remote sensing

MDPI

Article

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

Saeid Taleghanidoozdoozan [1,*,†], Linlin Xu [2] and David A. Clausi [1]

[1] Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada; dclausi@uwaterloo.ca

[2] Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada; lincoln.xu@ucalgary.ca

[*] Correspondence: stalegha@uwaterloo.ca

[†] Current address: University of Waterloo, East Campus 4, 295 Phillip St, Waterloo, ON N2L 3W8, Canada.

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-attention 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 (3 m) 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.

Keywords: RADARSAT Constellation Mission (RCM); synthetic aperture radar (SAR); compact polarimetry; attention; contextual information; feature learning; deep learning

check for updates

Academic Editor: Silvia Liberata Ullo

Received: 4 December 2024
Revised: 10 January 2025
Accepted: 16 January 2025
Published: 22 January 2025

Citation: Taleghanidoozdoozan, S.; Xu, L.; Clausi, D.A. Pyramid Fine and Coarse Attentions for Land Cover Classification from Compact Polarimetric SAR Imagery. *Remote Sens.* **2025**, *17*, 367. https://doi.org/10.3390/rs17030367

1. 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 the 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 [10–13]. 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 [14]. In contrast, vision transformer models are capable of capturing long-distance dependencies [15]. As an example, the vision transformer (ViT) [16] utilizes the idea of self-attention [17] to enable global receptive field processing of non-overlapping patches.

Despite successful performance on various computer vision tasks [18], 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 [19]. To enhance the accuracy and efficiency of ViT in different tasks, several transformer architectures have been introduced [19–22]. These approaches are global-based, such as the pyramid vision transformer (PVT) [19], or local-based, such as the Swin transformer [20]. 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 [21]. 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 [21].

Successful classification has been demonstrated by both global self-attention methods [19,21] and local self-attention methods [20,23]. However, these approaches impose limitations on the original full self-attention ability to concurrently capture short- and long-range dependencies [18]. Land cover exhibits high spatial heterogeneity [24]; 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 [18] is designed to integrate fine-grained and different scale coarse-grained spatial dependencies, but it requires a highly complex architecture with accompanying high computing requirements to accomplish this task.

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 [25].

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 the PFC transformer (pyramid of fine- and coarse-grained attention transformers), 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 are 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. To establish a robust comparison, we selected baseline methods that represent both local self-attention (e.g., Swin transformer) and global self-attention (e.g., PVT) approaches, as well as hybrid models such as Twins transformer [21] and SepViT [22], which aim to balance computational efficiency and feature extraction capability. The Focal transformer was included for its unique approach to integrating fine-grained and coarse-grained spatial dependencies, making it particularly relevant to the objectives of this study. Additionally, we included the CAT model [26] for its ability to aggregate context features and ResCNN [27] for its robust residual learning framework. These methods were chosen because they exemplify state-of-the-art advancements in both convolution-based and transformer-based architectures, offering diverse perspectives for benchmarking. Section 2 provides a literature review of land cover classification methods utilizing SAR data. Then, the fundamentals of CP SAR data are explained in Section 3. Section 4 describes the proposed method, and the study area as well as datasets are introduced in Section 5. Section 6 presents and analyzes the experimental results, and Section 7 provides the conclusions of this study.

2. Background

2.1. Land Cover Classification Using CP SAR Data

Most of the existing land cover classification methods using SAR data are based on QP (Quad-polarized) or DP. There are only a few known published papers on land cover classification using CP SAR data [28–30]. Robertson et al. [28] 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. [29] 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 were reported without a classified land cover map, so visual evaluation was not possible. Ghanbari et al. [30] introduced a region-based semi-supervised graph network land cover classification for 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.

The reliance on hand-crafted features across these methods limits their adaptability to other CP SAR datasets. Furthermore, the inability to effectively incorporate spatial and contextual information exacerbates the issue of signature ambiguity in CP SAR data, which is inherent in CP SAR data. These challenges necessitate designing a feature learning-based land cover classification method that minimizes reliance on hand-crafted features while effectively addressing the inherent limitations of CP SAR data.

2.2. Land Cover Classification Using CNNs

CNNs are widely used to generate SAR land cover maps [31]. Zhou et al. [32] 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 [3,5,33–37]. For example, Zhang et al. [36] 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. [37] 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 loss of useful spatial resolution detail [38]. These challenges are particularly pronounced in CP SAR data, where the reduced polarimetric information compared to QP data further exacerbates the difficulty of preserving fine-grained spatial details and distinguishing between land cover classes with similar polarimetric signatures. Moreover, despite their proficiency in organizing local features, CNNs encounter challenges in capturing spatial dependencies that extend over long distances [15,39] which limits their ability to fully leverage the contextual information necessary for accurate classification in complex CP SAR scenes.

In several recent studies [4,9,40–43], 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. 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. [41] 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. Li et al. [44] suggested the utilization of an FCN with a sliding window technique to alleviate the computational burden and minimize memory usage. 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.

2.3. Land Cover Classification Using Transformers

Recently, the effectiveness of transformer models in remote sensing applications has captured the attention of remote sensing researchers [31,39,45–50]. While several studies have employed transformer models to merge optical and SAR images and leverage the benefits of both data types [44,51–53], the absence of clear optical images of the same area due to cloud cover impedes their application to CP SAR data. Other studies have combined CNNs and ViT methods to utilize local and global information for land cover mapping [46,48,49]. To integrate the outputs of each branch, various fusion methods have been proposed [48,51,54,55]. 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 attention 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.

3. Compact Polarimetric SAR Basics

The backscattering field of single look complex (SLC) CP SAR data is defined by a 2×1 complex vector E. For the 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) [56]. 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 [57]. The coherency matrix of the RCM CP SAR data are a 2×2 semipositive-definite Hermitian matrix defined as [58,59]:

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

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.

4. Methodology

Figure 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 transformer block.

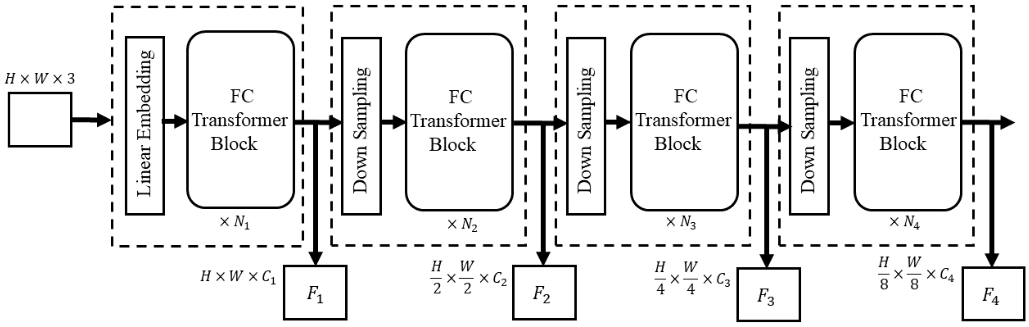


Figure 1. Architecture of the PFC transformer.

