Journal of Computer and Communications, 2024, 12, 22-38
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Research on PolSAR Image Classification Method Based on Vision Transformer Considering Local Information

Mingxia Zhang, Aichun Wang, Xiaozheng Du, Xinmeng Wang, Yu Wu

China Center for Resources Satellite Data and Application, Beijing, China
Email: 18236942851@163.com

How to cite this paper: Zhang, M.X., Wang, A.C., Du, X.Z., Wang, X.M. and Wu, Y. (2024) Research on PolSAR Image Classification Method Based on Vision Transformer Considering Local Information. *Journal of Computer and Communications*, **12**, 22-38.
https://doi.org/10.4236/jcc.2024.129002

Received: July 26, 2024
Accepted: September 7, 2024
Published: September 10, 2024

Abstract

In response to the problem of inadequate utilization of local information in PolSAR image classification using Vision Transformer in existing studies, this paper proposes a Vision Transformer method considering local information, LIViT. The method replaces image patch sequence with polarimetric feature sequence in the feature embedding, and uses convolution for mapping to preserve image spatial detail information. On the other hand, the addition of the wavelet transform branch enables the network to pay more attention to the shape and edge information of the feature target and improves the extraction of local edge information. The results in Wuhan, China and Flevoland, Netherlands show that considering local information when using Vision Transformer for PolSAR image classification effectively improves the image classification accuracy and shows better advantages in PolSAR image classification.

Keywords

Vision Transformer, PolSAR, Image Classification, LIViT

1. Introduction

As an important step in synthetic aperture radar (SAR) image processing, polarimetric synthetic aperture radar (PolSAR) image classification technique has long been widely valued appreciated [1]. In the early days, the PolSAR image classification process was mainly based on empirical manual selection of features for classification. Due to the existence of coherent spot noise and other effects of SAR images, so that manually selected SAR image features often lack of discriminative and robust, the classification accuracy obtained is low. In recent years, deep learning methods, based on the data-driven approach, automatically learn the intrinsic

laws of the sample data and the representation of the hierarchy. It can extract pixel high-level features and make feature extraction and classification result to achieve the overall optimal effect, which has been used by many scholars for PolSAR image classification.

Convolutional Neural Network (CNN)-based methods use convolutional kernels to extract features and make effective use of image local spatial context information [2], which have long been dominant, with the highest accuracy and the most extensive research. Zhou *et al.* [3] used CNN for PolSAR image classification for the first time, and achieved convincing results. Later, Gao *et al.* [4] proposed a two-branch convolutional neural network with polarimetric coherence matrix and Pauli-RGB color image as two-branch inputs to form joint features in polarization space context. Zhang *et al.* [5] proposed a complex-valued convolutional neural network to adapt the complex-valued characteristics of SAR images and better utilize the phase information. Dong *et al.* [6] introduced three-dimensional convolution to extract features from both spatial and channel dimensions. Tan *et al.* [7] used complex-valued three-dimensional convolution to incorporate the complex nature of SAR images into the network. Meanwhile, effective feature representations as network inputs can be obtained to better mine the image polarimetric information. Chen *et al.* [8] used classical roll-invariant features and hidden polarimetric features in rotation domain. Liu *et al.* [9] used scattering coding with scattering matrix. Zhang *et al.* [10] extracted the amplitude and phase of complex-valued PolSAR data and set a tailored network to match the improved input. Wang Lei investigated the rotated domain of the hidden polarimetric features [11], and several scholars have studied the feature combination and feature selection of different polarimetric decomposition features [12]-[16]. However, the structure of CNN-based networks is limited by the fixed convolutional kernel size, and the local information learned is not always optimal, which cannot make good use of the global information of the image.

In 2020, Dosovitskiy *et al.* [17] proposed the Vision Transformer (ViT) method, successfully applied Transformer to image data by dividing the image into multiple small localized image patches. Based on the Transformer model [18], ViT uses self-attention mechanism instead of convolutional structure for feature learning, which has a strong global feature learning capability and outperforms the classical ResNet-like CNN architecture [19] in several classification experiments. In the application of remote sensing image classification, Hong *et al.* [20] proposed Spectral Former network to learn the spectral sequence information of hyperspectral imagery from the perspective of order of the Transformer. Yakoub *et al.* [21] proposed a ViT-based optical remote sensing scene classification method, obtaining better classification results. Liu *et al.* [22] used a combination of lightweight CNN and ViT to learn local and global spatial features on high-resolution SAR imagery, mined complementary information through a fusion network, which has a strong feature extraction capability and anti-noise performance. In Land Cover Classification of PolSAR Imagery, Dong *et al.* [23] explored the long-range interaction

between each pixel for the first time by using ViT, which proved the potential and feasibility of the Transformer structure in PolSAR image processing. Wang *et al.* [24] used a larger sensory wild conditions to divide the patches, combined with Masked AutoEncoders (MAE) pre-trained with unlabeled data. The ViT network was used to better learn the spatial structure features of the image.

Existing studies have obtained good results when using ViT for PolSAR image classification, but there is still the problem of underutilizing the local information of objects. The ViT method regards the image as a series of small patches and flatten the patches into multiple one-dimensional vectors as inputs. This rough patch input process makes insufficient use of local detail information in the image, resulting in unclear delineation of image feature classification boundaries. In response to the above, this paper proposes the Vision Transformer PolSAR image classification method taking into account local information, Local Information Vision Transformer (LIViT). On the one hand, We use polarimetric feature sequences instead of patch sequences as the data embedding method to decrease the loss of local detail information within the image patch. On the other hand, adding the wavelet transform branch, which utilizes the discrete wavelet transform (DWT)'s ability to extract edge features, enables the network to focus more on the shape of the target and the feature edge information. In this way, the ViT network is enhanced to utilize local information and improve the clarity of feature boundary delineation, thus improving the accuracy of PolSAR image classification.

