ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

MPCR: Masked Autoencoder with Patch Merger Based on Convolutional Neural Network and Re-Attention for Polarimetric SAR Image Classification

Cui Yanyu[1], Li Yingying[1], Wang Jianlong[1,*]

[1]*School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454003, China*

Abstract: *In order to solve the deficiency of vision transformer (ViT) in local feature extraction and effectively integrate global and local information, the manuscript proposes a model called Masked autoencoder with Patch merger based on Convolutional neural network and Re-attention (MPCR) for polarimetric SAR image classification. The CNN is used to divide the input image into patches, which effectively extracts local features and enhances the ability of the model to capture details. However, as the number of transformer network layers increases, the traditional attention mechanism leads to information degradation, manifested by the attention maps of each layer tends to be consistent. To address this, re-attention is introduced. By dynamically adjusting the weights, the model can still maintain the diversified capture of information at a deep level, so as to better handle the complex input data. A simple merging operation is introduced between two consecutive transformer encoder layers to further improve computational efficiency. The operation effectively reduces the redundant calculation and reduces the computational complexity. The results show that the proposed method not only significantly improves the classification accuracy, but also effectively reduces the computational complexity when processing PolSAR images.*

Keywords: *Polarimetric Synthetic Aperture Radar; Masked Autoencoder; Vision Transformer; Convolutional Neural Network*

1. Introduction

Synthetic aperture radar (SAR) is an active microwave imaging remote sensing sensor, which transmits electromagnetic waves and receives echo signals to obtain surface information [1]. It has the ability to observe the earth all day and all weather, and can penetrate clouds and vegetation [2,3]. It can still provide stable and reliable observation data under complex meteorological conditions. Therefore, SAR plays an important role in both military and civilian fields, especially in battlefield surveillance and reconnaissance, target recognition, land use and disaster monitoring. With the continuous development of technology, the imaging resolution and data processing ability of SAR are continuously improved, and the application fields are also continuously expanded, providing more accurate ground information and decision support. Polarimetric SAR (PolSAR) is further developed from SAR and can work in different polarization combinations. By transmitting and receiving horizontally polarized H wave and vertically polarized V wave, the scattering echo obtained contains four polarization components, namely HH, HV, VH and VV [4]. The polarization combination of electromagnetic wave is relatively sensitive to the physical and geometric characteristics of ground objects, which improves the ability of PolSAR to obtain ground target information to a certain extent.

PolSAR image classification refers to the process of dividing all pixels in the image into a certain category according to the polarimetric scattering characteristics [5,6]. The process not only constitutes the core component of PolSAR image understanding and interpretation technology, but also lays a solid foundation for the subsequent ground object recognition and evaluation work. Traditional polsar image classification methods are mainly based on statistical distribution, polarization decomposition and machine learning. PolSAR image classification based on statistical distribution means to build a suitable statistical distribution model to realize the classification of different features. Wishart distribution is one of the outstanding results of statistical distribution method, but its classification performance is too dependent on the quality of the center pixel of each category, and the calculation of Wishart distance is also very time-consuming [7]. With the continuous deepening of research, researchers realized that only relying on a single statistical model was not enough to fully express the complex ground object scattering mechanism of polarimetric SAR, and the method based on

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

polarimetric target decomposition came into being. For example, Freeman decomposition, Cameron decomposition, Huynen decomposition have been proposed [8,9]. However, these methods can only determine the scattering type of pixels, and can not be divided into specific ground object categories. PolSAR image classification based on machine learning learns rich feature representations from massive PolSAR data, and models the relationship between data information and categories. As a large amount of PolSAR data is obtained, traditional classification methods cannot address the demand. At the same time, deep learning technology continues to develop, showing a strong ability of data processing and feature extraction. In recent years, deep learning methods have been gradually introduced into the field of PolSAR image classification, which has brought more efficient solutions for PolSAR image interpretation [10,11].

Convolutional neural network (CNN) is a widely used deep learning method in PolSAR image classification [12]. Zhou et al. [13] carried out pioneering research on PolSAR image classification method based on CNN. This method considers the spatial characteristics of POLSAR image, and captures the spatial information at different scales through multi-layer convolution operation. Zhang et al. [14] proposed a complex convolutional neural network for SAR image interpretation, which makes full use of the amplitude and phase information of SAR image. A feature selection algorithm was proposed by Yang et al., which innovatively combines 1-D CNN and Kullback-Leibler distances [15]. In addition, by considering the performance of feature combination rather than the contribution of a single feature, the method can better capture the potential relationship between different features. It enhances the effectiveness of the classification task while improving the efficiency of feature selection. Dong et al. [16] explored the application of neural architecture search in PolSAR for the first time on the basis of CNN and proposed a differentiable architecture search method. The method optimizes the architectural parameters and additionally introduces a search method in the complex domain to better fit the data form of polsar images. The innovation provides a new and effective way for polsar image classification, and promotes the technological progress in the field. However, CNN tends to learn the local structure of the data, and it is difficult to capture the global context relationships and long-distance dependencies in the input data, which limits the application of CNN-based methods in polsar image classification.

Transformer was originally applied in the field of natural language processing (NLP), which is a neural network based on self-attention (SA) mechanism [17]. Dosovitskiy et al. [18] proposed a vision transformer (vit) model based on transformer structure, which achieved the most advanced performance on multiple image recognition benchmarks. In view of the excellent performance of ViT in CV, it is also gradually applied to the field of PolSAR. Dong et al. [19] explored the application of ViT in PolSAR image classification for the first time, and its powerful global feature extraction capability can effectively improve the classification performance. Since then, the researches of ViT in PolSAR image classification have been more and more in-depth. Wang et al. [20] pointed out that the training data of ViT is too large and the model is too complex to be directly applied to PolSAR image classification tasks, so ViT is combined with hybrid convolutional tokenization and the same modules are recombined to form parallel blocks. The model reduces the complexity of parameters and the requirement of computing power, and significantly speeds up the training and prediction speed. ViT improves classification performance by effectively utilizing global information, which helps to solve some of the limitations of CNN networks. However, it typically relies on large amounts of labeled data for optimal performance, which poses a challenge in PolSAR image classification tasks with scarce labeled data. As a self-supervised learning method, masked autoencoder (MAE) based on ViT architecture provides a new idea to deal with the problem [21]. MAE can fully use unlabeled data for pre-training, and only requires a small amount of labeled data for fine-tuning, which significantly reduces the dependence on labeled data. Fuller et al. [22] applied MAE in the field of polsar image processing, which effectively alleviated the problem of insufficient labeled data and achieved excellent performance in downstream classification tasks. From the existing work on MAE, there are relatively a few researches in the polsar field. Nevertheless, the potential of MAE in polsar image classification tasks cannot be denied.

Polsar data is complex and contains a lot of information, which makes it difficult for most deep learning methods to fully capture its characteristics. With its unique convolution operation, CNN can sensitively capture the local features in the image. However, CNN has some limitations in capturing the global information of images, which limits its application performance in the field to some extent. As a new deep learning framework, ViT can better solve the problems faced by CNN in processing global information through self-attention mechanism. However, ViT is not as good at processing local features as CNN. Therefore, in PolSAR image classification task, how to use deep learning method to balance global information and local feature extraction has become a problem worth discussing. In addition, the

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

labeling of PolSAR data is highly dependent on expert knowledge, which also poses a limitation for deep learning models that require large amounts of labeled data for training.

Based on the above analysis, in order to better utilize the local feature extraction capability of cnn and the global information processing capability of vit to better capture polsar data characteristics under the limited polsar labeled data, a method called masked autoencoder with patch merging based on convolutional neural network and Re-attention (MPCR) is proposed. In the method, MAE is selected as the basic model, which not only makes effective use of unlabeled data, but also integrates the characteristics of ViT. By combining the advantages of CNN and MAE, it can realize more accurate classification of PolSAR images. The main contributions of the manuscript are as follows:

(1) Convolutional Neural Network for Image Blocking. CNN is used for image blocking to replace the tokenization of input data by linear projection in ViT. It can utilize the powerful local feature extraction capabilities of CNN to automatically capture more complex patterns from images. In addition, the hierarchical nature of CNN gives it a powerful ability to analyze image content on a multi-scale, and build increasingly abstract but content-rich feature representations layer by layer. It not only improves the ability of the model to perceive local features but also provides a more detailed information for the subsequent learning of the transformer architecture.