4.1. 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 [16,20]. 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 dimensions and C is the feature dimension, the linear embedding projects x_{in} into $z \in R^{H\times W\times C_1}$.

4.2. FC Transformer Block

The main core of the proposed method is the FC transformer block (see Figure 2a). This mechanism captures spatial dependencies at both local-scale and broad-scale. Each part of the block is described separately.

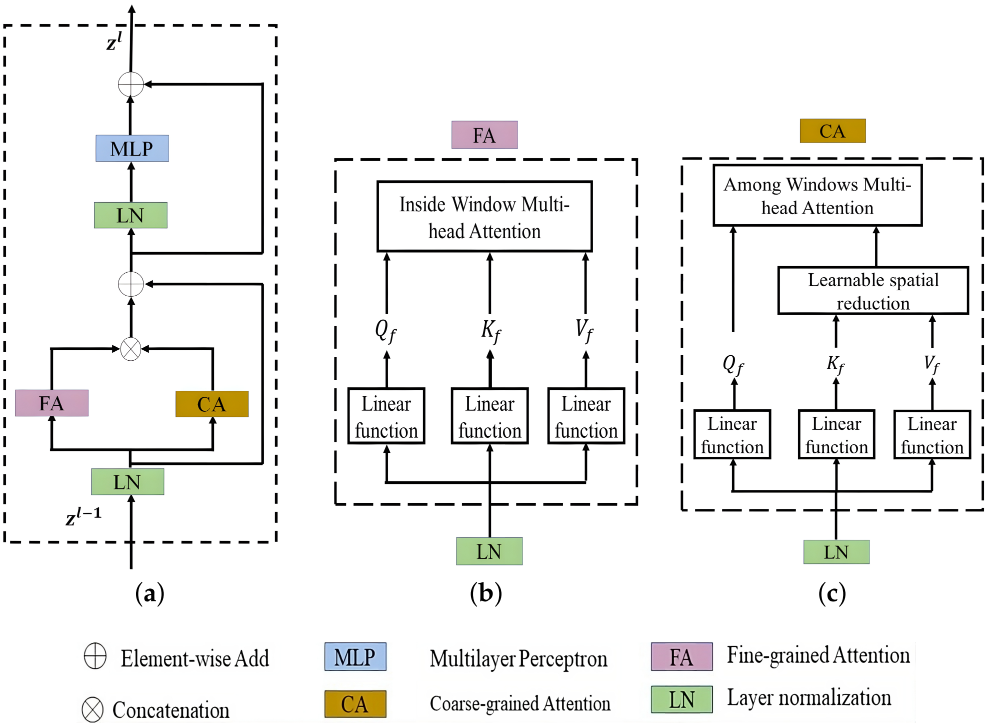


Figure 2. (**a**) Fine- and coarse-grained attention block; (**b**) fine-grained attention; (**c**) coarse-grained attention.

4.2.1. Fine-Grained Attention

Fine-grained attention captures local spatial dependencies by dividing an input feature map into small and non-overlapped windows. Each window is considered as an independent sub-region, and attention is computed within these localized regions.

Given an input feature map $z \in R^{H\times W\times C}$, it 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 [20].

To calculate fine-grained 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. [20], B_f is the learnable relative position bias which its values are taken from $\hat{B}_f \in R^{(2M-1)\times(2M-1)}$. This relative position bias ensures that the model encodes positional relationships explicitly. Figure 2b shows the structure of the F_{attn}.

4.2.2. Coarse-Grained Attention

In addition to the fine-grained attention, the PFC transformer method introduces an approach for calculating coarse-grained attention (C_{attn}). To model broad-scale dependencies, C_{attn} employs a learnable window pooling mechanism. This mechanism aggregates information from multiple windows, enabling the model to capture relationships across the entire feature map. The feature map is downsampled by averaging the features within each window, reducing its spatial dimensions from their original dimensions of $(M \times M) \times d$ to a compact $(1 \times 1) \times d$ 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, as well as coarse-grained K_c and V_c, is 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 sizes 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 [18]. By leveraging this coarse-grained attention mechanism, the model gains the ability to capture long-range spatial dependencies. The structure of C_{attn} is depicted in Figure 2c.

4.2.3. Combining Fine- and Coarse-Grained Attentions

Finally, to take advantage of both fine- and coarse-grained spatial dependencies and utilize them simultaneously, both attentions are concatenated along the channel dimension. 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. 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.

By combining fine-grained and coarse-grained attention, the model leverages local- and broad-scale contextual information simultaneously, leading to a more robust representation of spatial dependencies.

The rest of the FC transformer block is followed by a skip connection with the input feature map, an LN, and a two-layer multi-layer perceptron (MLP) with GELU nonlinearity in-between and again a skip connection, following the same procedure as [16,20]. 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.

4.3. 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.

4.4. Fusion

Unlike previous methods, which only utilized the output of the last stage, our proposed method employs a pyramid of the outputs of FC transformer blocks to aggregate information from all stages, allowing for more comprehensive characterization of land cover types (see Figure 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:

$$F_t = Concat(F_1W_1, F_2W_2, F_3W_3, F_4W_4) \tag{9}$$

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

5. Study Area and Dataset

Due to the limited availability of CP SAR data and the absence of labeled benchmark datasets, we selected two very high-resolution (3m) SLC RCM CP SAR scenes with similar land cover classes. These scenes, with sampled pixel and line spacings of 1.39 m and 2 m, were specifically chosen to ensure sufficient representation of the primary classes while maintaining diversity in their spatial and radiometric characteristics. Each land cover class was manually labeled based on visual inspection of both the SAR data and high-resolution Google Earth images to establish ground truth.

The first scene, captured on 9 August 2022, over Quebec City in Canada, covers approximately 43 km × 13 km and has a size of 10,954 × 8146 pixels. It has an incidence angle range of 47.50 to 48.67 degrees, and its corresponding Google Earth image is presented in Figure 3a. The second scene, acquired on 27 June 2020, has a size of 9344 × 21,942 pixels, covering around 43 km × 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 Figure 4a.

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 as bright as the first group.

