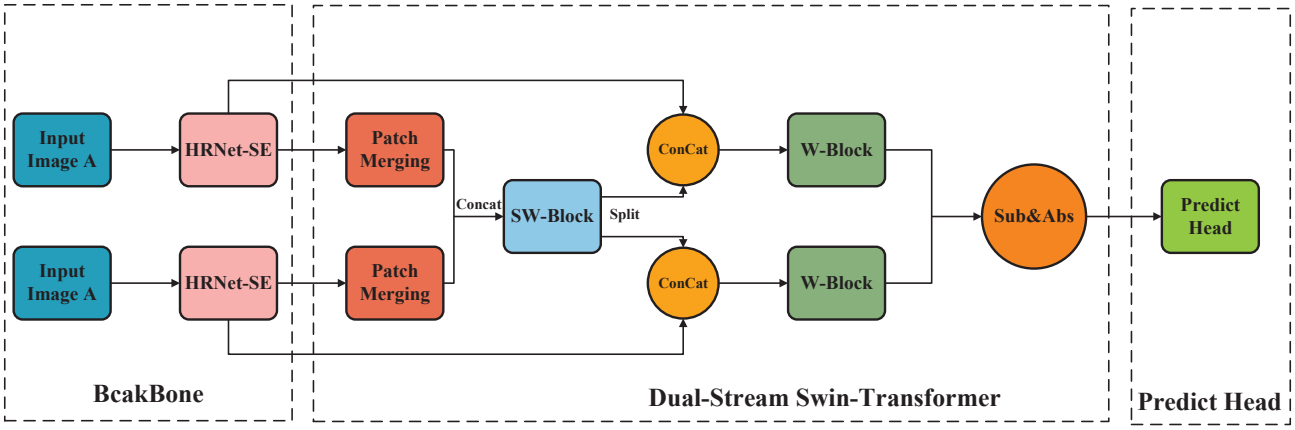

Fig. 3. Overview of the Proposed Model's Workflow

flow chart is in Fig. 4.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k + B}}\right)V \quad (8)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

Where Q, K, V represent the Query, Key, and Value matrices, d denotes the dimensionality of Query/Key, and B signifies the relative position encoding. Notably, SW-MSA utilizes relative position encoding B, while W-MSA does not.

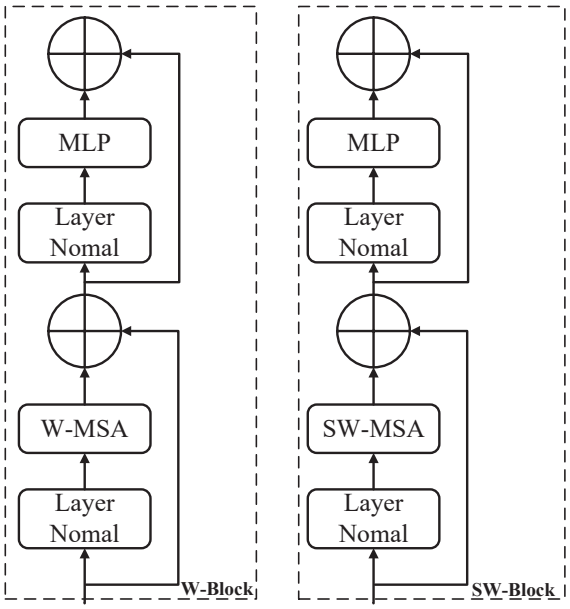

Fig. 4. Calculation Flow of W-Block and SW-Block

The core idea of W-Block is to divide the input image into multiple non-overlapping windows and perform multi-head self-attention operations within each region separately. This enables the model to focus on information within local regions, enhancing computational efficiency. Based on W-Block, SW-Block introduces a cyclic shift operation to enhance the model's ability to capture long-range dependencies. Specifically, SW-Block conducts a cyclic shift on the input feature map before performing W-Block. Each window can obtain information from its neighboring windows through this process, enlarging the model's receptive field. These characteristics of W-Block and SW-Block enable the Dual-Stream Swin Transformer to effectively capture local and global information while maintaining high computational efficiency.

The improved mechanism ensures that each window's attention is only related to its neighboring windows, further reducing computational complexity. The combination of this decomposition and window mechanism makes the Swin Transformer an ideal choice for handling high-resolution visual inputs. It can maintain high Precision while exhibiting higher computational efficiency and lower memory consumption.

D. Patch Fusion

The Patch Fusion approach adopted in this paper originates from Patch Merging, a downsampling method used to reduce spatial resolution and increase image channel capacity.

