Crop classification of multi-temporal PolSAR based on 3D attention module with ViT

Qiang Yin, Zhiyuan Lin, Wei Hu, Carlos López-Martínez, Jun Ni, Fan Zhang

***Abstract*—Multi-temporal polarimertic SAR is considered to be very effective in crop classification and cultivated land detection, which has received much attention from researchers. Currently, for most multi-temporal polarimetric SAR data classification methods, the simultaneous temporal-polarimetric-spatial feature extraction capability has not been exploited sufficiently. Also, the diversity of different time and different polarimetric features has not been taken into account sufficiently. In this paper, we propose a classification model that combines a dual-stream network as a temporal-polarimetric-spatial feature extraction module with Vision Transformer(ViT) called Temporal-Polarimetric-Spatial Transformer(TSPT) to address the above problems. Secondly, a 3 dimension(3D) convolutional attention module that enables the network to weight the temporal dimension, polarimetric feature dimension and spatial dimension is developed, according to their importance. Experimental results on both UAVSAR and RADARSAT-2 datasets show that the proposed method outperforms ResNet.**

***Index Terms*—multi-temporal PolSAR, crop classification, Vision Transformer, temporal-polarimetric-spatial characteristics.**

I. Introduction

Arable land is critical to human survival and socioeconomic development. Timely and accurate access to crop information can improve various food production and food security issues. There is a wide variation in the scattering exhibited by crops at different growth cycles. This leads to the fact that different crops may also exhibit similar scattering at certain time periods, which makes classification with mono-temporal polarimetric SAR data prone to misclassification. Fortunately, multi-temporal polarimetric SAR data which contain the variation of scattering mechanisms can distinguish different crops by finding the scattering regularity of similar crops over time. Several scholars have demonstrated that multi-temporal SAR data can better express the distributional characteristics of vegetation than single-temporal data[1, 2]. With the continuous development of SAR system, it provides more and more possibility of acquiring multi-temporal polarimetric SAR data.

In recent years, deep learning has flourished, so various Convolutional Neural Networks(CNNs) have become the mainstream networks in image processing due to their excellent feature extraction capabilities[3]. Usually, when using CNN to analyze the pixel-level features of SAR data, it is often necessary to consider the pixel of the data and its surrounding neighborhood information. Considering the computer resources and the compatibility of network models, SAR data are often cut into multiple small patches and input to the CNNs, but the number of layers and downsampling of the network is limited due to the often small size of the cut patches. In 2020, ViT(Vision Transformer) was proposed to provide a new option in the image classification field[4]. It cuts the image into multiple blocks of the same size, almost regardless of the size of the image, requiring only that the size of each block cut be divisible by the original image. Recently, ViT has been just demonstrated to show good results in polarimetric SAR image classification[5, 6].

As for the classification of multi-temporal polarimetric SAR data, a large number of scholars have already used deep learning methods[7, 8]. However, most of the common multi-temporal polarimetric SAR data interpretation models stitch together features of different time sequences or treat each pixel as a group of time series values and thus perform sequential classification[9, 10]. The former approach ignores the description of the variation of polarimetric scattering features in the time dimension, while the latter doesn't make full use of the spatial information in the images, where spatial relationship refers to the relationship between the distribution locations of various crops in the whole SAR data. These methods do not integrate the correlation between polarimetric features and temporal features and lack excellent temporal-polarimetric-spatial feature extraction capability. Three dimentional Convolutional Neural Networks(3DCNN) is a commonly used spatial-temporal feature extraction network that has been applied to human action recognition for a long time[11]. Hara *et al.* extended ResNet to 3D and demonstrated the effectiveness of 3DResNet on video classification[12]. However, using only 3DCNN for certain temporal image classification such as action recognition is still a great challenge. As of now, the dual-stream model is one of the most effective methods in the field of action recognition when the size of the training dataset is limited[13], which provides a brand new idea for multi-dimensional feature extraction.

In addition, the heterogeneity among polarimetric features and the real growth cycle of crops lead to complexity between different polarization and different time phases, and it is also a challenge in the classification model. The existing multi-

This work was supported in part by the National Natural Science Foundation of China under Grant 62271027, 62201034. (Corresponding author : Fan Zhang.)

Qiang Yin, Zhiyuan Lin, Wei Hu, Fan Zhang are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Jun Ni is with the College of Information Science and Engineering, Yunnan University, Yunnan 650500, China

Carlos López-Martínez is with the Signal Theory and Communication Department, Polytechnic University of Catalonia, Barcelona 08034, Spain.

