Local Window Attention Transformer for Polarimetric SAR Image Classification

Jamali, A.; Roy, S. K.; Bhattacharya, A.; Ghamisi, P.;

Originally published:

January 2023

IEEE Geoscience and Remote Sensing Letters 20(2023), 4004205

DOI: https://doi.org/10.1109/LGRS.2023.3239263

Local Window Attention Transformer for Polarimetric SAR Image Classification

Ali Jamali, Swalpa Kumar Roy, *Student Member, IEEE,*
Avik Bhattacharya, *Senior Member, IEEE,* and Pedram Ghamisi, *Senior Member, IEEE*

***Abstract*—Convolutional neural networks (CNNs) have recently found great attention in image classification since deep CNNs have exhibited excellent performance in computer vision. Owing to their immense success, of late, scientists are exploring the functionality of transformers in Earth observation applications. Nevertheless, the primary issue with transformers is that they demand significantly more training data than CNN classifiers. Thus, the use of these transformers in remote sensing is considered challenging, notably in utilizing polarimetric SAR (PolSAR) data, due to the insufficient number of existing labeled data. In this letter, we develop and propose a vision transformer-based framework that utilizes 3D and 2D CNNs as feature extractors and, in addition, local window attention for the effective classification of PolSAR data. Extensive experimental results demonstrated that the developed model `PolSARFormer` obtained better classification accuracy than the state-of-the-art vision Swin Transformer and FNet algorithms. The `PolSARFormer` outperformed the Swin Transformer and FNet by 7.79% and 6.94%, respectively, in terms of average accuracy in the San Francisco data benchmark. Moreover, the results over the Flevoland dataset illustrated that the `PolSARFormer` exceeds several other algorithms, including Swin Transformer (95.31%), AlexNet (97.93%), a 2D CNN (98.58%), FNet (98.63%), and ResNet (98.82%), with a kappa index of 98.93%. The code will be made available publicly at https://github.com/aj1365/PolSARFormer**

***Index Terms*—visual transformers, PolSAR image classification, convolutional neural networks (CNN), attention mechanism, local window attention (LWA).**

I. Introduction

SYnthetic aperture radar (SAR) as an active microwave imaging system perceives terrain without being constrained by illumination or the atmosphere [1]. Thus, it is extensively used within civil and military applications [2]–[5]. With the rapid growth of data, reliably characterizing SAR images has become an immediate demand [6]. The burst of deep learning algorithms has recently opened up a new opportunity for PolSAR classification tasks [2], [7], [8]. Nevertheless, the effectiveness of these methods depends on the availability of substantial quantities of class labels (i.e., reference data). It is well recognized that referencing PolSAR data is time-consuming and labor-intensive and requires domain-specific skills and knowledge, causing PolSAR annotations to be considerably challenging to acquire. Due to the lack of the desired quantity of labelled data and difficulties encountered by experts in PolSAR reference data creation and labelling, most research focus on the utilization of shallower CNNs (i.e., CNN models with less than five layers) [8].

On the other hand, given the considerable success of transformer models in language processing, scientists are now investigating the capabilities of these cutting-edge models in computer vision and Earth observation [9]–[11]. One should note that they have lately shown to be effective in a wide range of applications, including remote sensing imagery characterization [11]. Nevertheless, the main concern with transformers is that they demand more training data than CNNs. Consequently, using such transformers in remote sensing is regarded as challenging, particularly in PolSAR applications with a limited number of labeled data. As such, we develop and propose an efficient vision transformer that utilizes neighborhood attention to precisely classify PolSAR imagery. The objective is to develop a vision transformer capable of accurately classifying PolSAR data. The contributions of this paper can be explained as follows:

- We developed a deep learning based image classification framework that can effectively combine CNNs and vision transformers to classify PolSAR imagery accurately.
- The proposed model utilizes local window attention (LWA) instead of self-attention, which is computationally too expensive for improving the feature generalization capability in a local region by significantly decreasing the computation cost of vanilla ViTs.
- The integration of 3D and 2D CNNs with local window attention (LWA) resulted in much lower classification noises than the state-of-the-art vision transformers, i.e., Swin Transformer.