(2) Re-attention. Based on the structure of multi-head self-attention (MHSA), a learnable approach called Re-attention (RA) is introduced. The method aims to facilitate information exchange between different attention heads and effectively increase the diversity of the attention graph. RA enables the model to continuously optimize the feature representation at a deeper level, avoiding the learning stagnation caused by the simplification of attention distribution. It increases the ability of the model to capture complex patterns and offers the possibility of constructing deeper levels of vit.

(3) Patch merger. With the increase of model depth, processing a large number of tokens will lead to a great consumption of computing resources, and may also reduce the speed and effect of model training. To solve the problem, the patch merging module is adopted to dynamically evaluate and merge similar or redundant tokens in the early layers of transformer. The module aims to reduce the number of tokens entering the deeper network, so as to optimize the computational efficiency and performance of the model.

The rest of the literature is organized as follows. Section II briefly introduces the relevant technical background and describes the proposed method in detail. In Section III, the experimental results are fully presented and analyzed. Section IV gives a discussion of the factors that affect the effectiveness of the model. Finally, Section V summarizes the study and presents an outlook for future work.

2. The proposed method

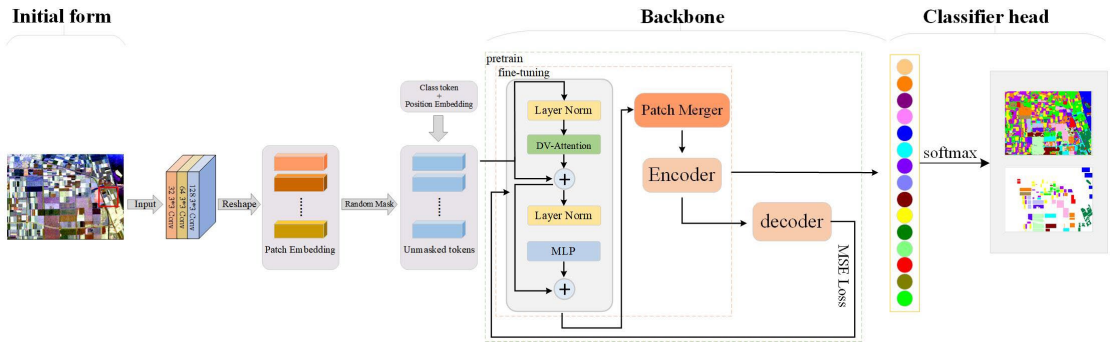

Fig.1 The schematic diagram of MPCR network

The manuscript proposes a hybrid framework designed to combine the advantages of CNN and ViT to capture local-global information in images. Fig 1 shows a schematic diagram of Masked autoencoder with Patch merger based on Convolutional neural network and Re-attention (MPCR). In the framework, firstly, the traditional linear projection method in ViT is replaced by the segmented operation of CNN, which is used for the tokenization of input data. By utilizing the excellent local feature extraction capability of CNN, the method can automatically capture complex and rich abstract feature representations from images, providing detailed information for subsequent processing. Secondly, in order to overcome the problem that the attention distribution tends to be uniform in the deep ViT model, which leads to the reduction of effective feature capture, Re-attention is introduced. It

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

can promote the exchange of information between different attention heads and ensure that the model continuously optimizes the feature representation at a deeper level, thus enhancing the ability to capture complex features. Finally, considering the computational overhead caused by the increase of model depth, the patch merger module is added. In the shallow layer of transformer, similar or redundant tokens are dynamically evaluated and merged, which reduces the number of tokens entering the deeper network and effectively optimizes the operation efficiency and model performance. In summary, the proposed model not only captures local-global information, but also optimizes computational efficiency while maintaining high performance.

2.1 Convolutional Neural Network for Image Blocking

As a deep learning model, CNN can automatically learn local features in images and gradually extract high-level semantic information. In recent years, more and more researches begin to apply CNN to PolSAR image classification. By using its unique convolution layer, CNN can effectively capture the spatial structure and local features of the image, which is very important for understanding the complex information in Polsar data. With the rise of vit, the self-attention mechanism exhibits great potential in processing polsar images. The original vit uses linear projection to segment the image into a series of image blocks with the same size and no overlap, and converts them into tokens as input data. However, the method has relatively weak ability to capture local features of the image, which may lead to the loss of some local details. In order to overcome the limitation and retain the global information processing ability in vit, CNN is applied to block the image instead of linear projection in MAE with vit as the backbone [23].

Fig 2 is a schematic diagram of image blocking using CNN. For the original polsar data, a polsar image block with the size of $f \times f \times 9$ is obtained by neighborhood extraction centered on pixel. The image blocks are subsequently fed as input data into a network composed of three convolutional layers. The network consists of 32, 64 and 128 convolutional layers designed to capture local features within each image block and map them into a high dimensional embedding vector. Firstly, the image blocks go through 32 convolutional kernels of size 3 × 3 in the first layer, and the initial feature extraction is carried out on the input image. The information of 9 channels is converted into 32 channel feature maps, and each feature map captures the local features of the input image in different aspects. Then more abstract features are extracted through the second layer of convolution, which enhances the ability of the model to express image features. The third layer of convolution converts the feature map of 64 channels into a feature map with d channels, where d denotes an embedding dimension of 128. After three layers of convolution, the $f \times f \times 9$ image block is divided into N^2 image blocks with the shape of (f/N), (f/N), 9). Each patch is expanded according to the spatial dimension to obtain a vector with the shape of ((f/N)·(f/N)·9)-d. These vectors are stacked, the input is reconstructed from $R^{f \times f \times 9}$ to $R^{N^2 \times ((f/N) \cdot (f/N) \cdot 9)}$, which completes the process from the original image block to tokenization. In order to ensure that the model correctly understands the location information of each patch, location embedding is added. As a special embedding vector, class token gradually aggregates information from all image blocks in the whole transformation process. Through the process, the model can not only capture the local features of the image blocks, but also understand the position and category information of the image blocks, enabling the model to handle the image data more comprehensively.

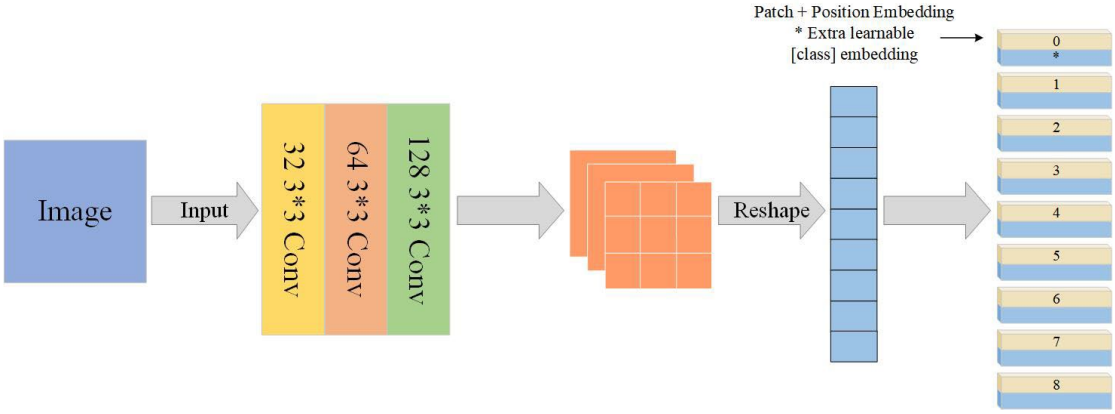

Fig.2 The schematic diagram of CNN for image blocking

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

2.2 Re-attention

ViT applies transformer architecture to visual tasks for the first time, and captures the global dependencies between various parts of an image through self-attention mechanism, which brings new perspectives and methods to image classification tasks. On large-scale datasets, ViT has shown comparable or even better performance than deep CNN. Although ViT has achieved remarkable results in image classification, its performance tends to be saturated with the increase of network depth. Unlike CNNs that gain consistent performance gains by stacking more layers, vit faces challenges when adding depth. Specifically, as the number of transformer blocks increases, the attention maps of the model gradually tend to be similar, resulting in the inability to effectively expand the representation ability of the model. The phenomenon is called attention collapse. To address this challenge, cross-head information exchange is introduced on the basis of multi-head self-attention, and the new attention mechanism is Re-attention [24].

