Pyramid Fine and Coarse Attentions for Land Cover Classification from Compact Polarimetric SAR Imagery

Saeid Taleghanidoozdoozan, *Student Member, IEEE,* Linlin Xu, *Member, IEEE,* and David A. Clausi, *Senior Member, IEEE*

***Abstract*—Land cover classification from compact polarimetry (CP) imagery captured by the launched RADARSAT Constellation Mission (RCM) is important but challenging due to class signature ambiguity issues and speckle noise. This paper presents a new land cover classification method to improve the learning of discriminative features based on a novel pyramid fine- and coarse-grained self-attentions transformer (PFC transformer). The fine-grained dependency inside a non-overlapping window and coarse-grained dependencies between non-overlapping windows are explicitly modeled and concatenated using a learnable linear function. This process is repeated in a hierarchical manner. Finally, the output of each stage of the proposed method is spatially reduced and concatenated to take advantage of both low- and high-level features. Two high-resolution (3m) RCM CP SAR scenes are used to evaluate the performance of the proposed method and compare it to other state-of-the-art deep learning methods. The results show that the proposed approach achieves an overall accuracy of 93.63% which was 4.83% higher than the best comparable method, demonstrating the effectiveness of the proposed approach for land cover classification from RCM CP SAR images.**

***Index Terms*—RADARSAT Constellation Mission (RCM), synthetic aperture radar (SAR), compact polarimetry, attention, contextual information, feature learning, deep learning.**

I. Introduction

SATELLITES comprising the RADARSAT Constellation Mission (RCM) provide synthetic aperture radar (SAR) data in various acquisition modes including compact polarimetric (CP). In contrast to dual-polarized (DP) mode, the CP mode imagery preserves the phase information between channels, making it more appropriate for various applications such as land cover classification [1]. Land cover classification is essential because it provides valuable information about the Earth's surface and its changes over time which are important for urban planning, natural resource management, and environmental monitoring [2], [3]. Due to the limited data availability, the potential of generating land cover maps using CP SAR data remains largely unexplored.

Land cover classification is challenging due to speckle noise [4] and ambiguities associated with backscatter and unique class discrimination [5]. To mitigate this, conventional land cover classification methods increase the number and type of hand-crafted features [6]. It is known that pixel-level features and spatially-based texture features have limited capabilities for scene classification [4].

Deep learning (DL) methods provide an advantage over shallow-structured machine learning tools (such as support vector machine [7]) by inherently extracting features [8], [9]. Due to the intrinsic 2-D structure of remote sensing images, convolutional neural networks (CNNs), as a DL approach, are widely used for image processing tasks [10]. While CNNs are able to extract local features, they do not inherently capture long-distance dependency among pixels which is important for land cover classification tasks due to spatial heterogeneity of targets [11]. In contrast, vision transformer models are capable of capturing long-distance dependencies [12]. As an example, the Vision Transformer (ViT) [13] utilizes the idea of self-attention [14] to enable global receptive field processing of non-overlapping patches.

Despite successful performance on various computer vision tasks [15], ViT has limitations of requiring high computational and memory costs, even for nominally-sized input images and keeping the dimensions of the produced feature maps consistent [16]. To enhance the accuracy and efficiency of ViT in different tasks, several transformer architectures have been introduced [16–19]. These approaches are local-based such as Swin Transformer [17] or global-based such as Pyramid Vision Transformer (PVT) [16]. The local-based approaches divide the input image patch into non-overlapping windows and calculate the self-attention inside of each window. The Swin Transformer uses a shifting window to describe the relationship among windows, which gradually moves the local window's boundaries. However, the window shifting technique lacks optimization for GPU usage and demonstrates inefficient memory utilization [18]. Global approaches such as PVT preserve the global receptive field of ViT but lower the resolution of the key and value feature maps to reduce complexity. However, despite this reduction, the model's complexity is frequently still quadratic in relation to the input image's resolution, posing issues for larger images [18].

Successful classification has been demonstrated by both the local self-attention methods [17], [20] and the global self-attention methods [16], [18]. However, these approaches impose limitations on the original full self-attention's ability to concurrently capture short- and long-range dependencies [15].

This work was supported in part by the Network of Centers of Excellence (ArcticNet), in part by the Natural Sciences and Engineering Research Council (NSERC), and in part by the Canadian Ice Service (CIS) of Environment Canada. (Corresponding author: Saeid Taleghanidoozdoozan.)

The authors are with the Vision and Image Processing (VIP) Research Group, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: stalegha@uwaterloo.ca).

Land cover exhibits high spatial heterogeneity [21]; therefore, capturing both fine-grained and coarse-grained spatial dependencies simultaneously is important because it allows for a comprehensive understanding of the relationships between different pixels in a given feature map. The Focal transformer [15] is designed to integrate fine-grained and different scale coarse-grained spatial dependencies, but to accomplish this task requires a highly complex architecture with accompanying high computing requirements.

In a DL model, the shallow layers primarily focus on capturing low-level and fine features. On the other hand, the deep layers of the model are responsible for extracting deeper, coarse, and semantic features that encapsulate higher-level features, including abstract representations and complex relationships within the data [9]. Consequently, by integrating both low-level and high-level features, the DL model can leverage the complementary nature of these features and achieve a more robust and accurate performance in classification tasks [22].

To the best of our knowledge, there is currently no published research specifically addressing the generation of land cover maps in CP SAR imagery using a self-attention method. As a result, this paper proposes a novel classification method called PFC transformer (Pyramid of Fine- and Coarse-grained attentions transformer), which utilizes a pyramid of window-based vision transformers to measure both fine-grained attention within a window and coarse-grained attention between windows. In summary, this study makes the following contributions in CP SAR land cover classification:

- Our proposed method simultaneously utilizes fine- and coarse-grained spatial dependencies, enabling the model to extract more discriminative and detailed features by capturing spatial relationships at different scales. This attribute effectively addresses spatial heterogeneity present in land covers, ultimately leading to more accurate land cover classification.
- Our proposed method incorporates the outputs of different stages and leverages information across multiple scales, resulting in enhanced accuracy for land cover classification. By addressing the challenges of signature ambiguity, this integration of low- and high-level features improves the accuracy of land cover classification.
- The potential of state-of-the-art (SOTA) DL methods in generating accurate land cover maps using CP SAR data is evaluated and compared with that of the proposed method. This thorough assessment not only advances the understanding of DL techniques in this domain but also provides valuable insights for decision-makers and researchers aiming to utilize SOTA DL method for land cover classification and monitoring in CP SAR data.

Experiments are based on a pair of high-resolution RCM CP SAR scenes. The proposed PFC transformer surpasses SOTA methods, including Swin, Focal, PVT, Twins [18], CAT [23], SepViT [19], and residual-based CNN (ResCNN) [24], in terms of generating accurate land-cover maps. This paper is organized as follows. Section II provides a literature review of land cover classification methods utilizing SAR data. Then, the fundamentals of CP SAR data is explained in Section III. Section IV describes the proposed method, and the study area as well as datasets are introduced in Section V. Section VI presents and analyzes the experimental results, and Section VII provides the conclusions of the study.

