IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10642190

POLSARCONVMIXER: A CHANNEL AND SPATIAL MIXING CONVOLUTIONAL ALGORITHM FOR POLSAR DATA CLASSIFICATION

Ali Jamali
Department of Geography
Simon Fraser University
8888 University Dr W, Burnaby,
BC V5A 1S6, Canada.
e-mail: alij@sfu.ca

Swalpa Kumar Roy
Department of Computer Science and Engineering
Alipurduar Government Engineering
and Management College
West Bengal 736206, India.
e-mail: swalpa@cse.jgec.ac.in

Bing Lu
Department of Geography
Simon Fraser University
8888 University Dr W, Burnaby,
BC V5A 1S6, Canada.
e-mail: alij@sfu.ca

Avik Bhattacharya
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

Pedram Ghamisi
Helmholtz-Zentrum Dresden-Rossendorf (HZDR)
Helmholtz Institute Freiberg for Resource Technology
09599 Freiberg, Germany.
e-mail: p.ghamisi@gmail.com

ABSTRACT

Given the exceptional effectiveness of deep Convolutional Neural Networks (CNNs) in computer vision, there has been a recent surge of interest in employing CNNs for various applications in image classification. Additionally, scientists are exploring the potential of vision transformers for Earth observation applications, owing to their recent tremendous success. However, a major challenge with vision transformers is their increased demand for training data compared to CNN classifiers. Furthermore, vision transformers exhibit quadratic complexity and necessitate substantial hardware resources. In the context of PolSAR image classification, we propose the PolSARConvMixer—a fundamental framework that segregates the mixing of spatial and channel dimensions, maintains uniform size and resolution across the network and directly processes PolSAR image patches as input. Our experiments on two PolSAR data benchmarks, namely Flevoland and San Francisco, demonstrate the significant superiority of the developed PolSARConvMixer over several other algorithms, including AlexNet, ResNet, FNet, a 2D CNN, and PolSARFormer.

Index Terms— PolSAR, CNN, image classification, land cover mapping, deep learning

1. INTRODUCTION

Urban planning, agricultural assessment, and environmental monitoring extensively utilize polarimetric synthetic aperture radar (PolSAR) images acquired through satellite and aircraft sensors. These images offer rich insights into the Earth's surface [1]. A comprehensive understanding and interpretation of PolSAR images are essential for various applications, with image classification being a significant area of focus. Recent advancements in SAR target detection and classification leverage deep convolutional neural networks (CNN) models [2]. In this regard, the deep learning technique proposed by Zhou et al., [3] has demonstrated remarkable success across diverse image processing domains. Deep CNN algorithms show substantial potential for enhancing PolSAR classification accuracy compared to traditional classifiers like Random Forest [4]. They incorporate neural network-based classifiers, capturing translationally-invariant spatial features. However, a key challenge lies in maintaining effective generalization with limited training samples for PolSAR classification [5]. In parallel, vision transformers (ViTs) exhibit significant capability in classifying complex and high-dimensional remote sensing data [6]. Yet, their major drawback is the

demand for a substantial quantity of ground truth data compared to CNN classifiers. Additionally, vision transformers require significant hardware resources and exhibit quadratic complexity.

We introduce a fundamental framework called PolSARConvMixer for PolSAR image classification to address these challenges. This framework operates particularly on PolSAR image patches as data input, maintains a uniform size and resolution across the network, and segregates the mixing of spatial and channel dimensions. This is inspired by MLP-Mixer [7].