II. Proposed Classification Framework

Convolutional neural network (CNN) has already been proven to be a high-level feature extractor and has been successfully applied in many computer vision tasks. In this section, we introduce PolSarFormer, an effective, reliable, and scalable hierarchical vision transformer (ViT)-based encoder network for PolSAR imagery classification. An adaptable

This research was partially funded by the Institute of Advanced Research in Artificial Intelligence (IARAI).(Corresponding author: *Pedram Ghamisi*)

A. Jamali is with the the Department of Engineering, Karabuk University, Karabuk, Turkey; (e-mail: alijamali@karabuk.edu.tr).

S. K. Roy is with the Department of Computer Science and Engineering, Jalpaiguri Engineering College, West Bengal 735102, India (e-mail: swalpa@cse.jgec.ac.in).

A. Bhattacharya is with the the Microwave Remote Sensing Lab, Centre of Studies in Resources Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India; (e-mail: avikb@csre.iitb.ac.in).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and is also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

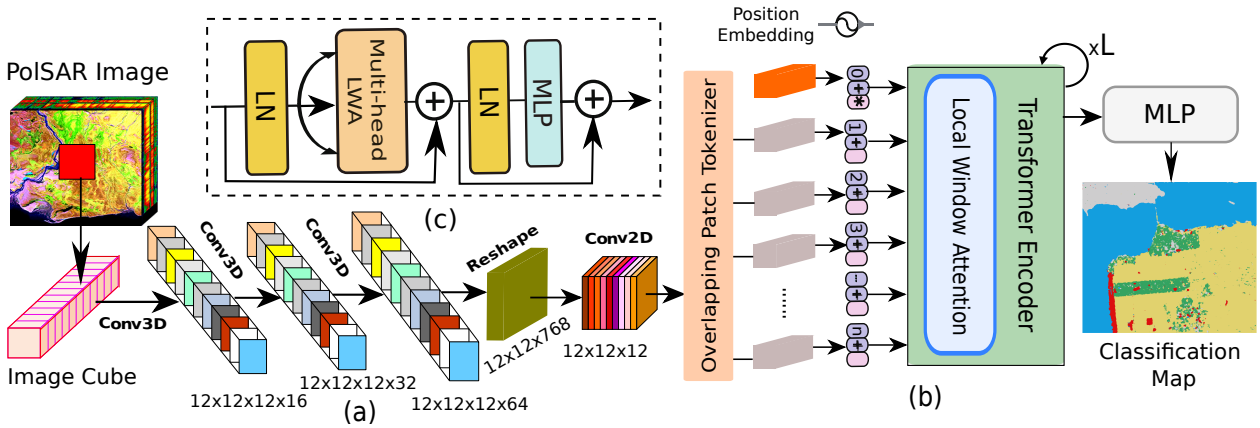

Fig. 1: Graphical representation of the proposed PolSARFormer network for Polarimetric SAR Image Classification where (a) 3D-2D CNN as feature extractor, (b) local window attention transformer (LWAT) and (c) visualization of multihead local window attention, respectively.

and simple attention mechanism called local attention transformer (LAT) is explored for the effective utilization of extracted features in classification. Hence, the proposed network comprises two crucial parts, i.e., feature extractor and self attention mechanism. Fig. 1 shows the proposed framework for PolSAR image classification.

Feature Extractor: A PolSAR imagery can be represented as $\mathbf{X} \in \mathcal{R}^{W \times H \times D}$, which contains two spatial dimensions, i.e., the width W and the height H, and a polarimetric dimension D. All the pixels under the region of interest are classified into c land-cover classes denoted by $Y = (y_1, y_2, \ldots y_c)$. The class-wise land-cover regions of size 12×12 are sampled from the PolSAR data $\mathbf{X_{orig}}$ to create the training and validation data. To utilize the capabilities of CNNs as feature extractors, we employed a 3D and 2D CNN hierarchical architecture as the backbone network [12]. The aim is to utilize the ability of Conv3D to extract both the polarimetric-spatial features. In contrast, Conv2D helps to refine the prominent spatial feature among the backscattering and polarimetric data so that the backbone feature extractor is not too computationally intensive. The feature extractor has three 3D convolutional layers with the number of kernels 16, 32, and 64 with the size of $12 \times 12 \times 12$ followed by a 2D CNN convolutional layer with 12 kernels with the size of 12×12. To create long range dependencies among the extracted feature maps, a simple attention mechanism, local window attention (LWA), is utilized to effectively localize each query's receptive field to its closest neighboring pixels within a local window.