2. Methods

As shown in Figure 1, the PolSAR image classification framework mainly consists

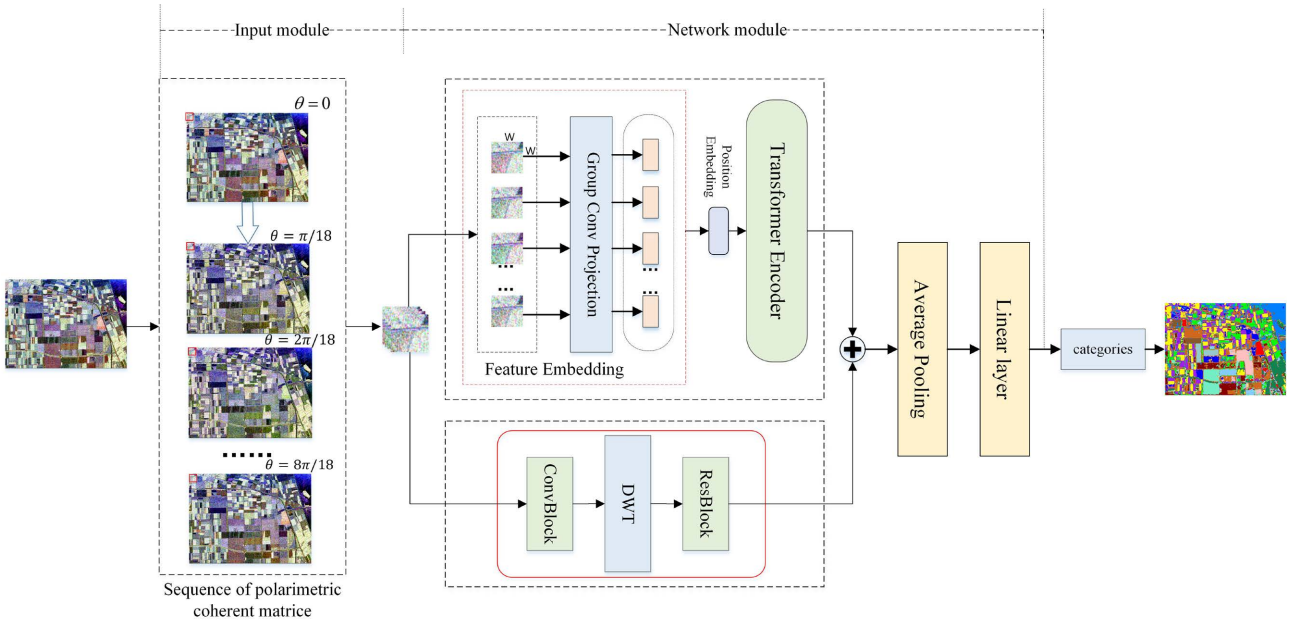

Figure 1. The architecture of LIViT method.

of an input module and a LIViT network module, the input module is the polarimetric coherence matrix sequence input, and the network module contains the Transformer branch and the wavelet transform branch, as well as linear classification layer. The Transformer branch feature embedding part uses polarimetric feature sequence embedding instead of image patch sequence embedding, the wavelet transform branch part uses discrete wavelet transform (DWT) to extract the image edge features, and use of convolution to transform the data form and integrate the information. The different features learned by the two branches are fused by element summing and average pooling to achieve feature fusion, and finally the classification results are obtained using softmax linear classifier.

2.1. Representation and Preprocessing of PolSAR Image

The radar transmits electromagnetic sigals to the ground target, and the ground object responds, scattering the echo signal according to its own characteristics, and recording it in the PolSAR image in complex form. Polarimetric response strongly depends on orientation of the target, and the polarimetric matrices of the same object from different orientations can be different. This effect makes images contain rich hidden information. Si-Wei Chen [25] introduces the concept of rotation domain to understand the diversity of target scattering orientation. The method is to extend the acquired polarimetric matrix to the rotation domain of the radar line of sight given geometry. Then, using suitable interpretation tools, rich and hidden information can be discovered from polarimetric matrix in rotation domain.

Polarimetric information of each pixel in a PolSAR image can be represented by polarimetric coherency matrix $\boldsymbol{T}$, under the reciprocity condition, which is defined as:

$$\boldsymbol{T} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \tag{1}$$

The polarimetric coherency matrix $\boldsymbol{T}$ is a Hermitian matrix, with all complex elements except the diagonals. Its upper triangular elements are usually used as inputs to the network. Rotating the region along the radar line of sight, the polarimetric coherency matrix is as follows:

$$\boldsymbol{T}(\theta) = R_3(\theta) T R_3^{\mathrm{T}}(\theta), \ \theta \in [-\pi, \pi) \tag{2}$$

where the rotation matrix $R_3(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 2\theta & \sin 2\theta \\ 0 & -\sin 2\theta & \cos 2\theta \end{bmatrix}$.

Expanding Equations (2) yields the value of each element of $T_{ij}(\theta)$, as shown in Equations (3), as follows:

$$T_{11}(\theta) = T_{11}$$

$$T_{12}(\theta) = T_{12}\cos 2\theta + T_{13}\sin 2\theta$$

$$T_{13}(\theta) = -T_{12}\sin 2\theta + T_{13}\cos 2\theta \tag{3}$$

$$T_{23}(\theta) = \frac{1}{2}(T_{33} - T_{22})\sin 4\theta + Re[T_{23}]\cos 4\theta + jIm[T_{23}]$$

$$T_{22}(\theta) = T_{22}\cos^2 2\theta + T_{33}\sin^2 2\theta + Re[T_{23}]\sin 4\theta$$

$$T_{33}(\theta) = T_{22}\sin^2 2\theta + T_{33}\cos^2 2\theta - Re[T_{23}]\sin 4\theta$$

where $Re[\cdot]$ and $Im[\cdot]$ denote the real part and imaginary part of a complex T_{ij}, respectively.

On the one hand, in order to make full use of the hidden polarimetric properties of the rotation domain, and on the other hand to better adapt the adopted network to the processing of the sequence data. The study chooses to use the sequence of polarimetric coherency matrices used by Wang Lei [11] 2020 to express the PolSAR images, and as the network input.

As in Equations (3), the polarization azimuth (θ) is stepped by π/18 so that θ changes from 0 to π/2, and nine polarimetric coherence matrices at different polarization orientation angles are obtained to form a sequence of polarimetric coherency matrices. For each pixel in the PolSAR image, its polarimetric information can be characterized as a vector t_p:

$$\begin{aligned} t_p(\theta) = \big[&T_{11}(\theta), T_{22}(\theta), T_{33}(\theta), Re[T_{12}(\theta)], Im[T_{12}(\theta)], \\ &Re[T_{13}(\theta)], Im[T_{13}(\theta)], Re[T_{23}(\theta)], Im[T_{3}(\theta)]\big] \end{aligned} \tag{4}$$

The Pauli pseudo-color maps for different rotation angles of the Flevoland image are shown in **Figure 2**.

For each pixel, the use of neighborhood window data not only suppresses the noise but also better reserve the spatial information. Thus, in this study, the classification result of the neighborhood window image of each pixel is used as the classification result of the center pixel. The size of each sample is 9 × 9 × w × w,