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

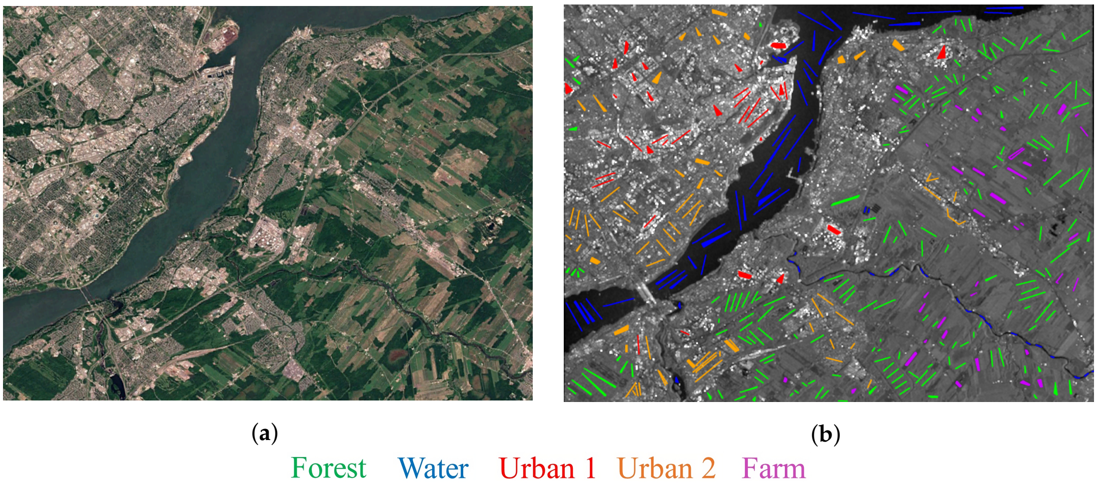


Figure 3. (**a**) Google Earth image of Quebec City scene; (**b**) the first element of CP coherency matrix along with manually selected samples.

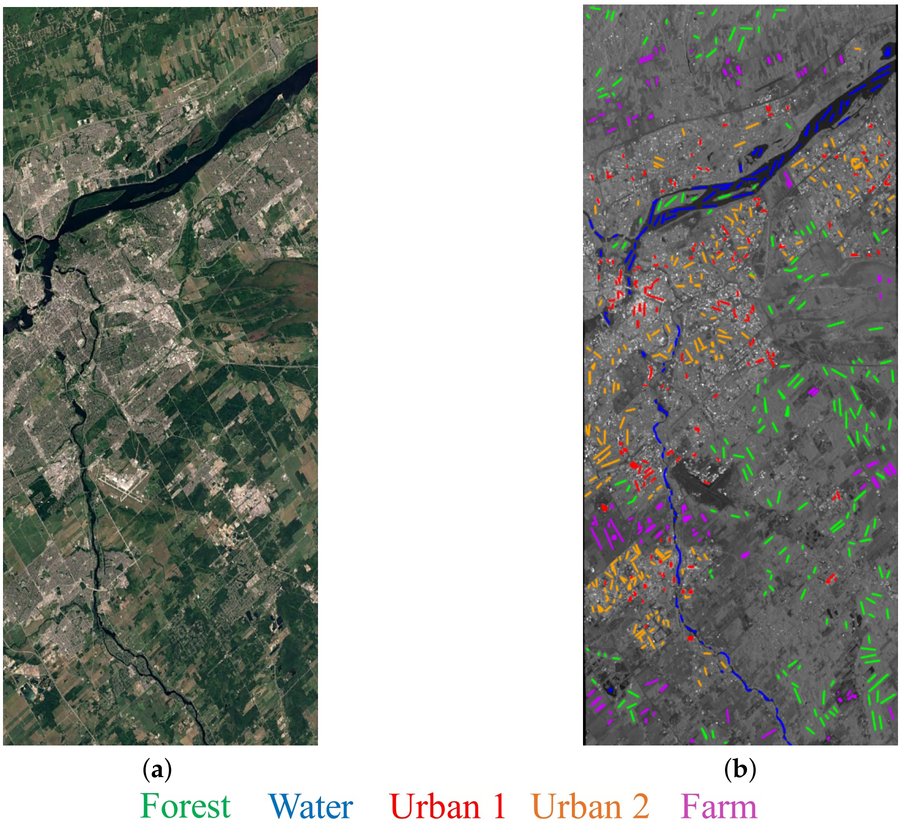


Figure 4. (**a**) Google Earth image of the city of Ottawa; (**b**) the first element of CP coherency matrix along with manually selected samples.

6. 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 contributions of the proposed components, we conducted ablation experiments focusing on key aspects of the model:

- Fine-grained + coarse-grained attention (FC transformer): This configuration evaluates the combination of fine-grained and coarse-grained attention mechanisms without the incorporation of feature fusion.

- Full model (PFC transformer): This configuration includes all three core components, integrating fine-grained attention, coarse-grained attention, and fusion of features from different levels (the pyramid of features).

We note that the effectiveness of using only fine-grained attention, limited to attention within windows (local attention), has already been evaluated in the Swin transformer paper. The Swin authors demonstrated that while fine-grained attention is effective, extending attention across windows further improves model performance. Therefore, the comparison between the FC transformer and the PFC transformer highlights the importance of combining both attention mechanisms, as well as the added value of multi-level feature fusion. These experiments provide a comprehensive understanding of how each component contributes to the overall performance of the model. Table 1 indicates the structure of the proposed method.

Table 1. Detailed architecture of the proposed PFC attention method.

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

It includes four 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 experiments were conducted on a system with a 13th Gen Intel(R) Core(TM) i5-13400 CPU, an NVIDIA GeForce RTX 4060 GPU, and 32 GB of RAM. The implementation code of the proposed method is publicly available at https://github.com/saeidtaleghani23/PFC_Attention.git. Table 2 presents the number of parameters for each model. The parameters used in the proposed method have been set identically in the other models for consistency. All other parameters specific to each model are kept at their default values. As shown in Table 2, the proposed methods (FC and PFC) exhibit a computational cost that is comparable to models such as Swin and SepViT, with a training time of 1.61 h, respectively. Among all models, ResCNN demonstrates the shortest training time (0.67 h), followed by

PVT (0.98 h) and Twins (1.02 h). In contrast, CAT requires the longest training time (7.64 h), indicating a significantly higher computational cost.

Table 2. Comparison of computational cost for models.