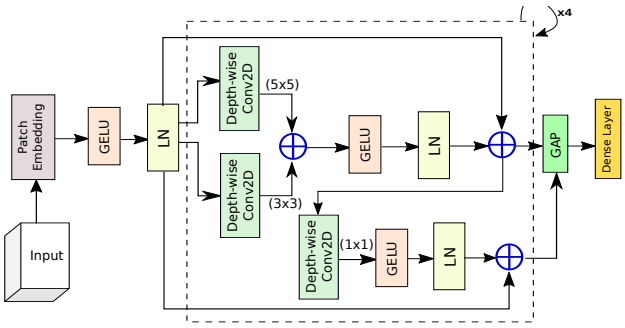

Fig. 1: Proposed PolSARConvMixer model for accurate PolSAR data classification where $\oplus$ represents element wise addition.

2. PROPOSED MODEL

While convolutional networks have long been the preferred architecture for vision tasks, recent studies indicate that in certain situations, Transformer-based models—most notably the Vision Transformer (ViT)—may outperform traditional approaches [8]. However, ViTs must be augmented with $patch\ embeddings$ when applied to larger image sizes. These embeddings consolidate small image regions into single input features, addressing the quadratic complexity of self-attention layers in Transformers [9]. We introduce the PolSARConvMixer, an inherently fundamental framework that operates directly on PolSAR image patches as input. It divides the mixing of spatial and channel dimensions while maintaining a uniform size and resolution across the network. This design resembles the ViT and the more basic MLP-Mixer [7]. The PolSARConvMixer achieves the mixing steps using typical convolutions. Our model comprises a patch embedding layer followed by repeated applications of a simple, fully convolutional block, as illustrated in Fig. 1. The concept of mixing, inspired by Tolstikhin et al. (2021) [7], serves as the foundation of our architecture. The patch embedding, with a patch size p and embedding dimension h, can be achieved through convolution with c_{in} input bands, o output bands, kernel size k, and stride s:

$$z_0 = LN(\alpha(Conv_{c_{in}\rightarrow o(X,\ stride=s,\ kernel\ size=k)})) \quad (1)$$

where α illustrates the activation function of $GELU$. The Convolutional block consists of two depth-wise convolution (i.e., grouped convolution with groups equal to the number of bands, h) followed by point-wise (i.e., kernel size 1×1) convolution. Convolutional performs optimally when using large kernel sizes for the depth-wise convolution (i.e., kernel size 5×5). After each convolutional layer, we used an activation layer of $GELU$ followed by a Layer normalization (LN) as defined by:

$$z_l^{'} = LN(\alpha \sum_{k=3,5} \mathrm{DWConv2D}_{(\mathrm{k}\times\mathrm{k})}(\mathrm{z}_{\mathrm{l}-1})) + \mathrm{z}_{\mathrm{l}-1} \quad (2)$$

$$z_{l+1} = LN(\alpha(ConvPointWise(z_l^{'}))) + z_{l-1} \quad (3)$$

After repeatedly using this Convolutional block (i.e., 4 blocks), we obtain a feature vector of size h by performing average global pooling, which we then feed to a $softmax$ classifier.

3. RESULTS

The evaluation of the proposed PolSARConvMixer model, in comparison with Swin Transformer [10], AlexNet [2], ResNet [11], FNet [12], a 2D CNN [4], and PolSARFormer [8], for PolSAR data classification across the Flevoland and San Francisco benchmarks, highlights its exceptional performance.

In the Flevoland dataset, PolSARConvMixer exhibited an average accuracy of 99.81 $\pm$ 0.04 with a kappa index of 99.73 $\pm$ 0.06 and an overall accuracy of 99.75 $\pm$ 0.05 shown in Table 1 and Fig. 2. Notably, the vision transformer of PolSARFormer demonstrated comparable classification accuracy to PolSARConvMixer, while AlexNet and FNet classifiers exhibited comparatively lower accuracy. For example, with an average accuracy of 99.81 $\pm$ 0.04, the PolSARConvMixer significantly enhanced the PolSAR classification accuracy of PolSARFormer, ResNet, Swin Transformer, 2D CNN, AlexNet, and FNet by about 1, 2, 3, 5, 8, and 8 percentage points, respectively, as shown in Table 1.