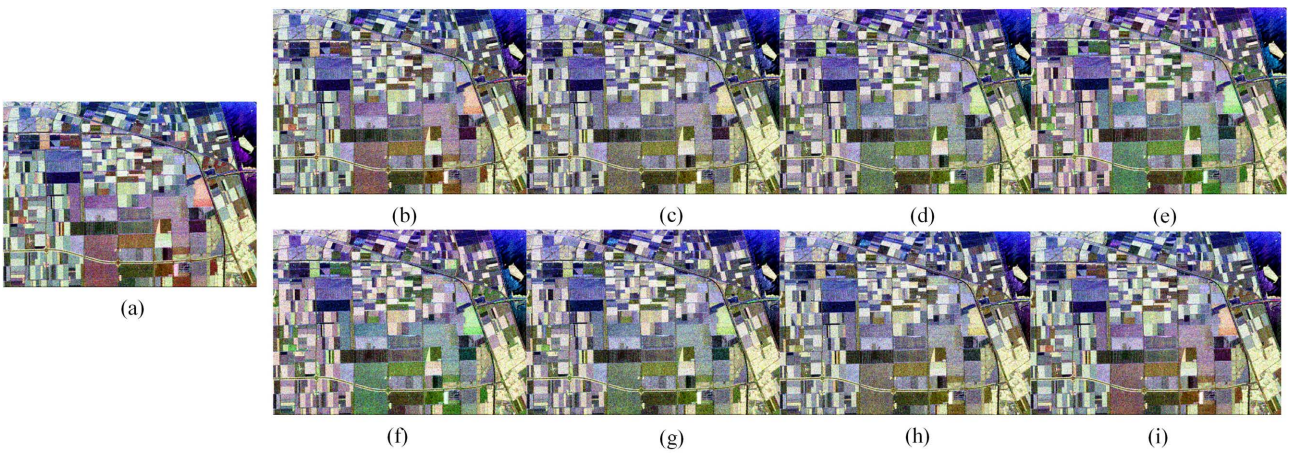

Figure 2. Pauli pseudo-color images of polarimetric coherent matrices of AIRSAR Flevoland. (a) $\theta = 0$. (b) $\theta = \frac{\pi}{18}$. (c) $\theta = \frac{2\pi}{18}$. (d) $\theta = \frac{3\pi}{18}$. (e) $\theta = \frac{4\pi}{18}$. (f) $\theta = \frac{5\pi}{18}$. (g) $\theta = \frac{6\pi}{18}$. (h) $\theta = \frac{7\pi}{18}$. (i) $\theta = \frac{8\pi}{18}$.

where w denotes the window size, the front 9 is the number of rotation angles, and the back 9 is the number of channels under each rotation angle, as shown in Figure 3.

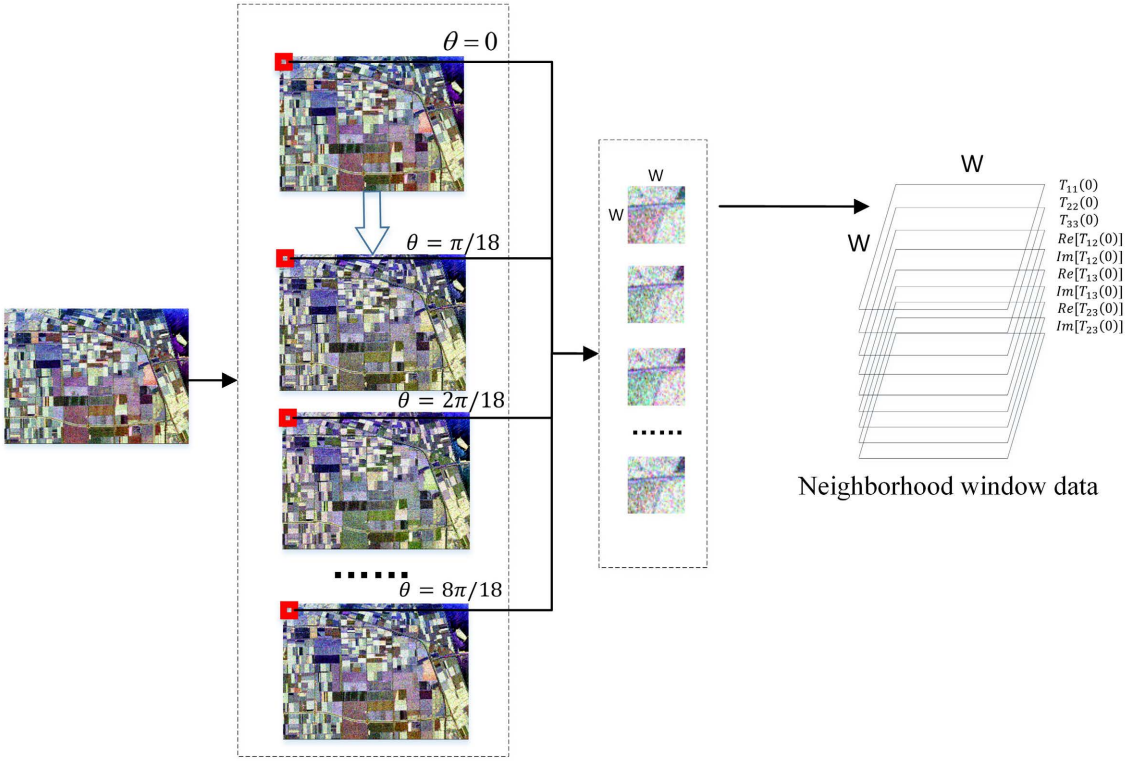

Figure 3. Sequence of polarimetric coherent matrices of the pixel's neighborhood window.

2.2. Polarimetric Feature Sequence Embedding

The original ViT method, based on the linear embedding of the patch sequence. The data to be classified is divided into small image patches in the form of a grid, each small patch is flattened into a one-dimensional vector in a linear manner through the embedding matrix E, and the multiple small patches form a sequence of vectors that can be processed by transformer encoder. In this way, the rough division of the image destroys the spatial structure of the image, and the simple flattening operation of each small patch makes the detailed information within the small patches neglected.

Considering that the PolSAR image itself is composed of multi-polarimetric features, from the perspective of feature dimension; the polarimetric coherence matrices under different polarization orientation angles are selected to form a sequence of vectors as the network input. And the polarimetric information under each polarization orientation angle is transformed into the vector for the subsequent network learning, avoiding spatial segmentation of the image. At the same time, for the polarimetric information under each polarization orientation angle, this paper does not use the linear way to spread into vector form, but uses the convolution to carry out the convolution mapping. This way changes into the

form of a vector sequence that can be processed by the network, and protects the image spatial detail information well and avoids the loss of local details.

The rotation domain polarimetric coherence matrix can be considered as a 4-dimensional matrix 9 × 9 × w × w, where the first 9 represents 9 polarization orientation angle and the second 9 represents the number of channels. Take each polarimetric coherence matrix of the rotation domain within the neighboring window, to form a sequence of polarimetric coherent matrices T as LIViT network inputs, 9 polarization orientation angles, and get 9 feature vectors. The input form is shown in Figure 4.

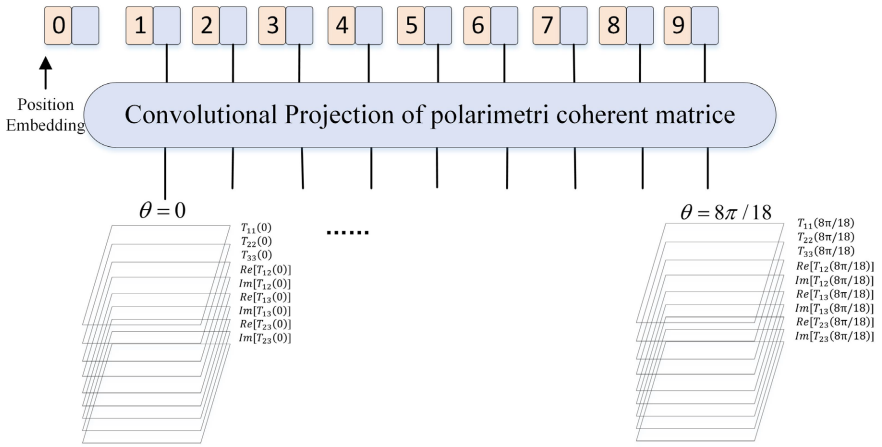

Figure 4. Feature embedding of polarimetric coherent matrice sequence in LIViT network.