Model Name	Number of Parameters (Millions)	Training Time (Hours)	Efficiency Metric (Hours/Million Params)
CAT [26]	1.45 M	7.64	5.26
Focal [18]	0.59 M	2.99	5.06
PVT [19]	0.69 M	0.98	1.42
ResCNN [27]	0.70 M	0.67	0.95
SepViT [22]	0.72 M	1.37	1.91
Twins [21]	0.67 M	1.02	1.52
Swin [20]	0.76 M	1.25	1.64
FC	0.78 M	1.61	2.06
PFC	0.78 M	1.61	2.06

To evaluate the robustness of our model, we utilized metrics such as overall accuracy (OA), f-1 score ($F1$) for each class, average f-1 score ($F1_{avg}$), and kappa coefficient (κ), which account for class imbalances and provide a more detailed performance evaluation. 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 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. Additionally, testing the model on three distinct subregions demonstrated its ability to generalize across variations in data distribution.

6.1. 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 Figure 5. Regions A and B show agricultural, forest and urban areas, while Region C includes forest and agricultural areas.

Table 3 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 [60] with the learning rate, weight decay, and beta parameters set to 1×10^{-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 [61], while the remaining 20% of training samples were used for validation purposes. The weight values of the model that achieved the highest validation accuracy were selected.

Table 3. The number of training and testing pixels for each class selected from the Quebec and Ottawa scenes, respectively. Note that # refers to the number of samples.

Class	# of Train	# of Test
Forest	15,381	11,690
Water	14,853	8093
Urban 1	12,032	11,263
Urban 2	15,098	10,206
Farm	10,773	20,022

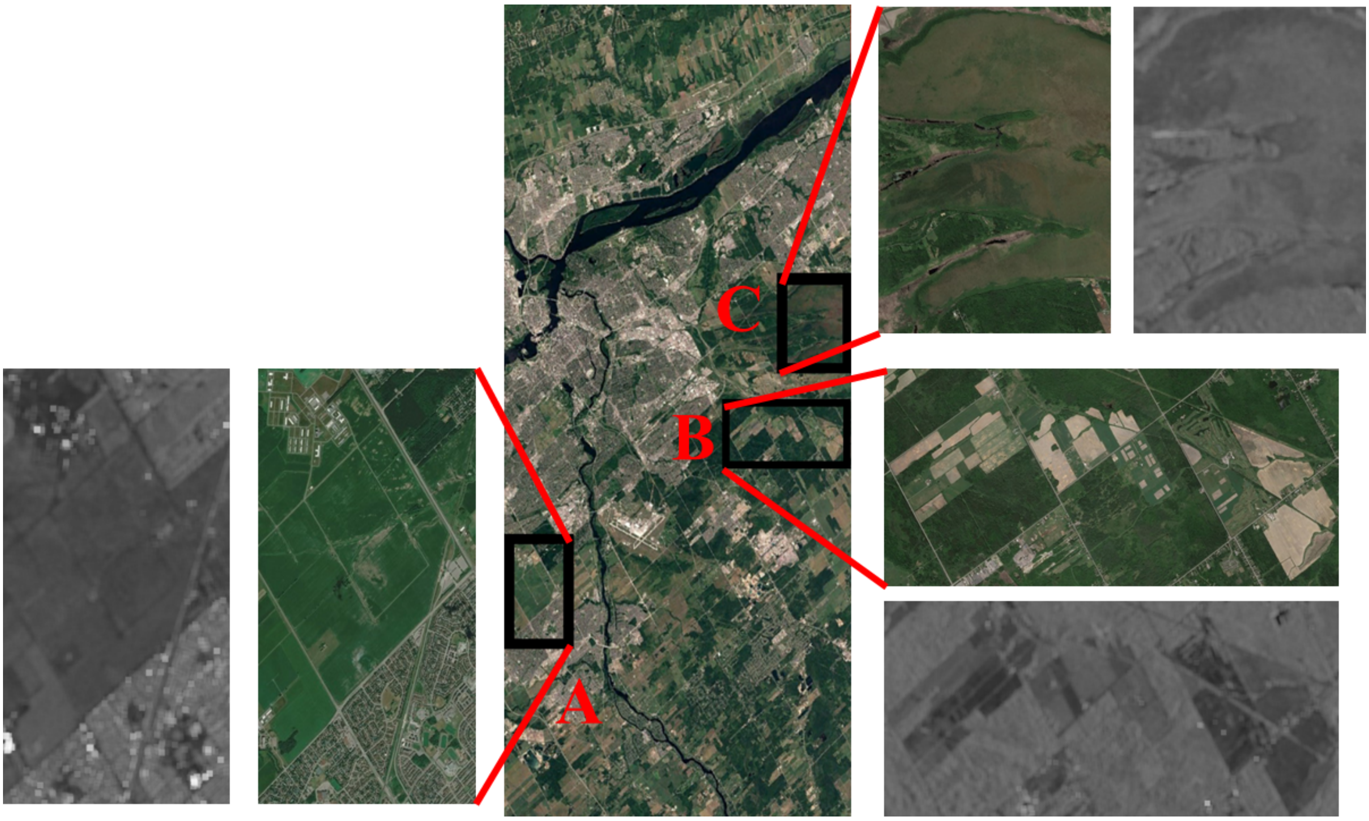
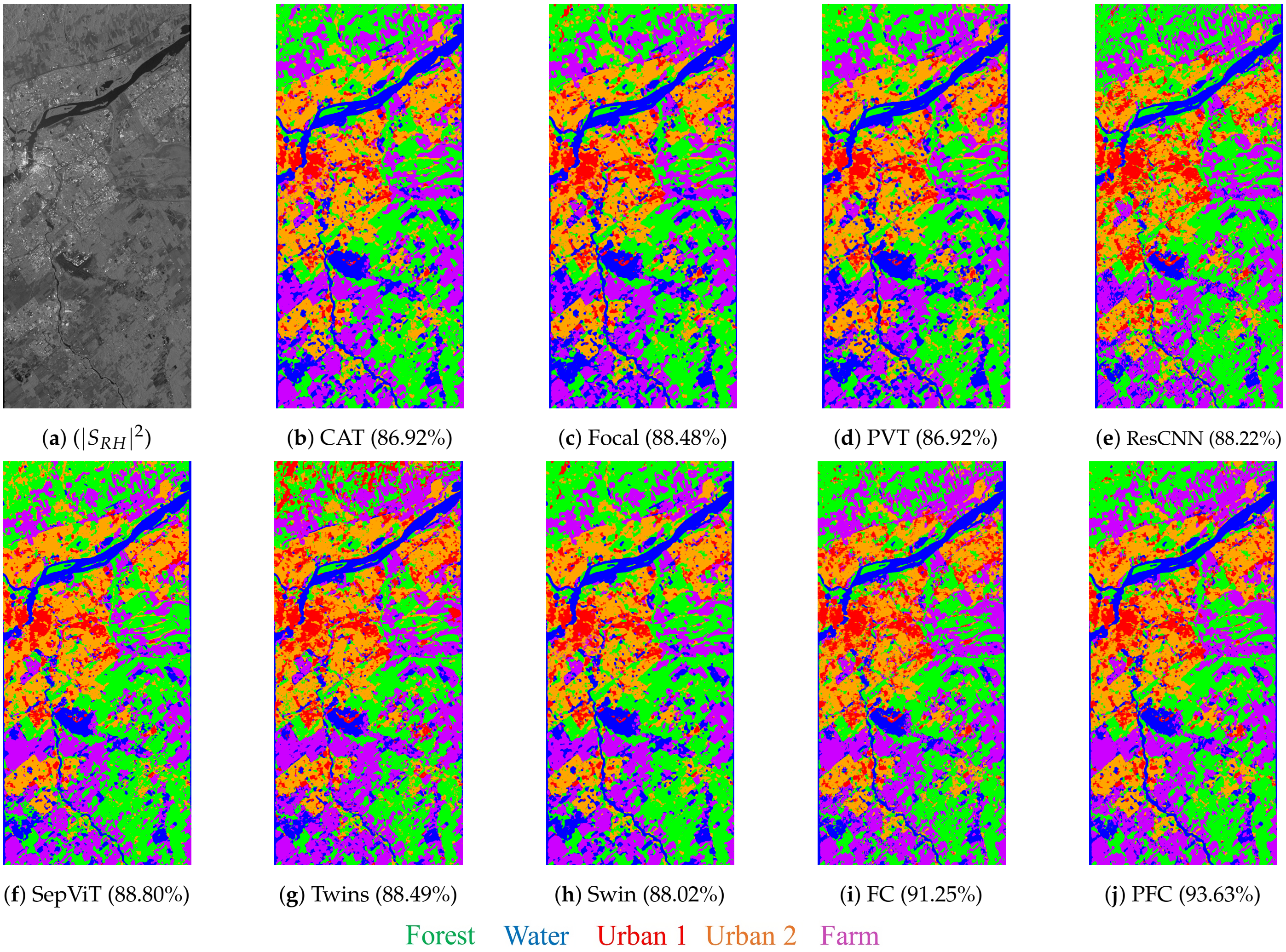