Local window attention: It is considered as a localized self-attention that involves inductive biases similar to convolution like operations, removing the requirement for additional overhead like pixel shifts explored in the advanced models of ViT such as Swin Transformer [13]. The LWA limits the receptive field of each query token to fixed-sized neighboring pixels. The motivation behind the LWA is to create the local neighborhood window; the smaller neighboring region receives greater local attention, whereas the larger neighboring region receives greater global attention. Thus, the LWA mechanism better controls the receptive fields while balancing translational invariance and equivariance properties compared to other vision transformers.

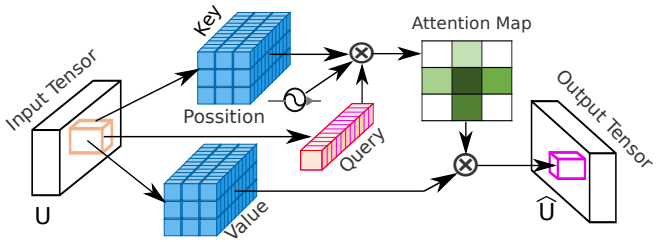

Fig. 2: local window attention module

The neighborhood of a pixel at a spatial position (i, j) in a lth feature map, i.e., (i, j, l) within a local window, is expressed by $\tau(i, j, l)$, which contains a finite set of indices belonging to pixels close to the location (i, j) of l^{th} feature map. For a local window of size $S \times S$, a neighborhood of $\tau(i, j, l)$ is calculated as $||\tau(i, j, l)|| = S^2$. The LWA on a single pixel at location (i, j) in the lth feature map can be expressed in terms of linear projections of the extracted features X. The queries $\mathbf{q}$ of size $(1 \times D)$ whereas keys $\mathbf{k}$ and values $\mathbf{v}$, both work on the entire patch matrix of size $(N \times D)$:

$$\mathbf{q} = \mathbf{W_q X}, \quad \mathbf{k} = \mathbf{W_k X}, \quad \mathbf{v} = \mathbf{W_v X}$$

The pairwise dot product applied between queries ($\mathbf{q}$) and keys ($\mathbf{k}$) inside a `softmax` function to calculate the attention map ($\mathbf{A}$) as shown in Fig. 2.

$$\mathbf{A} = \texttt{softmax}\left(\mathbf{q_{(i,j,l)}} \mathbf{k^T_{\tau(i,j,l)}} + \mathbf{b_{(i,j,l)}} / \sqrt{\mathbf{D}}\right)$$

where the relative positional bias is denoted by $b_{(i,j,l)}$, which is added to each attention map depending upon its relative position. $1/\sqrt{D}$ represents the scale and helps to resist the small gradient propagation of the `softmax` function. `LWA` is then computed as:

$$\texttt{LWA}(\mathbf{X}) = \mathbf{A} \mathbf{v}_{\tau(\mathbf{i},\mathbf{j},\mathbf{l})} \tag{1}$$

One should note that self-attention enables each token to interact with all the other tokens, whereas LWA restricts each token's receptive field to an area surrounding itself. As a result,

the LWA has added the benefit of directly limiting each pixel to its neighboring area at no additional computational expenses, eliminating the need for pixel transitions to incorporate cross-window interrelations. Moreover, unlike window attention, the LWA is not constrained to operate on inputs by the window size. It should be noted that if the size of the pixel's neighborhood is greater than or equal to the size of the feature map, self-attention and neighborhood attention will produce a similar result to that of the input map.