Projection part, this paper chooses group convolution to caputure polarimetric information of differernt polarization orientation angles, spatial information was extracted and retained by convolution. The network sets 9 convolutional groups based on 9 polarization orientation angles, and sets three convolutional layers. In the first convolutional layer, each group uses 16 convolution kernel of 5 × 5, step is 1, padding is 2, the second layer uses 32 convolution kernel of 3 × 3, step is 1, padding is 1, two layers of convolution are used in the RelU activation function, as well as maximum pooling layer of 2 × 2. Finally, 64 convolutional kernels of 3 × 3 with a step length of 3 are used to map the features to one dimension. The order of the features is adjusted according to the Transformer encoder input form, resulting in data with dimensions [9, 64], where 9 represents 9 differently polarization orientation angles and 64 is the number of dimensions after feature embedding.

After the feature embedding, increase the class token, and position encoding, fed into the Transformer encoding block, here set up a layer of Transformer Encorder, using the multi-head self-attention mechanism to learn the relationship of polarimetric information under different polarization orientation angles.

2.3. DWT Information Extraction

When Vision Transformer performs PolAR image classification, the embedding

of image patches and the global feature learning approach based on the self-attention mechanism, makes the local detail information under-utilized and unclearly delineated at the feature boundaries. Discrete wavelet transform (DWT) helps to accurately extract the edge features in the image, making the network more attentive to the shape of the target and the feature edge information.

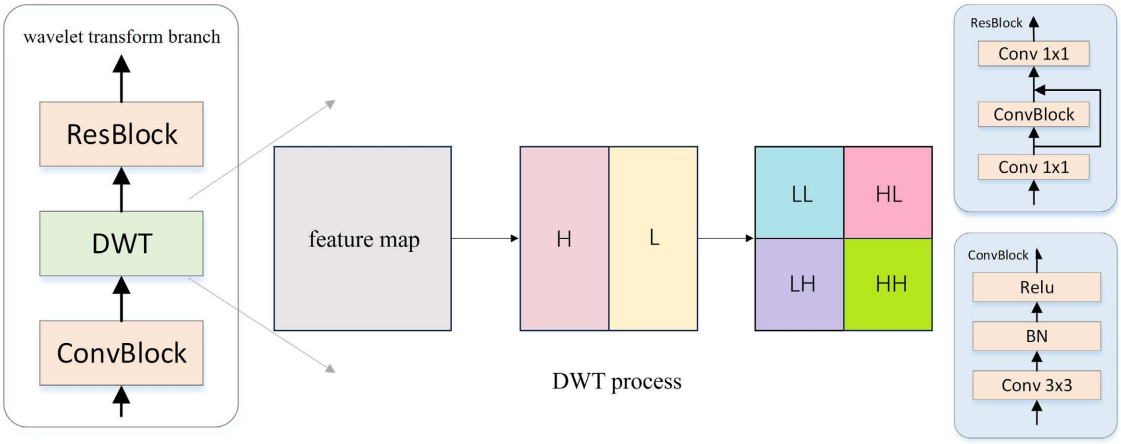

Figure 5. DWT information extraction.

The process of feature extraction using DWT is shown in Figure 5, where the shallow polarimetric information is first extracted using convolution to generate a feature map, which facilitates the DWT to make changes. The input feature map is decomposed into low and high frequency components by DWT and the network can learn from the high and low frequency components. In 2D discrete wavelet transform, there are four filters namely low pass filter fLL and high pass filters fLH, fHL, fHH. By convolving with each filter, the image or feature map can be decomposed into four sub bands namely LL, LH, HL and HH. By using DWT, the knowledge of the frequency domain that preserves the details of the blurred image can be gained, especially from the LH, HL and HH. By converting the DWT extracted features then go through the ResBlock module so that the network can further integrate the information from the spatial and frequency information.

3. Results and Discussion

3.1. Data Description

In order to consider the classification effect in different scenarios, we choose Wuhan area in China for the urban scenario and the Flevoland area in the Netherlands for the agricultural scenario as the experimental datasets.

1) GF-3 WuHan: The C-band full polarimetric image of Wuhan urban area in China, acquired by Gaofen-3 (GF-3) sensor. It covers the local area of Wuhan with a spatial resolution of 5.20 × 2.25 m and an image size of 605 × 923 pixels. There are five feature classes, buildings, agriculture, water bodies, vegetation 1 and vegetation 2. Figure 6 shows its Pauli RGB map and ground truth map, which was

manually produced by Jiang Tao's team at Shandong University of Science and Technology through high-resolution optical remote sensing images.

2) AIRSAR Flevoland: The L-band full polarimetric image of the agricultural region in the Netherlands, acquired with the NASA/JPL Laboratory AIRSAR sensor in the United States, with a spatial resolution of 6.6 × 12.1 m and an image size of 750 × 1024 pixels. Since Lee's use [26], this dataset has been widely used in studies of feature classification due to its well labeled map. The data has 15 feature classes, namely stem bean, pea, alfalfa, wheat, sugar beet, potato, oilseed rape, barley, wheat 2, wheat 3, forest, water body, building, grassland and bare ground. Its Pauli RGB map and ground truth map are shown in Figure 7.

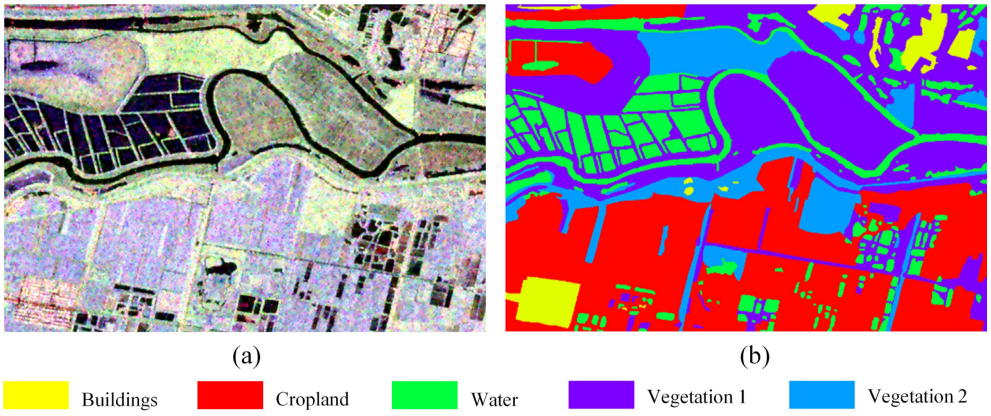

Figure 6. GF-3 WuHan dataset. (a) Pauli RGB map. (b) Ground truth map.