As a key operation in the Swin Transformer model, Patch Merging aims to merge local feature maps according to certain rules into larger feature maps. This operation is typically performed after a series of windowed self-attention operations[22, 23]s. In the Swin Transformer, the input image is first divided into a set of non-overlapping local feature map blocks (patches), each containing a set of feature vectors. Then, W-MSA operations are performed on these feature map blocks to capture local self-attention relationships. However, the Swin Transformer introduces the patch merging operation to capture global information and long-range dependencies better. Its basic idea is to merge adjacent feature map blocks into larger ones, such that each merged feature map block contains information from neighboring regions.

This paper proposes an innovative approach, Patch Fusion, as illustrated in Fig. 5. Unlike Patch Merging, Patch Fusion maintains the same channel count regarding feature map size but reduces spatial resolution by half. The improved Patch Fusion enhances the detail and quality of low-resolution images.

E. Predict Head

In change detection, the output head is pivotal in generating a change map delineating differences between two images. The design of the output head leverages the feature maps generated by the W-Block, which inherently encapsulates differential information between the two images. Following Fully Convolutional Network (FCN) principles,

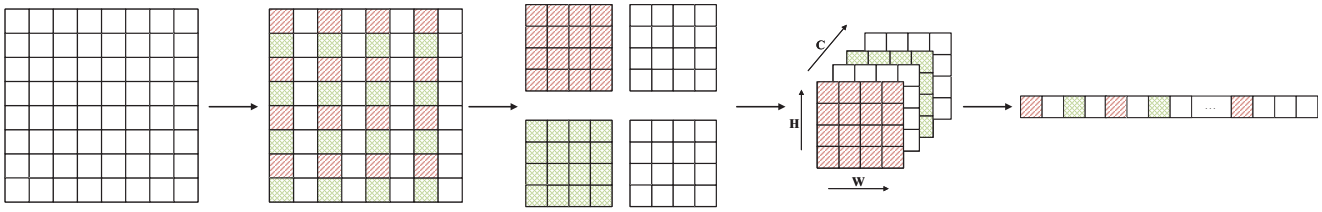

Fig. 5. The Calculation of Patch Fusion

the output head involves up-sampling and convolution on the feature maps to preserve spatial and semantic information.

The primary objective of the output head is to execute pixel-wise subtraction between corresponding positions in the two images, yielding a change map. Each pixel value in the change map signifies the degree of difference between related positions in the two images, facilitating tasks such as analyzing changes in remote sensing images, detecting object appearance or disappearance, and more.

Visualization of the change map provides insights into alterations between the two images. Larger pixel difference values denote significant changes, while smaller values indicate less conspicuous alterations. The change map furnishes a visual means to scrutinize and comprehend patterns and trends in image changes.

Consequently, the output head assumes a crucial role in change detection, processing feature maps generated by the W-Block to produce a change map. This provides an intuitive understanding and analytical capability to interpret image changes.

IV. Dataset and Baseline

This section is organized into subheadings to provide a succinct and precise depiction of experimental results, their interpretation, and the empirical conclusions drawn.

A. Dataset

1) LEVIR-CD (Land-Use and Vegetation-Change Detection Dataset): LEVIR-CD is a publicly available dataset specifically designed for change detection in remote sensing imagery, with the primary goal of supporting land use and vegetation change analysis. The dataset comprises multi-temporal remote sensing images from various regions and is primarily applied to change detection tasks in urban and forested areas. The imagery in LEVIR-CD is sourced from multiple remote sensing platforms, including Google Earth, Sentinel-2, and Landsat, with high spatial resolution (30 meters or higher). It provides rich annotations of changes, covering categories such as buildings, roads, forests, and vegetation. By including multi-temporal imagery, the dataset enables researchers to perform time-series analyses to detect land cover changes. A key feature of LEVIR-CD is its high-quality change annotations, making it well-suited for the training and evaluation of change detection algorithms, with particular relevance to urban development and forest monitoring.

2) SpaceNet: SpaceNet is an open-access dataset for object detection in remote sensing images, with a particular focus on the detection and analysis of buildings. Initiated through collaborations among several organizations, the dataset provides annotated data derived from high-resolution satellite imagery, with spatial resolutions reaching up to 0.3 meters, covering multiple countries and regions. The images in SpaceNet are primarily sourced from DigitalGlobe satellites, and annotations include detailed building footprints and locations. One of the distinguishing features of the dataset is its highly detailed building annotations, which make it a critical resource for advancing object detection, semantic understanding, and deep learning algorithm development in the domain of remote sensing.

