



Cross-domain PolSAR terrain classification using a lightweight dual-stream vision transformer

A. Rega & E. Fantin Irudaya Raj

To cite this article: A. Rega & E. Fantin Irudaya Raj (2025) Cross-domain PolSAR terrain classification using a lightweight dual-stream vision transformer, International Journal of Remote Sensing, 46:22, 8640-8674, DOI: [10.1080/01431161.2025.2571234](https://doi.org/10.1080/01431161.2025.2571234)

To link to this article: <https://doi.org/10.1080/01431161.2025.2571234>



Published online: 14 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 51



View related articles [↗](#)



View Crossmark data [↗](#)



Cross-domain PolSAR terrain classification using a lightweight dual-stream vision transformer

A. Rega ^a and E. Fantin Irudaya Raj ^b

^aDepartment of Electronics and Communication Engineering, Jayaraj Annapackiam CSI College of Engineering, Nazareth, Thoothukudi, India; ^bDepartment of Electrical and Electronics Engineering, Dr. Sivanthi Aditanar College of Engineering Tiruchendur, Thoothukudi, India

ABSTRACT

Accurate terrain classification using polarimetric synthetic aperture radar (PolSAR) imagery is important for several remote sensing applications. However, conventional models struggle to generalize across domains due to variations in sensor type, acquisition conditions, and geographic context. In this work, we propose a dual-stream vision transformer framework for unsupervised domain adaptation in PolSAR terrain classification. The architecture combines a SimPool-based global attention stream with a ResMLP local stream, enabling robust modelling of both global semantic context and local spectral – spatial structures. The two streams are fused via element-wise integration, and the model is trained using labelled source data without requiring target domain labels. The framework is evaluated on four benchmark PolSAR datasets over ten units of domain adaptation consisting of sensor, region, and combined shifts. The proposed model achieves better performance than state-of-the-art models consistently. Further, ablation and per class analyses further confirm its effectiveness and its capability to generalize. We establish a new state-of-the-art for domain adaptive PolSAR terrain classification and demonstrate the advantages of combining global and local modelling streams within a unified transformer architecture.

ARTICLE HISTORY

Received 23 June 2025

Accepted 30 September 2025

KEYWORDS

PolSAR classification; unsupervised domain adaptation; vision transformer; dual-stream architecture; cross-domain learning; remote sensing; terrain segmentation

1. Introduction

PolSAR has emerged as a pivotal modality in remote sensing, offering rich backscatter information that is independent of lighting and weather conditions. This makes PolSAR highly effective for earth observation tasks, including terrain classification, environmental monitoring, and urban mapping. The encoded scattering mechanisms – surface, double-bounce, and volume scattering – enable fine-grained discrimination among land cover types such as vegetation, water, and urban structures (Parikh, Patel, and Patel 2020).

Despite these advantages, accurate terrain classification using PolSAR remains challenging due to several inherent complexities. Firstly, PolSAR data is complex-valued and multi-channel, making it difficult to process using conventional feature extraction techniques. Secondly, variations in radar incidence angle, polarization orientation, and sensor

configurations introduce significant diversity in scattering behaviour across scenes. Thirdly, the limited availability of labelled data – particularly in newly observed or geographically distinct regions – further complicates model training and generalization. Several early efforts tackled these challenges using statistical models, feature-based segmentation, or filter-based preprocessing. For example, Gabor filters, wavelet transforms, and expectation-maximization clustering have been widely used to extract spatial and texture features in segmentation tasks (Akbarizadeh and Rahmani 2015, 2017; Tirandaz, Akbarizadeh, and Kaabi 2020). Although effective at low-level representation, these methods may lack generalizability under cross-domain conditions and require handcrafted parameter tuning.

Deep learning methods have shown promise by automating feature extraction, but they also suffer from limitations when applied to PolSAR data. CNN-based architectures struggle with overfitting in few-shot settings, catastrophic forgetting, and domain shift due to their locality bias and dependence on large annotated datasets (N. Wang et al. 2024). To overcome these limitations, recent works have explored meta-learning, multi-modal fusion, and attention-based models with promising results. For instance, Fang et al. (2024) introduced a polarization orientation angle-aligned transformer with cyclic channel attention to align polarization features across spatial domains.

However, most existing methods fall short in two critical aspects. On the one hand, they struggle to jointly capture global semantic context alongside local spatial – spectral structures essential for interpreting complex PolSAR data. On the other hand, they often fail to achieve robust domain adaptation across varying sensor types and geographic regions, particularly when labelled data in the target domain is unavailable. Furthermore, standard vision transformers (ViTs) are often computationally heavy and data inefficient, limiting their utility for PolSAR data with limited annotations and small datasets (Y. Wang et al. 2025). In this context, there is a pressing need for a lightweight, modular, and generalizable architecture that can bridge this gap.

1.1. Motivation

Accurate classification of PolSAR imagery across different domains remains a significant and unresolved challenge in remote sensing. Although deep learning models, particularly CNNs, have achieved notable success in image-based classification tasks, their performance on PolSAR data is often hindered by the inherent domain sensitivity of learned features (N. Wang et al. 2024). However, in most cases, training and testing data for PolSAR applications have drastically different distributions, which creates significant domain shifts; the models commonly assume that the training and testing data distributions are similar, which can rarely be achieved in real-world PolSAR applications. This issue is especially critical in the context of unsupervised domain adaptation, where the model is trained on a labelled source domain but must generalize to an unlabelled target domain. Despite the success of existing Unsupervised Domain Adaptation (UDA) methods in aligning marginal and conditional distributions between domains, single stream architectures have limitations in capturing global semantic context and localized spatial – spectral structure required for performing effective PolSAR interpretation (Ren et al. 2024).