In the transformer model, a multi-head self-attention mechanism is adopted. Each head has different parameters and is capable of capturing different features of the input sequence. Therefore, within the same transformer block, the similarity of attention maps of different heads is relatively low. Based on this, the Re-attention mechanism dynamically fuses the information of different attention heads by introducing a learnable transformation matrix $\theta \in R^{r\times r}$, so as to generate a new and more diverse attention map.

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right), A = \left[A_1, A_2, ..., A_r\right]^T, \\ \text{RA}\left(Q, K, V\right) = \text{Norm}\left(\theta^T A\right)V \tag{1}$$

Where A_i represents the attention weight of the i-th head, and A represents the splicing of the attention matrices of r heads. In the way, the model is able to integrate information from different heads and effectively enhance the flow of information between individual head. Specifically, matrix θ, as a learnable transformation matrix, can fuse the attention outputs of multiple heads into a new representation through linear transformation according to different input data, so that the information of different heads can be better combined. After transformation, the obtained matrix is standardized to ensure that the output has a stable distribution. Then it is multiplied by the value matrix V to generate the final attention output. As shown in Fig 3, the self-attention of the original ViT is directly replaced by the Re-attention. In contrast, the Re-attention mechanism can effectively enhance the information exchange between the multi-head, which enables the model to better capture the interrelationships between different features, and improves the expression ability and performance of the model.

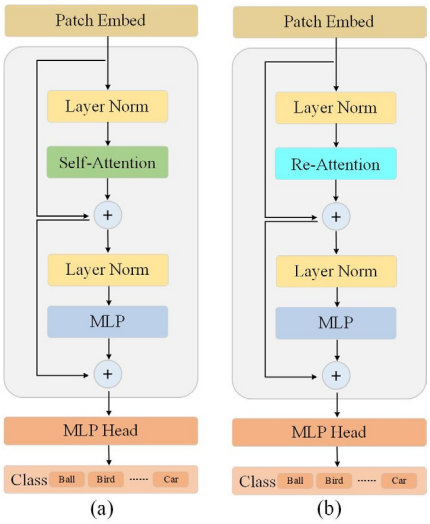

Fig.3 The schematic diagram of Re-attention and multi head attention: (a) ViT based on multi-head self-attention (b) ViT based on Re-attention

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

2.3 Patch merging module

With the rapid development of deep learning, ViT has shown its excellent performance and unique advantages in many visual tasks. ViT divides the input image into a number of fixed size image blocks, which are tokenized by linear projection, and then sent to the transformer block for processing. The innovative approach opens up a new perspective for image classification, and gives full play to the powerful efficiency of transformer in processing sequence data. Through the self-attention mechanism, it can effectively capture the global information and show a stronger ability in dealing with long-distance dependencies and complex structures. Although ViT performs well in performance, its computational overhead has also become a problem that can not be ignored. The computational complexity of the ViT architecture grows quadratically with the number of input image blocks, and thus the computational cost rises sharply as the image resolution increases. How to balance the relationship between performance and computational overhead has become an important research direction. To solve the problem, Renggli et al. [25] proposed a patch merging module that can effectively reduce the computational cost of ViT while retaining the powerful performance of ViT.

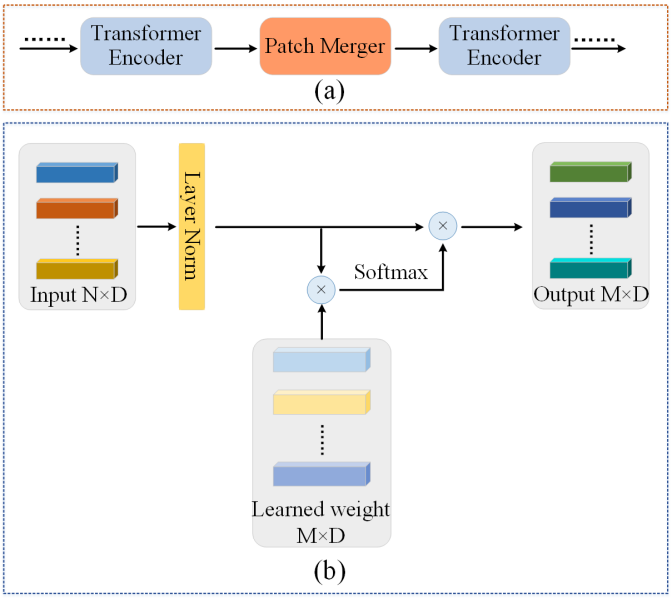

Fig.4 The schematic diagram of patch merging module: (a) Overall schematic diagram of patch merging module (b) Detailed process of patch merging

The overall process is displayed in Fig 4, the core idea of patch merging is to reduce the computational burden of ViT by introducing a simple merging operation between two consecutive transformer encoder layers. To be specific, the module performs a weighted combination of input tokens by introducing a learnable matrix. First, the input token sequence $X \in R^{N\times d}$ is linearly transformed, and each d-dimensional token is mapped to the M-dimensional space using matrix W. Subsequently, it is transposed and a weight distribution is computed for each output token with softmax function. Finally, the normalized weight matrix is multiplied with the original input tokens sequence to obtain the merged token sequence. The specific operation is as follows.

$$PM = \text{softmax}\left(XW\right)^{T} X \tag{2}$$

Where N is the number of input tokens, d is the embedded dimension of each token, and $PM \in R^{M\times d}$ is the combined token sequence.

The main significance of patch merging operation is to reduce the computational burden of the model, while trying to maintain the performance of the model. By combining multiple input tokens into a smaller number of output tokens, the model can process fewer tokens in the subsequent transformer encoder layer, thus significantly reducing the computational complexity. Since the merge operation is carried out through the learned weights, the model can adaptively determine which input tokens contribute more to the final output tokens. It helps to retain important information while reducing the impact of redundant or unimportant information, achieving a balance of optimized performance and efficiency.

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

3. Experimental results and analysis

3.1 Datasets description

(1) AIRSAR Flevoland: The Flevoland image was acquired by the AIRSAR platform and is a sub-image of the L-band multi view PolSAR dataset, which has a size of 750 × 1024. The ground resolution of the image is 6.6 m × 12.1 m, containing 15 types of ground objects, and each type is presented through different colors. The image has 167,712 pixels manually labeled by expert knowledge. Fig 5(a) illustrates the Pauli-based pseudo-color map, and Fig 5(b) and (c) show the corresponding ground truth map and legend for this dataset, respectively.

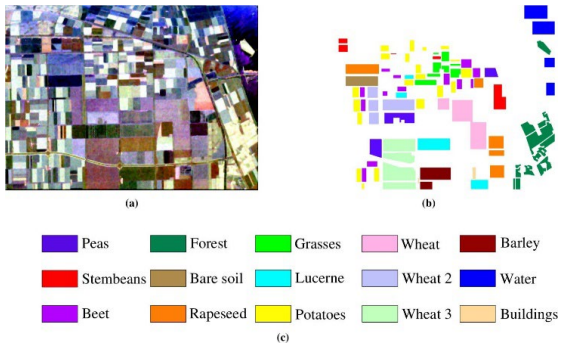

Fig.5 AIRSAR Flevoland dataset and the color code. (a) Pauli-RGB map (b) Ground truth map (c) Legend

(2) RADARSAT-2 San Francisco Bay: The second image is a C-band image of the San Francisco Bay area acquired by the RADARSAT-2 satellite. The size of the image is 1380 × 1800, and there are 1804087 pixels with known label information. Fig 6(a) shows the Pauli-RGB image in the scene, which mainly covers five types of land cover: High-Density Urban, Water, Vegetation, Developed, and Low-Density Urban. The ground truth map and the legend of land cover type are shown in Fig 6(b) and (c), respectively.