3) RESISC45 (Remote Sensing Image Scene Classification 45): RESISC45 is a benchmark dataset designed for scene classification tasks in remote sensing imagery. Developed by the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, the dataset contains 45 scene categories, each represented by 700 images with a resolution of 256×256 pixels. These scenes encompass a wide range of land cover types, including forests, cities, farmlands, deserts, and water bodies. The images are sourced from various geographic regions across the globe, reflecting strong geographical diversity and environmental complexity. RESISC45 is characterized by its balanced class distribution and is suitable for training and evaluating deep learning models in multi-class scene classification problems. It is widely used in remote sensing image classification research, particularly for tasks involving automatic recognition of diverse land cover scenes, and presents notable challenges in boundary recognition and class discrimination.

Collectively, these three datasets hold significant value in the field of remote sensing image analysis. LEVIR-CD focuses on change detection, SpaceNet emphasizes object detection, and RESISC45 is dedicated to scene classification. Together, they provide rich data resources and evaluation benchmarks that drive the advancement of automated detection and classification techniques in remote sensing.

B. Evaluation Indicators

Several performance metrics were employed when evaluating deep learning models for data classification, including accuracy, recall, intersection over union (IoU), F1 score, overall accuracy (OACC), and Kappa coefficient. Before formally describing the metrics mentioned above, it is necessary to introduce some prerequisite concepts, as shown in Table 1, which include the following concepts for classification problems:

1) True Positive (TP): Samples originally positive and classified as positive.
2) False Negative (FN): Samples originally positive but classified as negative.
3) False Positive (FP): Samples originally negative but classified as positive.
4) True Negative (TN): Samples originally negative and classified as negative.

After introducing the prerequisite concepts, a better understanding of the Precision and Recall metrics can be achieved. Precision measures the proportion of samples predicted as positive by the model that are truly positive. Its calculation process is depicted in Equation (10). A high precision value indicates that the model has few false positives among the samples labeled as positive, meaning the model rarely misclassifies negatives as positives. This is particularly important for tasks sensitive to false positives. Recall measures the proportion of all true positives successfully predicted as positive by the model. Its calculation process is illustrated in Equation (11). A high value of Recall indicates that the model can capture more true positives, meaning the model misses fewer true positives. This is particularly important for tasks sensitive to missed detections.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

After analyzing the fundamental evaluation metrics, the following section introduces the evaluation metrics used in this paper.

1) F1 Score: As evident from the preceding formula, a trade-off relationship exists between Precision and Recall. If strict judgment regarding the negative class is crucial, Precision is given more emphasis; conversely, if comprehensive coverage of the positive class is paramount, Recall is prioritized. A common method is to utilize the F1-score, which considers the balance between Precision and Recall, thereby enabling a more comprehensive evaluation of model performance. Its calculation process is depicted in Equation (12).

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{12}$$

When both the model's Precision and Recall are high, the F1 score tends to be high as well. However, when either Precision or Recall is low, the F1 Score is significantly affected, especially when there is a substantial difference between the two.

2) IoU: IoU is a commonly used metric for evaluating tasks such as object detection and semantic segmentation, measuring the degree of overlap between the model's predicted region and the target region. In classification tasks, IoU is often used to assess the localization performance of the model for each class. Referring to Fig. 6, its calculation process is depicted in Equation (13):

$$IoU = \frac{Intersection(A, B)}{Union(A, B)} \tag{13}$$

The IoU value ranges between 0 and 1. IoU = 0 indicates no overlap between the predicted region and the actual region, while IoU = 1 indicates complete overlap. Typically, when IoU exceeds a certain threshold, the prediction result is considered correct. This threshold can be adjusted based on the task's requirements and specific application scenarios.

3) OAAC: OAAC is a metric used to evaluate the performance of multi-class classification models, considering the overall classification accuracy of the model across all classes. Its calculation process is depicted in Equation (14):

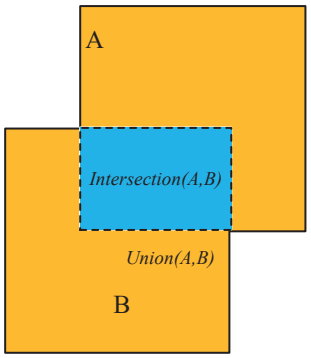

Fig. 6. Illustration of IoU

$$\text{OAAC} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \tag{14}$$

TP_i represents the True Positives of the $i-th$ class, FN_i represents the False Negatives of the $i-th$ class, and n is the total number of classes. OAAC represents the proportion of correctly classified samples across all classes to the total number of samples. It is a comprehensive performance metric reflecting the overall classification capability of the model across all classes. Unlike single accuracy metrics, OAAC considers the importance of each class and is more discriminative. For example, suppose a model has a high classification accuracy in one class but performs poorly in others. In that case, a single accuracy metric may give this model a high evaluation, but OAAC can better reflect the overall performance of the model.