II. Background

A. Land cover classification using CP SAR data

Most of the existing land cover classification methods using SAR data are based on QP or DP. There are only a few known published papers on land cover classification using CP SAR data [25–27]. Robertson *et al.* [25] utilized hand-crafted features derived from CP SAR data and employed a random forest (RF) classifier for producing crop maps. Nonetheless, the creation of efficient hand-crafted features necessitates expertise in the field and a deep comprehension of the particular domain. Furthermore, the RF classifier does not consider spatial information. Roy *et al.* [26] proposed a MapReduce-based multi-layer perceptron algorithm to distinguish different land cover classes. However, the algorithm did not utilize contextual information, and only numerical results are reported without a classified land cover map, so visual evaluation is not possible. Ghanbari *et al.* [27] proposed a region-based semi-supervised graph network land cover classification using RCM CP SAR data. Despite achieving reliable outcomes, the utilization of hand-crafted features and uncertainty in the homogeneity of generated regions may impact the results. Therefore, it is imperative to focus on designing a feature learning-based land cover classification method for large CP SAR scenes that reduces reliance on hand-crafted features and effectively addresses the issue of signature ambiguity by incorporating spatial information.

B. Land cover classification using CNNs

CNNs are widely used to generate SAR land cover maps [28]. Zhou *et al.* [29] applied a CNN for QP SAR land cover classification, employing a model that included two convolutional layers and two fully connected layers. Then, several methods for land cover classification based on CNNs were proposed [3], [5], [30–34]. For example, Zhang *et al.* [30] proposed a complex-valued CNN that was tailored to accommodate the arithmetic features of complex data. To extract both spatial- and channel-wise information, Dong *et al.* [31] utilized 3-D convolution. Liu *et al.* [5] considered the statistical distribution of the mid-level features generated by a CNN model to increase the generalization of the model. Although CNNs reached reliable results, they can introduce artifacts along the edges of adjacent patches, leading to the over-smoothing of object boundaries and losing of useful spatial resolution detail [35]. Moreover, despite their proficiency in organizing local features, CNNs encounter challenges in capturing spatial dependencies that extend over long distances [12], [36].

In several recent studies [4], [9], [37–40], fully convolutional networks (FCNs) have been identified as another common approach that exhibits promising land cover results. Wang *et al.* [4] proposed an integration of FCN with sparse and low-rank subspace features network to classify

QP SAR images. Li *et al.* [41] suggested the utilization of an FCN with a sliding window technique to alleviate the computational burden and minimize memory usage. Mohammadimanesh *et al.* [9] proposed an FCN network including inception and skip connection to utilize richer contextual information and more detailed information in QP SAR data to classify. Henry *et al.* [38] evaluated the potential of three FCNs in extracting roads from high-resolution SAR images. However, the utilization of FCN models faces a significant hurdle due to the requirement of whole or dense labeled scenes for their training. The scarcity of labeled SAR data, especially in RCM CP data, makes it infeasible to utilize FCN models [4]. Given the limitations of CNNs and FCNs in capturing fine- and coarse-grained spatial dependencies and the requirement for dense labeled scenes, it is necessary to explore a method that can effectively capture both levels of spatial dependencies in CP SAR data without relying on whole labeled scenes.

C. Land cover classification using transformers

Recently, the effectiveness of transformer models in remote sensing applications has captured the attention of remote sensing researchers [28], [36], [42–47]. While several studies have employed transformer models to merge optical and SAR images and leverage the benefits of both data types [41], [48–50], the absence of clear optical images of the same area due to cloud cover impedes progress. Other studies have combined CNNs and ViT methods to utilize local and global information for land cover mapping [43], [45], [46]. To integrate the outputs of each branch, various fusion methods have been proposed [45], [48], [51], [52]. However, these methods have certain limitations, such as increased complexity compared to individual models, requiring more time and data for training

To address the limitations discussed, we propose a hierarchical fine- and coarse-grained attentions transformer for land cover classification. Our approach integrates fine and coarse attentions, capturing spatial dependencies, within the same layer using a learnable mechanism. This integration leads to richer information integration. Additionally, our method leverages a pyramid of low- and high-level features to accommodate varying levels of complexity. By combining these techniques, we aim to overcome the limitations of existing CP SAR land cover methods and improve accuracy.

III. Compact polarimetric SAR basics

The backscattering field of a single look complex (SLC) CP SAR data is defined by a 2×1 complex vector E. For RCM CP mode, in which a right circular polarized wave (R) is transmitted, and both horizontal (H) and vertical (V) polarizations are received, E is defined as:

$$E = \begin{bmatrix} E_{RH} \\ E_{RV} \end{bmatrix} = S\hat{u}_t = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \hat{u}_t \tag{1}$$

where $\hat{u}_t$ is a unit Jones vector associated with a canonical polarization. S_{ij} is a complex number where the polarizations are represented by i (transmitted) and j (received) [53]. To utilize polarimetric SAR data, the coherency matrix is often used instead of the scattering matrix due to several reasons, including enhancing information content and mitigating the adverse impact of speckle noise [54]. The coherency matrix of the RCM CP SAR data is a 2×2 semipositive-definite Hermitian matrix defined as [55], [56]:

$$J = \frac{1}{n}\sum_{i=1}^{n} E\cdot E^{*^T} = \begin{bmatrix} \langle \mid S_{RH} \mid^2 \rangle & \langle S_{RH}S_{RV}^* \rangle \\ \langle S_{RV}S_{RH}^* \rangle & \langle \mid S_{RV} \mid^2 \rangle \end{bmatrix} \tag{2}$$

where n is the number of looks for averaging. The term T represents the transpose, $*$ represents the complex conjugate, and $<\cdots>$ defines spatial ensemble averaging. The diagonal elements describe the intensities and the non-diagonal describe the intensities and phase between polarizations.

IV. Methodology

Fig 1 shows the architecture of the proposed PFC transformer method. The proposed method consists of four stages that produce four feature maps of varying scales. The structure of all stages is similar, comprising of a downsampling layer, except for the stage 1 which includes linear embedding, and N_i times FC block attention. Each part of the architecture is described separately.

A. Linear Embedding

The linear embedding is a linear transformer that is applied to reduce the spatial size of the image patch and increase the dimension of the raw-valued features into an arbitrary dimension [13], [17]. Since, in this study, the size of the input image patch is not very big, linear embedding is used to increase the dimension of features. Assume that $x_{in} \in R^{H\times W\times C_0}$ is the input image patch where H and W are the spatial dimension and C is the feature dimension, the linear embedding projects x_{in} into $z \in R^{H\times W\times C_1}$.

B. FC Transformer Block

The main core of the proposed method is the FC transformer block (see Fig 2 (a)). Since the proposed method is window-based, an input feature map (z) is divided into non-overlapping $M\times M$ windows, and a layer normalization (LN) is applied. Then, by using a linear function, query (Q_f), key (K_f), and value (V_f)$\in R^{(M\times M)\times d}$ matrices are calculated where f stands for fine-grained and d is the depth equals to the feature dimension of z divided by the number of heads [17].

To calculate fine-coarse attention (F_{attn}), similar to the approach employed by the Swin transformer, the self-attention within each window is computed as follows:

$$F_{attn} = softmax(Q_f K_f^T/\sqrt{d} + B_f)V_f \tag{3}$$

As described by Liu *et al.* [17], B_f is the learnable relative position bias which its values are taken from $\hat{B}_f \in R^{(2M-1)\times(2M-1)}$. Fig 2 (b) shows the structure of the F_{attn}.