Additionally, standard ViT models, while powerful in capturing long-range dependencies, tend to suffer from data inefficiency and high computational demands, particularly

when applied to smaller remote sensing datasets. However, purely multilayer perceptrons (MLP) based models like ResMLP are not capable of modelling global relationship (Takahashi et al. 2024). On the other hand, object classification in PolSAR images is particularly challenging due to the presence of speckle noise and complex scattering mechanisms, which often obscure class boundaries and reduce feature separability. While MLPs can capture non-linear relationships, they lack the inductive bias required to effectively model long-range dependencies and are therefore prone to overfitting and noise sensitivity in high-dimensional PolSAR data. To address these limitations, we adopt an ensemble-inspired dual-stream strategy. Instead of relying solely on either MLPs or attention-based mechanisms, the proposed framework integrates their complementary strengths. Although the ensemble design may appear conceptually simple, its novelty lies in the synergistic fusion of global and local streams, ensuring robustness against noise and improved generalization under domain shifts.

1.2. Contributions

To address the challenges of domain-adaptive PolSAR terrain classification, we propose a novel dual-stream vision transformer framework that introduces architectural innovations tailored specifically for polarimetric SAR data. The main contributions of this work are:

- We propose SimPool+, an enhanced version of SimPool, which incorporates adaptive kernel selection, multi-scale contextual pooling, and context-aware token fusion. This significantly improves global semantic modelling efficiency while reducing complexity, as validated in our ablation studies.
- We design ResMLP+, which extends ResMLP by integrating squeeze-and-excitation blocks and gated residual modulation, enabling the network to dynamically recalibrate channel importance and control residual flow – thereby improving spatial discrimination across complex terrain classes.
- We introduce a novel element-wise fusion mechanism that combines SimPool+ and ResMLP+ outputs, ensuring complementary global – local feature alignment under domain shifts. This architecture demonstrates superior generalization in both sensor and region shifts.
- The proposed model is evaluated across ten domain adaptation units spanning four benchmark PolSAR datasets, covering both cross-sensor and cross-region shifts. It achieves new state-of-the-art performance, outperforming existing models.

Collectively, these innovations form a lightweight yet highly expressive transformer architecture tailored for robust and data-efficient domain adaptation in PolSAR classification tasks.

The remainder of this paper is organized as follows: [Section 2](#) reviews the related literature on PolSAR terrain classification, vision transformer architectures in remote sensing, and domain adaptation strategies. [Section 3](#) details the proposed dual-stream vision transformer framework, including its architectural components and domain adaptation mechanisms. [Section 4](#) describes the PolSAR datasets used, experimental setup, and evaluation protocol. [Section 4](#) presents the experimental results, including ablation

studies and performance under various domain shifts. Finally, [Section 5](#) concludes the paper and discusses future directions.

2. Related works

2.1. PolSAR terrain classification

PolSAR provides rich backscattering information by recording the amplitude and phase of electromagnetic waves in multiple polarization channels. This allows for Parcefer (frequency, wavelength) discrimination of terrain types (e.g. vegetation and water) according to scattering mechanisms, making PolSAR an effective tool for land cover classification, environmental monitoring, and urban analysis. Nonetheless, the inherent nature of PolSAR data is complex valued and multi-dimensional, which make terrain classification with PolSAR data, especially with varied acquisition condition, very challenging (Zhang et al. 2024). Traditional approaches to PolSAR classification primarily relied on statistical models and physical decomposition techniques. For example, the Wishart classifier, a covariance matrix modelling approach, and the widely used polarimetric decomposition methods have been employed to interpret the scattering behaviour. However, these methods are sensitive to noise and assume prior knowledge that restricts their applicability in different settings (Parida and Mandal 2020).

With the advent of deep learning, particularly CNNs, PolSAR classification has experienced a paradigm shift. In many supervised settings, CNN-based models can automatically learn hierarchical spatial features, which surpasses the traditional methods. However, early CNNs applied directly to polarimetric channels or their decompositions, but were limited in being able to capture long-range dependencies and dealing with label scarcity. In order to exploit the intrinsic structure of PolSAR data, subsequent enhancements such as 3D-CNNs, multi-scale feature extractors, and spectral – spatial fusion networks offered to enhance the properties of a SAR image. Unfortunately, these models can consume large amount of labelled data and are vulnerable to overfitting in domain shifted conditions (Shang et al. 2022).

Beyond terrain classification, a number of works have explored object detection, segmentation, and environmental monitoring in Synthetic Aperture Radar (SAR) and PolSAR imagery using deep learning, highlighting the broader applicability of hybrid and end-to-end models. Recent advancements in object detection and classification using SAR imagery have highlighted the potential of hybrid and deep learning models. For instance, Sharifzadeh, Akbarizadeh, and Seifi Kavian (2019) introduced a CNN – MLP hybrid classifier for ship detection in SAR images, effectively reducing false alarms by combining deep feature extraction with statistical decision models. Similarly, Samadi, Akbarizadeh, and Kaabi (2019) proposed a supervised change detection method based on deep belief networks, leveraging morphological operators to enhance training quality and computational efficiency. These studies demonstrate the growing relevance of deep architectures in robust SAR interpretation under complex conditions.