4) Kappa: Kappa is a statistical metric used to evaluate the performance of classification models, particularly suitable for handling situations of class imbalance. It measures the consistency between the classifier's predictions and the actual situation, considering the impact of accuracy caused by chance. Its calculation process is depicted in Equation (15):

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{15}$$

P_o is the observed accuracy, and P_e is the expected accuracy. When calculating P_e, it is typically assumed that the probability of randomly selecting labels for each class is proportional to their occurrence in the dataset. Therefore, the calculation formula for P_e is the sum of the product of the predicted probabilities of each class in the real data. The value of the Kappa coefficient typically ranges from -1 to 1. A value of 1 indicates perfect agreement between the classifier's predictions and the actual situation. A value of 0 indicates agreement between the classifier's predictions and random chance, meaning no ability beyond random prediction. A value less than 0 indicates that the classifier's predictions are worse than random chance, possibly due to the classifier's erroneous predictions exceeding random predictions.

5) Mean Average Precision (mAP): mAP is a widely used performance evaluation metric in object detection tasks, particularly in multi-class object detection. It serves as a key indicator for assessing the overall performance of a detection model. The mAP metric incorporates both precision and recall across different object categories. For each class, the

Average Precision (AP) is computed, reflecting the model's ability to balance precision and recall. The mAP is then obtained by averaging the AP values across all classes. The calculation process is formally defined in Equation (16).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{16}$$

The aforementioned evaluation metrics approach 1, indicating better model performance.

C. Baseline

To test the performance of the SA-HRNet in this paper, multiple mainstream image recognition models were selected in the experimental section, namely:

1) Spatial–Temporal Attention Neural Network(STANet)[24]: STANet is a neural network model designed for image segmentation tasks. It incorporates techniques such as attention mechanisms and dilated convolutions to enhance image segmentation performance. The attention mechanism enables the network to automatically focus on image regions that are more important for specific tasks. At the same time, dilated convolutions expand the receptive field of convolutional operations, improving the accuracy of image segmentation.

2) SNUNet[25]: SNUNet is based on the U-Net architecture but has been improved and optimized, including mechanisms such as multimodal feature fusion and structured interaction. While SNUNet is a neural network model used for medical image segmentation tasks, specifically designed to address precise segmentation of organs and lesions in medical images, it can also be trained specifically for remote sensing image recognition.

3) Deep Structured Interconnected Fusion Network(DSIFN)[26]: DSIFN is a deep neural network model designed for image recognition tasks aimed at addressing feature fusion and interaction issues in multimodal image recognition. The DSIFN model features characteristics such as multimodal feature fusion and structured interaction, making it suitable for scenarios requiring the fusion of multiple modalities of information and efficient interaction processing.

4) Big Transfer(Bit)[27]: The Bit model is an image classification model proposed by Google, employing the Transformer architecture and pre-trained on large-scale data to achieve outstanding image classification and detection performance. The Bit model can apply features learned on large-scale data to specific image recognition tasks through transfer learning, exhibiting advantages such as parameter efficiency and versatility.

In addition to evaluating the overall performance of the models, this paper also aims to assess the performance of the improved HRNet-SE as a backbone. Therefore, the experimental section also selects the following mainstream backbone networks for comparison:

1) Residual Network (ResNet)[28]: ResNet is a highly popular deep convolutional neural network commonly used as a backbone in image recognition tasks. The fundamental unit of ResNet is the residual block, which includes skip connections. Skip connections allow the input to bypass one or more layers and then be added to the output. This design enables the network to learn residuals, i.e., the differences between the input and output, rather than directly learning the output. The number following ResNet typically denotes the depth of the network, i.e., the number of residual blocks in the network. For example, ResNet-18 used in this experiment refers to a ResNet network containing 18 residual blocks, while ResNet-64 refers to a ResNet network containing 64 residual blocks. Generally, ResNet-18 to ResNet-34 are considered shallow ResNet models suitable for small-scale datasets and computational resources. On the other hand, ResNet-50 and higher models are more suitable for handling more complex tasks and larger datasets.