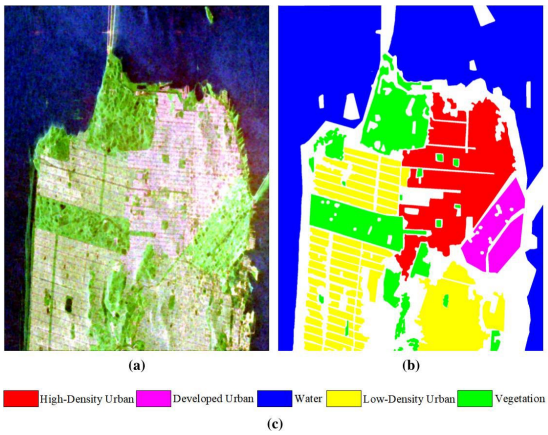

Fig.6 RADARSAT-2 San Francisco Bay dataset and the color code. (a) Pauli-RGB map (b) Ground truth map (c) Legend

(3) ESAR Oberpfaffenhofen: The Oberpfaffenhofen dataset was derived from the ESAR airborne platform provided by the German Aerospace Center and belongs to the L-band. An image of Pauli-RGB is given in Fig 7(a). The scene is relatively simple and includes only three feature classes: Built-up Area, Wood Land and Open Area. The image size is 1300 × 1200, and Fig 7(b) shows the corresponding ground truth map, which contains 1,374,298 labeled data. Fig 7(c) shows the corresponding feature types.

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

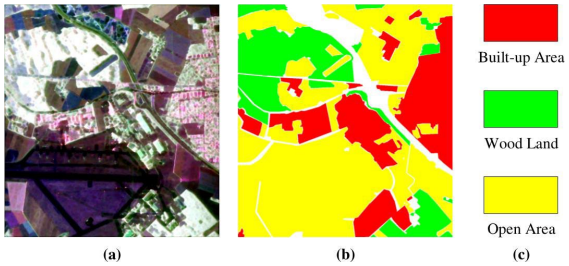

Fig.7 RADARSAT-2 San Francisco Bay dataset and the color code. (a) Pauli-RGB map (b) Ground truth map (c) Legend

3.2 Experimental setup

ViT extracts a field of size 14 × 14 centered on the pixel. Therefore, the input space size is set to 15. MAE is a self-supervised learning model, 150,000 unlabeled data were selected for pre-training, and the number of training rounds was 1,600. In the fine-tuning phase, the Flevoland image was randomly selected with 300 labeled data from each category. Since the San Francisco Bay and Oberpfaffenhofen dataset have fewer categories, 1000 and 1500 labeled data were adopted for each category. The Adam optimizer is used for optimization, and the weight attenuation is set to 1×10^{-3}, and the initial learning rate is set to 1×10^{-4}.

Table 1 Network architecture settings for encoder and decoder

Network architecture	Encoder	Decoder
Depth	12	8
OutputDimension	128	64
Number of heads	4	4
Dimension of Each Head	32	16
Number of Hidden Nodes	512	256

Table 1 shows the parameter settings of the network architecture. Specifically, in the pre-training phase, the depth of the encoder part is 12 and the output dimension is 128. The number of heads in the self-attention is set to 4, where the dimension of each head is 32. 512 is the number of hidden nodes. In the decoder part, the depth is set to 8 and the output dimension is 64. The number of heads in self-attention is set to 4, where the dimension of each head is 16. The number of hidden nodes is 256. During the fine-tuning stage, only the trained encoder is removed and its parameter Settings remain consistent with those of the encoder in the pre-training stage.

The manuscript explores the combined form of CNN and vit, so in order to further demonstrate the effectiveness of the method, six polsar image classification methods based on CNN, vit, or a combination of the two are selected for comparative experiments. The CNN-based methods include ResNet [26], SKNet [27], and MobileNetV3 [28]. The ViT-based methods include CCT [29], MCPT [30], and MAPM [30], where CCT and MCPT are based on the combination of CNN and ViT.

3.3 Classification metrics

The commonly used PolSAR image classification metrics are Overall Accuracy (OA), Average Accuracy per Class (AA) and kappa Coefficient. OA is the most direct performance measure, which indicates the proportion of the number of correctly differentiated samples to the total number of samples. AA is used to evaluate the performance of the classification model in dealing with multi classification problems. By calculating the average classification accuracy across all categories, the overall performance of the model is evaluated to ensure that sufficient attention is given to each category. The Kappa coefficient is a statistic to measure the label consistency in a classification task, which usually takes a range of values between 0 and 1. As the coefficient gets closer to 1, it indicates a higher level of consistency, which indicates that the model is more capable of categorizing. The formulas for OA, AA and Kappa are given below:

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

$$AA = \frac{1}{U}\sum_{i=1}^{U}\frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

Where *TP* refers to true positive, *FN* refers to false negative, *FP* refers to false positive, and *TN* refers to true negative, *i* denotes the category and *U* denotes the total number of categories. Given the number of true samples per class $a_1, a_2, ..., a_c$ and the number of predicted samples per class $b_1, b_2, ..., b_c$, and the total number of samples n, the probability of accidental consistency can be expressed as follows:

$$P_e = \sum_{i=1}^{c}\left(\frac{a_i}{n}\right)\left(\frac{b_i}{n}\right) \tag{6}$$

3.4 Classification results

(1) The results and analysis of the Flevoland dataset

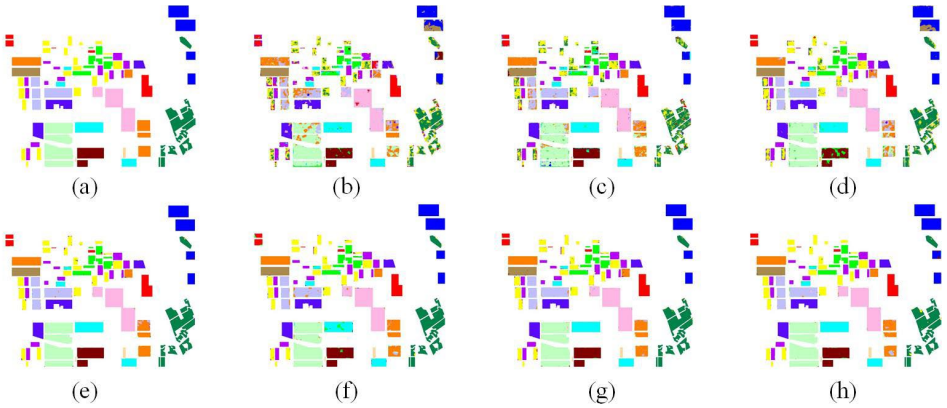

Fig.8 Predicted images about ground truth of the AIRSAR Flevoland dataset. (a) Ground truth map (b) ResNet (c) SKNet (d) MobileNetV3 (e) CCT (f) MCPT (g) MAPM (h) The proposed method

Fig 8 presents the prediction results of different methods on the Flevoland ground truth map. Fig 8(a) is a standard ground truth map, which provides an important reference for evaluating the classification accuracy and prediction effect of the model. Fig 8(b)-(d) show the classification images based on CNN method, where each region contains a large amount of noises. The classification result of the ResNet method in Fig 8(b) is the most severely affected by noise, with misclassification occurring in most areas. The SKNet method can reduce the interference of noise to a certain extent and improve the discrimination of categories. It is shown in Fig 8(c), but there are still some inaccurate categorized regions. The brown part at the bottom of Fig 8(d) contains a lot of green, and in addition, other color classifications also appear in the blue part in the upper right corner. It indicates that the MobileNetV3 method has a large misclassification when dealing with individual regions. In contrast, the methods based on ViT or the combination of both CNN and ViT can effectively mitigate the interference of noise, and the corresponding classification results are given in Fig 8(e)-(h). Fig 8(e) shows excellent performance, and the CCT method can handle most areas well, but there are still shortcomings in predicting the edge parts. Fig 8(f) demonstrates the results of the MCPT method, which is able to predict the categories correctly, however, there is significant noise in most of the regions. Fig 8(g) presents a relatively clear category boundary, but the MAPM method has slight errors in the classification of some regions. Fig 8(h) exhibits the prediction results of the proposed method, which shows that the method divides the region better and is less affected by noise. In general, the proposed method shows significant advantages in the prediction of labeled data, providing more stable and accurate predictions.