PolSARFormer: The backscattering and polarimetric features in the developed model will be passed to the feature extractor as described above. The PolSAR data input size will remain unchanged through the feature extractor part. Afterward, the resulting output of the feature extractor will be passed to the LWA. The LWA embeds the output of the feature extractor with two successive 3×3 convolution layers using strides of 2×2, yielding a spatial size one-fourth of the PolSAR data input. The LWA employs overlapping convolutions rather than non-overlapping ones. The developed model consists of two levels, wherein the first level has three LWA blocks, while we used four blocks in the second level. It should be noted that there can be several levels consisting of different/similar numbers of blocks of LWA, identical to that of Swin Transformer [13]. It is worth noting that the results of each level will be passed to the next existing level.

TABLE I: Classification results of Flevoland dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, ST = Swin Transformer, and PolSF= PolSARFormer, respectively.

Class	ST	AlexNet	FNet	2DCNN	ResNet	PolSF
Rapeseed	0.96	0.98	0.99	0.98	0.99	0.99
Beet	0.96	0.99	0.99	0.99	0.99	1
Stembeans	0.99	1	0.99	1	0.99	0.99
Peas	0.99	1	0.99	1	1	1
Forest	0.92	0.98	0.98	0.97	0.98	0.99
Lucerne	0.98	0.99	0.99	0.99	1	0.99
Wheat	0.96	0.97	0.99	0.99	0.99	0.98
Barley	0.98	1	0.99	1	1	1
Potatoes	0.92	0.97	0.98	0.97	0.98	0.99
Bare Soils	0.91	0.99	0.98	1	0.99	1
Wheat3	0.97	0.98	0.99	0.99	0.99	0.99
Water	0.98	1	1	1	1	1
Grass	0.88	0.93	0.97	0.97	0.98	0.98
Building	0.94	0.99	0.92	0.99	0.95	0.96
Wheat2	0.97	0.97	0.99	0.99	0.99	0.97
OA$\times$100	95.70	98.1	98.75	98.69	98.92	**99.02**
AA$\times$100	96.28	97.96	97.77	**98.84**	98.53	98.78
$\kappa\times$100	95.31	97.93	98.63	98.58	98.82	**98.93**

III. Experimental Results

A. Experimental Data

NASA/JPL AIRSAR recorded the data of Flevoland, situated in the Netherlands, on August 16, 1989. The Flevoland image is 750×1024 pixels in size. The other dataset illustrates a NASA/JPL AIRSAR L-band image of the San Francisco area. The resolution of the data of the San Francisco is 900×1024 pixels. It is worth remembering that from both PolSAR benchmarks, we have extracted 12×12 image patches. In San Francisco, we have used 5% of the data for model training and the remaining 95% as test data. In contrast, in the Flevonad, due to the availability of less labeled data, only 10% of the labeled data is used for training, whereas 90% of the data is used to perform test experiments.

TABLE II: Classification results of San Francisco dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, ST = Swin Transformer, and PolSF=PolSARFormer, respectively.

Class	ST	AlexNet	FNet	2DCNN	ResNet	PolSF
Bare Soil	0.84	0.82	0.88	0.88	0.83	0.85
Building	0.96	0.97	0.96	0.97	0.96	0.98
Water	0.99	0.99	0.99	0.99	0.99	0.99
Vegetation	0.68	0.77	0.68	0.8	0.71	0.86
Mountain	0.95	0.94	0.94	0.95	0.92	0.97
OA$\times$100	95.67	96.11	95.71	96.72	95.18	**97.39**
AA$\times$100	86.83	87.77	87.68	90.99	86.89	**94.62**
$\kappa\times$100	93.12	93.89	93.19	94.85	92.41	**95.93**

B. Classification Results

The developed model, PolSARFormer, is compared with several other models, including a state-of-the-art visual Swin Transformer [13], an advanced multi-layer perceptron, i.e., FNet, which was developed by Google that uses Fourier Transforms [14], AlexNet [15], a 2D CNN [7], and ResNet [16] for the classification of PolSAR imagery in two widely used datasets, i.e., Flevoland and San Francisco, respectively. The evaluated results over the Flevoland dataset are shown in Table I. The reported results show that the PolSARFormer model exceeds the other algorithms, including that of the Swin Transformer (95.31%), AlexNet (97.93%), 2D CNN (98.58%), FNet (98.63%), and ResNet (98.82%), and achieves the kappa index of 98.93%. In the Flevoland region, the proposed model outperforms the cutting-edge Swin Transformer in terms of overall accuracy (OA) by 3.62%. Fig. 3 shows the class-wise land-cover classification map using various methods for the Flevoland region. As seen in Figs. 5(a)-(b), the PolSARFormer network produces a better homogeneous land cover map with less noise when compared with the other visual Swin Transformer.