In addition to the fine-grained attention, the PFC transformer method introduces an approach for calculating the coarse-grained attention (C_{attn}). To compute C_{attn}, a learnable window pooling is applied to reduce the size of K_f and V_f matrices from their original dimensions of $(M\times M)\times d$ to a compact

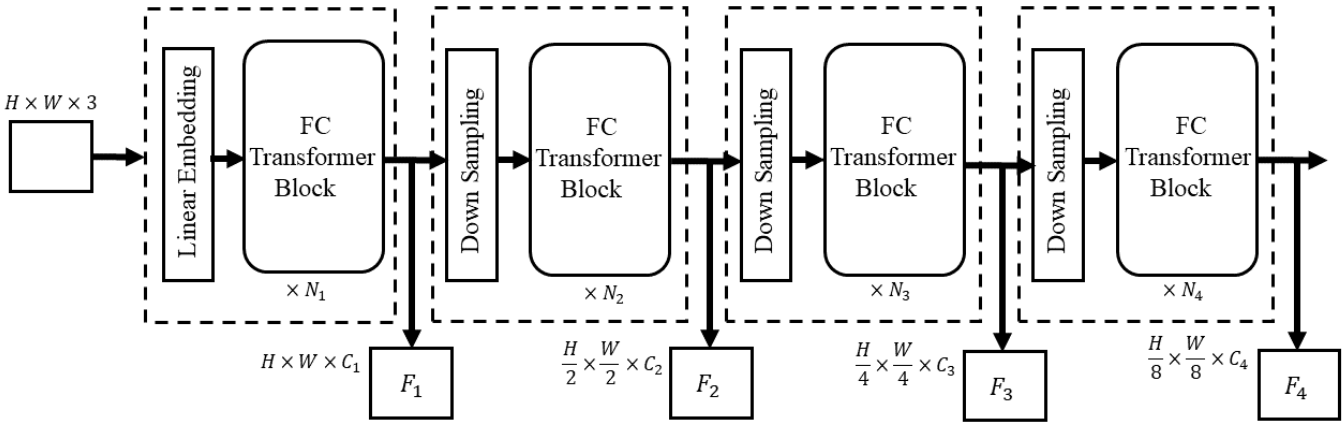


Fig. 1: Architecture of the PFC Transformer

$(1 \times 1) \times d$ representation, called K_c and V_c where c stands for coarse-grained. This reduction in size not only helps to alleviate the complexity and computational costs associated with the model but also enables the consideration of far spatial dependencies. Then, the attention between each fine-grained query matrix, Q_f, and coarse-grained K_c and V_c are calculated as follows:

$$C_{attn} = softmax(Q_f K_c^T / \sqrt{d} + B_c)V_c \tag{4}$$

B_c is the relative position bias among fine- and coarse-grained windows; however, since the size of the K_c and V_c are not the same as Q_f, to represent the relative position bias between them, values in B_c are taken from $\hat{B}_c \in R^{(NW+M-1)\times(NW+M-1)}$ where NW stands for the number of coarse-grained windows [15]. By leveraging this coarse-grained attention mechanism, the model gains the ability to capture long-range spatial dependencies. The structure of the C_{attn} is depicted in Fig 2 (c).

Finally, to take advantage of both fine- and coarse-grained spatial dependencies and utilize them simultaneously, both attentions are concatenated. Nevertheless, this concatenation operation leads to a doubling of the feature dimension. As a result, to restore the number of features to its original value in the input, a projection step becomes necessary. This projection ensures compatibility and coherence in subsequent stages of the computation. The fine- and coarse- attention (FC_{attn}) is computed as:

$$FC_{attn} = Concat(F_{attn}, C_{attn})W_{fc} \tag{5}$$

where W_{fc} is the learnable linear projection.

The rest of the FC transformer block is followed by a skip connection with the input feature map, an LN, and a 2-layer multi-layer perceptron (MLP) with GELU nonlinearity in between and again a skip connection, following the same procedure as [13], [17]. In general, the FC transformer block is computed as

$$\alpha = LN(z^{l-1}) \tag{6}$$

$$\hat{z}^l = FC_{attn}(F_{attn}(\alpha), C_{attn}(\alpha)) + z^{l-1} \tag{7}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \tag{8}$$

in which z^{l-1} is the input feature map from the previous layer.

C. Downsampling

As the network becomes deeper, reducing the spatial dimensions of the feature maps to produce a hierarchical representation is necessary. Therefore, the downsampling layer which is a convolutional operator compromised of a 2×2 kernel with stride 2 along with an adjustable number of output features is employed to reduce the spatial size of feature maps by a factor of 2. The downsampling reduces the computational cost and allows the network to learn a hierarchical representation of the input.

D. Fusion

Unlike previous methods which only utilized the output of the last stage, our proposed method employs a pyramid of FC transformer blocks' outputs to aggregate information from all stages, allowing for more comprehensive characterization of land cover types (see Fig. 1). Given an input image patch of size $H \times W \times 3$, the linear embedding is applied to increase the number of features to C_1. Then, it is passed through an FC transformer block with N_1 layers resulting in F_1 with the shape of $H \times W \times C_1$. Next, F_1 is used as the input of the next stage and this process is repeated to obtain feature maps of F_2, F_3, and F_4. To combine F_1, F_2, F_3, F_4, a learnable linear function is applied to decrease their spatial sizes to that of the final stage's output, which is $H/8 \times W/8$, as follows:

$$F_t = Concat(F_1 W_1, F_2 W_2, F_3 W_3, F_4 W_4) \tag{9}$$

in which W_1, W_2, W_3, and W_4 are convolutional operators with strides of 8, 4, 2, and 1 respectively. The size of F_t is $H/8 \times W/8 \times (C_1 + C_2 + C_3 + C_4)$. Then, a global average pooling layer is applied to F_t followed by a fully connected