The numerical prediction results of different methods on the Flevoland dataset are shown in Table

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

2. The overall performance of the ResNet method is poor compared to the other methods and has a large standard deviation. It indicates that the method has significant fluctuations, especially with lower accuracy on Wheat 2 and Rapeseed, further proving its unsatisfactory performance in these categories. The sknet method has a relatively good performance among the CNN-based methods, but does not show significant prominence in predicting the individual feature categories. The performance of MobileNetV3 method is between ResNet and SKNet methods, but its classification performance on Forest is the worst among all methods. The CCT method has outstanding performance on several categories, showing strong robustness and feature learning ability. The lower standard deviation indicates that it is effective in reducing the fluctuation of the prediction results. The MCPT and MAPM methods perform very close to each other in terms of performance, with both showing high accuracy and stability. The proposed method performs relatively well, especially achieving the highest classification accuracy in most land cover categories. In addition, the proposed method exhibits a small standard deviation, which indicates that the method has good stability.

Table 2 Objective evaluation indicators of seven methods on the AIRSAR Flevoland dataset

	ResNet	SKNet	MobileNetV3	CCT	MCPT	MAPM	Proposed
Water	0.8746±0.0973	0.9405±0.0650	0.8882±0.0736	0.9855±0.0126	0.9648±0.0267	0.9916±0.0071	**0.9992±0.0018**
Forest	0.7262±0.0966	0.7756±0.0513	0.6803±0.1066	**0.9942±0.0017**	0.9871±0.0131	0.9872±0.0027	0.9907±0.0054
Lucerne	0.9023±0.0799	0.9831±0.0187	0.9170±0.0891	0.9897±0.0045	0.9769±0.0189	0.9871±0.0047	**0.9944±0.0033**
Grass	0.7907±0.0848	0.9463±0.0295	0.8207±0.0441	0.9810±0.0060	0.9406±0.0190	0.9723±0.0089	**0.9860±0.0078**
Peas	0.9176±0.0416	0.9574±0.0218	0.9220±0.0336	0.9907±0.0041	0.9941±0.0023	0.9773±0.0097	**0.9971±0.0035**
Barley	0.9180±0.0557	0.8976±0.0845	0.8660±0.1080	**0.9948±0.0020**	0.9872±0.0080	0.9887±0.0053	0.9923±0.0066
Bare Soil	0.9552±0.0587	0.9961±0.0027	0.9707±0.0270	**0.9975±0.0028**	0.9890±0.0130	0.9872±0.0098	0.9954±0.0025
Beet	0.8271±0.0737	0.9583±0.0046	0.9005±0.0371	0.9772±0.0085	0.9861±0.0044	0.9825±0.0049	**0.9883±0.0024**
Wheat 2	0.6621±0.1393	0.9078±0.0358	0.7561±0.0827	**0.9767±0.0125**	0.9465±0.0561	0.9647±0.0128	0.9704±0.0241
Wheat 3	0.8864±0.0391	0.9464±0.0314	0.8746±0.0567	**0.9967±0.0012**	0.9882±0.0050	0.9931±0.0024	0.9919±0.0056
Stem beans	0.8644±0.0532	0.9245±0.0305	0.8897±0.0466	**0.9976±0.0026**	0.9811±0.0061	0.9892±0.0043	0.9967±0.0033
Rapeseed	0.6384±0.1009	0.8446±0.1198	0.7252±0.1199	0.9633±0.0118	0.9200±0.0547	0.9612±0.0138	**0.9723±0.0195**
Wheat	0.8736±0.0435	0.9350±0.0279	0.8670±0.0396	0.9801±0.0086	0.9795±0.0047	0.9748±0.0073	**0.9846±0.0065**
Buildings	0.9344±0.0580	0.8684±0.0078	0.9546±0.0280	0.9965±0.0013	0.9965±0.0016	0.9888±0.0078	**0.9968±0.0012**
Potatoes	0.6715±0.0681	0.8013±0.0241	0.6836±0.0716	0.9803±0.0044	0.9685±0.0143	0.9511±0.0119	**0.9939±0.0018**
AA	0.8295±0.0218	0.9122±0.0138	0.8478±0.0182	0.9868±0.0007	0.9737±0.0023	0.9798±0.0019	**0.9900±0.0014**
Kappa	0.7921±0.0235	0.8947±0.0151	0.8081±0.0240	0.9844±0.0008	0.9699±0.0031	0.9768±0.0014	**0.9881±0.0014**
OA	0.8089±0.0217	0.9032±0.0139	0.8234±0.0222	0.9856±0.0007	0.9723±0.0028	0.9787±0.0013	**0.9891±0.0013**

In addition to the prediction of labeled regions, more emphasis is placed on the prediction results of unlabeled regions in the PolSAR image classification task. Fig 9 shows the overall prediction results of various methods on the Flevoland dataset, and compares them with the Pauli-RGB image in Fig 9(a). The comparison not only reveals the differences in image detail processing between different algorithms, but also demonstrates their performance in unlabeled areas. Fig 9 (b) shows the prediction figure of the ResNet method, and it can be observed that it performs better in handling large-scale regions. However, it is generally affected by a large amount of noises, which leads to obvious misclassification in most of the regions. The prediction result of SKNet method in Fig 9(c) displays good adaptability, but the prediction result still has some deviation in the part with relatively complex background. MobileNetV3 method can effectively recognize the feature types in most areas, but due to its relatively low accuracy, the classification effect is rough when dealing with the detail part. The corresponding prediction image is shown in Fig 9(d). Fig 9(e) performs excellent as a whole, but for some small or irregular feature types, the CCT method is still insufficient. Fig 9(f) is the prediction image of MCPT method, which is able to clearly distinguish different objects when dealing with unlabeled areas. However, it is more ambiguous in the processing of the boundary. The MAPM method exhibits a strong classification ability with relatively pure areas, but there will be some adhesion in Fig 9(g). Fig 9(h) shows the classification result image of the proposed method. Compared with other methods, the boundary of various ground objects is clearer, and each area is purer.

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

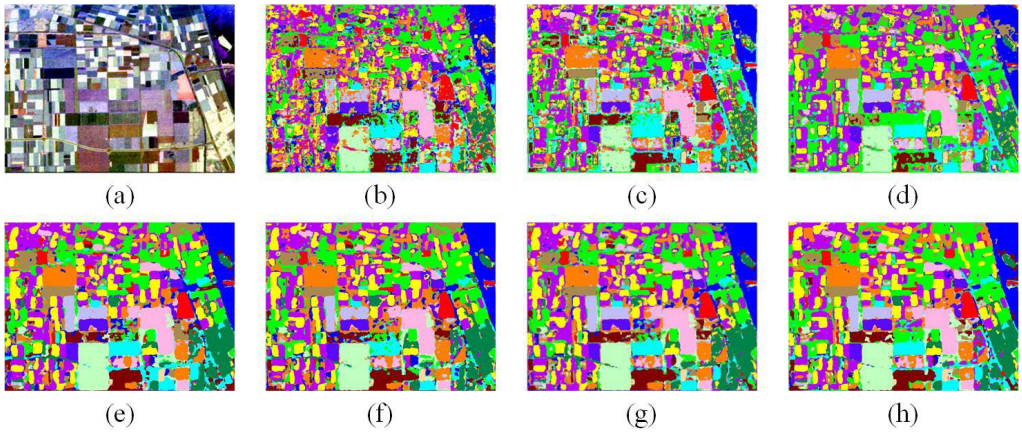

Fig.9 Predicted images of the AIRSAR Flevoland dataset. (a) Pauli-RGB image (b) ResNet (c) SKNet (d) MobileNetV3 (e) CCT (f) MCPT (g) MAPM (h) The proposed method

(2) The results and analysis of the San Francisco Bay dataset.