2) Residual Networks with Extremely-Deep Networks (ResNeXt)[29]: ResNeXt is a convolutional neural network structure proposed by Microsoft Research, which is an improvement upon ResNet. In ResNeXt, the branches within each residual block are divided into multiple groups, with shared parameters within each group and independent parameters between different groups. This design allows the network to more efficiently utilize parameters, thus enhancing the network's expressive power. The number following ResNeXt represents both the depth and width of the network. Unlike ResNet, where complexity is increased by adding depth, in ResNeXt, complexity is increased by adding the number of branches, thereby increasing the network's width. Like ResNet, this experiment employs both shallow and deep variants of ResNeXt, namely ResNeXt-18 and ResNeXt-64.

3) High-Resolution Network(HRNet)[8]: HRNet is a high-resolution network used for image recognition and other computer vision tasks. Unlike traditional CNNs, HRNet is dedicated to effectively capturing multiscale information while maintaining high-resolution features. This makes it perform exceptionally well in tasks requiring retaining details and being sensitive to multiscale information. The number following HRNet typically indicates the depth of the network. Like other networks, this number represents the residual blocks included in HRNet, i.e., the number of network layers. For example, the HRNet-18 used in this experimental section represents an HRNet network with a depth 18, consisting of 18 residual blocks. This number is commonly used to indicate the complexity and depth of the network; deeper networks may have more parameters and stronger representation capabilities, but they also require more computational resources and longer training times.

D. Experimental Setup

The experimental setup comprised an Intel(R) Xeon(R) Bronze 3104 CPU @ 1.70GHz processor, 128GB of memory, and two NVIDIA GeForce GTX TITAN XP GPUs. The operating system utilized was Ubuntu 22.04, with experiments conducted using PaddleRS 1.0 based on PaddlePaddle 2.4. The training involved a learning rate scheduler with uniformly spaced fixed-rate decay, warm-up operations, and a learning rate 0.0004. Adam optimizer was employed with a batch size of 32. Momentum optimizer, linear learning rate decay, and Exponential Moving Average (EMA) enhanced training. The training spanned 100 epochs to enhance model performance and generalization. Data augmentation strategies included random cropping, flipping, rotation, blurring, adjacent image swapping, and color jittering to enhance

data diversity and model generalization. The presented experimental data represents the average of five independent experiments, with the best results highlighted in bold and the second-best results underscored.

V. Result and Analysis

A. Comparative Analysis of Backbone Networks

The first group of experiments focuses on analyzing the performance differences of various backbone network architectures on three representative remote sensing change detection datasets: LEVIR-CD, SpaceNet, and RESISC45. The result is show in Tab I. Fig. 7 shows the individual Backbone results in percentage for a better demonstration of the model effect using the LEVIR-CD dataset as an example. The comparison includes popular architectures such as ResNet, ResNeXt, HRNet, and HRNet-SE. Five evaluation metrics—F1-score, IoU, OAAC, Kappa, and mAP—are used to comprehensively assess model performance from multiple perspectives. Specifically, F1-score reflects the balance between precision and recall, IoU measures the accuracy of spatial overlap between predictions and ground truth, OAAC represents the total classification accuracy, Kappa indicates model consistency and robustness, and mAP evaluates the model's detection performance across varying thresholds. On the LEVIR-CD dataset, HRNet-18 slightly outperforms in F1-score (92.43) and Kappa (91.12), achieving improvements of 0.20% and 0.16% over HRNet-SE, respectively. This suggests that HRNet-18 is better at enhancing the recall capability and maintaining prediction consistency for changed regions. In contrast, HRNet-SE shows superiority in IoU (90.62), OAAC (98.19), and mAP (97.89), with respective gains of 0.14%, 0.48%, and 0.36%. The notably high OAAC close to 98.2% demonstrates that HRNet-SE is more effective at distinguishing changed from unchanged areas, showing stronger discrimination capabilities at the full-image level. On the SpaceNet dataset, the performance