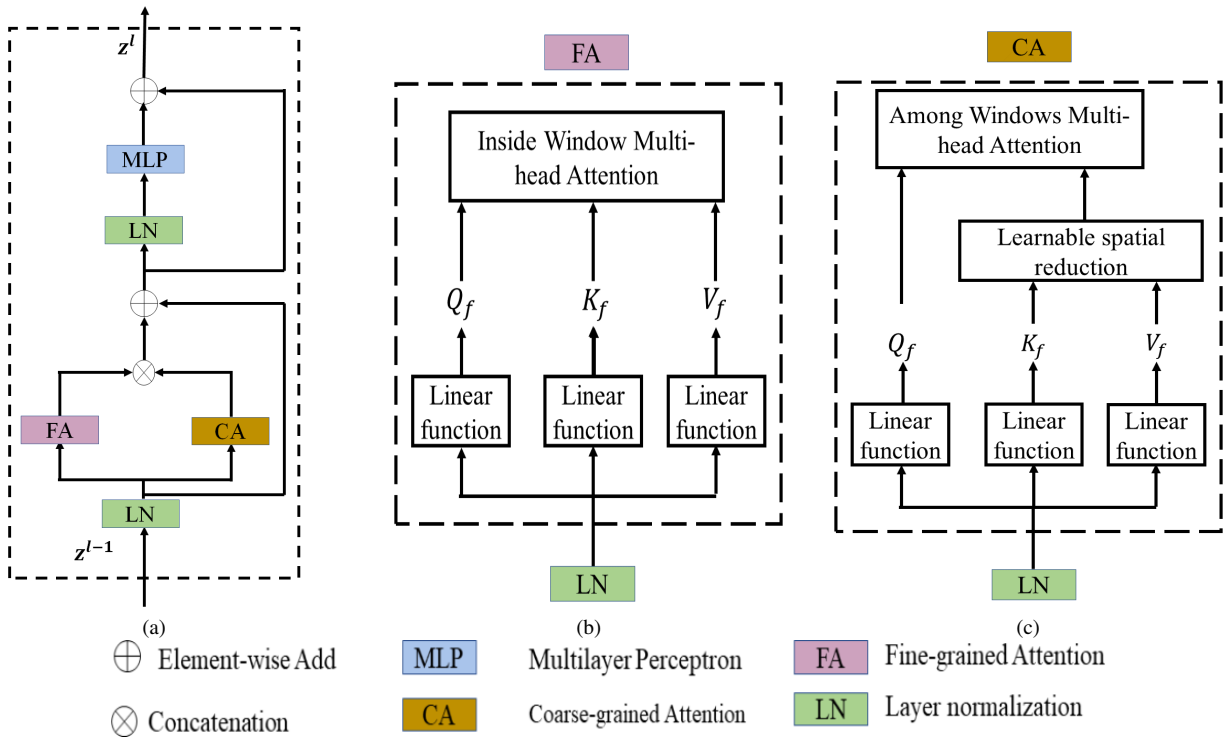


Fig. 2: (a) Fine- and Coarse-grained Attention Block, (b) Fine-grained Attention, (c) Coarse-grained Attention.

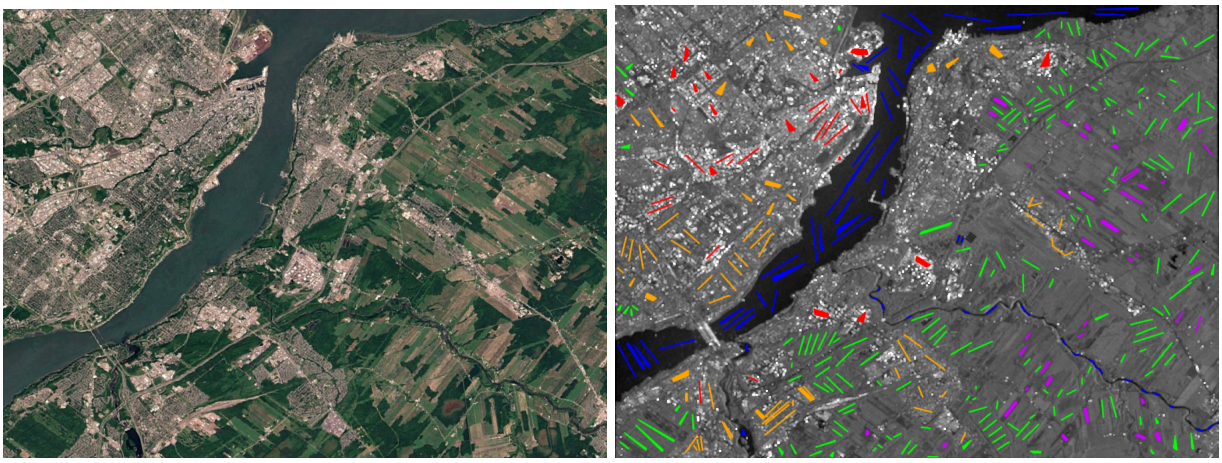


Fig. 3: (a) Google Earth image of Quebec City scene, (b) the first element of CP coherency matrix along with manually selected samples.

layer with nodes equal to the number of classes to determine the land cover class.

V. Study Area and Dataset

Two very high-resolution (3m) SLC RCM CP SAR scenes with the sampled pixel and line spacing of 1.39 and 2 meters were used to evaluate the performance of the proposed method and compare it with other methods. Captured on August 9th, 2022, over Quebec City in Canada, the first scene covers approximately 43 km $\times$ 13 km and has a size of 10954 $\times$ 8146 pixels, with an incidence angle range of 47.50 to 48.67 degrees. Fig 3 (a) presents the Google Earth image of this scene. The second scene, acquired on June 27th, 2020, has a size of 9344$\times$21942 pixels, covering around 43 km$\times$130 km over the city of Ottawa in Canada. Its incidence angle ranges from 38.48 to 39.90 degrees, and its corresponding Google Earth image is shown in Fig 4 (a).

The study area has five primary classes: forest, water, two distinct urban areas, and agricultural lands (farms). The urban areas are divided into two groups because some buildings appear bright (Urban 1) while other ones are a mixture of trees and buildings (Urban 2) and their backscattering is not

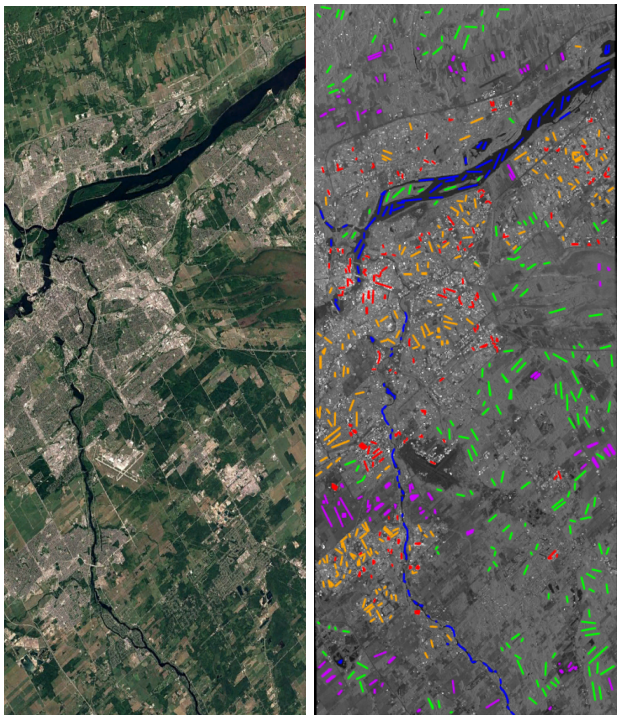


Fig. 4: (a) Google Earth image of the city of Ottawa, (b) the first element of CP coherency matrix along with manually selected samples.

as bright as the first group. The samples were chosen manually by visually examining the SAR scenes and the Google Earth images.

A 7×7 boxcar filter is applied on both datasets to reduce the impact of speckle noise. Since the images are large and this leads to exceptional computational cost, we reduce this cost by taking a 4×4 non-overlapping block-wise average of the pixels.

VI. Experiments

In this section, the performance of the proposed method in classifying land type covers is discussed and compared to that of the SOTA methods. To assess the efficacy of combining features with different levels, the proposed model was applied both with (PFC transformer) and without (FC transformer) using the pyramid of features. Table I indicates the structure of the proposed method. It includes 4 stages where the FC transformer block is repeated twice in each stage. The number of feature maps in each stage is set to 16, 32, 64, and 128, respectively. The size of non-overlapping windows is set to 4×4, and the number of heads for each stage is 1, 4, 4, and 8, respectively.