Deep learning has also played a significant role in semantic segmentation tasks involving environmental monitoring. Aghaei, Akbarizadeh, and Kosarian (2022a) proposed OSDES-Net, a ShuffleNet-based architecture for oil spill detection in SAR images, integrating group and atrous convolutions to enhance feature extraction efficiency.

Davari, Akbarizadeh, and Mashhour (2021) explored power equipment classification and corona defect detection using a GoogleNet – AlexNet pipeline, combining object tracking and intelligent classification in video sequences. In a separate contribution, Aghaei, Akbarizadeh, and Kosarian (2022b) introduced GreyWolfLSM, a level-set-based oil spill detector that fuses clustering, moment-based features, and SVM classification for enhanced performance on noisy SAR data. Ghara, Shokouhi, and Akbarizadeh (2022) further compared CNN architectures like U-Net and DeepLabV3 for oil spill segmentation, concluding that U-Net outperformed DeepLabV3 in both accuracy and robustness when trained on augmented SAR datasets. Together, these studies underscore the diverse applications of deep learning in SAR and PolSAR analysis, ranging from object detection to environmental hazard monitoring. However, most existing methods are tailored to supervised settings or rely heavily on labelled data, which limits their adaptability under domain shift conditions.

Recently, efforts have been made to incorporate attention mechanisms and transformer-based architectures into PolSAR classification. Although these approaches have demonstrated increased capability to model global context, their performance is still limited by high data requirements and lack of architectural adaptation to polarimetric data. Additionally, most of the existing models are trained and tested in a domain-specific manner and do not generalize when used across different sensors or geographic regions (Chen et al. 2025). However, this approach is limited by the difficulty in capturing general contextual relationships or local scattering characteristics, which are inherent in PolSAR data.

2.2. Transformer architectures in remote sensing

Transformer-based models have rapidly gained traction in remote sensing due to their capacity to model long-range dependencies and capture complex spatial relationships across image patches. The vision transformer repurposed the transformer architecture for visual task by flattening images into non-overlapping patches, and processed as token sequence as in multi head self-attention based on the observation that transformers are originally developed for natural language processing. With this shift from convolutional to attention-based modelling, global context aggregation, a property that is highly useful for high-resolution earth observation imagery, is possible (Tao et al. 2025).

In the context of remote sensing, ViTs have shown strong performance in tasks such as land use classification, hyperspectral image analysis, and change detection. Since they can reason over wide spatial extents, they are ideal to apply to places where spatial patterns and semantic relationships span large areas. Yet, standard ViTs are subject to a few limitations that prevent their use in specialized domains such as PolSAR. For example, these have data inefficiency, high computational complexity, and lack of inductive biases such as translation invariance and locality that are naturally present in convolutional architectures (Huang et al. 2025).

To address these limitations, a range of transformer variants have been proposed. Hierarchical designs and local window attention-based lightweight architectures, pyramid vision transformer, and swin transformer are introduced to reduce computational load. There are others, such as MixFormer and Hybrid CNN-Transformer models, that merge convolutional backbones with attention modules to exploit the advantages of both

paradigms (Huo et al. 2025). Although these techniques commemorate optical and hyperspectral data, their fusion to PolSAR, in which the polarimetric structure and spatial content, is not well understood yet. Recent efforts have also investigated simplified transformer alternatives. As an example, SimPool substitutes the MHSA (multi-head self-attention) block with an attention mechanism based on global pooling and matches the performance of the existing models with lower complexity. Likewise, ResMLP architectures remove attention altogether and instead use token mixing MLP layers to capture inter patch dependencies (X. Wang et al. 2025). Although these models have promise in reducing overhead while preserving key global features, they have not been systematically evaluated for PolSAR terrain classification and in domain adaptation settings. Given the inherent challenges of feature discrimination and domain adaptation in PolSAR imagery, such as domain shifts among sensors and regions, as well as the need to balance local features with global context information, there is a strong need for data efficient, domain adapted transformer-based architectures.

2.3. Domain adaptation for remote sensing

In real-world remote sensing applications, models often encounter significant distribution shifts between the training (source) domain and the testing (target) domain. The variations in these shifts may be due to the sensor type, acquisition geometry, atmospheric conditions, or geographic characteristics. To overcome this challenge, domain adaptation techniques attempt to transfer knowledge from a labelled source domain, to an unlabelled or sparsely labelled target one, and hence improve generalization without exhaustive annotation in any new environment (Yan, Han, and Hou 2025).

In the remote sensing community, unsupervised domain adaptation has garnered particular attention due to the high cost and limited availability of labelled data in diverse target domains. Early UDA in remote sensing followed statistical alignment strategies such as aligning the marginal distributions by minimizing maximum mean discrepancy or aligning the conditional distributions by using class-specific metrics. In smooth cases, these methods are effective, but they struggle in high-dimensional remote sensing data where semantic boundaries are not well separated (Yang et al. 2025).

More recent advancements have incorporated deep neural networks with domain adaptation objectives. The Domain-Adversarial Neural Network (DANN) framework brought adversarial learning to encourage feature extraction invariant with respect to the domains. Attention mechanism was integrated into extensions like domain adaptive attention network to focus the adaptation at more transferable regions. In addition, the entropy-based methods (minimum class confusion) and self-training techniques with pseudo label refinement have been successfully applied, in particular to hyperspectral and optical image classification (Lang et al. 2024).