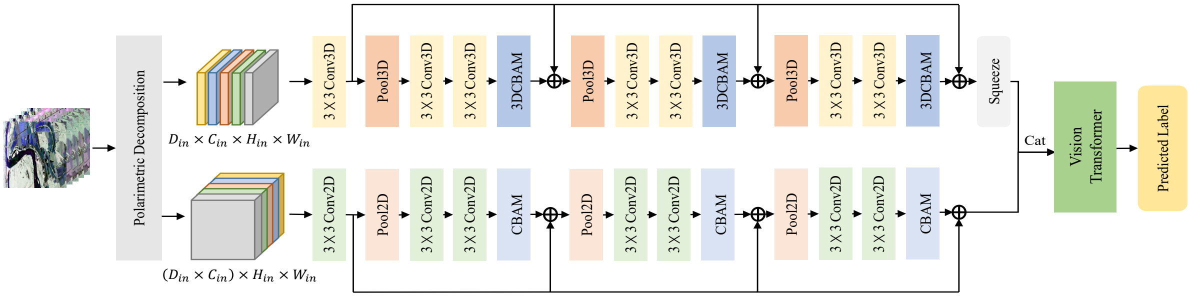

Fig. 1. The structure of TPST.

temporal polarimetric SAR data interpretation models can neither consider three dimensions, namely: time, polarization and space, nor take into account the importance of different polarization features and different time phases in classification [8].Therefore, the 3D convolutional attention module is needed to address this existing problem.

The contributions of the work in this paper are summarized as follows.

- To address the problem of small input size of classification network for SAR data which is not beneficial for feature extraction, ViT is introduced for the first time in multi-temporal polarimetric SAR data classification;
- To propose an image interpretation model called TPST combining 3D convolution with ViT, using 3D convolution for temporal-polarimetric-spatial feature extraction and then as the input into ViT for classification;
- To develop a 3D convolutional attention module, which adds temporal dimensional attention to traditional attention modules. It could weight them according to their importance from three dimensions: time, polarization and space.

II. Methodology

In this paper, we design a Vision Transformer network structure, namely Temporal-Polarimetric-Spatial Transformer(TPST) which can extract temporal-polarimetric-spatial features. In this method, a 3D convolutional attention module is proposed to expand the scope to three dimensions. The overall flow of the method is shown in Fig. 1.

A. *Polarimetric feature selection*

Under the horizontal and vertical polarimetric basic (H, V), the polarimetric SAR system acquires the full polarimetric information of the target which is characterized by the polarimetric scattering matrix,

$$\mathbf{S} = \left[\begin{array}{cc} \boldsymbol{S}_{HH} & \boldsymbol{S}_{HV} \\ \boldsymbol{S}_{VH} & \boldsymbol{S}_{VV} \end{array} \right] \tag{1}$$

where H represents the horizontally polarimetric emitted or received electromagnetic wave signal and V is the vertically emitted or received electromagnetic wave signal. In the monostatic case, the polarimetric coherence matrix commonly used in polarimetric SAR information processing can be obtained when the reciprocity assumption condition ($S_{HV} = S_{VH}$) is satisfied.

$$\begin{aligned} \mathbf{T}_3 &= \left\langle k_P k_P^H \right\rangle \\ &= \left[\begin{array}{ccc} \left\langle \left| S_{HH}' + S_{VV} \right|^2 \right\rangle & \left\langle \left(S_{HH}' + S_{VV}' \right) \left(S_{HH}' - S_{VV}' \right)^* \right\rangle & 2 \left\langle \left(S_{HH}' + S_{VV}' \right) S_{HV} \right\rangle \\ \left\langle \left(S_{HH}' - S_{VV}' \right) \left(S_{HH}' + S_{VV}' \right)^* \right\rangle & \left\langle \left| S_{HH}' - S_{VV}' \right| \right\rangle^2 & 2 \left\langle S_{HV}' \left(S_{HH}' + S_{VV}' \right)^* \right\rangle \\ 2 \left\langle S_{HV}' \left(S_{HH}' + S_{VV}' \right)^* \right\rangle & 2 \left\langle S_{HV}' \left(S_{HH}' - S_{VV}' \right)^* \right\rangle & 4 \left\langle S_{HV}'^2 \right\rangle \end{array} \right] \end{aligned} \tag{2}$$

$$k_P = \frac{1}{\sqrt{2}} \left[\begin{array}{ccc} S_{HH} + S_W & S_{HH} - S_W & 2S_{HV} \end{array} \right]^{\mathrm{T}} \tag{3}$$

In Eq. 2, k_p is the Pauli scattering vector, and $(.)^H$ denotes the conjugate transpose. $\langle . \rangle$ is the ensemble mean.