The performance of the methods was examined using several metrics, namely, overall accuracy (OA), kappa coefficient (κ), f-1 scores of each class ($F1$), and averaged f-1 score ($F1_{avg}$). OA is determined by dividing the number of correctly classified test samples by the total number of test samples. κ measures the level of agreement between the test samples and

TABLE I: Detailed architecture of the proposed PFC Attention method.

	Output	PFC Attention Method
Stage 1	32×32×16	Linear Embedding, LN
		$\begin{Bmatrix}\text{window size}: 4\times4\\ \#\text{heads}: 1\end{Bmatrix}\times 2$
Stage 2	16×16×32	Downsampling, LN
		$\begin{Bmatrix}\text{window size}: 4\times4\\ \#\text{heads}: 4\end{Bmatrix}\times 2$
Stage 3	8×8×64	Downsampling, LN
		$\begin{Bmatrix}\text{window size}: 4\times4\\ \#\text{heads}: 4\end{Bmatrix}\times 2$
Stage 4	4×4×128	Downsampling, LN
		$\begin{Bmatrix}\text{window size}: 4\times4\\ \#\text{heads}: 8\end{Bmatrix}\times 2$
Global Average	1×1×128	4×4 average pool
Classification	5	128×5 fully connected
Softmax	5	

the final labeled map [9]. $F1$ is a harmonic mean of precision and recall, which is particularly useful for imbalanced classes [9]. The highest and lowest possible values of $F1$ are 1 and 0.

A. Training and Testing

In this study, the labeled pixels of the Quebec scene were used for training the models that were evaluated using the labeled pixels chosen from the Ottawa scene. Moreover, to better evaluate the performance of the methods, three different regions have been selected and shown in Fig. 5. Regions A and B show agricultural, forest and urban areas, while Region C includes forest and agricultural areas.

Table II represents the number of training and testing samples. The training samples were used to standardize the Quebec and Ottawa scenes. To train the models, patches of size 32×32×3 were extracted around each labeled pixel, where 3 represents the absolute value of the coherency matrix elements in (2). In addition, the models were trained using ADAMW optimization [57] with the learning rate, weight decay, and beta parameters set to 1e-5, 0.05, 0.9, and 0.999 as well as the batch size and training epochs are 32 and 100, respectively. In the training step, 80% of the training samples were utilized to adjust the model's weight values by minimizing the multi-class cross-entropy lost function [58], while the remaining 20% were used for validation purposes. The weight values of the model that achieved the highest validation accuracy were selected.

B. Results

Fig. 6 shows the results obtained by the different methods along with their OA. Due to resizing the images to fit the page,

Fig. 5: The Google Earth image of the test scene with three regions of interest along with their corresponding $|S_{RH}|^2$. Regions A and B primarily consist of urban, farm and forest classes, while Region C displays both forest and farm classes.

TABLE II: The number of training and testing pixels for each class selected from the Quebec and Ottawa scenes, respectively.

Class	# of train	# of test
Forest	15381	11690
Water	14853	8093
Urban 1	12032	11263
Urban 2	15098	10206
Farm	10773	20022

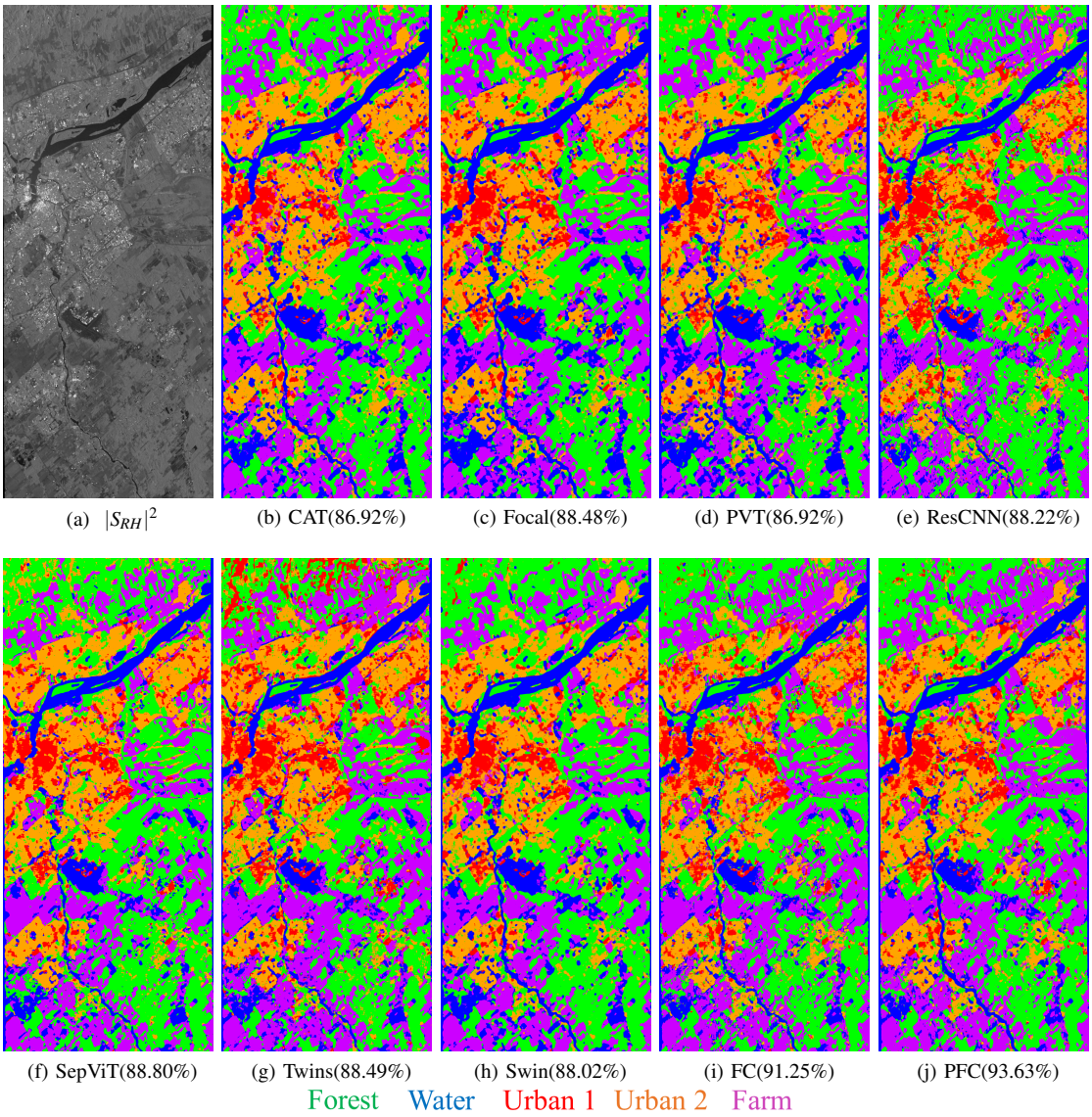
finer details present in the original images are not apparent. Upon visual inspection of the outputs, the CAT, Focal, PVT, Swin, and ResCNN methods appear to overestimate the water class in the lower portion of the scene. Twins misclassifies many forest and farm samples into Urban 1 class in the upper part of the scene. SepViT and Twins exhibited poor detection of the river in the middle of the scene and it is narrow than that detected by the other methods as well as the proposed methods.