The challenges are more pronounced for PolSAR data, since scattering characteristics are domain sensitive and polarimetric information is complex. CDFNet is one of the most significant contributions in this space as a PolSAR UDA model, introducing the art in PolSAR domain adaptation (S. Wang et al. 2025). Two key innovations included in CDFNet are the Domain balanced sampling (DBS) for dealing with class imbalance and cross-domain feature fusion to align the source and target distribution without introducing noisy pseudo labels. In cross sensor and cross region, CDFNet handled these settings very

well as it outperformed all other methods in traditional UDA and was established as a strong model for further works.

Although promising results have been achieved, most of the existing UDA methods leverage a single stream feature extractor and do not make the best use of the complementary nature of global and local representations, which is crucial to learn both the large-scale scene context and the fine-grained scattering patterns in PolSAR imagery. Furthermore, there are only a few works that utilize transformer-based architectures for domain adaptation, which are naturally capable of learning domain robust features through self-attention and token level modelling (Fu et al. 2024). In order to bridge these gaps, our proposed framework adopts a dual-stream transformer backbone specifically designed for UDA settings.

2.4. Summary and gap analysis

The task of PolSAR terrain classification presents unique challenges due to the complex nature of polarimetric scattering mechanisms and the frequent occurrence of domain shifts across geographic regions and sensor modalities. Early methods were based on physical decompositions or statistical modelling, where recently deep learning methods, most especially deep CNNs, have drastically improved the classification performance. However, their inherent locality and domain sensitivity limit their effectiveness in cross-domain scenarios.

Transformer-based architectures have introduced new possibilities for capturing long-range dependencies in remote sensing imagery. However, their application to PolSAR is limited mainly due to computational complexity, lack of inductive bias, and insufficient adaptation to polarimetric inputs. However, simplified transformers such as SimPool and ResMLP have not been fully explored in PolSAR tasks or domain adaptation cases, and the third option is in fact an efficient alternative.

In parallel, several unsupervised domain adaptation techniques have been developed for remote sensing, ranging from adversarial alignment to entropy minimization. Though suitable in certain circumstances, such methods are usually carried out through single stream architectures and fail to capture the entire spatial and spectral complexity of PolSAR data. CDFNet has become a strong model for PolSAR-specific UDA with novel sampling and feature fusion strategies. However, it is lacking architectural diversity and may not perform well where the class boundaries are fine-grained and there is multi scale spatial variability.

Taken as a whole, these insights reveal a key research gap: existing models fall short in unifying global contextual modelling and local spatial sensitivity within a single, adaptable framework for PolSAR domain adaptation.

3. Proposed framework

3.1. Model overview

The overall workflow of the proposed method is illustrated in Figure 1. The process begins with patch-level tokenization of both labelled source and unlabelled target PolSAR images, where each image is divided into fixed-size patches and linearly projected into

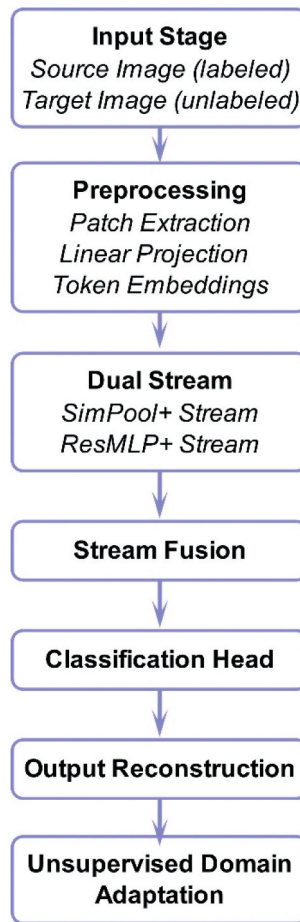


Figure 1. End-to-end workflow of the proposed dual-stream PolSAR classification framework.

high-dimensional embeddings. These embeddings are then processed in parallel through two complementary streams: the SimPool+ stream, which models global semantic relationships using adaptive kernel pooling and context-aware token fusion; and the ResMLP + stream, which captures local spectral – textural patterns through channel-wise MLP layers enhanced with Squeeze-and-Excitation (SE) blocks and gated residual connections. The outputs from both streams are fused and passed to a lightweight classification head to predict terrain classes. This dual-stream design enables effective unsupervised domain adaptation by learning shared representations across domains with different imaging characteristics (Ahmad et al. 2025). Figure 2 shows the architecture of the proposed model.

3.2. Stream 1: SimPool+

To mathematically describe the original SimPool attention mechanism (Psomas et al. 2023), consider an input token matrix $X' \in \mathbb{R}^{p \times d}$, where $p = W \times H$ denotes the number of patches and d is the feature dimension. SimPool computes a global query vector as:

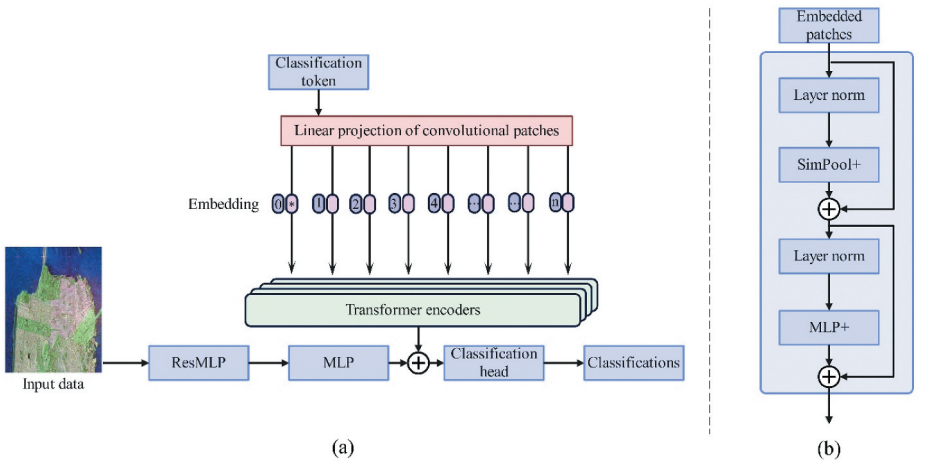


Figure 2. Architecture of the proposed model. (a) structure and components of the vision transformer module, (b) internal configuration of the transformer encoder layer.

$$I_0 = \mu_A(X') = \frac{1}{p} \sum_{i=1}^p x_i \in \mathbb{R}^d. \quad (1)$$

Linear mappings are used to get the query and key projections.

$$q = F_Q I_0, K = F_K X' \in \mathbb{R}^{d \times p} \quad (2)$$

Attention scores $a \in \mathbb{R}^p$ are derived using scaled dot-product attention:

$$a = \text{softmax} \left(\frac{K^T q}{\sqrt{d}} \right). \quad (3)$$

Next, the value matrix V is formed by subtracting the minimum of X :

$$V = X' - \min(X'). \quad (4)$$

A generalized pooling function $f_a(\cdot)$ is applied to adaptively mix-max and average pooling:

$$f_a(x) = \begin{cases} x^{\frac{1-a}{2}} & a \neq 1 \\ \ln(x) & a = 1. \end{cases} \quad (5)$$

Finally, the pooled attention output is computed as:

$$I = \mu_{Sp}(x) = f_a^{-1}(f_a(V) \cdot a). \quad (6)$$

This attention vector $I \in \mathbb{R}^d$ acts as a global descriptor, replacing MHSA and feeding into an MLP for classification.

The proposed SimPool+ module in Stream 1 is designed to replace traditional MHSA with a computationally efficient yet semantically rich attention mechanism. MHSA has quadratic computational complexity $\mathcal{O}(N^2 \cdot d)$, where N is the number of tokens and d is the embedding dimension. In contrast, SimPool+ achieves linear complexity by employing generalized pooling and attention-weighted aggregation. Instead of computing all

pairwise interactions, SimPool+ first reduces the token sequence length through adaptive pooling into a fixed number of representative prototypes M (with $M \ll N$). Attention is then computed only between tokens and these prototypes, resulting in a reduced complexity of $\mathcal{O}(N \cdot d + M \cdot d)$. Since M is constant with respect to input size, the overall complexity scales linearly with the number of input tokens. The proposed SimPool+ module enhances the original SimPool by incorporating adaptive kernel selection, multi-scale pooling, and context-aware token fusion.

3.2.1. Global query initialization

A global context vector is initialized via average pooling across tokens:

$$Q_g = \frac{1}{N} \sum_{i=1}^N x_i. \quad (7)$$

Linear projections produce:

$$Q = XW_Q, K = XW_K, V = XW_V \quad (8)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times d}$ are learnable weight matrices and d is the projected feature dimension.

3.2.2. Adaptive kernel selection

To identify contextually important tokens, we compute attention weights between the global query and all keys:

$$\alpha = \text{softmax}\left(\frac{Q_g K^T}{\sqrt{d}}\right) \quad (9)$$

where $\alpha \in \mathbb{R}^{1 \times N}$ captures token-level relevance and dynamically guides subsequent pooling.

3.2.3. Multi-scale contextual pooling

To preserve both fine and coarse spatial features, SimPool+ applies multi-scale pooling to the value matrix V , using window sizes $s \in \{2, 4, 8\}$. At each scale s , pooled features are computed as:

$$P_s = \beta \cdot \text{maxpool}_s(V) + (1 - \beta) \cdot \text{avgpool}_s(V) \quad (10)$$

where $\beta \in [0, 1]$ is a learnable scalar balancing between max and average pooling.

3.2.4. Context-aware token fusion

Instead of directly combining pooled features, SimPool+ introduces context-aware fusion. A scale-specific gating coefficient $\gamma_s \in [0, 1]$ is computed from the global query using a small MLP:

$$\gamma_s(Q_g) = \text{MLP}_s(Q_g). \quad (11)$$

The final fused representation is:

$$Z = \sum_s \gamma_s(Q_g) \cdot (\alpha \cdot P_s) \quad (12)$$

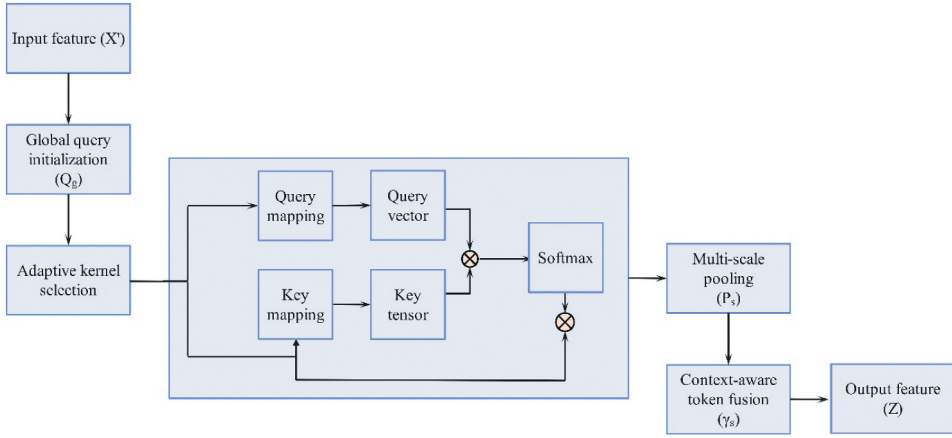


Figure 3. Architecture of the SimPool+ module.

where $\alpha \cdot P_s$ applies attention weighting to pooled features, and the sum adaptively fuses multi-scale outputs based on semantic relevance.