Table II shows the classification results, whereas Fig4 depicts the visual classification maps for the San Francisco region dataset. It has been observed from the results in Table II that the proposed PolSARFormer model achieves an average accuracy (AA) of 94.62%, illustrating better classification results as compared to the 2D CNN (90.99%), AlexNet (87.77%), FNet (87.68%), ResNet (86.89%), and Swin Transformer (86.83%). The Swin Transformer and FNet algorithms are considered state-of-the-art vision transformers. Still, both methods illustrated a lower level of classification accuracy due to their need for a much higher number of training data than the CNN classifiers. However, the proposed PolSARFormer classifier demonstrated much better PolSAR imagery classification accuracy when compared with the state-of-the-art vision Swin Transformer and FNet. In more detail, the PolSARFormer outperforms the Swin Transformer and FNet by the margin of 7.79% and 6.94%, respectively, in terms of average accuracy.

C. Ablation study

To better understand the significance of each part of the developed PolSARFormer, we have drawn an ablation study.

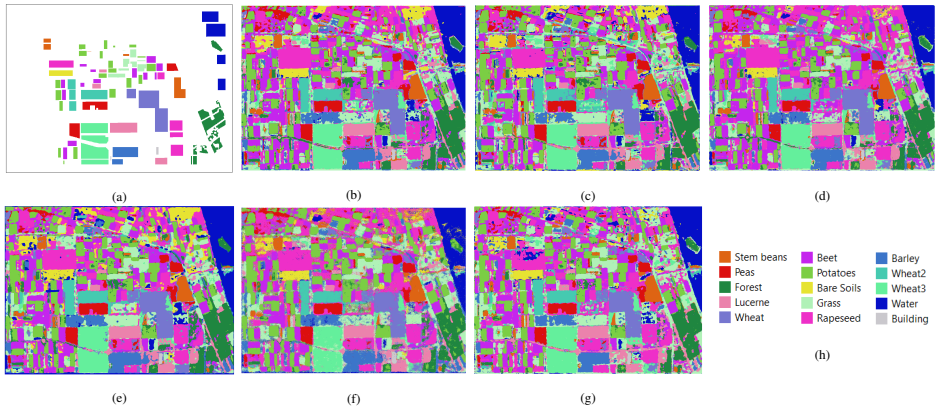

Fig. 3: Classification Maps over the Flevoland dataset using a) 2D CNN, b) AlexNet,c) FNet, d) ResNet e) Swin Transformer, and f) the PolSARFormer.

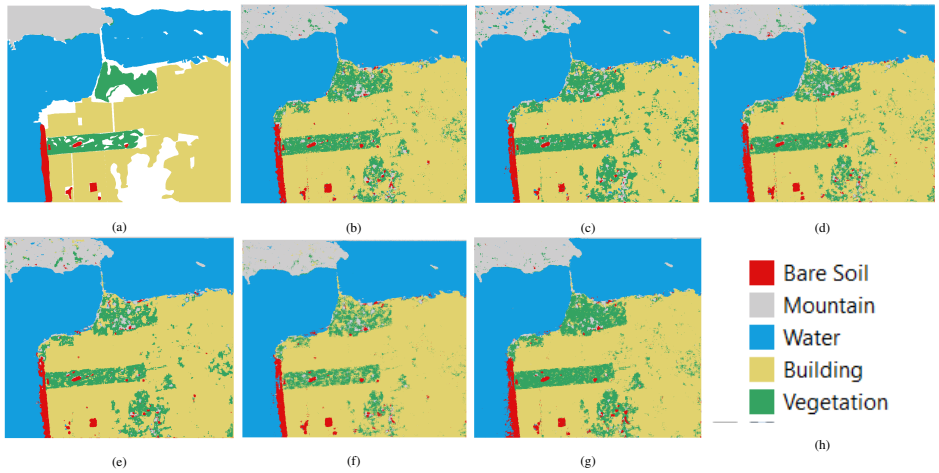

Fig. 4: Classification maps over the San Francisco dataset using a) 2D CNN, b) AlexNet, c) FNet, d) ResNet e) Swin Transformer, and f) the PolSARFormer.