Table 3 demonstrates the objective evaluation metrics of the different methods on the San Francisco Bay dataset. ResNet and MobileNetV3 perform relatively close to each other, with small differences in several evaluation metrics. However, the ResNet method has a larger standard deviation and may be subject to larger fluctuations in classification, especially in the delineation of High-Density Urban. In contrast, MobileNetV3 has better stability. The performance of SKNet method is better than ResNet and MobileNetV3 method, especially in AA, kappa coefficient and OA. However, compared with the ViT based method, SKNet method has no advantages. The CCT method performs well in Developed areas. In addition, small variances in various categories can effectively reduce fluctuations. The MCPT method shows a performance close to that of the CCT method, with less large fluctuations and a more balanced performance. The MAPM method is able to better classify High-Density Urban, and the higher Kappa coefficient further proves its better classification consistency and robustness. The performance of the proposed method on the San Francisco Bay dataset is better than other methods, and has achieved high accuracy in Water and Low-Density Urban, and has made significant improvements in various evaluation indicators. In addition, the smaller standard deviation indicates better robustness and stronger classification ability, which can effectively handle complex feature classification tasks.

Table 3 Objective evaluation indicators of seven methods on the San Francisco Bay dataset

	ResNet	SKNet	MobileNetV3	CCT	MCPT	MAPM	Proposed
Water	0.9899±0.0058	0.9968±0.0029	0.9838±0.0107	0.9982±0.0007	0.9932±0.0051	0.9985±0.0007	**0.9989±0.0012**
Vegetation	0.9019±0.0306	**0.9506±0.0134**	0.9163±0.0139	0.9102±0.0053	0.9034±0.0138	0.9342±0.0060	0.9340±0.0065
High-Density Urban	0.7865±0.0740	0.9381±0.0077	0.7903±0.0403	0.9604±0.0032	0.9568±0.0063	**0.9617±0.0051**	0.9611±0.0058
Developed	0.8735±0.0356	0.9053±0.0399	0.8479±0.0256	**0.9617±0.0041**	0.9506±0.0103	0.9511±0.0052	0.9505±0.0067
Low-Density Urban	0.8873±0.0475	0.9093±0.0093	0.8740±0.0246	0.9469±0.0052	0.9411±0.0081	0.9490±0.0039	**0.9617±0.0064**
AA	0.8878±0.0127	0.9400±0.0063	0.8825±0.0091	0.9555±0.0013	0.9490±0.0019	0.9589±0.0007	**0.9612±0.0014**
Kappa	0.8868±0.0106	0.9431±0.0019	0.8812±0.0090	0.9556±0.0016	0.9478±0.0039	0.9605±0.0004	**0.9641±0.0017**
OA	0.9212±0.0073	0.9604±0.0014	0.9171±0.0063	0.9691±0.0011	0.9636±0.0027	0.9725±0.0003	**0.9750±0.0012**

The classification results of different comparison methods on San Francisco Bay dataset are shown in Fig 10. Fig 10(a) shows the Pauli-RGB map of the dataset. Fig 10(b) presents the results of ground classification using the resnet method, which can divide most areas, but in the yellow and green areas, the results contain more other feature types. It can be seen from Fig 10(c) that the SKNet method performs relatively well in the CNN based method, and a small blue area in the lower left corner is also accurately recognized. In contrast, the MobileNetV3 method performs poorly and contains a lot of noises in each region, which appears to be more cluttered, as reflected in Fig 10(d). The CCT method achieves a better delineation performance, but it can be seen from Fig 10(e) that its boundary is not as regular as other images. The predicted images of MCPT and MAPM methods are shown in Fig 10(f) and (g), respectively. The two methods perform well in the division of yellow and green areas, and can better handle the transition between urban and natural areas. The classification result of the proposed method Fig 10(h) exhibits significant superiority. Each region is purer and possesses clearer

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

boundaries. It further enhances the usability and reliability of the classification results.

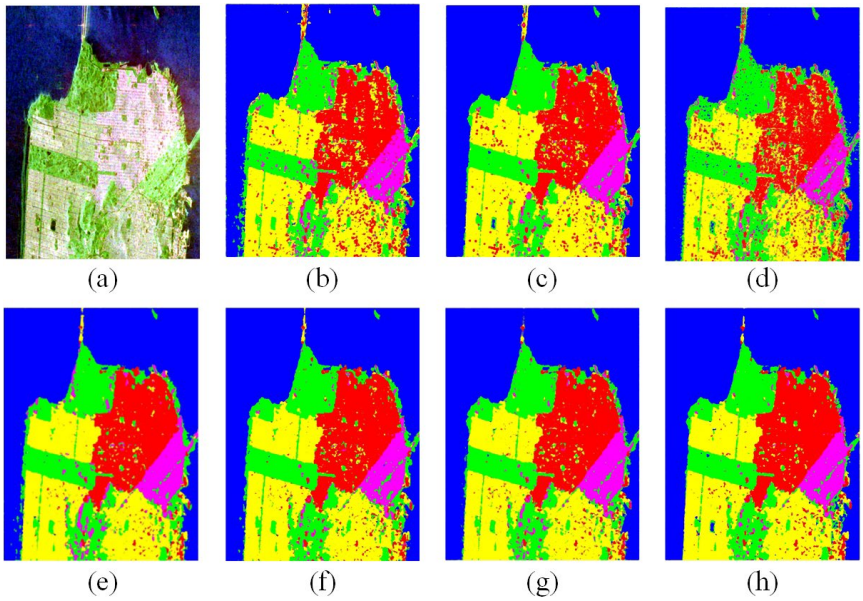

Fig.10 Predicted images of the San Francisco Bay dataset. (a) Pauli-RGB image (b) ResNet (c) SKNet (d) MobileNetV3 (e) CCT (f) MCPT (g) MAPM (h) The proposed method

(3) The results and analysis of the Oberpfaffenhofen dataset

Table 4 Objective evaluation indicators of seven methods on the Oberpfaffenhofen dataset

	ResNet	SKNet	MobileNetV3	CCT	MCPT	MAPM	Proposed
Built-up Area	**0.9720±0.0342**	0.9390±0.0556	0.9546±0.0300	0.9176±.0059	0.9298±0.0110	0.9060±0.0252	0.9338±0.0204
Wood Land	0.6984±0.1260	0.8228±0.0371	0.7014±0.0865	0.7733±0.0241	0.7535±0.0213	0.7606±0.0470	**0.8303±0.0240**
Open Area	0.2264±0.2923	0.5569±0.3749	0.3746±0.2330	0.9599±0.0072	**0.9631±0.0072**	0.9547±0.0125	0.9386±0.0099
AA	0.6323±0.0784	0.7729±0.1184	0.6769±0.0810	0.8836±0.0069	0.8821±0.0042	0.8738±0.0075	**0.9009±0.0050**
Kappa	0.3113±0.1734	0.5705±0.2543	0.3930±0.1464	0.8385±0.0071	0.8366±0.0055	0.8242±0.0061	**0.8496±0.0050**
OA	0.4885±0.1467	0.6972±0.2064	0.5685±0.1313	0.9055±0.0039	0.9048±0.0031	0.8972±0.0031	**0.9109±0.0031**

Table 4 demonstrates the objective evaluation metrics of the seven different methods on the Oberpfaffenhofen dataset. The results reveal that there is a significant gap between the CNN based method and the ViT based method or a combination of the two. The ResNet method performs well in the Built-up Area with a high level of accuracy. But the performance in other categories is very weak, especially in the Open Area, the classification accuracy is far lower than other methods. However, the method has a large standard deviation, which indicates that its classification result may be affected by large fluctuations, thus reducing the stability of the classification result. The MobileNetV3 method has a similar performance to the ResNet method, with a weaker and less stable performance on the Open Area. In comparison, CCT, MCPT, MAPM and the proposed method display a more balanced performance, with significant advantages in accuracy and stability in each category. These methods improve the robustness of classification and reduce the fluctuations. In general, the proposed method shows the best classification ability and strong anti-interference ability, which proves its effectiveness in the task of complex ground object classification.