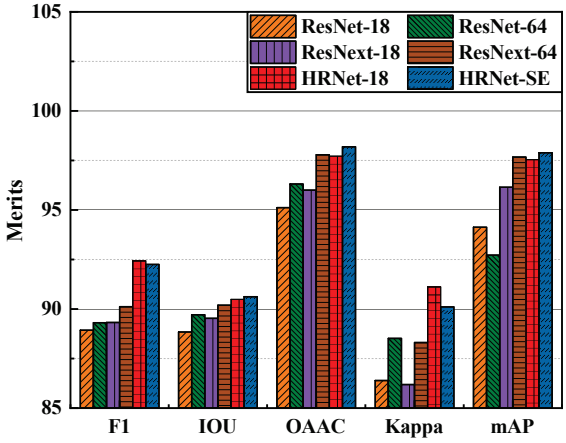

Fig. 7. Proportional Evaluation Metrics of HRNet-SE and Other Backbone Model

of HRNet-18 and HRNet-SE is closely matched. HRNet-18 takes the lead in F1-score (90.03) and Kappa (87.77), with improvements of 0.20% and 0.55%, indicating its advantage in edge discrimination and prediction stability. Meanwhile, HRNet-SE performs better in OAAC (95.64) and mAP (94.32), with respective increases of 0.40% and 0.12%, highlighting its stronger ability to focus on targets in complex backgrounds. On the RESISC45 dataset, which contains diverse land-cover types and large-scale variations, HRNet-SE consistently achieves superior performance across all metrics. It improves F1-score to 96.29% (+0.85%), IoU to 93.61% (+0.14%), OAAC to 96.43% (+1.57%), Kappa to 93.08% (+2.25%), and mAP to 95.34% (+0.17%). The substantial improvement in Kappa suggests that HRNet-SE offers higher consistency and generalization in multi-class

TABLE I
BackBone Network Comparison

DataSets	Backbone	Merits				
		F1	IoU	OAAC	Kappa	mAP
LEVIR-CD	ResNet-18	88.93	88.84	95.12	86.39	94.13
	ResNet-64	89.31	89.71	96.31	88.52	92.72
	ResNext-18	89.32	89.53	96	86.19	96.15
	ResNext-64	90.12	90.19	97.79	88.31	97.67
	HRNet-18	92.43	90.49	97.72	91.12	97.54
	HRNet-SE	92.25	90.62	98.19	90.11	97.89
SpaceNet	ResNet-18	86.62	86.53	92.65	84.14	91.24
	ResNet-64	86.99	87.38	93.81	86.22	93.25
	ResNext-18	87.00	87.20	93.50	83.95	93.56
	ResNext-64	87.78	87.85	95.25	86.01	94.21
	HRNet-18	90.03	88.14	95.18	88.75	95.32
	HRNet-SE	89.85	88.26	95.64	88.77	94.32
RESISC45	ResNet-18	91.86	91.77	93.26	89.24	91.34
	ResNet-64	92.26	92.67	93.49	91.44	93.02
	ResNext-18	92.27	92.48	94.17	89.03	94.23
	ResNext-64	93.09	93.17	95.02	91.22	94.80
	HRNet-18	95.48	93.48	95.94	94.13	95.12
	HRNet-SE	96.29	93.61	96.43	93.08	95.34

TABLE II
DIFFERENT MODEL COMPARISON

DataSets	Model	Merits				
		F1	IoU	OAAC	Kappa	mAP
LEVIR-CD	STANet	88.21	88.73	96.12	87.69	94.37
	SNUNet	89.61	89.91	97.01	88.51	95.26
	DSIFN	90.00	90.64	96.57	88.54	94.33
	Bit	91.02	91.59	98.71	89.63	97.1
	SA_HRNet	91.52	91.22	99.21	92.12	96.74
SpaceNet	STANet	81.51	81.99	88.81	81.03	87.20
	SNUNet	82.80	83.08	89.64	81.78	88.02
	DSIFN	83.16	83.75	89.23	81.81	87.16
	Bit	84.10	84.63	91.21	82.82	89.72
	SA_HRNet	84.56	84.29	91.67	85.12	89.39
RESISC45	STANet	82.48	82.97	89.88	82.00	88.24
	SNUNet	83.79	84.07	90.71	82.76	89.08
	DSIFN	84.16	84.76	90.30	82.79	88.21
	Bit	85.11	85.64	92.30	83.81	90.80
	SA_HRNet	85.58	85.30	92.77	86.14	90.46

recognition. An OAAC over 96% further indicates excellent pixel-level classification performance, making HRNet-SE suitable for change detection systems where classification accuracy is critical. In conclusion, HRNet-SE exhibits stable and superior performance across all three datasets. It demonstrates significant improvements in OAAC and Kappa, indicating not only higher classification accuracy but also enhanced consistency under complex conditions. This makes HRNet-SE a highly promising backbone network for change detection tasks in remote sensing.