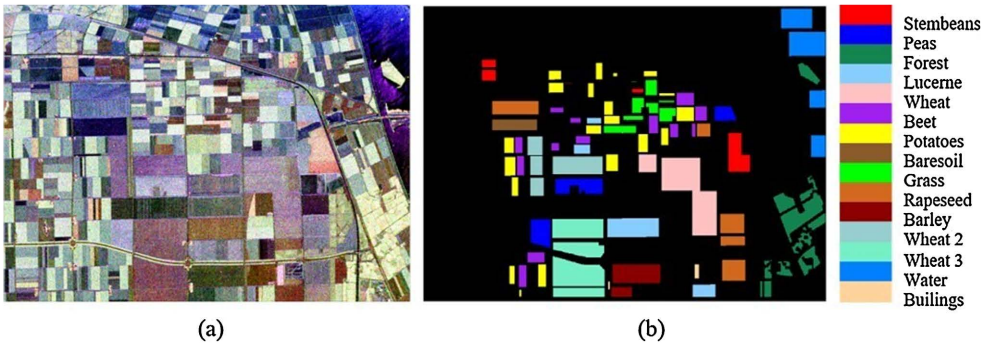

Figure 7. AIRSAR Flevoland dataset. (a) Pauli RGB map. (b) Ground truth map.

3.2. Experimental Setup

1) Evaluation Indicators: overall accuracy (OA), average accuracy (AA), kappa coefficient (kappa) are selected as evaluation criteria in this paper. They can be computed as follows:

$$\mathrm{OA} = \frac{\sum_{i=1}^{c} M_i}{\sum_{i=1}^{c} N_i} \tag{5}$$

$$\mathrm{AA} = \frac{1}{c}\sum_{i=1}^{c} \frac{M_i}{N_i} \tag{6}$$

In the formula, c is the number of categories, M_i, N_i denote the total number of correctly categorized samples in category i and the total number of labeled samples in category i, respectively.

$$\text{kappa} = \frac{\text{OA} - P}{1 - P}, \quad P = \frac{1}{N^2} \sum_i^c S_{i:} S_{:i} \tag{7}$$

Among them, N is the total number of samples, and $S_{i:}$ and $S_{:i}$ are the sum of the i-th row and i-th column elements of the confusion matrix respectively.

Here, OA is used to represent the proportion of all correct results in total data and AA is used to represent the accuracy for a given category. In contrast to their involvement with only correctly predicted samples, *kappa* takes into account a variety of missing and misclassified samples, measuring the consistency of the predicted output with the ground truth [27].

2) Configuration: Experimental environment: a PC with Intel (R) Xeon (R) Gold 5320 CPU, Nvidia RTX-A4000 GPU (16 GB RAM) and 32 GB RAM.

3) Parameter Analysis: For both images, all pixels with ground truth are taken as sample sets, where 4% of each object category is randomly selected as the training set, 1% as the validation set, and the remaining sample pixels are the test set. After the classification network is trained by the training set and validation set, the whole image is input to get the prediction result of each pixel, where the test set is used to verify the classification effect. In order to ensure the comparability of the methods, the method comparison is conducted under the same randomly selected samples, and specific information for the dataset division is shown in **Table 1**.

Table 1. The detail information of three datasets used in experiments.

Dataset	Sample set	Training num	Validation num	Testing num
WuHan	558415	22336 (4%)	5584 (1%)	530495
Flevoland	157296	6492 (4%)	1573 (1%)	149431

Considering the effect of neighborhood window size on classification results, it is set to vary from 7 * 7 to 21 * 21, and the classification effect is examined on two images with LIViT network as an example. **Figure 8** shows that the classification accuracy of the images increases as the neighborhood window increases, but the increase in accuracy is not obvious after increasing to a certain size. Flevoland image presents the best classification effect under the neighborhood window size of 15 * 15, and Wuhan image shows a slow increase in accuracy after the neighborhood window size reaches 15 * 15. Considering the influence of neighborhood window size on classification accuracy and space complexity, all classification methods in this paper use 15 * 15 neighborhood window data to represent the center pixel.

3.3. Classification Results of PolSAR Datasets

To verify the applicability of the proposed LIViT method over other methods, a

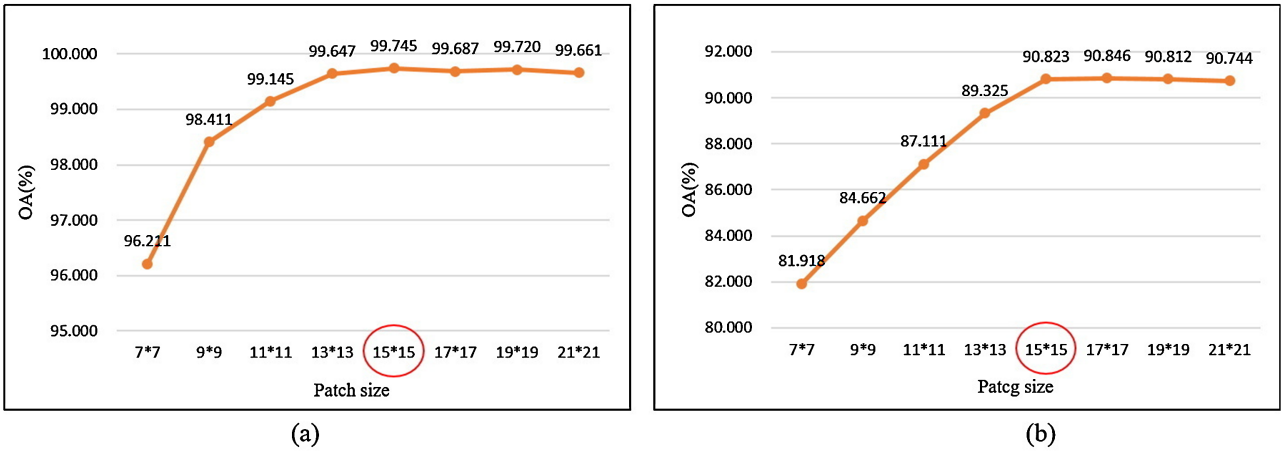

Figure 8. OA under patches of different size. (a) AIRSAR Flevoland. (b) GF3 Wuhan.

comparison is made with existing advanced algorithms for PolSAR image classification, including support vector machine (SVM), classical convolution-based two-dimensional convolutional neural network (2D-CNN), three-dimensional convolutional neural network (3D-CNN), attention-based polarimetric feature selection convolutional network (AFS-CNN) [13], and Convolutional Long Short Term Memory network (ConvLSTM) [28] that is good at dealing with spatial-temporal data, and the network inputs are all using polarimetric coherent matrix sequences. Meanwhile, in order to test the role of the polarimetric feature sequence embedding module and the two-branch module, the ViT network using only polarimetric coherent matrix T sequences for embedding is named TViT, which participates in the comparison of the results.