Target decomposition theory is an important tool for polarimetric SAR data characterization. It decomposes the polarimetric matrix into a weighted sum of multiple independent components to decode the scene scattering characteristics of polarimetric SAR images. Using target decomposition theory, the scattering types of features can be expressed quantitatively, which helps the efficient unfolding of classification tasks, and thus is widely used in related fields such as soil moisture estimation, plant growth prediction, and building feature extraction etc.

Specifically, the approach of polarimetric decomposition is usually to process the covariance matrix, coherence matrix or Kennaugh matrix to decompose the scattering mechanism of the target and thus obtain more polarimetric features. Previous studies have demonstrated the effectiveness of the H/A/α and the Freeman-Durden decompositions for land classification, therefore, in this paper, nine of the decomposition components obtained from these two methods were screened to construct feature combination based on independent experiments in a previous work[14].

B. 3D convolution and feature extraction

Unlike the single temporal data, the correlation information seen in adjacent times of temporal data is very important. The proposed model in this paper attempts to extract features from temporal, polarimetric and spatial dimensions in order to capture information about the changes of images in multiple adjacent times. The specific approach is a combination of 3D convolution and dual-stream network. We arrange the data in temporal order and use this order relationship to describe the temporal relationship, thus enabling 3D convolution to extract spatial and polarimetric information along with the temporal information. And the process of each polarimetric channel is shown in Fig. 2. Its input layer can be expressed as $D_{in} \times C_{in} \times H_{in} \times W_{in}$, where D, C, H and W represent the depth, number of channels of the data, height and width, respectively. For multi-temporal polarimetric SAR data, where D represents the number of time and C depicts the number of selected polarimetric features.

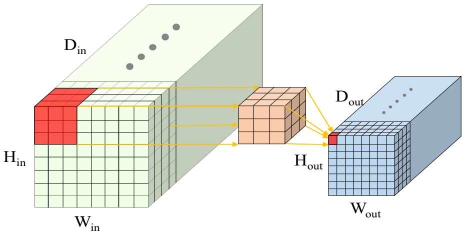

Fig. 2. The schematic of 3D convolution.

The size of the 3D convolution kernel can also be expressed in terms of length, width and depth. Suppose k convolution kernels of size $d \times h \times w$ are used, "Stride" is s and "Padding" is p. Then the relationship between the output layer and the input layer can be described as

$$\begin{cases} D_{\text{out}} = \left(D_{\text{in}} + 2p - d\right)/s + 1 \\ C_{\text{out}} = k \\ H_{\text{out}} = \left(H_{\text{in}} + 2p - h\right)/s + 1 \\ W_{\text{out}} = \left(W_{\text{in}} + 2p - w\right)/s + 1 \end{cases} \tag{4}$$

The size of the output layer is $D_{out} \times C_{out} \times H_{out} \times W_{out}$.

Dual-stream networks are commonly used for spatial-temporal feature extraction, but unlike temporal information for motion recognition, multi-temporal polarimetric SAR data has large time intervals for each data set, often up to several tens of days. Even for images at adjacent times, the scattering of certain crops may have large variations due to changes in the growth cycle. Hence, following the example of video analysis to extract motion information from optical flow cannot achieve good results. Therefore, this paper replaces the original optical flow layer with 3D convolution and designs a spatial-temporal feature extraction module applicable to multi-temporal polarimetric SAR data as shown in the left part of Fig. 1. As shown in upper stream of the figure, the input to the 3D convolutional branch is composed of 4D tensors for each sample and the shape can be expressed as number of time × number of polarimetric feature channels × height × width. The input of the lower stream of 2D convolutional branch, on the other hand, is that the polarimetric features of all times are concated in the dimension of the polarimetric feature channel. The different network structures of the two branches and the different input shapes allow the two branches to learn distinct features.

C. 3D Convolutional Block Attention Module(3DCBAM)

In this paper, we make an improvement to the existing channel attention module and spatial attention module to make them applicable to multi-temporal polarimetric SAR data. Its overall structure is shown in Fig. 3.