In the San Francisco dataset, the PolSARConvMixer CNN model achieved the highest PolSAR classification accuracy. Particularly, it produced an average accuracy of 90.71 $\pm$ 1.69 with a kappa index of 99.73 $\pm$ 0.06 and an overall accuracy of 99.75 $\pm$ 0.05 shown in Table 2 and Fig. 3. For example, in terms of average accuracy, the developed PolSARConvMixer substantially outperformed PolSARFormer, ResNet, AlexNet, 2D CNN, Swin Transformer, and FNet by approximately 1, 3, 3, 5, 7, and 18 percentage points, respectively, with and average accuracy of 90.71 $\pm$ 1.69. Notably, the FNet algorithm displayed the least accuracy, with an average accuracy of 72.43 $\pm$ 3.10 with a kappa index of 87.55

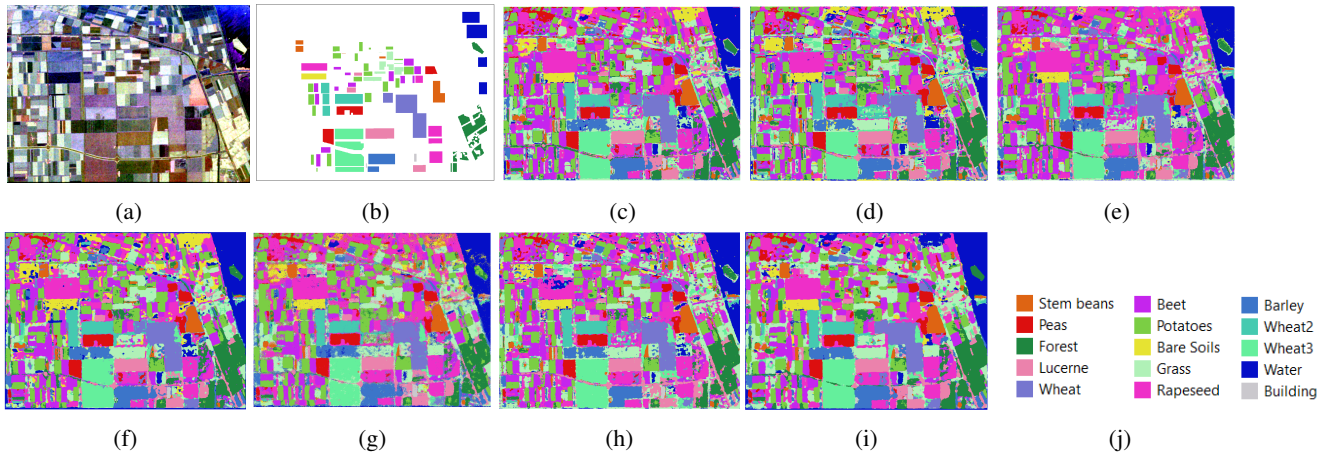

Fig. 2: LULC maps of the Flevoland data benchmark using a) PauliRGB image, b) Ground Truth, c) 2D CNN, d) AlexNet, e) FNet, f) ResNet g) Swin Transformer, h) PolSARFormer and i) PolSARConvMixer.