SVM method is easy to understand and implement, and has better performance in image classification. Two-dimensional convolution is adept at learning the local context information of the image. In this paper, we set up two layers of convolution, two layers of pooling, three layers of full connected, and the convolution part is the same as the single-group convolution setup of LIViT convolutional embedding. Three-dimensional CNN is good at learning spatial polarimetric information; the framework structure is set the same as two-dimensional CNN, and only convolutional layers are replaced by three-dimensional. AFS-CNN is the network structure set by Dong scholars. Attention-based selection structure added to CNN classifier, enables feature selection at multidimensional feature inputs. ConvLSTM network adopts the network structure set by Lei Wangto realize the classification of spatial-temporal sequence data.

Figure 9 and Table 2 show the classification results of Wuhan images using different methods are demonstrated. Firstly, it can be seen that the SVM method has more misclassification points, and the deep learning method has obvious advantages over the traditional machine learning method SVM, and the classification performance has obvious improvement. Compared with different deep learning methods, ViT-based methods have more advantages, and LIViT network has the best classification effect. From the results, it can be seen that 2D

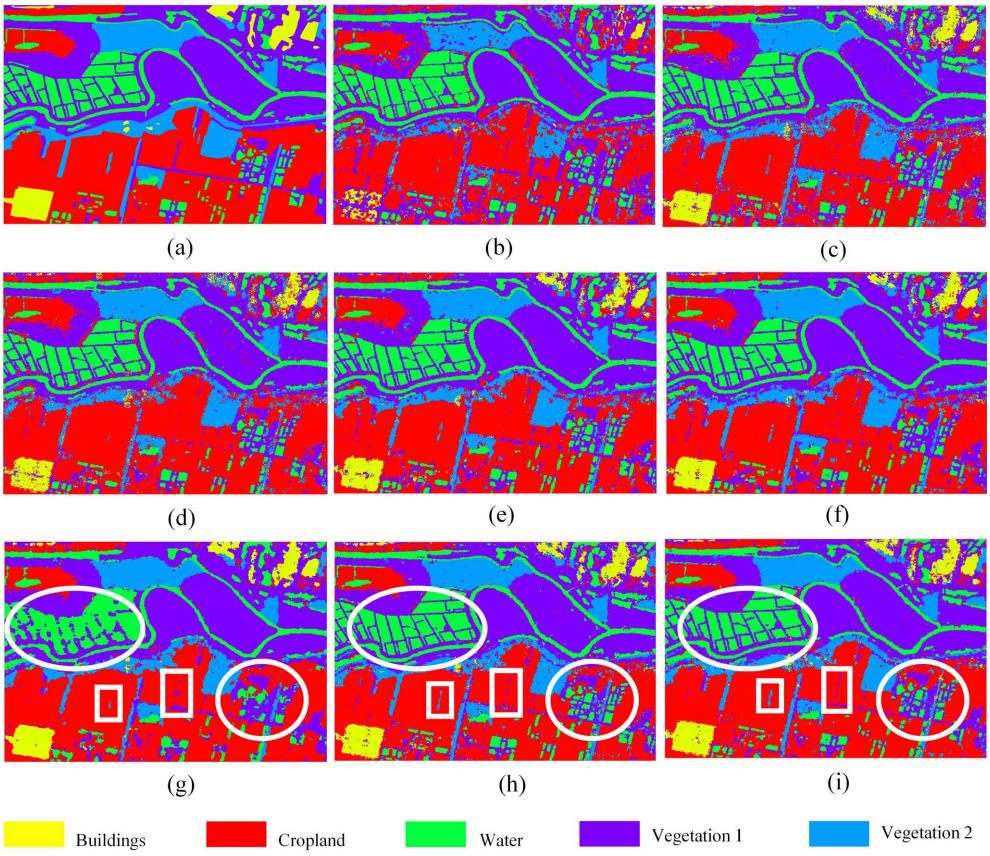

Figure 9. Classification results of whole map on GF 3 WuHan data with different methods. (a) Ground truth map. (b) SVM. (c) CNN. (d) 3D CNN. (e) AFSCNN. (f) ConvLSTM. (g) ViT. (h) TViT. (i) LIViT.

Table 2. Classification results of different networks on WuHan dataset (%).

Category	SVM	CNN	3DCNN CNN	AFS-CNN	ConvLSTM	ViT	TViT	LICViT
Buildings	24.860	74.981	61.813	69.795	73.786	84.837	86.483	85.458
Cropland	79.659	87.016	90.076	88.952	88.977	93.026	91.059	93.176
Water	93.641	91.064	89.331	92.957	90.262	83.350	90.889	91.102
Vegetation 1	82.318	85.963	81.991	85.372	87.657	87.996	90.631	90.544
Vegetation 2	60.908	80.424	75.019	81.236	82.992	86.147	84.155	86.374
OA	78.519	86.071	84.341	86.723	87.448	88.680	89.914	90.823
AA	68.277	83.890	79.646	83.662	84.735	87.091	88.643	89.311
Kappa	69.811	80.753	78.203	81.622	82.585	84.310	86.045	87.297

convolution, 3D convolution, and AFS-CNN networks use convolution to extract information on high dimensional polarimetric SAR images with a weak effect and a somewhat lower accuracy. ConvLSTM method has a good ability to process sequence features, it makes better use of the polarimetric features, and obtains higher accuracy compared to the convolutional method with the polarimetric

coherence matrix sequence as input. However, the ViT method also has excellent sequence feature learning ability and obtains higher accuracy on Wuhan images.

Figure 9 compares the ViT, TViT and LIViT methods, as in the white circle, it can be seen that although the ViT method obtains better result. There is a certain amount of unclear demarcation of the detail boundaries, which is most obvious in the water body part. In the result map, the details of the intersection between water body and vegetation 1 are blurred. At the same time, due to the excessive global learning ability of the ViT network, small land parcels of vegetation 1 and vegetation 2 in the farmland are partially obliterated. While the resultant maps classified by TViT method and LIViT, the classification boundaries between water bodies and vegetation 2 are kept intact, the small patches of vegetation 1 and vegetation 2 in the farmland are well presented as a whole, and the local detail information of the images is well preserved. The results of TViT demonstrate the advantage of polarimetric feature sequence embedding over image patches embedding for detail preservation. LIViT adds the DWT feature extraction module, which makes the network focus more on the image edge information compared to TViT network, further enhances the edge extraction, and improves the classification accuracy of the network.

In Wuhan images, LIViT network obtains the best classification results compared to other methods, its overall classification accuracy is 12.304% higher compared to SVM, 4.752% higher compared to 2D CNN, 6.482% higher compared to 3D CNN, 4.1% higher compared to AFS-CNN, 3.375% higher compared to ConvLSTM, and its accuracy is improved by 2.143% on the basis of ViT and 0.909% accuracy improvement on TViT. Compared to the ViT method, LIViT improves 7.752 in water bodies and 2.548% in vegetation 1, and the accuracy improvement is obvious in features with more boundary detail information.