The FC and PFC transformer methods have a higher accuracy and improved spatial representation in specifying the type of land covers than the SOTA ones. This is because the proposed methods, unlike the other approaches, utilize close and far dependency among pixels simultaneously and combine different feature levels resulting in reducing the rate of misclassification.

As shown in Table III, we compared the quantitative results obtained by the proposed methods to those of the SOTA ones. The CAT and PVT methods were found to have the lowest *OA* among the SOTA methods, with both achieving of 86.92%. In contrast, the SepViT method achieved a reliable overall accuracy of 88.80%. While the Swin, CAT, and PVT methods showed comparable κ and $F1_{avg}$, the Focal, ResCNN, SepViT, and Twins methods obtained higher accuracies, albeit still lower than those achieved by the FC and PFC transformer methods.

The proposed FC transformer method achieved an *OA* of 91.25%, which is about 3-4% higher than those achieved by the SepViT and CAT methods. The higher values of κ and $F1_{avg}$ obtained by the FC transformer method provide additional evidence of the effectiveness of the proposed attention mechanism in improving the accuracy of generating land cover type maps. These findings identify that the SOTA models have limitations that prevent them from achieving the same level of performance as the FC transformer.

When comparing the performance of the FC and PFC transformer methods, we found that the PFC transformer outperformed the former, with a higher accuracy. By fusion of the different feature levels in a learnable manner, the *OA* value reached 2% higher. Moreover, the higher values of κ and $F1_{avg}$ obtained by the PFC transformer suggest that leveraging the outputs of all stages can lead to improved accuracy of the balanced and imbalanced classes [6].



(a) $|S_{RH}|^2$ (b) CAT(86.92%) (c) Focal(88.48%) (d) PVT(86.92%) (e) ResCNN(88.22%)

(f) SepViT(88.80%) (g) Twins(88.49%) (h) Swin(88.02%) (i) FC(91.25%) (j) PFC(93.63%)

Fig. 6: (a) shows the $|S_{RH}|^2$ image of the test scene and (b)-(j) are the results obtained by each method along with their *OA*.

The FC and PFC transformers yield higher $F1$ for the forest, water, and farm classes than the SOTA methods, demonstrating the significance of fine- and coarse-grained dependencies among pixels and the benefits of utilizing features at different levels. The ResCNN and Focal methods achieved slightly higher $F1$ for the urban classes compared to the proposed methods, but the difference is negligible.

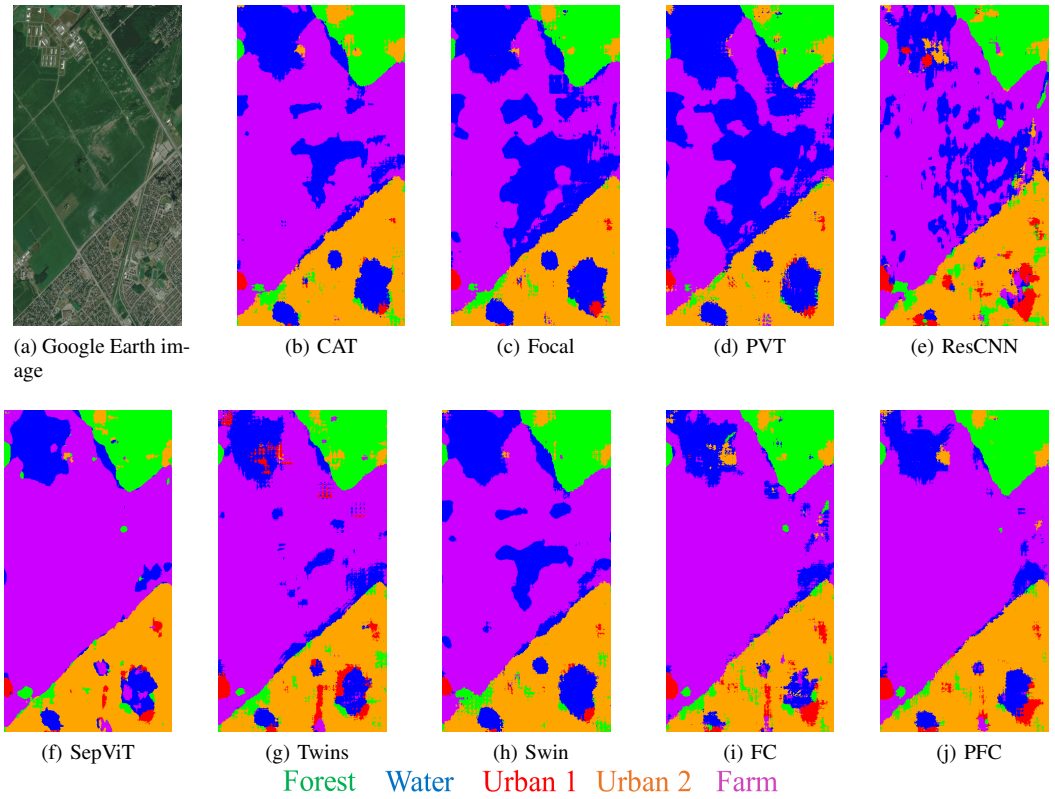
Fig. 7 shows the outputs of the methods on Region A which is a mixture of buildings, forest, and agricultural areas. Notably, the CAT, Focal, PVT, ResCNN, and Swin methods exhibited a higher rate of misclassifying water in this region, while the SepViT, Twins, and proposed methods yielded more accurate outcomes. The output of the methods for Region B is shown in Fig. 8. Among the SOTA methods, the CAT, Focal, PVT, SepViT, and Swin methods misclassified a significant portion of the agricultural lands as water class while the ResCNN and Twins performed better. Moreover, the FC transformer method exhibited performance over ResCNN and Twins, but the PFC transformer method achieved the best classification performance in Region B. Using fine- and coarse-grained spatial information decreased the rate of water misclassification in particular for the left agricultural land. By adding the pyramid of low- and high-level features to the FC transformer method, the rate of misclassification was reduced significantly. This is because the integration of different level features enables the model to capture a wide range of features across different scales. Fig. 9 shows the output of the methods on Region C, which includes forest and farm classes. All methods, except the proposed ones, had a high rate of misclassifying agricultural areas as forests. The proposed methods exhibited significantly better classification performance for forests.

VII. Conclusion

A new transformer approach was introduced in this paper for generating land cover maps using high-resolution CP SAR

TABLE III: Assessment of the results obtained by the different methods by using overall accuracy (OA), kappa coefficient (κ), averaged f-1 score ($F1_{avg}$), and f-1 score of each class. the **bold** numbers indicate the highest accurate results.