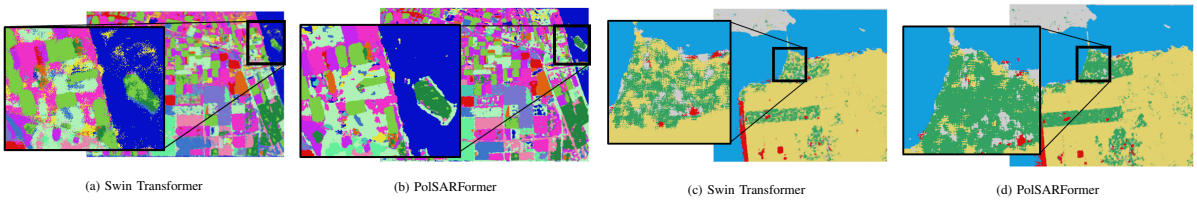

Fig. 5: A comparison of noises by the vision transformers of Swin Transformer and the PolSARFormer.

The results of the Flevoland dataset are shown in Table. III illustrates the utilization of the capabilities for both CNNs and transformers achieved the highest classification accuracy with an average accuracy of 98.78% whereas CNNs and transformers achieve 96.87% and 98.56%, respectively. Moreover, the results of San Francisco demonstrated that the integration of CNNs and transformers improved the accuracy of PolSAR classification by 1.93% and 5.36%, respectively, as compared to CNNs and transformers individually, as shown in Table IV.

D. Impact of training samples on the PolSARFormer

To better understand how the size of training data affects the performance of the proposed PolSARFormer model, we evaluated the PolSAR classification accuracy obtained by the proposed model for varying training ratios in Flevoland dataset, as shown in Fig 6. The results illustrated that the PolSARFormer classifier achieves a high classification accuracy with significantly less training data (i.e., training ratio of 0.5% considered from the reference data) in terms of average accuracy (86.03%) as compared to the Swin Transformer with an average accuracy of 75.95%. This demonstrates the efficiency of the PolSARFormer for classifying PolSAR imagery, contradictory to the current state-of-the-art vision transformers that demand a higher number of labelled samples. In addition, the results show that the classification accuracy of the PolSARFormer algorithm in terms of average accuracy, kappa index, and overall accuracy improved by approximately 13%, 8%, and 7% by utilizing 0.5% to 10% training data ratio, respectively, as seen in Fig 6.

IV. Conclusion

This letter presents a vision transformer-based framework for PolSAR image classification that uses local window atten-

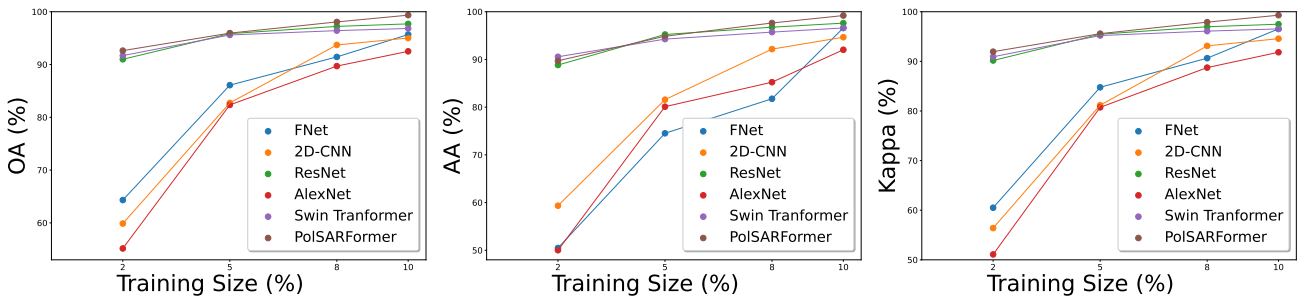

Fig. 6: A comparison of (a) OA (b) AA and (C) Kappa performance achieved by different classifiers over the various training ratio on the Flevoland dataset.