Figure 5. The Google Earth image of the test scene with three regions of interest along with their corresponding $|S_{RH}|^2$. Regions A and B primarily consist of urban, farm and forest classes, while Region C displays both forest and farm classes.

6.2. Results

Figure 6 shows the results obtained by the different methods along with their OA. Due to resizing the images to fit the page, 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 narrower than the one detected by the other and 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. The improvement in performance with the proposed transformer method can be attributed to two key features that enhance its ability to capture spatial and radiometric information in a way that existing methods cannot. One significant factor is the simultaneous utilization of close and far dependencies among pixels, which directs the model's focus to the most relevant areas of the scene, improving its ability to correctly classify challenging regions. By assigning different levels of importance to pixels based on their contextual relationships, the attention mechanism enhances the precision of the model's predictions, particularly in regions with complex spatial patterns such as agricultural areas. Moreover, the proposed method benefits from the learnable fusion of features at multiple levels. This allows the model to combine low- and high-level features, capturing both local details and global context. The flexible and learnable integration of these features results in a more robust scene representation, which enables the model to better differentiate between land cover types with similar spectral characteristics, such as forests and farmlands. This fusion mechanism not only improves the model's overall accuracy but also enhances its ability to make precise predictions, especially in areas where both fine-grained details and broader contextual information are critical for accurate classification.



(**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%)

Figure 6. Part (**a**) shows the $|S_{RH}|^2$ image of the test scene, and parts (**b**–**j**) show the results obtained by each method along with their OA.

As shown in Table 4, 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 SOTA models have limitations that prevent them from achieving the same level of performance as the FC transformer.

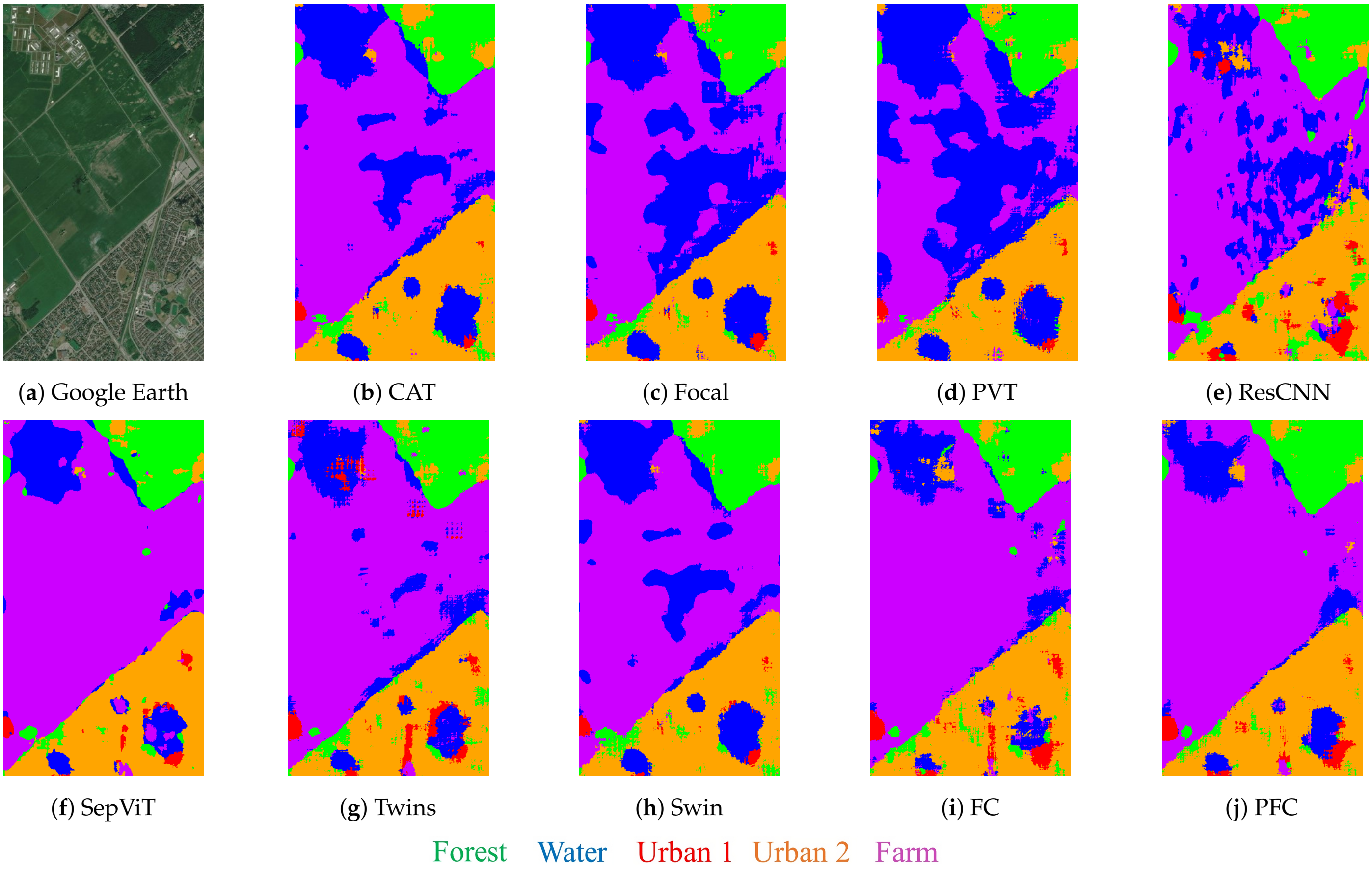
When comparing the performance of the FC and PFC transformer methods, we found that the PFC transformer outperformed the former, with a higher accuracy. By fusion of the different feature levels in a learnable manner, the OA value reached a value that was 2% higher. Moreover, the higher values of κ and $F1_{avg}$ 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].