The final output Z replaces standard MHSA and serves as a global descriptor, summarizing semantically meaningful content with reduced computational cost. Figure 3 shows the architecture of the SimPool+ module. This stream captures high-level context efficiently, and its output is passed through an MLP head for classification.

3.3. Stream 2: ResMLP+

The transformer stream uses attention mechanisms to process the same patch embeddings as the ResMLP stream. It does not rely on such a structure, but instead model intra patch dependencies using affine transformations, MLP layers, and residual connections (Roy et al. 2025). Let the input patches be of size $P \times P$, each mapped to a d -dimensional embedding vector $X \in \mathbb{R}^{p \times d}$, where p is the number of tokens. Finally, the ResMLP module does the following (Touvron et al. 2022).

3.3.1. Affine transformation

An affine transformation is applied directly to replace conventional normalization layers.

$$f_{\theta, \kappa}(x') = \text{Diag}(\theta)x + \kappa \quad (13)$$

Here, $\theta \in \mathbb{R}^d$ and $\kappa \in \mathbb{R}^d$ are learnable parameters, and $\text{Diag}(\theta)$ denotes a diagonal matrix that scales each dimension independently.

3.3.2. First ResMLP layer

An activation function f_a , such as *GELU*, is applied within a residual block:

$$Z = X + f_a(f[X]). \quad (14)$$

It allows the network to perform nonlinear transformation while maintaining input identity.

3.3.3. Second ResMLP layer

A second MLP is applied on Z , incorporating additional linear layers and activation:

$$Y = Z + f[\gamma(\text{GELU}(\beta(f[Z])))]. \quad (15)$$

In the above equation, $f[\cdot]$ is a linear transformation, α, β, γ are learnable weight matrices, Gaussian Error Linear Unit (GELU) is linear unit activation of Gaussian error, which combines the properties of ReLU and sigmoid activations. Unlike standard ReLU that applies a hard threshold, GELU applies a smooth gating mechanism based on the Gaussian cumulative distribution function. GELU is particularly suitable for transformer-based models due to its nonlinearity and smooth gradient behaviour, promoting better convergence and generalization.

In this research work, we propose ResMLP+ which enhances ResMLP with SE Blocks and gated residuals. The SE block recalibrates feature channels by applying a global context-aware weighting. It is inserted after the MLP block within each ResMLP layer, as shown in Figure 4.

3.3.4. SE block integration

For an input feature matrix $X \in \mathbb{R}^{N \times D}$:

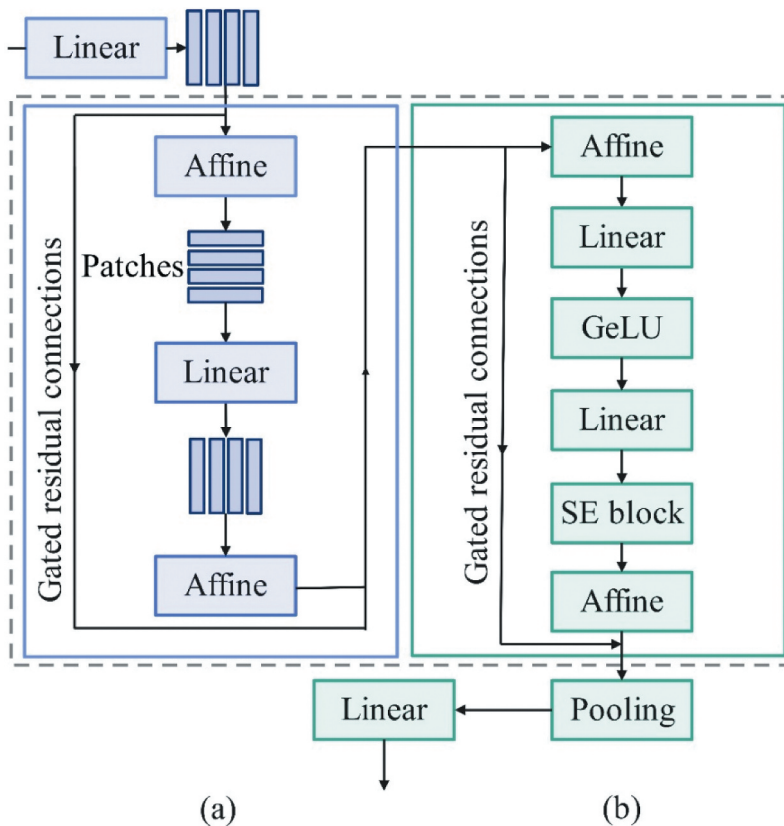


Figure 4. Architecture of the proposed ResMLP+, (a) cross-patch linear sublayer, (b) cross-channel two-layer MLP.