TABLE III: Classification results of Flevoland dataset in terms of F-1 score (κ = Kappa index, OA = Overall Accuracy, and AA = Average Accuracy).

Class	CNN	Transformer	CNN+Transformer
Rapeseed	0.99	0.98	0.99
Beet	0.99	0.99	1
Stembeans	0.99	0.99	0.99
Peas	1	1	1
Forest	0.98	0.98	0.99
Lucerne	0.99	0.99	0.99
Wheat	0.99	0.98	0.98
Barley	0.99	1	1
Potatoes	0.98	0.98	0.99
Bare Soils	1	0.96	1
Wheat3	0.99	1	0.99
Water	0.99	0.99	1
Grass	0.97	0.98	0.98
Building	0.84	0.96	0.96
Wheat2	0.99	0.99	0.97
OA×100	98.85	98.77	**99.02**
AA×100	96.87	98.56	**98.78**
κ×100	98.74	98.66	**98.93**

TABLE IV: Classification results of San Francisco dataset in terms of F-1 score (κ = Kappa index, OA = Overall Accuracy, and AA = Average Accuracy).

Class	CNN	Transformer	CNN+Transformer
Bare Soil	0.87	0.87	0.85
Building	0.98	0.97	0.98
Water	0.99	0.99	0.99
Vegetation	0.83	0.76	0.86
Mountain	0.96	0.93	0.97
OA×100	97.01	96.12	**97.39**
AA×100	92.69	89.26	**94.62**
κ×100	95.34	93.9	**95.93**

tion (LWA) to improve the feature representation capabilities locally while drastically reducing annotation costs and hardware requirements. The results on two PolSAR benchmark datasets revealed that the developed model, PolSARFormer, outperforms the existing state-of-the-art vision transformers of the Swin Transformer, FNet, and other models. In the San Francisco benchmark, the PolSARFormer surpassed the Swin Transformer and FNet by 7.79% and 6.94%, respectively, in terms of average accuracy. Furthermore, the PolSARFormer exceeds several other algorithms on the Flevoland dataset, such as the Swin Transformer (95.31%), AlexNet (97.93%), a 2D CNN (98.58%), FNet (98.63%), and ResNet (98.82%), with a Kappa index of 98.93%.

REFERENCES

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[2] W. Zhang, Z. Pan, and Y. Hu, "Exploring polsar images representation via self-supervised learning and its application on few-shot classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[3] Y. Zuo, J. Guo, Y. Zhang, Y. Hu, B. Lei, X. Qiu, and C. Ding, "Winner takes all: A superpixel aided voting algorithm for training unsupervised polsar cnn classifiers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[4] J. Ye, C. Wang, H. Gao, H. Fan, T. Song, and L. Ding, "A novel unsupervised object-level crop rotation detection with time-series dual-polarimetric sar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[5] G. Parrella, I. Hajnsek, and K. P. Papathanassiou, "Model-based interpretation of polsar data for the characterization of glacier zones in greenland," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 593–11 607, 2021.

[6] Z. Ren, B. Hou, Q. Wu, Z. Wen, and L. Jiao, "A distribution and structure match generative adversarial network for sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3864–3880, 2020.

[7] A. Jamali, M. Mahdianpari, F. Mohammadimanesh, A. Bhattacharya, and S. Homayouni, "Polsar image classification based on deep convolutional neural networks using wavelet transformation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[8] W. Hua, W. Xie, and X. Jin, "Three-channel convolutional neural network for polarimetric sar images classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4895–4907, 2020.

[9] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "Scvit: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[10] D. Cai and P. Zhang, "T^{3}SR: Texture transfer transformer for remote sensing image super-resolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–14, 2022.

[11] L. Luo, J.-X. Wang, S.-B. Chen, J. Tang, and B. Luo, "Bdtnet: Road extraction by bi-direction transformer from remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[12] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2020.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[14] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontañón, "Fnet: Mixing tokens with fourier transforms," *CoRR*, vol. abs/2105.03824, 2021. [Online]. Available: https://arxiv.org/abs/2105.03824

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[16] 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 (CVPR)*, June 2016.