Table 4. Assessment of the results obtained by the different methods by using overall accuracy (OA), kappa coefficient (κ), average 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	86.92	0.8343	0.8696	0.9116	0.7723	0.8843	0.9234	0.8574
Focal	88.48	0.8544	0.8852	0.9248	0.7812	0.9208	**0.9278**	0.8715
PVT	86.92	0.8351	0.8694	0.9218	0.7596	0.8955	0.9138	0.8561
ResCNN	88.22	0.8500	0.8780	0.8835	0.8246	**0.9215**	0.8820	0.8984
SepViT	88.80	0.8579	0.8850	0.9049	0.8219	0.9046	0.8971	0.8970
Twins	88.49	0.8538	0.8788	0.9181	0.8153	0.8531	0.8931	0.9144
Swin	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**

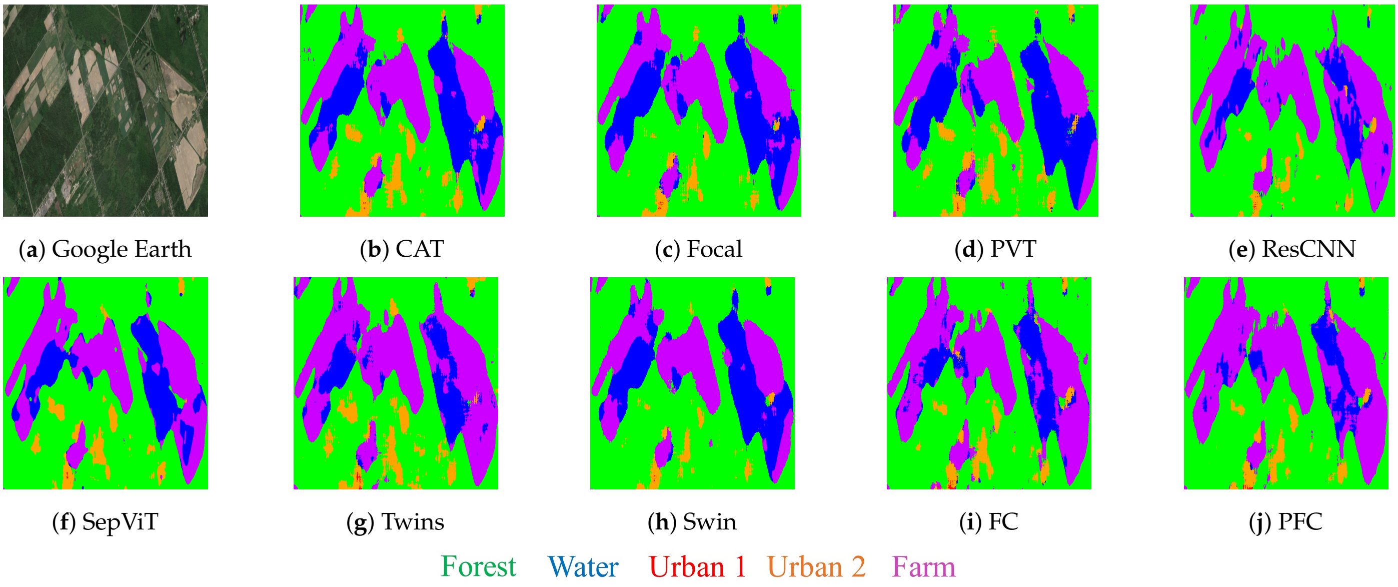
The FC and PFC transformers yield higher $F1$ for the forest, water, and farm classes than the SOTA methods, demonstrating the significance of fine- and coarse-grained dependencies among pixels and the benefits of utilizing features at different levels. The ResCNN and Focal methods achieved slightly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

Figure 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 Figure 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. Figure 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.


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

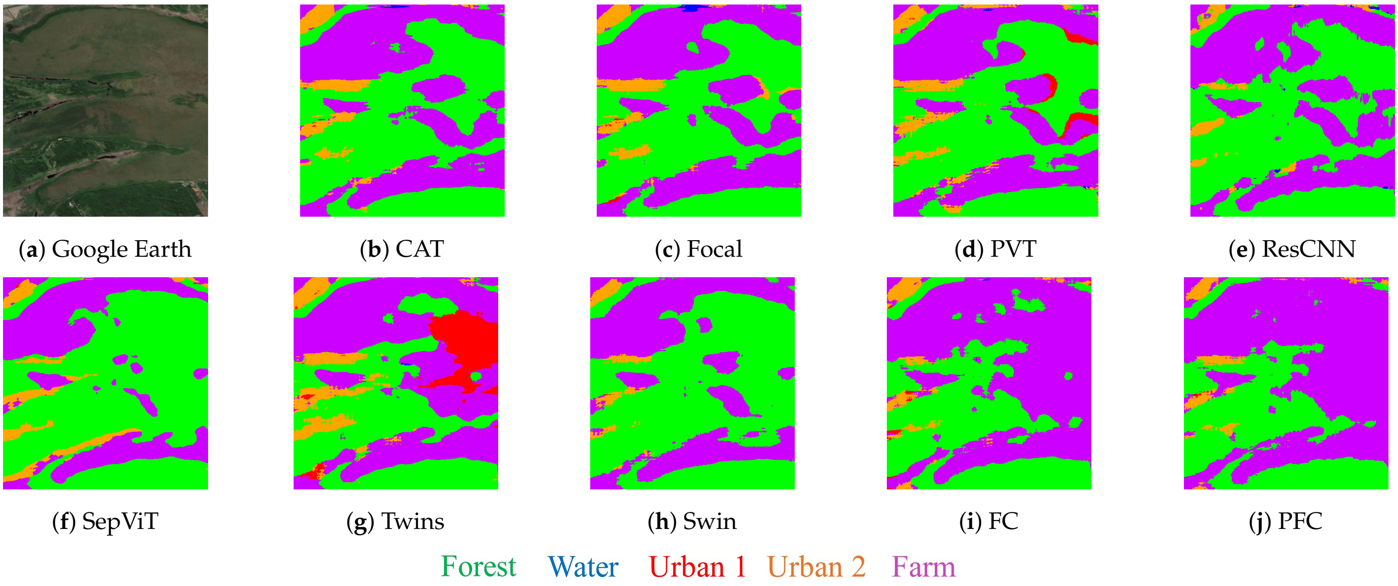
(**f**) SepViT (**g**) Twins (**h**) Swin (**i**) FC (**j**) PFC