Name	$OA(\%)$	κ	$F1_{avg}$	Forest	Water	Urban1	Urban2	Farm
CAT [23]	86.92	0.8343	0.8696	0.9116	0.7723	0.8843	0.9234	0.8574
Focal [15]	88.48	0.8544	0.8852	0.9248	0.7812	0.9208	**0.9278**	0.8715
PVT [16]	86.92	0.8351	0.8694	0.9218	0.7596	0.8955	0.9138	0.8561
ResCNN [24]	88.22	0.8500	0.8780	0.8835	0.8246	**0.9215**	0.8820	0.8984
SepViT [19]	88.80	0.8579	0.8850	0.9049	0.8219	0.9046	0.8971	0.8970
Twins [18]	88.49	0.8538	0.8788	0.9181	0.8153	0.8531	0.8931	0.9144
Swin [17]	87.14	0.8372	0.8707	0.9049	0.7881	0.9076	0.8866	0.8661
FC	91.25	0.8885	0.9054	0.9346	0.8564	0.9087	0.8844	0.9428
PFC	**93.63**	**0.9185**	**0.9285**	**0.9491**	**0.8864**	0.9191	0.9179	**0.9701**



(a) Google Earth image (b) CAT (c) Focal (d) PVT (e) ResCNN

(f) SepViT (g) Twins (h) Swin (i) FC (j) PFC

Fig. 7: (a) shows the Google Earth image of Region A including urban, farm, and forest classes. (b)-(j) are the results obtained by each method.

scenes. To the best of our knowledge, this is the first study that leverages spatial attention information in CP SAR data for land type classification. The proposed attention mechanism captures both fine- and coarse-grained dependencies among pixels within a feature map, resulting in richer information. This attribute endows the method with the ability to consider the spatial relationship among the pixels resulting in more accurate outputs. The qualitative and quantitative comparison among the results obtained by the proposed transformer method and the well-known SOTA methods confirm the efficiency of the long dependency in increasing the accuracy of the generated land cover maps.

Furthermore, we take into account the outputs from all stages and exploit the information across various scales to utilize more detailed information. The comparison of the outputs from the proposed method, both with and without feature fusion, highlights the importance of incorporating low-level features. This fusion approach improves the proposed method's ability to identify different land cover types.

The limited availability of RCM data has led to a shortage of annotated CP scenes. As training deep learning methods demand a large number of samples, it is essential to consider semi-supervised techniques in studies. The proposed method can potentially be applied for dense semantic segmentation

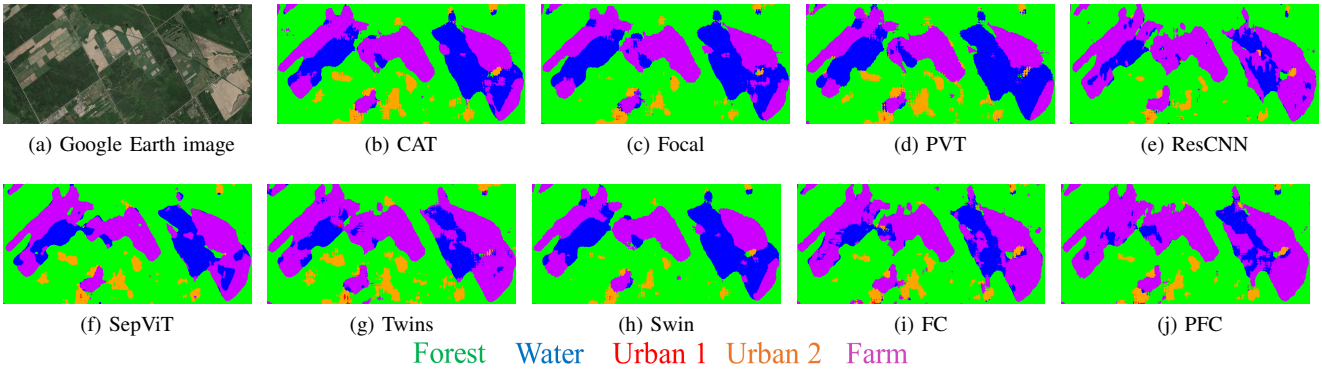


Fig. 8: (a) displays a Google Earth image of Region B, which includes agricultural lands, forests, and a few buildings. (b)-(j) are the results obtained by each method.

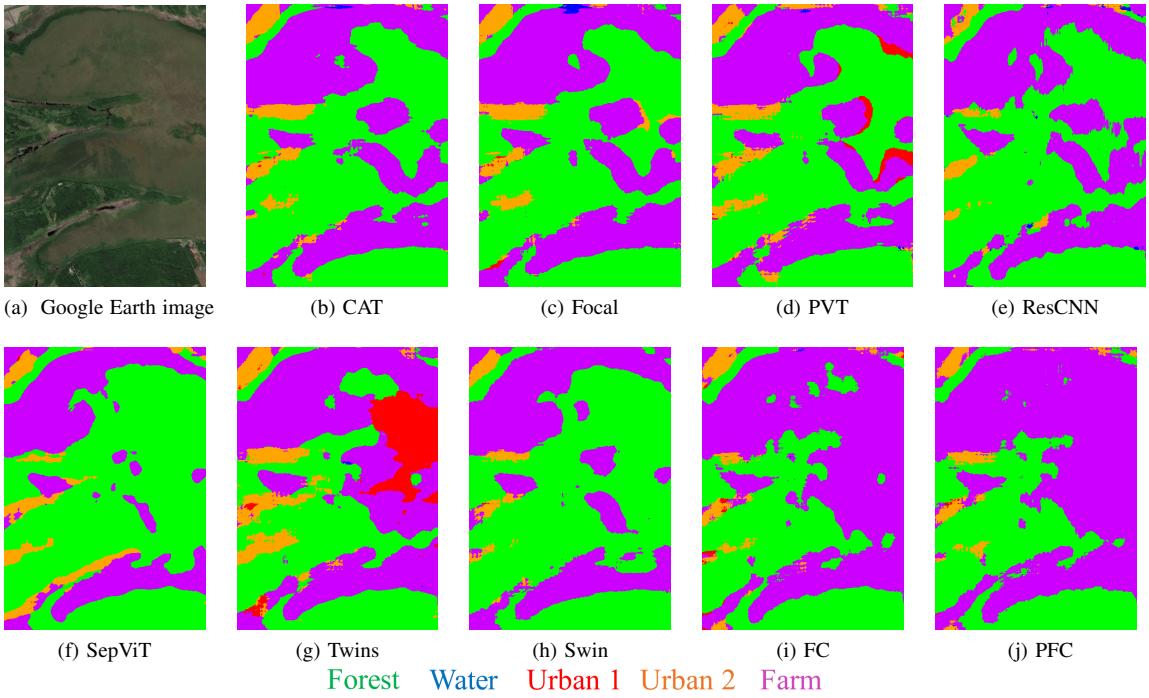


Fig. 9: (a) shows the Google Earth image of Region C including forest and farm classes. (b)-(j) are the results obtained by each method.

purposes by increasing the availability of RCM CP SAR scenes and ground truth samples in the future.

REFERENCES

[1] M. Dabboor, S. Iris, and V. Singhroy, "The RADARSAT constellation mission in support of environmental applications," in *Proceedings*, vol. 2, no. 7. MDPI, 2018, p. 323.

[2] Z. Qi, A. G.-O. Yeh, X. Li, and Z. Lin, "A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data," *Remote Sensing of Environment*, vol. 118, pp. 21–39, 2012.