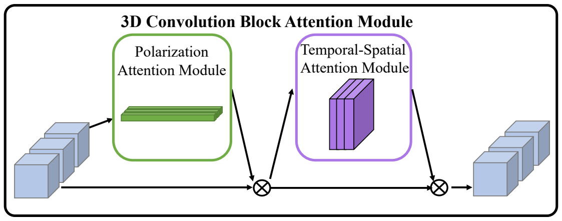
3D Convolution Block Attention Module
Polarization Attention Module
Temporal-Spatial Attention Module

Fig. 3. The schematic of 3DCBAM.

We replace the original two 2D tensors with the corresponding 3D maximum pooling and 3D average pooling so that they can assign corresponding weights to the different temporal dimensions of the image. In addition, we replace the original 7×7 convolution with 3×3 to make it more suitable for the small input sample size of our SAR dataset. The Polarization Attention Module $\mathbf{M_P}$ and the Temporal-Spatial Attention Module $\mathbf{M}_{\text{TS}}$ can be expressed separately as:

$$\mathbf{M_P}(\mathbf{F}) = \sigma(\text{MLP}(\text{AvgPool}_{3\text{D}}(\mathbf{F})) + \text{MLP}(\text{MaxPool}_{3\text{D}}(\mathbf{F}))) \tag{5}$$

$$\mathbf{M}_{\text{TS}}(\mathbf{F}) = \sigma\left(f^{3\times3}([\text{AvgPool}_{3\text{D}}(\mathbf{F}); \text{MaxPool}_{3\text{D}}(\mathbf{F})])\right) \tag{6}$$

The Polarization Attention module compresses the feature map in both temporal and spatial dimensions to obtain a one-dimensional vector, which is then manipulated. 3D Average Pooling and 3D Max Pooling are used to aggregate the spatial information of the feature graph, send them to a shared network, compress the spatial-temporal dimensions of the input feature graph, and sum and merge element by element to produce a channel attention graph. Polarization attention, on a single graph, is concerned with which polarimetric features of this graph are important. The structure is given by Fig. 4.

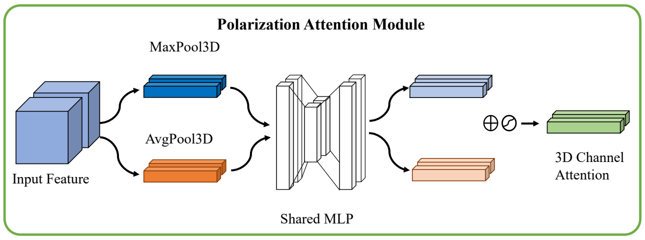

Fig. 4. The schematic of Polarization Attention Module.

The Temporal-Spatial Attention Module takes the feature map output from the Polarization Attention Module as the input feature map of this module. The specific process is shown in Fig. 5. First, we do 3D global AvgPool and 3D global MaxPool according to the channel, and then do the join operation on these two results according to the channel. The temporal-spatial attention feature map is generated by sigmoid. Finally, this feature map is multiplied with the input features of the module to get the final generated features.

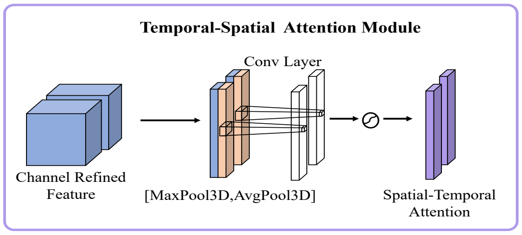

Fig. 5. The schematic of Temporal-Spatial Attention Module.

This improvement of ours expands the scope of the attention mechanism from space and channel to space dimension, channel dimension and time dimension, and for the multi-temporal polarimetric SAR classification task, it enables the network to find the optimal polarimetric feature as well as time during the training process.

D. Vision Transformer

We introduce Vision Transformer into multi-temporal polarimetric SAR data for crop classification. Its main structure is shown in Fig. 6.

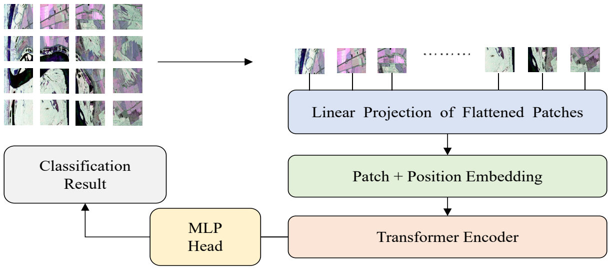
Linear Projection of Flattened Patches
Patch + Position Embedding
Transformer Encoder
MLP Head
Classification Result

Fig. 6. The structure of Vision Transformer.