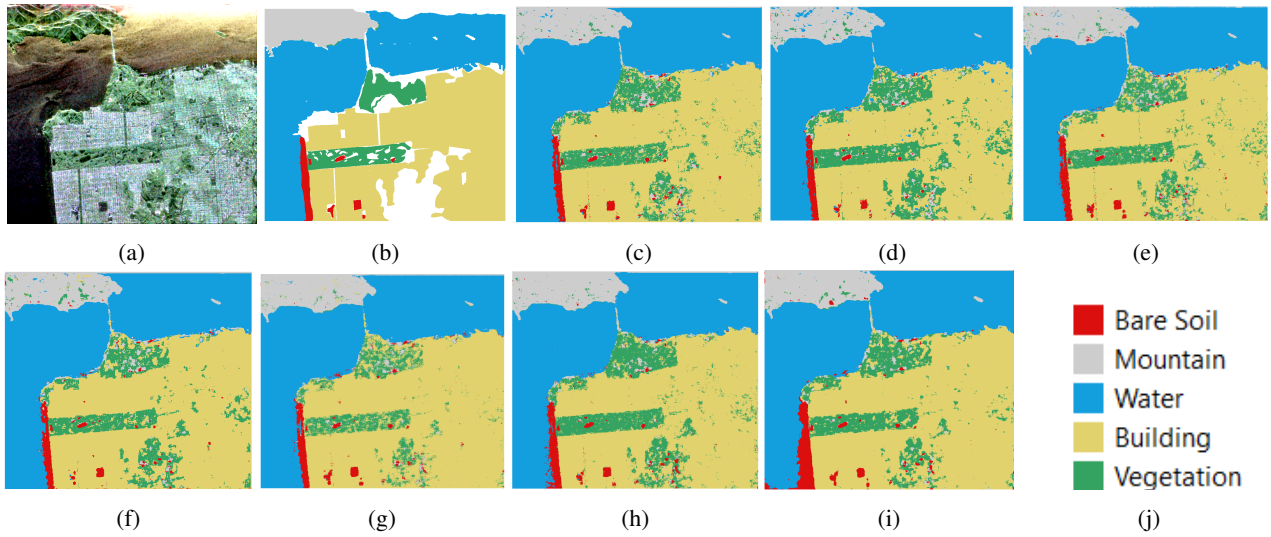

Fig. 3: LULC maps of the San Francisco data benchmark using a) PauliRGB image, b) Ground Truth, c) 2D CNN, d) AlexNet, e) FNet, f) ResNet g) Swin Transformer, h) PolSARFormer and i) PolSARConvMixer.

$\pm$ 0.78 and an overall accuracy of 92.25 $\pm$ 0.47 as shown in Table 2.

These results underscore the superior performance of PolSARConvMixer in both datasets and emphasize its potential as an effective model for PolSAR data classification tasks. Additionally, the comparison with diverse models provides valuable insights into the strengths of PolSARConvMixer over existing algorithms.

4. CONCLUSION

In this study, we developed the PolSARConvMixer algorithm for accurately classifying PolSAR imagery in two data benchmarks, San Francisco and Flevoland. Specifically, we utilized point-wise convolution for mixing channel locations and depth-wise convolution for mixing spatial locations. An essential concept from earlier research is that MLPs and self-attention can effectively mix disparate spatial locations, resulting in an infinitely broad receptive field. To achieve this, the PolSARConvMixer operates directly on PolSAR image patches as input with convolutions of large kernel sizes, such as 5×5, to mix distant spatial locations. Our experiments demonstrated that the developed model outperformed several other CNNs and vision transformers, showcasing its superior ability for land use land cover mapping using PolSAR data with typical convolutional functions.

Table 1: PolSAR data classification results of the Flevoland benchmark data in terms of F-1 score, OA = Overall Accuracy, κ = Kappa index, and AA = Average Accuracy. ST and PolSF refer to Swin Transformer and PolSARFormer, respectively.