B. Comparative Analysis of Model Architectures

The second set of experiments builds upon the best-performing backbone each dataset to further compare different change detection architectures, including STANet, SNUNet, DSIFN, Bit, and SA-HRNet.The result is show in TabII. Fig. 8, again using LEVIR-CD as an example, shows how SA-HRNet compares to other baseline models by percentage. The evaluation is again based on five metrics, with a particular focus on the impact of the spatial attention mechanism introduced in SA-HRNet.

On the LEVIR-CD dataset, SA-HRNet achieves the best results in F1-score (91.52), OAAC (99.21), and Kappa (92.15), outperforming the Bit model by 0.55%, 0.51%, and 2.81%, respectively. Notably, the nearly 3% increase in Kappa highlights the superior stability and consistency of SA-HRNet in predicting change regions. The OAAC of 99.2% indicates that SA-HRNet achieves almost pixel-perfect classification. Although the Bit model slightly surpasses in IoU (91.59) and mAP (97.10), with improvements of 0.41% and 0.13%, it remains strong in contour fitting and the detection of difficult targets.

On the SpaceNet dataset, SA-HRNet again outperforms all other models in F1-score, OAAC, Kappa, and mAP, with gains of 0.55%, 0.14%, 1.56%, and 1.23%, respectively. The mAP of 89.29% particularly reflects SA-HRNet's effectiveness in identifying small targets and fine-grained changes. While the Bit model leads in IoU (84.64), surpassing SA-HRNet by 0.61%, this advantage illustrates its better geometric fitting of change regions.

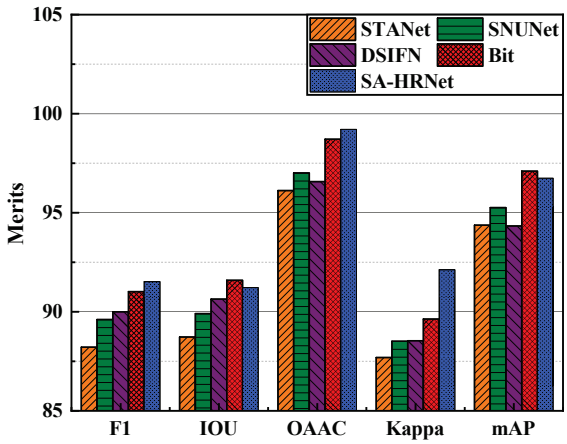

Fig. 8. Proportional Evaluation Metrics of SA-HRNet and Other Baseline$_{M}odel$

On the RESISC45 dataset, SA-HRNet maintains its leading performance across all metrics except IoU. It improves F1-score to 85.58% (+0.55%), OAAC to 92.77% (+0.51%), Kappa to 86.14% (+2.78%), and mAP to 90.46% (+0.51%). The significant increase in Kappa again emphasizes the model's stability and generalization in complex multi-class change scenarios. These results make SA-HRNet particularly well-suited for remote sensing imagery with overlapping land cover types and complex spatial structures.

Overall, SA-HRNet demonstrates the most comprehensive performance among all tested models. In all three datasets, it achieves Kappa improvements exceeding 1.5% and OAAC values consistently approaching or surpassing 99%, reflecting its high classification stability and strong adaptability to various scenes. Its robust handling of complex boundary changes, spatial texture variations, and multi-scale targets makes it an ideal architecture for real-world remote sensing

TABLE III
ABLATION EXPERIMENT

Model	Merits				
	F1	IoU	OAAC	Kappa	mAP
DS-HRNet	86.23	90.39	94.61	82.72	92.18
SA-PM-HRNet	91.16	90.72	97.45	91.33	95.22
SA-O-HRNet	91.82	90.92	99.52	92.42	89.32
SA-HRNet	91.55	91.32	99.14	91.90	97.60

applications.

C. Ablation experiments

Finally, in the experimental design, this paper conducted ablation experiments to compare the effectiveness of various mechanisms in SpaceNet. Three variants of SA-HRNet were designed for the improvements:

1) SA-O-HRNet: HRNet-SE was replaced with the original HRNet to verify the performance of HRNet-SE.
2) DS-HRNet: The improved SW-Block and W-Block were replaced with the original SW-MSA and W-MSA to verify the effectiveness of SW-Block and W-Block.
3) SA-PM-HRNet: The proposed Patch Fusion was replaced with the original Patch Merging to verify the effectiveness of Patch Fusion.