First, ViT divides the image into N “patches”, e.g. 7×7. But since each patch is itself 3D data with shape height × width × number of polarimetric channels, they cannot be processed directly by the Transformer proposed originally for the processing language (2D) data, so they need to be flattened and linearly projected to 2D data before classification.

However, both ViT and ResNet need to transform the image into the form of a three-dimensional tensor. For multi-temporal polarimetric SAR data, we need to superimpose the polarimetric features of multiple times in the same dimension to form a three-dimensional tensor. This approach actually lacks independent descriptions of temporal and polarimetric features, and for this reason, we express the multi-temporal polarimetric data as a four-dimensional tensor of the number of time × number of polarimetric features × length × width, and use 3D convolution to extract spatial-temporal features.

III. Experiment

A. Datasets

In this paper, two sets of data are selected to validate our method. The first one is five fully polarimetric SAR data of UAVSAR from July 1, 2019 to September 23, 2019. We intercepted the part of the image that contains a large amount of farmland areas. The image size is 15000×9900 pixels, and we combined these 5 scene images with Google Maps for manual annotation. It contains 16 categories with a total of 9047044 pixels in the labeled area.

The second dataset is RARDARSAT-2 data of the Flevoland scene, which was produced by the European Space Agency(ESA) in the “AgriSAR 2009” project. It consists of 8 scenes of multi-temporal full polarimetric SAR images and 21 crop categories, which were collected from April 14 to September 29 in 2009, with a 24-day interval between each scene. We took data from one of the time phases in each of the two datasets, and their Pauli images and ground truth are shown in Fig. 7.

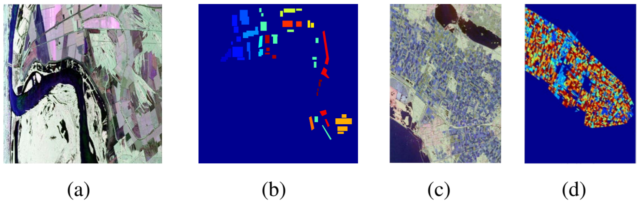

Fig. 7. The Pauli image and Ground Truth of both datasets. (a) 2019-07-01. (b) Ground Truth of UAVSAR. (c) 2009-04-14 (d) Ground Truth of Radarsat-2.

B. Classification results and analysis

We conducted experiments on these two kinds of data separately, and obtained two metrics of Overall Accuracy(OA) and Kappa Coefficient, and the results are shown in Table I.

TABLE I
The classification accuracy of each model.

	UAVSAR		Radarsat-2	
	OA(%)	Kappa	OA(%)	Kappa
ResNet [3]	95.24	0.9474	87.99	0.8603
3DResNet [12]	95.96	0.9553	89.91	0.8828
ViT [4]	96.19	0.9579	89.35	0.8763
Two-Stream Model [13]	97.54	0.9728	89.58	0.8789
CTNet [6]	97.65	0.9737	89.66	0.8967
PolSARFormer[15]	96.86	0.9615	89.73	0.8802
Conformer[16]	97.15	0.9685	90.61	0.8910
TPST	97.68	0.9744	91.38	0.8999
ResNet	95.66	0.9527	88.43	0.8745
CTNet	98.02	0.9753	90.45	0.8907
PolSARFormer	97.23	0.9681	90.31	0.8829
Conformer	97.46	0.9703	90.78	0.8926
3DResNet*	96.48	0.9591	90.45	0.8873
*Two-Stream Model**	98.11	0.9744	91.26	0.9033
*PolSARFormer**	97.81	0.9779	90.68	0.8897
*TPST**	**98.23**	**0.9804**	**91.97**	**0.9068**

1 Italics indicate the addition of the CBAM module.
2 ()* indicates the addition of the 3DCBAM module.

We compare the proposed network TPST in this paper with several popular networks, it can be seen that TPST achieves the best accuracy in both datasets, and without adding the