Figure 10 and Table 3 show the classification results of different networks for Flevoland images are demonstrated. It can also be seen in Flevoland images that the ViT method is more advantageous than SVM and other deep learning methods. ConvLSTM obtains the result close to the ViT method, but the complexity and time cost are too high compared to the ViT method. Meanwhile, LIViT obtains a certain accuracy improvement based on both ViT and TViT.

A detailed comparison of ViT, TViT, and LIViT is presented in Figure 10. As the features in the red box, it can be seen that in the ViT method, the forest, which is in the form of a long strip, is much overwhelmed by the surrounding wheat 3. In TViT and LIViT methods the forest of thin strips is better preserved due to the fact that local detail information is retained by switching the learning of the spatial global to the learning of a sequence of polarimetric features. At the same time, it can be seen on the Flevoland image that the delineation on the boundaries between parcels is more regular after taking into account the local information on the basis of ViT. After adding DWT, the extraction of feature boundaries is clearer.

The LIViT method also obtains the best classification results in Flevoland

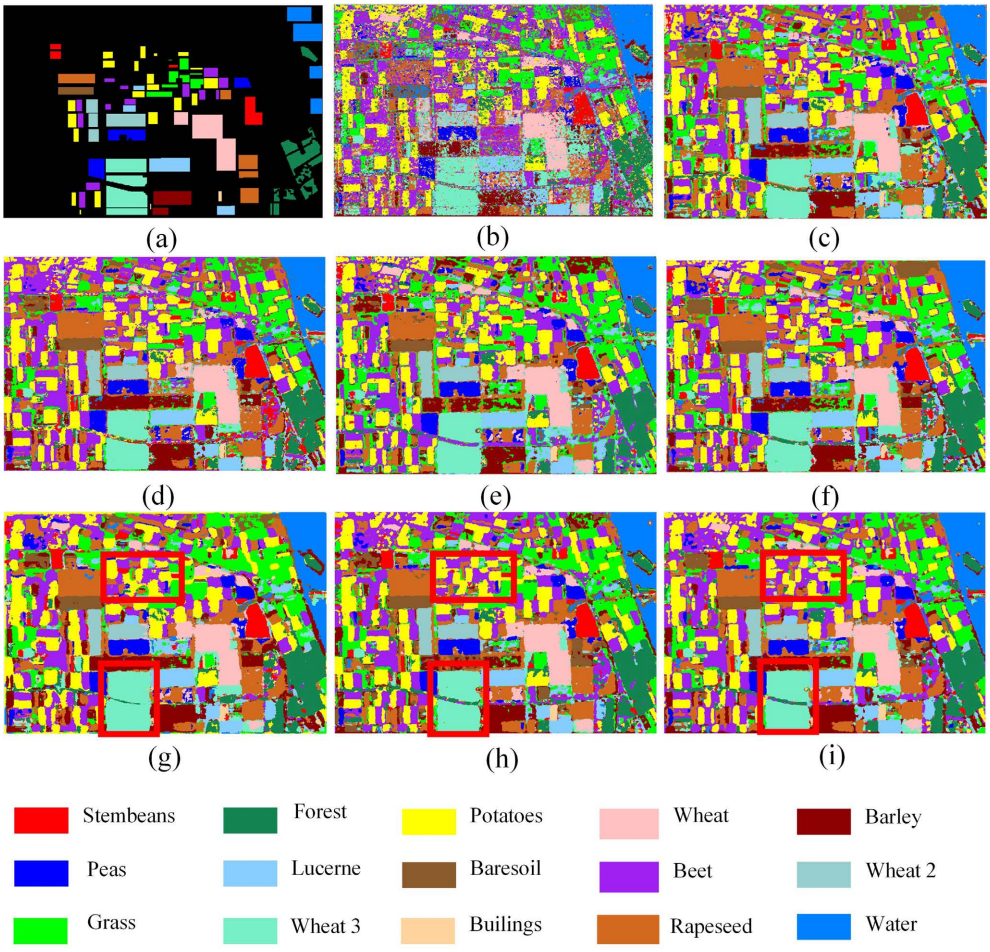

Figure 10. Classification results of whole map on AIRSAR Flevoland data with different methods. (a) Ground truth map. (b) SVM. (c) CNN. (d) 3D CNN. (e) AFSCNN. (f) ConvLSTM. (g) ViT. (h) TViT. (i) LIViT.

Table 3. Classification results of different networks on Flevoland dataset (%).

Category	SVM	CNN	3DCNN	AFS-CNN	ConvLST	ViT	TViT	LIViT
Stembean	79.177	99.828	99.776	99.551	**99.983**	99.601	99.603	99.551
Peas	87.946	99.861	98.544	99.041	99.700	**99.988**	99.803	**99.988**
Forest	92.129	99.380	98.626	99.894	99.732	99.211	99.837	**99.943**
Lucerne	86.343	98.778	98.789	98.334	**99.889**	99.789	99.300	99.611
Wheat 1	86.969	99.074	97.083	98.867	98.703	99.208	99.677	**99.780**
Beet	78.870	97.047	98.932	98.586	97.759	95.256	99.434	**99.581**
Potatoes	84.323	98.520	99.635	99.525	99.842	99.656	99.786	**99.903**
Baresoil	57.682	99.761	99.316	98.290	**100.000**	98.564	**100.000**	**100.000**
Grass	81.497	96.021	95.803	96.255	98.170	98.237	97.71	**98.707**
Rapeseed	58.208	99.112	97.014	97.934	98.922	98.930	99.452	**99.651**

Continued

Barley	80.482	98.338	98.514	95.837	**99.868**	99.706	99.117	99.720
Wheat 2	51.414	99.115	98.251	99.145	97.615	97.525	**99.562**	99.443
Wheat 3	92.800	99.600	99.456	99.609	99.832	99.634	99.822	**99.896**
Water	96.178	99.992	99.930	99.984	99.984	99.649	**100**	**100**
Buildings	76.821	79.470	87.859	90.728	87.417	93.819	94.701	**98.896**
OA	82.138	98.931	98.556	98.877	99.260	99.022	99.578	**99.745**
AA	79.389	97.593	97.835	98.105	98.494	98.607	99.187	**99.645**
Kappa	80.444	98.833	98.424	98.774	99.192	98.933	99.539	**99.722**

images. Its overall classification accuracy is 17.354% higher compared to SVM, 0.814% higher compared to 2D CNN, 1.189% higher compared to 3D CNN, 0.868% higher compared to AFS-CNN, 0.485% higher compared to ConvLSTM, and improves by 0.723% on top of ViT, and 0.167% on top of TViT.

The classification results of Wuhan image and Flevoland image demonstrate that the LIViT network, which takes into account local information, has a good advantage in PolSAR image classification.