Figure 7. Part (**a**) shows the Google Earth image of Region A including urban, farm, and forest classes. Parts (**b**–**j**) show the results obtained by each method.


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

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

Figure 8. Part (**a**) displays a Google Earth image of Region B, which includes agricultural lands, forests, and a few buildings. Parts (**b**–**j**) show the results obtained by each method.


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

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

Figure 9. Part (**a**) shows the Google Earth image of Region C including forest and farm classes. Parts (**b**–**j**) show the results obtained by each method.

7. Conclusions

A new transformer approach was introduced in this paper for generating land cover maps using high-resolution CP SAR 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 purposes by increasing the availability of RCM CP SAR scenes and ground truth samples in the future.

Author Contributions: Conceptualization, S.T. and L.X.; methodology, S.T. and L.X.; software, S.T.; validation, S.T.; formal analysis, S.T.; investigation, S.T.; resources, S.T.; data curation, S.T.; writing—original draft preparation, S.T.; writing—review and editing, L.X. and D.A.C.; visualization, S.T.; supervision, L.X. and D.A.C.; project administration, D.A.C.; funding acquisition, D.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant DGDND-2017-00078, Grant RGPAS2017-50794 and Grant RGPIN-2019-06744.

Data Availability Statement: Data available in a publicly accessible repository: The original data presented in the study are openly available on the EODMS website (https://www.eodms-sgdot.nrcan-rncan.gc.ca/index-en.html) for users with a vetted account.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dabboor, M.; Iris, S.; Singhroy, V. The RADARSAT Constellation Mission in Support of Environmental Applications. *Proceedings* **2018**, *2*, 323. [CrossRef]
2. Qi, Z.; Yeh, A.G.-O.; Li, X.; Lin, Z. A Novel Algorithm for Land Use and Land Cover Classification Using RADARSAT-2 Polarimetric SAR Data. *Remote Sens. Environ.* **2012**, *118*, 21–39. [CrossRef]
3. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Multimodal Bilinear Fusion Network with Second-Order Attention-Based Channel Selection for Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1011–1026. [CrossRef]
4. Wang, Y.; He, C.; Liu, X.; Liao, M. A Hierarchical Fully Convolutional Network Integrated with Sparse and Low-Rank Subspace Representations for PolSAR Imagery Classification. *Remote Sens.* **2018**, *10*, 342. [CrossRef]
5. Liu, X.; He, C.; Zhang, Q.; Liao, M. Statistical Convolutional Neural Network for Land-Cover Classification from SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1548–1552. [CrossRef]
6. Mohammadimanesh, F.; Salehi, B.; Mahdianpari, M.; Brisco, B.; Gill, E. Full and Simulated Compact Polarimetry SAR Responses to Canadian Wetlands: Separability Analysis and Classification. *Remote Sens.* **2019**, *11*, 516. [CrossRef]
7. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
8. Song, W.; Li, M.; Gao, W.; Huang, D.; Ma, Z.; Liotta, A.; Perra, C. Automatic Sea-Ice Classification of SAR Images Based on Spatial and Temporal Features Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9887–9901. [CrossRef]
9. Mohammadimanesh, F.; Salehi, B.; Mahdianpari, M.; Gill, E.; Molinier, M. A New Fully Convolutional Neural Network for Semantic Segmentation of Polarimetric SAR Imagery in Complex Land Cover Ecosystem. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 223–236. [CrossRef]
10. Ma, X.; Fu, A.; Wang, J.; Wang, H.; Yin, B. Hyperspectral Image Classification Based on Deep Deconvolution Network with Skip Architecture. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4781–4791. [CrossRef]
11. Taleghanidoozdoozan, S.; Xu, L.; Clausi, D.A. Region-Based Sea Ice Mapping Using Compact Polarimetric Synthetic Aperture Radar Imagery with Learned Features and Contextual Information. *Remote Sens.* **2023**, *15*, 3199. [CrossRef]
12. Taleghanidoozdoozan, S.; Xu, L.; Clausi, D.A. Sea Ice Mapping from Compact Polarimetric SAR Imagery Using Contextual Information and Learned Features. *J. Comput. Vis. Imaging Syst.* **2023**, *8*, 64–66.
13. Stokholm, A.; Buus-Hinkler, J.; Wulf, T.; Korosov, A.; Saldo, R.; Pedersen, L.T.; Arthurs, D.; Dragan, I.; Modica, I.; Pedro, J.; et al. The AutoICE Challenge. *Cryosphere* **2024**, *18*, 3471–3494. [CrossRef]
14. Ghimire, B.; Rogan, J.; Miller, J. Contextual Land-Cover Classification: Incorporating Spatial Dependence in Land-Cover Classification Models Using Random Forests and the Getis Statistic. *Remote Sens. Lett.* **2010**, *1*, 45–54. [CrossRef]
15. Peng, Z.; Guo, Z.; Huang, W.; Wang, Y.; Xie, L.; Jiao, J.; Tian, Q.; Ye, Q. Conformer: Local Features Coupling Global Representations for Recognition and Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9454–9468. [CrossRef] [PubMed]
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, arXiv: 1706.03762.
18. Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; Gao, J. Focal Self-Attention for Local-Global Interactions in Vision Transformers. *arXiv* **2021**, arXiv:2107.00641.
19. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
21. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9355–9366.
22. Li, W.; Wang, X.; Xia, X.; Wu, J.; Xiao, X.; Zheng, M.; Wen, S. Sepvit: Separable Vision Transformer. *arXiv* **2022**, arXiv:2203.15380