The results of the ablation experiments are shown in Table III. The IoU metrics for all variants remained at the same level, indicating that selecting HRNet and its improvements as the Backbone was reasonable. A robust backbone can provide more discriminative feature representations, aiding the model in accurately identifying object boundaries or segmentation regions and thus improving the IoU metric. The performance drop of DS-HRNet was the most significant, especially in the OAAC metric, indicating that SW-Block and W-Block can enhance the model's ability to obtain information from neighboring windows, thereby improving classification performance. Although SA-O-HRNet performed best among all variants, the original HRNet has a relatively high FLOPs metric, as mentioned earlier. At the same time, the improved HRNet-SE in this paper reduced the number of parameters without significant performance degradation, thereby reducing the computational overhead of the model.

VI. CONCLUSIONS

This paper presents a deep learning approach for remote sensing change detection, quantitatively analyzing and identifying surface changes in two distinct time-period remote sensing images. Conventional change detection methods, typically relying on differences between two frames, are susceptible to noise, occlusions, and intricate changes. This paper introduces a change detection method based on the Dual-Stream Swin-Transformer network to address these challenges. Employing a dual-stream architecture and techniques like the Swin block enhances the extraction of change information between images, improving feature extraction capability and accuracy. Compared to traditional methods, this proposed technique demonstrates superior adaptability and Precision and is more suitable for complex change detection scenarios. The experimental results affirm the significant performance enhancements of this method compared to other change detection networks, highlighting improved feature extraction, adaptability, and accuracy. Its effective applications include updating geospatial data, disaster trend assessment, land cover/land use monitoring and advanced intelligent Earth observation satellite endeavors.

REFERENCES

[1] F. Van der Meer, "Remote-sensing image analysis and geostatistics," *International Journal of Remote Sensing*, vol. 33, no. 18, pp. 5644–5676, 2012.

[2] H. Ghassemian, "A review of remote sensing image fusion methods," *Information Fusion*, vol. 32, pp. 75–89, 2016.

[3] J. A. Richards, J. A. Richards *et al.*, *Remote sensing digital image analysis*. Springer, 2022, vol. 5.

[4] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 173, pp. 24–49, 2021.

[5] J. Song, S. Gao, Y. Zhu, and C. Ma, "A survey of remote sensing image classification based on cnns," *Big earth data*, vol. 3, no. 3, pp. 232–254, 2019.

[6] V. D. Marri, V. N. R. P, and C. M. R. S, "Rnn-based multispectral satellite image processing for remote sensing applications," *International Journal of Pervasive Computing and Communications*, vol. 17, no. 5, pp. 583–595, 2021.

[7] M. I. Lakhal, H. Çevikalp, S. Escalera, and F. Ofli, "Recurrent neural networks for remote sensing image classification," *IET Computer Vision*, vol. 12, no. 7, pp. 1040–1045, 2018.

[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[9] ——, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[10] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[11] S. Ghaffarian, J. Valente, M. Van Der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sensing*, vol. 13, no. 15, p. 2965, 2021.

[12] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[13] 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.

[14] 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.

[15] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khan, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, p. 1860, 2023.

[16] A. M. Ali, B. Benjdira, A. Koubaa, W. El-Shafai, Z. Khan, and W. Boulila, "Vision transformers in image restoration: A survey," *Sensors*, vol. 23, no. 5, p. 2385, 2023.

[17] D. Yu, Q. Li, X. Wang, Z. Zhang, Y. Qian, and C. Xu, "Dstrans: Dual-stream transformer for hyperspectral image restoration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3739–3749.

[18] M. Mao, R. Zhang, H. Zheng, T. Ma, Y. Peng, E. Ding, B. Zhang, S. Han *et al.*, "Dual-stream network for visual recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 346–25 358, 2021.

[19] Y. Li, N. Miao, L. Ma, F. Shuang, and X. Huang, "Transformer for object detection: Review and benchmark," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107021, 2023.

[20] S. Zuo, Y. Xiao, X. Chang, and X. Wang, "Vision transformers for dense prediction: A survey," *Knowledge-Based Systems*, vol. 253, p. 109552, 2022.

[21] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[22] 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.
[23] X. Gu, S. Li, S. Ren, H. Zheng, C. Fan, and H. Xu, "Adaptive enhanced swin transformer with u-net for remote sensing image segmentation," *Computers and Electrical Engineering*, vol. 102, p. 108223, 2022.
[24] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.
[25] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
[26] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.
[27] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.
[28] 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.
[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.