3DCBAM module in either case, TPST can improve the OA by 2.45% and 3.39%, respectively, compared to ResNet. The OA of TPST without 3DCBAM also has about 1%-2% improvement compared to ViT, which proves the effectiveness of our Temporal-Polarimetric-Spatial feature extraction module. In addition to this, We added CBAM and 3DCBAM modules to the networks containing 2D and 3D convolution respectively to verify the effectiveness of the module. For these four models with only CBAM added, their OA and Kappa improved in average on the two datasets are 0.49%, 0.0035, 0.50%, and 0.0031, respectively. For 3DResNet, which can only add 3DCBAM, the OA and Kappa in the two datasets are improved by 0.52%, 0.0038, 0.54%, and 0.0045, respectively. As for the three models added to both attention modules, their results improved in average by 0.63%, 0.0080, 1.07%, 0.0136 on both datasets. This demonstrates the effectiveness of the 3DCBAM module for classification of multi-temporal polarimetric SAR data. In order to show the improvement of the classification effect more intuitively, we present the classification results in Fig. 8.

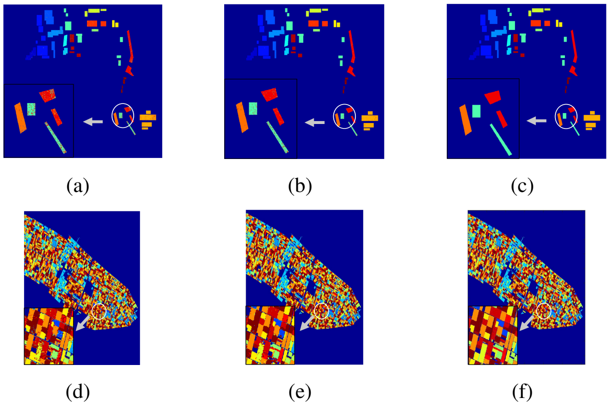

Fig. 8. Classification maps and ground truth on ResNet and TPST. (a) UAVSAR on ResNet. (b) UAVSAR on TPST. (c) UAVSAR ground truth. (d) RADARSAT-2 on ResNet. (e) RADARSAT-2 on TPST. (f) RADARSAT-2 ground truth.

IV. Conclusion

For multi-temporal polarimetric SAR data, this paper proposes a network structure TPST that combines a temporal-polarimetric-spatial feature extraction module with ViT, which is a two-stream network formed by a 2D residual module and a 3D residual module. In addition, this paper improves the common attention module to extend the domain of this mechanism to three dimensions: time, polarization and space. The effectiveness of the network structure and attention module is validated on two multi-temporal polarimetric SAR datasets, UAVSAR and RADARSAT-2. However, due to the addition of 3D convolution, the number of parameters and computation are increased, and how to solve this problem and how to optimize the feature fusion between 2D convolution and 3D convolution is still a problem that needs to be solved in the future.

V. Acknowledgment

The authors would like to thank the European Space Agency(ESA) for providing the temporal dataset through the AgriSAR 2009 campaign and PolSARPro software, and also like to thank Macro Lavalle to select the UAVSAR data.

References

[1] A. Alonso González, C. López Martínez, K. Papathanassiou, and I. Hajnsek, "Polarimetric SAR time series change analysis over agricultural areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7317–7330, 2020.

[2] J. Ni, C. López-Martínez, Z. Hu, and F. Zhang, "Multitemporal SAR and Polarimetric SAR Optimization and Classification: Reinterpreting Temporal Coherence," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[5] H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric SAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[6] P. Deng, K. Xu, and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[7] M. Teimouri, M. Mokhtarzade, N. Baghdadi, and C. Heipke, "Fusion of time-series optical and SAR images using 3D convolutional neural networks for crop classification," *Geocarto International*, pp. 1–18, 2022.

[8] C. Silva-Perez, A. Marino, J. M. Lopez-Sanchez, and I. Cameron, "Multitemporal polarimetric SAR change detection for crop monitoring and crop type classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12 361–12 374, 2021.

[9] X. Jiao, J. M. Kovacs, J. Shang, H. McNairn, D. Walters, B. Ma, and X. Geng, "Object-oriented crop mapping and monitoring using multi-temporal polarimetric RADARSAT-2 data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 96, pp. 38–46, 2014.

[10] B. K. Kenduiywo, D. Bargiel, and U. Soergel, "Higher order dynamic conditional random fields ensemble for crop type classification in radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4638–4654, 2017.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[12] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[14] Z. Lin, Q. Yin, Y. Zhou, J. Ni, and F. Ma, "Time-Series PolSAR Crop Classification Based on Joint Feature Extraction," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 831–834.

[15] A. Jamali, S. K. Roy, A. Bhattacharya, and P. Ghamisi, "Local window attention transformer for polarimetric sar image classification," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.