4. Conclusion

In this paper, we propose a PolSAR image classification method that takes into account the local information of Vision Transformer, LIViT network. On the one hand, the method uses group convolution to map the polarimetric feature sequences composed of different polarization orientation angles, which retains the spatial detail information of the image well. On the other hand, the use of the wavelet transform branch enhances the ability to extract edge information about the feature target. Compared with the ViT network, this method obtains a great improvement in local detail retention, and at the same time, it obtains the best classification effect compared with many methods, which demonstrates the strong classification ability.

Acknowledgements

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA26010201).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jafari, M., Maghsoudi, Y. and Valadan Zoej, M.J. (2015) A New Method for Land Cover Characterization and Classification of Polarimetric SAR Data Using Polarimetric Signatures. *IEEE Journal of Selected Topics in Applied Earth Observations*

and Remote Sensing, **8**, 3595-3607. https://doi.org/10.1109/jstars.2014.2387374

[2] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[3] Zhou, Y., Wang, H., Xu, F. and Jin, Y. (2016) Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, **13**, 1935-1939. https://doi.org/10.1109/lgrs.2016.2618840

[4] Hua, W., Wang, S., Xie, W., Guo, Y. and Jin, X. (2019) Dual-Channel Convolutional Neural Network for Polarimetric SAR Images Classification. *IGARSS* 2019-2019 *IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, 28 July-2 August 2019, 3201-3204. https://doi.org/10.1109/igarss.2019.8899103

[5] Zhang, Z., Wang, H., Xu, F. and Jin, Y. (2017) Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **55**, 7177-7188. https://doi.org/10.1109/tgrs.2017.2743222

[6] Zhang, L., Chen, Z., Zou, B. and Gao, Y. (2018). Polarimetric SAR Terrain Classification Using 3D Convolutional Neural Network. *IGARSS* 2018-2018 *IEEE International Geoscience and Remote Sensing Symposium*, Valencia, 22-27 July 2018, 4551-4554. https://doi.org/10.1109/igarss.2018.8519557

[7] Tan, X., Li, M., Zhang, P., Wu, Y. and Song, W. (2020) Complex-Valued 3-D Convolutional Neural Network for PolSAR Image Classification. *IEEE Geoscience and Remote Sensing Letters*, **17**, 1022-1026. https://doi.org/10.1109/lgrs.2019.2940387

[8] Chen, S. and Tao, C. (2018) PolSAR Image Classification Using Polarimetric-Feature-Driven Deep Convolutional Neural Network. *IEEE Geoscience and Remote Sensing Letters*, **15**, 627-631. https://doi.org/10.1109/lgrs.2018.2799877

[9] Cui, Y., Liu, F., Jiao, L., Guo, Y., Liang, X., Li, L., *et al.* (2022) Polarimetric Multipath Convolutional Neural Network for PolSAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1-18. https://doi.org/10.1109/tgrs.2021.3071559

[10] Zhang, L., Dong, H. and Zou, B. (2019) Efficiently Utilizing Complex-Valued PolSAR Image Data via a Multi-Task Deep Learning Framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, **157**, 59-72. https://doi.org/10.1016/j.isprsjprs.2019.09.002

[11] Wang, L. (2020) Polarimetric SAR Image Information Representation and Classification Based on Deep Learning.

[12] Wang, J., Hou, B., Ren, B., Zhang, Y., Yang, M., Wang, S., *et al.* (2022) Parameter Selection of Touzi Decomposition and a Distribution Improved Autoencoder for PolSAR Image Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, **186**, 246-266. https://doi.org/10.1016/j.isprsjprs.2022.02.003

[13] Dong, H., Zhang, L., Lu, D. and Zou, B. (2022) Attention-Based Polarimetric Feature Selection Convolutional Network for PolSAR Image Classification. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1-5. https://doi.org/10.1109/lgrs.2020.3021373

[14] Guo, J., Li, H., Ning, J., Han, W., Zhang, W. and Zhou, Z. (2020) Feature Dimension Reduction Using Stacked Sparse Auto-Encoders for Crop Classification with Multi-Temporal, Quad-PolSAR Data. *Remote Sensing*, **12**, Article No. 321. https://doi.org/10.3390/rs12020321

[15] Yang, C., Hou, B., Ren, B., Hu, Y. and Jiao, L. (2019) CNN-Based Polarimetric Decomposition Feature Selection for PolSAR Image Classification. *IEEE Transactions*

on Geoscience and Remote Sensing, **57**, 8796-8812.
https://doi.org/10.1109/tgrs.2019.2922978

[16] Zhang, W., Wang, M., Guo, J. and Lou, S. (2021) Crop Classification Using MSCDN Classifier and Sparse Auto-Encoders with Non-Negativity Constraints for Multi-Temporal, Quad-PolSAR Data. *Remote Sensing*, **13**, Article No. 2749.
https://doi.org/10.3390/rs13142749

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[18] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[19] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., *et al.* (2018) Exploring the Limits of Weakly Supervised Pretraining. In: Ferrari, V., *et al.*, Eds., *Computer Vision—ECCV* 2018, Springer International Publishing, 185-201.
https://doi.org/10.1007/978-3-030-01216-8_12

[20] Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A. and Chanussot, J. (2022) Spectral-Former: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1-13.

[21] Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G., *et al.* (2023) Transformers in Remote Sensing: A Survey. *Remote Sensing*, **15**, Article No. 1860.
https://doi.org/10.3390/rs15071860

[22] Liu, X., Wu, Y., Liang, W., Cao, Y. and Li, M. (2022) High Resolution SAR Image Classification Using Global-Local Network Structure Based on Vision Transformer and CNN. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1-5.
https://doi.org/10.1109/lgrs.2022.3151353

[23] Dong, H., Zhang, L. and Zou, B. (2022) Exploring Vision Transformers for Polarimetric SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1-15. https://doi.org/10.1109/tgrs.2021.3137383

[24] Wang, H., Xing, C., Yin, J. and Yang, J. (2022) Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer. *Remote Sensing*, **14**, Article No. 4656. https://doi.org/10.3390/rs14184656

[25] Chen, S.-W., Wang, X.-S. and Sato, M. (2014) Uniform Polarimetric Matrix Rotation Theory and Its Applications. *IEEE Transactions on Geoscience and Remote Sensing*, **52**, 4756-4770. https://doi.org/10.1109/tgrs.2013.2284359

[26] Lee, J.-S., Grunes, M.R. and Pottier, E. (2001) Quantitative Comparison of Classification Capability: Fully Polarimetric versus Dual and Single-Polarization SAR. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 2343-2351.
https://doi.org/10.1109/36.964970

[27] Yang, C. and Wang, G. (2022) Research on Agricultural Image Classification Method Based on Transformer.

[28] Wang, L., Xu, X., Gui, R., Yang, R. and Pu, F. (2020) Learning Rotation Domain Deep Mutual Information Using Convolutional LSTM for Unsupervised PolSAR Image Classification. *Remote Sensing*, **12**, Article No. 4075.
https://doi.org/10.3390/rs12244075