Class	ST	AlexNet	FNet	2DCNN	ResNet	PolSF	PolSARConvMixer
Grass	0.87	0.97	0.9	0.96	0.98	1	1
Rapeseed	0.99	0.87	0.93	0.91	0.99	0.99	1
Beet	0.99	0.98	0.95	0.97	0.99	1	1
Water	0.99	0.99	0.97	1	1	1	1
Wheat2	0.98	0.83	0.94	0.87	0.91	1	1
Stembeans	1	0.92	0.98	0.97	1	1	1
Peas	0.99	0.99	0.98	0.93	1	1	1
Lucerne	0.99	1	0.97	0.99	1	1	1
Wheat	0.98	0.91	0.95	0.9	0.95	1	1
Forest	0.95	0.97	0.96	0.92	0.96	0.98	0.99
Barley	0.99	1	0.97	0.98	1	1	1
Bare Soils	0.98	0.98	0.88	1	1	1	1
Wheat3	0.96	0.98	0.98	0.97	0.99	1	1
Potatoes	0.94	0.95	0.95	0.92	0.95	0.98	0.99
Building	0.98	0.26	0.75	0.93	0.99	1	1
AA×100	96.60 ± (0.29)	92.03 ± (1.69)	92.19 ± (1.12)	94.66 ± (1.91)	97.61 ± (2.04)	99.20 ± (0.28)	**99.81 ± (0.04)**
OA×100	96.83 ± (0.20)	92.50 ± (2.22)	95.68 ± (0.12)	95.02 ± (2.26)	97.70 ± (1.91)	99.35 ± (0.03)	**99.75 ± (0.05)**
κ×100	96.54± (0.22)	91.83± (2.42)	95.28± (0.13)	94.57± (2.47)	97.49 ± (2.08)	99.30 ± (0.03)	**99.73 ± (0.06)**

Table 2: PolSAR data classification results of the San Francisco benchmark data in terms of F-1 score, OA = Overall Accuracy, κ = Kappa index, and AA = Average Accuracy. ST and PolSF refer to Swin Transformer and PolSARFormer, respectively.

Class	ST	AlexNet	FNet	2DCNN	ResNet	PolSF	PolSARConvMixer
Building	0.96	0.96	0.94	0.95	0.97	0.97	0.94
Bare Soil	0.81	0.82	0.39	0.79	0.91	0.89	0.79
Mountain	0.9	0.92	0.83	0.9	0.92	0.94	0.83
Water	0.99	0.99	0.98	0.99	0.99	0.99	0.99
Vegetation	0.61	0.72	0.52	0.55	0.76	0.76	0.98
OA×100	94.82 ± (0.29)	94.29 ± (0.16)	92.25 ± (0.47)	93.93 ± (0.38)	95.58 ± (0.13)	95.77 ± (0.21)	**96.01 ± (0.40)**
AA×100	84.20 ± (2.42)	87.54 ± (1.95)	72.43 ± (3.10)	85.84±(1.73)	87.92 ± (0.65)	90.06 ± (0.55)	**90.71 ± (1.69)**
κ×100	91.80 ± (0.54)	91.07 ± (0.3)	87.55 ±(0.78)	90.50±(0.55)	93.02±(0.18)	93.36±(0.30)	**93.76±(0.61)**

5. REFERENCES

[1] Hongquan Wang, Ramata Magagi, and Kalifa Goita, "Comparison of different polarimetric decompositions for soil moisture retrieval over vegetation covered agricultural area," *Remote Sensing of Environment*, vol. 199, pp. 120–136, 2017.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] Yu Zhou, Haipeng Wang, Feng Xu, and Ya-Qiu Jin, "Polarimetric sar image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[4] Ali Jamali, Masoud Mahdianpari, Fariba Mohammadimanesh, Avik Bhattacharya, and Saeid 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.

[5] Xu Liu, Licheng Jiao, Xu Tang, Qigong Sun, and Dan Zhang, "Polarimetric convolutional network for polsar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 3040–3054, 2019.

[6] Hongwei Dong, Lamei Zhang, and Bin Zou, "Exploring vision transformers for polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[7] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 24261–24272, Curran Associates, Inc.

[8] Ali Jamali, Swalpa Kumar Roy, Avik Bhattacharya, and Pedram Ghamisi, "Local window attention transformer for polarimetric sar image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[9] Asher Trockman and J. Zico Kolter, "Patches are all you need?," 2022.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón, "Fnet: Mixing tokens with fourier transforms," *CoRR*, vol. abs/2105.03824, 2021.