Fig 11 illustrates the classification results of the different methods on the Oberpfaffenhofen dataset, which performs in line with the numerical results in Table 4. Fig 11(a) shows the Pauli-RGB diagram of the dataset. There are prediction results of the three methods based on CNN in Fig 11(b)-(d), with a wide range of misclassification and irregular color distribution. The classification result of ResNet method Fig 11(b) contains a large number of green areas, while Fig 11 (c) and (d) are covered by a large number of red areas, and the overall performance is messy. The CCT method performs well in most areas, but it still exhibits ambiguity in the boundary regions of Fig 11(e), particularly at the intersection of yellow and red areas, leading to erroneous classification of some regions. Fig 11(f) shows the classification results of the MCPT method, which performs well in classifying yellow and green regions. However, in the details section, particularly in the lower-right corner of the image, the classification results of the MCPT method still appear somewhat ambiguous, and the classification in certain small areas lacks sufficient precision. The overall performance of the MAPM method is relatively good. As can be seen from Fig 11(g), it exhibits clearer boundaries and is capable of

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

capturing detailed changes between regions effectively. The classification results of the proposed method are shown in Fig 11(h), which demonstrates significant superiority. In the classification of each region, the proposed method clearly divides the color regions and has very clear boundaries. It is able to effectively handle transitions in complex regions, reducing misclassifications, and demonstrates the most stable performance among all methods.

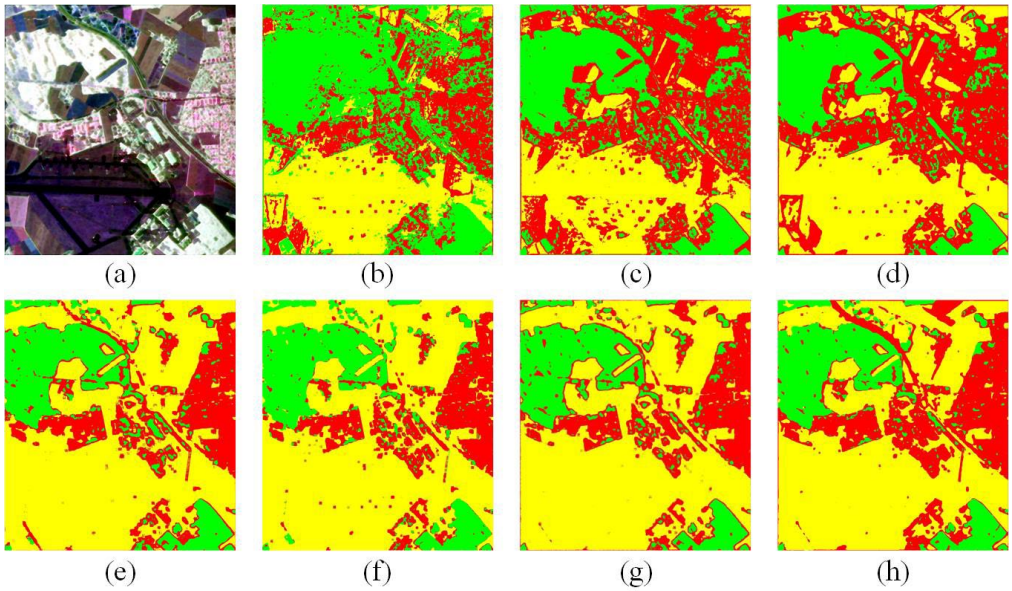

Fig.11 Predicted images of the Oberpfaffenhofen dataset. (a) Pauli-RGB image (b) ResNet (c) SKNet (d) MobileNetV3 (e) CCT (f) MCPT (g) MAPM (h) The proposed method

The experimental results on these three datasets demonstrate that the proposed method exhibits higher classification accuracy compared to other methods. The regions of each land type are purer, and the classification boundaries between different land types are more distinct. In summary, the proposed method delivers satisfactory overall performance.

4. Discussions

4.1 Ablation experiment

Ablation experiments were conducted on the Flevoland dataset by progressively introducing different mechanisms to optimize the baseline model MAE, and the performance of each protocol was analyzed. The specific results are presented in Table 5. Scheme 1 is an unimproved baseline model, which serves as the baseline for the overall ablation experiments. Scheme 2 introduces CNN to extract more detailed local information, significantly enhancing the feature extraction capability of the model and improving classification performance. However, when processing local features, the CNN requires greater computational effort, which increases the computational complexity of the model. In order to avoid the singularity of attention distribution, Scheme 3 uses re-attention instead of self-attention, further improving the performance of the model. In Scheme 4, a patch merging module is introduced to reduce redundant information and effectively lower computational complexity. Schemes 5-7 demonstrates the complementary effects between different mechanisms by combining them pairwise. Scheme 5 combines CNN and re-attention mechanism to capture local information while maintaining the attention of the model on important features. At this point, FLOPs has seen a significant increase, but it is understandable. Scheme 6 combines the advantages of CNN in feature extraction with the effect of patch merging to optimize the computational complexity, which can effectively reduce the computational overhead while maintaining the accuracy. Re-attention and patch merging enable Scheme 7 to optimize the attention distribution while reducing redundant computations. Scheme 8 achieves optimal accuracy by combining three mechanisms, with well-balanced computational cost and parameter count. The results of the ablation experiments demonstrate the effectiveness of the introduced mechanism. By enhancing the feature extraction ability and classification accuracy, the classification performance of ground objects has been significantly improved.

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

Table 5 Results of ablation experiments of the proposed method on the AIRSAR Flevoland dataset

Scheme	MAE	CNN	Re-attention	Patch merging	OA	AA	Kappa	FLOPs (M)	Params (M)
1	√				0.9791	0.9780	0.9760	85.81	3.17
2		√			0.9857	0.9849	0.9836	92.12	3.26
3			√		0.9895	0.9896	0.9886	85.84	3.17
4				√	0.9856	0.9873	0.9861	55.70	3.17
5		√	√		0.9870	0.9850	0.9837	92.16	3.26
6		√		√	0.9872	0.9866	0.9854	62.02	3.26
7			√	√	0.9868	0.9867	0.9856	55.72	3.17
8		√	√	√	0.9893	0.9882	0.9872	62.04	3.26