3.3.4.1. Squeeze. Global average pooling over tokens:

$$z = \frac{1}{N} \sum_{i=1}^N X_i \in \mathbb{R}^D. \quad (16)$$

3.3.4.2. Excitation. Bottleneck MLP with non-linearity:

$$s = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \quad (17)$$

where $W_1 \in \mathbb{R}^{D/r \times D}$, $W_2 \in \mathbb{R}^{D \times D/r}$, and r is the reduction ratio.

3.3.4.3. Recalibration.

$$\tilde{X} = X \cdot s \quad (18)$$

where $s \in \mathbb{R}^D$ is broadcast across all tokens.

This enables the model to emphasize informative channels and suppress irrelevant ones adaptively.

3.3.5. Residual gate modulation

Instead of a fixed residual connection $Y = X + f(X)$, use a learnable gate to control the contribution of the residual path.

3.3.5.1. Gated residual formula. Let $f(X)$ be the output of the MLP + SE block:

$$Y = X + \sigma(g(X)) \cdot f(X) \quad (19)$$

where $g(X)$ is a gating function (single-layer MLP), σ is the sigmoid function ensures gating in $[0,1]$.

This allows the model to adaptively scale residual contributions based on input content – helping training stability and improving representational control.

The ResMLP+ module (Figure 4), which is responsible for capturing fine-grained spectral spatial dependencies. In contrast to MLP, ResMLP alternates between a cross-patch linear layer and a cross-channel two layer MLP, which makes efficient communication within and across the PolSAR patches possible (Jamali et al. 2025). Unlike transformer layers that consist of layer normalization and attention matrices, ResMLP does not require batch-specific normalization layers like BatchNorm or LayerNorm but only linear transformations with GELU activations. A highly efficient pipeline with robust convergence and low parameter overhead is obtained from this design.

The two streams are processed in parallel and their outputs are fused element-wise via addition, which is less computationally intensive than concatenation and allows gradient flow during training. This fusion combines the original feature dimensionality with complementary spatial and spectral information. The final MLP head with a softmax activation takes the resultant feature vector, and outputs class probabilities for terrain types.

To overcome the complexity of deep and wide transformer models, the proposed dual stream structure uses lightweight and modular components. Both models (SimPool+ and ResMLP+) exhibit linear or near linear scaling properties, making the models train efficiently on limited hardware and (small scale) datasets of relevance in the PolSAR domain. The trade-off is balanced between

performance, interpretability, and requirements on computational resources, which all make the combined design fit for real-world deployment and domain transfer tasks in remote sensing.

3.4. Adaptation to PolSAR data

In order to adapt the proposed dual stream vision transformer framework to PolSAR data effectively, we designed preprocessing and embedding strategy, which is adapted to physical and statistical properties of polarimetric information. It consists of choosing good input features, creating spatial tokens, and combining them into a single feature space for classification.

3.4.1. Input feature representation

The PolSAR data is inherently complex valued and multi-dimensional, which is usually represented by a coherency matrix or scattering matrix. In this work, we extract this real-valued feature vectors using Freeman Durden decomposition, a well-established polarimetric decomposition method, where the scattering behaviour is decomposed into physically interpretable components (i.e. surface, double bounce and volume scattering). The resulting multi-channel feature maps capture spatially distributed scattering mechanisms and are stacked to form a 3D tensor $X \in \mathbb{R}^{H \times W \times C}$, where H and W are the spatial dimensions and C is the number of feature channels (this study uses nine feature channels).

3.4.2. Patch generation and tokenization

To process the PolSAR feature maps within a transformer framework, we divide X into a grid of non-overlapping or partially overlapping patches of size $P \times P$. In each patch, it is flattened and passed through a 2D convolutional embedding layer which projects it to a fixed length token vector with dimension d . This generates a sequence of tokens $\{x_1, x_2, \dots, x_p\}$, where $p = \frac{H \cdot W}{P^2}$ denotes the total number of patches.

We add learnable positional embeddings to each token to retain spatial context.

$$Z_i = X_i + P_i \text{ for } i = 1, \dots, p(20)$$

where $p_i \in \mathbb{R}$ denotes the positional encoding for the i -th patch.

3.4.3. Fusion mechanism and classification head

The SimPool+ and ResMLP+ streams process the token sequences in parallel and learn complementary representations of the PolSAR scene. After stream-wise processing, the resulting features are fused by element-wise addition.

$$h_{\text{fused}} = h_{\text{SimPool}} + h_{\text{ResMLP}}. \quad (20)$$

The fused embedding includes both global contextual dependencies and localized spatial – spectral patterns. The fused representation is mapped to the terrain class probabilities by a final multi-layer perceptron (MLP) head, with a softmax layer.

$$\hat{y} = \text{softmax}(\text{MLP}(h_{\text{fused}})). \quad (21)$$

This provides a modular design that makes it easy for domain-specific PolSAR features to be seamlessly integrated to the dual stream transformer and its further extension to multimodal data.

3.5. State-of-the-art models

To thoroughly analyse the performance of the proposed dual stream vision transformer framework for PolSAR terrain classification, it is benchmarked against a collection of established and diverse state-of-the-art models. Traditional convolutional architectures, self-attention-based transformers along with state-of-the-art domain adaptation techniques for remote sensing and PolSAR data are considered.