23. Zhang, P.; Dai, X.; Yang, J.; Xiao, B.; Yuan, L.; Zhang, L.; Gao, J. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
24. Xing, H.; Zhu, L.; Feng, Y.; Wang, W.; Hou, D.; Meng, F.; Ni, Y. An Adaptive Change Threshold Selection Method Based on Land Cover Posterior Probability and Spatial Neighborhood Information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 11608–11621. [CrossRef]
25. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
26. Lin, H.; Cheng, X.; Wu, X.; Shen, D. CAT: Cross Attention in Vision Transformer. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 21–26 July 2016.
28. Robertson, L.D.; McNairn, H.; McNairn, C.; Ihuoma, S.; Jiao, X. Compact Polarimetry for Operational Crop Inventory. In Proceedings of the IGARSS IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022.
29. Roy, S.; Das, A.; Omkar, S.N. 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. *Int. J. Remote Sens.* **2023**, *44*, 1510–1532. [CrossRef]
30. Ghanbari, M.; Xu, L.; Clausi, D.A. Local and Global Spatial Information for Land Cover Semi-Supervised Classification of Complex Polarimetric SAR Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**,*20*, 4004205.
31. Wang, H.; Xing, C.; Yin, J.; Yang, J. Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer. *Remote Sens.* **2022**, *14*, 4656. [CrossRef]
32. Zhou, Y.; Wang, H.; Xu, F.; Jin, Y.-Q. Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1935–1939. [CrossRef]
33. Chen, S.-W.; Tao, C.-S. PolSAR Image Classification Using Polarimetric-Feature-Driven Deep Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 627–631. [CrossRef]
34. Yang, C.; Hou, B.; Ren, B.; Hu, Y.; Jiao, L. CNN-Based Polarimetric Decomposition Feature Selection for PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8796–8812. [CrossRef]
35. Xie, W.; Ma, G.; Zhao, F.; Liu, H.; Zhang, L. PolSAR Image Classification via a Novel Semi-Supervised Recurrent Complex-Valued Convolution Neural Network. *Neurocomputing* **2020**, *388*, 255–268. [CrossRef]
36. Zhang, Z.; Wang, H.; Xu, F.; Jin, Y. Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 7177–7188. [CrossRef]
37. Dong, H.; Zhang, L.; Zou, B. PolSAR Image Classification with Lightweight 3D Convolutional Networks. *Remote. Sens.* **2020**, *12*, 396. [CrossRef]
38. Zhang, C.; Pan, X.; Li, H.; Gardiner, A.; Sargent, I.; Hare, J.; Atkinson, P.M. A Hybrid MLP-CNN Classifier for Very Fine Resolution Remotely Sensed Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 133–144. [CrossRef]
39. Dong, H.; Zhang, L.; Zou, B. Exploring Vision Transformers for Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–15. [CrossRef]
40. Wu, W.; Li, H.; Li, X.; Guo, H.; Zhang, L. PolSAR Image Semantic Segmentation Based on Deep Transfer Learning—Realizing Smooth Classification with Small Training Sets. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *16*, 977–981. [CrossRef]
41. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images with Deep Fully Convolutional Neural Networks. *IEEE Geosci. Remote. Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
42. Mullissa, A.G.; Persello, C.; Tolpekin, V. Fully Convolutional Networks for Multi-Temporal SAR Image Classification. In Proceedings of the IGARSS IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018.
43. Mullissa, A.G.; Persello, C.; Stein, A. PolSARNet: A Deep Fully Convolutional Network for Polarimetric SAR Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 5300–5309. [CrossRef]
44. Li, Y.; Chen, Y.; Liu, G.; Jiao, L. A Novel Deep Fully Convolutional Network for PolSAR Image Classification. *Remote Sens.* **2018**, *10*, 1984. [CrossRef]
45. Jamali, A.; Roy, S.K.; Bhattacharya, A.; Ghamisi, P. Local Window Attention Transformer for Polarimetric SAR Image Classification. *IEee Geosci. Remote. Sens. Lett.* **2023**. [CrossRef]
46. Liu, X.; Wu, Y.; Liang, W.; Cao, Y.; Li, M. High Resolution SAR Image Classification Using Global-Local Network Structure Based on Vision Transformer and CNN. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 4505405. [CrossRef]
47. Cai, J.; Zhang, Y.; Guo, J.; Zhao, X.; Lv, J.; Hu, Y. ST-PN: A Spatial Transformed Prototypical Network for Few-Shot SAR Image Classification. *Remote Sens.* **2022**, *14*, 2019. [CrossRef]
48. Wang, C.; Huang, Y.; Liu, X.; Pei, J.; Zhang, Y.; Yang, J. Global in Local: A Convolutional Transformer for SAR ATR FSL. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]

49. Yang, Z.-A.; Zheng, N.-R.; Wang, F. 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 Accurac, Singapore, 30 November 2022.
50. Ramathilagam, A.B.; Natarajan, S.; Kumar, A. TransCropNet: A Multichannel Transformer with Feature-Level Fusion for Crop Classification in Agricultural Smallholdings Using Sentinel Images. *J. Appl. Remote Sens.* **2023**, *17*, 024501. [CrossRef]
51. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Collaborative Attention-Based Heterogeneous Gated Fusion Network for Land Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3829–3845. [CrossRef]
52. Wang, Y.; Albrecht, C.M.; Zhu, X. Self-Supervised Vision Transformers for Joint SAR-Optical Representation Learning. In Proceedings of the IGARSS IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022.
53. Li, K.; Zhao, W.; Peng, R.; Ye, T. Multi-Branch Self-Learning Vision Transformer (MSViT) for Crop Type Mapping with Optical-SAR Time-Series. *Comput. Electron. Agric.* **2022**, *203*, 107497. [CrossRef]
54. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
55. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI, 24th International Conference, Strasbourg, France, 27 September–1 October 1 2021.
56. Lee, J.-S.; Pottier, E. *Polarimetric Radar Imaging: From Basics to Applications*; CRC Press: Boca Raton, FL, USA, 2009.
57. Jafari, M.; Maghsoudi, Y.; Zoej, M.J.V. A New Method for Land Cover Characterization and Classification of Polarimetric SAR Data Using Polarimetric Signatures. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3595–3607. [CrossRef]
58. Cloude, S. *Polarisation: Applications in Remote Sensing*; OUP Oxford: Oxford, UK, 2009.
59. Raney, R.K. Hybrid-Polarity SAR Architecture. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3397–3404. [CrossRef]
60. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.
61. Bahri, A.; Majelan, S.G.; Mohammadi, S.; Noori, M.; Mohammadi, K. Remote Sensing Image Classification via Improved Cross-Entropy Loss and Transfer Learning Strategy Based on Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1087–1091. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.