[3] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1011–1026, 2020.

[4] Y. Wang, C. He, X. Liu, and M. Liao, "A hierarchical fully convolutional network integrated with sparse and low-rank subspace representations for PolSAR imagery classification," *Remote Sensing*, vol. 10, no. 2, p. 342, 2018.

[5] X. Liu, C. He, Q. Zhang, and M. Liao, "Statistical convolutional neural network for land-cover classification from SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 9, pp. 1548–1552, 2019.

[6] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, B. Brisco, and E. Gill, "Full and simulated compact polarimetry SAR responses to Canadian wetlands: Separability analysis and classification," *Remote Sensing*, vol. 11, no. 5, p. 516, 2019.

[7] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[8] W. Song, M. Li, W. Gao, D. Huang, Z. Ma, A. Liotta, and C. Perra, "Automatic sea-ice classification of SAR images based on spatial and temporal features learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 9887–9901, 2021.

[9] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS journal of photogrammetry and remote sensing*, vol. 151, pp. 223–236, 2019.

[10] X. Ma, A. Fu, J. Wang, H. Wang, and B. Yin, "Hyperspectral image classification based on deep deconvolution network with skip architec-

ture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4781–4791, 2018.

[11] B. Ghimire, J. Rogan, and J. Miller, "Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the getis statistic," *Remote Sensing Letters*, vol. 1, no. 1, pp. 45–54, 2010.

[12] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye, "Conformer: Local features coupling global representations for recognition and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[13] 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.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," *arXiv preprint arXiv:2107.00641*, 2021.

[16] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[17] 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.

[18] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.

[19] W. Li, X. Wang, X. Xia, J. Wu, X. Xiao, M. Zheng, and S. Wen, "Sepvit: Separable vision transformer," *arXiv preprint arXiv:2203.15380*, 2022.

[20] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2998–3008.

[21] H. Xing, L. Zhu, Y. Feng, W. Wang, D. Hou, F. Meng, and Y. Ni, "An adaptive change threshold selection method based on land cover posterior probability and spatial neighborhood information," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 608–11 621, 2021.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[23] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[24] 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.

[25] L. D. Robertson, H. McNairn, C. McNairn, S. Ihuoma, and X. Jiao, "Compact polarimetry for operational crop inventory," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 4423–4426.

[26] S. Roy, A. Das, and S. N. Omkar, "A distributed land cover classification of FP and CP SAR observation using MapReduce-based multi-layer perceptron algorithm over the Mumbai mangrove region of India," *International Journal of Remote Sensing*, vol. 44, no. 5, pp. 1510–1532, 2023.

[27] M. Ghanbari, L. Xu, and D. A. Clausi, "Local and global spatial information for land cover semi-supervised classification of complex polarimetric SAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[28] H. Wang, C. Xing, J. Yin, and J. Yang, "Land cover classification for polarimetric SAR images based on vision transformer," *Remote Sensing*, vol. 14, no. 18, p. 4656, 2022.

[29] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[30] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7177–7188, 2017.

[31] H. Dong, L. Zhang, and B. Zou, "PolSAR image classification with lightweight 3d convolutional networks," *Remote Sensing*, vol. 12, no. 3, p. 396, 2020.

[32] S.-W. Chen and C.-S. Tao, "PolSAR image classification using polarimetric-feature-driven deep convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 627–631, 2018.

[33] C. Yang, B. Hou, B. Ren, Y. Hu, and L. Jiao, "CNN-based polarimetric decomposition feature selection for PolSAR image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8796–8812, 2019.

[34] W. Xie, G. Ma, F. Zhao, H. Liu, and L. Zhang, "PolSAR image classification via a novel semi-supervised recurrent complex-valued convolution neural network," *Neurocomputing*, vol. 388, pp. 255–268, 2020.

[35] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 133–144, 2018.

[36] 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.

[37] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, "PolSAR image semantic segmentation based on deep transfer learning—Realizing smooth classification with small training sets," *IEEE Geoscience and remote sensing letters*, vol. 16, no. 6, pp. 977–981, 2019.

[38] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in SAR satellite images with deep fully convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 12, pp. 1867–1871, 2018.

[39] A. G. Mullissa, C. Persello, and V. Tolpekin, "Fully convolutional networks for multi-temporal SAR image classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 6635–6638.

[40] A. G. Mullissa, C. Persello, and A. Stein, "PolSARNet: A deep fully convolutional network for polarimetric SAR image classification," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 12, no. 12, pp. 5300–5309, 2019.

[41] Y. Li, Y. Chen, G. Liu, and L. Jiao, "A novel deep fully convolutional network for PolSAR image classification," *Remote Sensing*, vol. 10, no. 12, p. 1984, 2018.

[42] 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.

[43] X. Liu, Y. Wu, W. Liang, Y. Cao, and M. Li, "High resolution SAR image classification using global-local network structure based on vision transformer and CNN," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[44] J. Cai, Y. Zhang, J. Guo, X. Zhao, J. Lv, and Y. Hu, "ST-PN: a spatial transformed prototypical network for few-shot SAR image classification," *Remote Sensing*, vol. 14, no. 9, p. 2019, 2022.

[45] C. Wang, Y. Huang, X. Liu, J. Pei, Y. Zhang, and J. Yang, "Global in local: A convolutional transformer for SAR ATR FSL," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[46] Z.-A. Yang, N.-R. Zheng, and F. Wang, "SAR image classification by combining transformer and convolutional neural networks," in *Proceedings of the 8th China High Resolution Earth Observation Conference (CHREOC 2022) High Resolution Earth Observation: Wide Horizon, High Accuracy*. Springer, 2022, pp. 193–200.

[47] A. B. Ramathilagam, S. Natarajan, and A. Kumar, "TransCropNet: a multichannel transformer with feature-level fusion for crop classification in agricultural smallholdings using Sentinel images," *Journal of Applied Remote Sensing*, vol. 17, no. 2, p. 024501, 2023.

[48] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3829–3845, 2020.

[49] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Self-supervised vision transformers for joint SAR-optical representation learning," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 139–142.

[50] K. Li, W. Zhao, R. Peng, and T. Ye, "Multi-branch self-learning Vision Transformer (MSViT) for crop type mapping with optical-SAR time-series," *Computers and Electronics in Agriculture*, vol. 203, p. 107497, 2022.

[51] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled cnn and transformer network for crop segmentation of remote sensing images," *Remote Sensing*, vol. 14, no. 9, p. 1956, 2022.
[52] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.
[53] J.-S. Lee and E. Pottier, *Polarimetric radar imaging: from basics to applications*. CRC press, 2009.
[54] M. Jafari, Y. Maghsoudi, and M. J. V. Zoej, "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*, vol. 8, no. 7, pp. 3595–3607, 2015.
[55] S. Cloude, *Polarisation: applications in remote sensing*. OUP Oxford, 2009.
[56] R. K. Raney, "Hybrid-Polarity SAR architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 11, pp. 3397–3404, 2007.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[58] A. Bahri, S. G. Majelan, S. Mohammadi, M. Noori, and K. Mohammadi, "Remote sensing image classification via improved cross-entropy loss and transfer learning strategy based on deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1087–1091, 2019.