4.2 Impact of the amount of training data

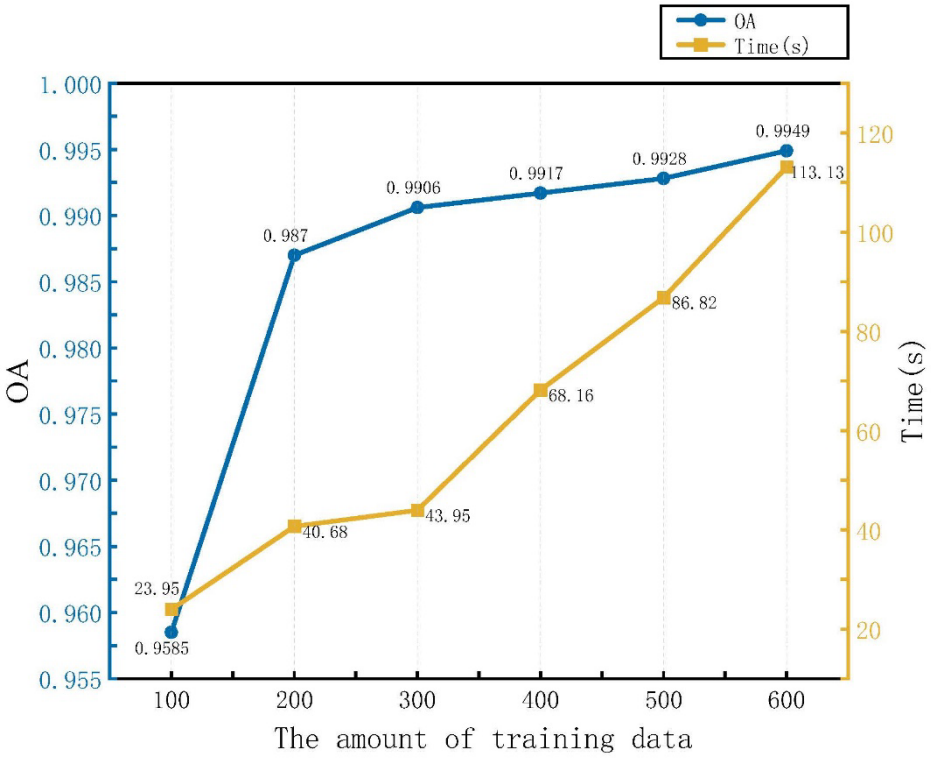

Fig.12 Impact of the amount of training data on Flevoland dataset

In deep learning, training data is the core foundation of model performance, which can help the model better capture the features and potential patterns in the input data, and enhance its generalization ability for unknown data. Fig 12 illustrates the performance of different training data volumes on the Flevoland dataset. With the increase of training data, the performance of the model shows different trends. When the data volume increases from 100 to 200, OA shows a significant improvement. However, there is still a large upside in the performance at this point. At the same time, the yellow curve in the figure reveals that the training time also increases significantly with the increase of data volume. It indicates that the model needs to process more information, which leads to the extension of training time. The trend reveals the direct relationship between data volume and training time. When the data volume reaches 300, OA continues to improve, but the growth rate relatively slows down, indicating that the improvement of model performance gradually stabilizes. Meanwhile, the training time only increased slightly, and the computing time and performance tended to balance. With the further increase of data volume, the performance of the model is basically stable. The improvement speed of OA gradually slowed down, while the training time showed an obvious growth trend. It indicates that after the data volume reaches a certain scale, the contribution of further increasing the data volume to the model performance has gradually decreased, while the computational cost and time have increased more significantly. Therefore, it is crucial to select an appropriate data volume for enhanced model performance and optimized computational efficiency. In addition, considering that the Flevoland dataset contain a relatively large number of land cover categories, a data volume of 300

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

samples per category was selected for training. It can not only effectively improve the performance of the model, but also obtain better classification results within a reasonable calculation time range, balancing the accuracy of the model and training efficiency.

5. Conclusions

The manuscript discusses the problem of insufficient ability to capture local features in PolSAR image classification. CNN is used instead of the original linear projection operation to block the image and enhance the extraction of local features. Although the process increases computational load, the dynamic merging of redundant patches in the early transformer layers effectively reduces the number of input tokens, thereby compensating for additional computational overhead while maintaining the computational efficiency of the model. To solve the problem of information degradation caused by the attention mechanism in deep transformer models, which manifests as the attention distribution of each layer tending to become consistent with increasing model depth, a method of reordering attention weights is introduced. The model can effectively capture diverse and complex information and promote information exchange among different attention heads. The proposed method enhances the performance of the model, which helps the model achieve higher accuracy and better generalization ability in PolSAR image classification.

Acknowledgments

This research was funded partly by the National Natural Science Foundation of China under Grant 62201201; the Double First-Class Construction Project of Henan Polytechnic University under Grant XJBS202510; the Doctoral Foundation of Henan Polytechnic University under Grant B2025-48.

References

[1] Brown W M. Synthetic aperture radar[J]. IEEE Transactions on Aerospace and Electronic Systems, 2010 (2): 217-229.
[2] Wang H, Xu F, Jin Y Q. A review of polsar image classification: From polarimetry to deep learning[C]//IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019: 3189-3192.
[3] Moreira A, Prats-Iraola P, Younis M, et al. A tutorial on synthetic aperture radar[J]. IEEE Geoscience and remote sensing magazine, 2013, 1(1): 6-43.
[4] Gomez L, Alvarez L, Mazorra L, et al. Fully PolSAR image classification using machine learning techniques and reaction-diffusion systems[J]. Neurocomputing, 2017, 255: 52-60.
[5] West R D, Riley R M. Polarimetric Interferometric SAR Change Detection Discrimination[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(6): 3091-3104.
[6] Xie W, Ma G, Zhao F, et al. PolSAR image classification via a novel semi-supervised recurrent complex-valued convolution neural network[J]. Neurocomputing, 2020, 388: 255-268.
[7] Jiao L, Liu F. Wishart deep stacking network for fast POLSAR image classification[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3273-3286.
[8] Cameron W L, Leung L K. Feature motivated polarization scattering matrix decomposition[C]// IEEE International Conference on Radar. IEEE, 1990: 549-557.
[9] Cameron W L, Rais H. Derivation of a signed Cameron decomposition asymmetry parameter and relationship of Cameron to Huynen decomposition parameters[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(5): 1677-1688.
[10] Parikh H, Patel S, Patel V. Classification of SAR and PolSAR images using deep learning: A review[J]. International Journal of Image and Data Fusion, 2020, 11(1): 1-32.
[11] Takizawa Y, Shang F, Hirose A. Adaptive land classification and new class generation by unsupervised double-stage learning in Poincare sphere space for polarimetric synthetic aperture radars[J]. Neurocomputing, 2017, 248: 3-10.
[12] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.
[13] Zhou Y, Wang H, Xu F, et al. Polarimetric SAR image classification using deep convolutional neural networks[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(12): 1935-1939.
[14] Zhang Z, Wang H, Xu F, et al. Complex-valued convolutional neural network and its application in polarimetric SAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing,

ISSN 2616-5775 Vol. 8, Issue 10: 108-123, DOI: 10.25236/AJCIS.2025.081014

2017, 55(12): 7177-7188.
[15] Yang C, Hou B, Ren B, et al. CNN-based polarimetric decomposition feature selection for PolSAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 8796-8812.
[16] Dong H, Zou B, Zhang L, et al. Automatic design of CNNs via differentiable neural architecture search for PolSAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(9): 6362-6375.
[17] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.
[18] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[19] Dong H, Zhang L, Zou B. Exploring vision transformers for polarimetric SAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-15.
[20] Wang W, Wang J, Lu B, et al. MCPT: mixed convolutional parallel transformer for polarimetric SAR image classification[J]. Remote Sensing, 2023, 15(11): 2936.
[21] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.
[22] Fuller A, Millard K, Green J R. Satvit: Pretraining transformers for earth observation[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.
[23] Jeevan P, Sethi A. Vision Xformers: Efficient attention for image classification[J]. arXiv preprint arXiv:2107.02239, 2021.
[24] Zhou D, Kang B, Jin X, et al. Deepvit: Towards deeper vision transformer[J]. arXiv preprint arXiv:2103.11886, 2021.
[25] Renggli C, Pinto A S, Houlsby N, et al. Learning to merge tokens in vision transformers[J]. arXiv preprint arXiv:2202.12015, 2022.
[26] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.
[27] Jeny AA, Sakib ANM, Junayed MS, et al. SkNet: A convolutional neural networks based classification approach for skin cancer classes[C]//2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020: 1-6.
[28] Koonce B. MobileNetV3. In: Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. Ed. by Koonce B. Berkeley, CA: Apress, 2021:125–44.
[29] Hassani A, Walton S, Shah N, et al. Escaping the big data paradigm with compact transformers[J]. arXiv preprint arXiv:2104.05704, 2021.
[30] Wang J, Li Y, Quan D, et al. MAPM: PolSAR image classification with masked autoencoder based on position prediction and memory tokens[J]. Remote Sensing, 2024, 16(22): 4280.