3.5.1. Complex-valued convolutional Neural networks (CV-CNN) (Alkhatib et al. 2023)

The CV-CNN is a multi-branch, feature-fusion network tailored for PolSAR image classification and interpretation. Each branch ingests a distinct set of polarimetric features (e.g. complex channel/descriptor groups) and processes them with complex-valued convolutions, preserving amplitude – phase relationships that are essential to polarimetric scattering physics. Along the depth of each branch, attention is used to progressively refine representations from shallow textures to deeper semantics. The branches are then fused to form a complementary, discriminative feature space.

3.5.2. Domain adaptive attention network (DAAN) (Ganin et al. 2016)

Besides generic architectures, we leverage a set of unsupervised domain adaptation methods that tackle domain shift explicitly. In this work, the DAAN aligns both marginal and conditional distributions between source and target domains with a multikernel maximum mean discrepancy criterion, with the addition of attention weighted feature transfer. One of the early models that successfully utilized attention mechanisms for UDA was DAAN.

3.5.3. Minimum class confusion (MCC) (Jin et al. 2020)

MCC is an entropy-based UDA approach that regularizes the predictions by penalizing class confusion. In contrast to pseudo-labels, MCC brings a class aware alignment term to focus the model to produce confident and well-separated output distributions on the unlabelled target domain. This makes it particularly well suited to situations where target supervision is limited or the domains are greatly divergent.

3.5.4. Progressive semantic context-aware network (PSCAN) (Dong et al. 2023)

Then, we consider PSCAN, another domain adaptation method, that applies the progressive learning strategy to match the semantic features between domains. PSCAN starts with global domain alignment and then successively incorporates local semantic consistency in a memory-based feature bank. Such an approach allows more reliable pseudo label refinement and more robust semantic discrimination over heterogeneous PolSAR terrains.

3.5.5. Cross-domain feature fusion network (CDFNet) (S. Wang et al. 2025)

We then compare our model to the CDFNet, which is currently the state-of-the-art in PolSAR domain adaptation. Two key innovations included in CDFNet are the DBS module

to offset label distribution shift through class awareness balanced sampling and the cross-domain feature fusion module, which provides a way to effectively fuse source and target features without including label noise. To the best of our knowledge, CDFNet is the most relevant and competitive model for this study since it is tailored for PolSAR UDA tasks and achieves strong performance on SF-RS2, SF-GF3, and Fle-RS2 benchmarks.

4. Datasets and experimental setup

4.1. PolSAR datasets

To rigorously evaluate the performance of the proposed dual-stream vision transformer framework under various domain shift scenarios, we utilize four benchmark PolSAR datasets with distinct characteristics in terms of geographic region, imaging sensor, frequency band, and semantic class labels (Liu et al. 2022). SF-RS2, SF-GF3, SF-ALOS2, and Fle-RS2 are the datasets with which this problem is addressed, and all of them pose different challenges to domain adaptation due to the different resolution, sensor configuration, and land cover distribution.

The SF-RS2 dataset (Figure 5), acquired using the RADARSAT-2 satellite, captures a portion of the San Francisco metropolitan area. It operates in the C-band and provides fully polarimetric SAR data with a spatial resolution of 8 metres. The typical urban land cover classes of water bodies, vegetation, and generalized building areas are included in the image. The San Francisco region is also covered by SF-GF3 dataset (Figure 6) from the Gaofen-3 (GF-3) satellite, which also operates in C-band. Although it greatly differs from SF-RS2 in terms of sensor geometry and

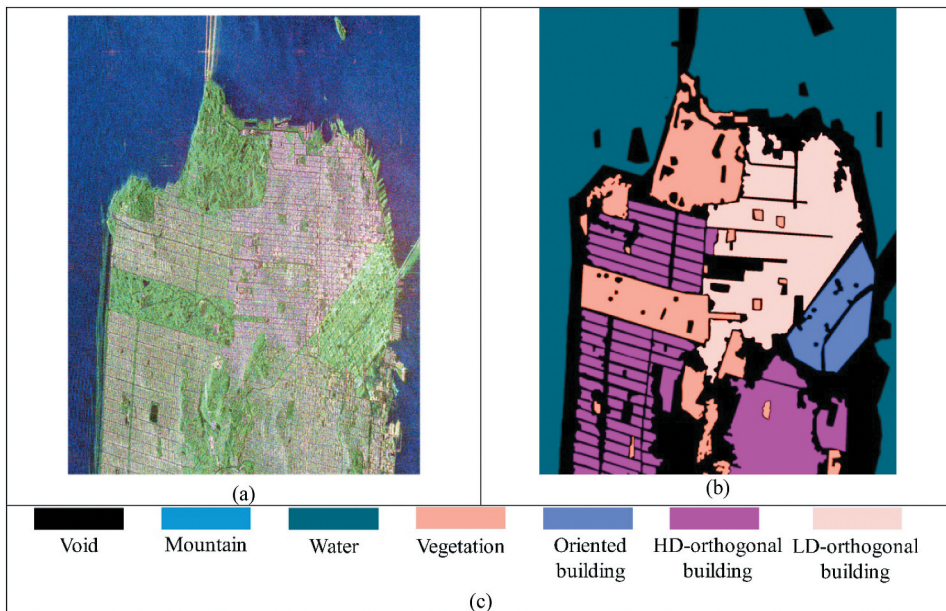


Figure 5. Visual representation of the SF-RS2 dataset. (a) Pseudo-color PolSAR image, (b) ground truth land cover map, and (c) class color legend.

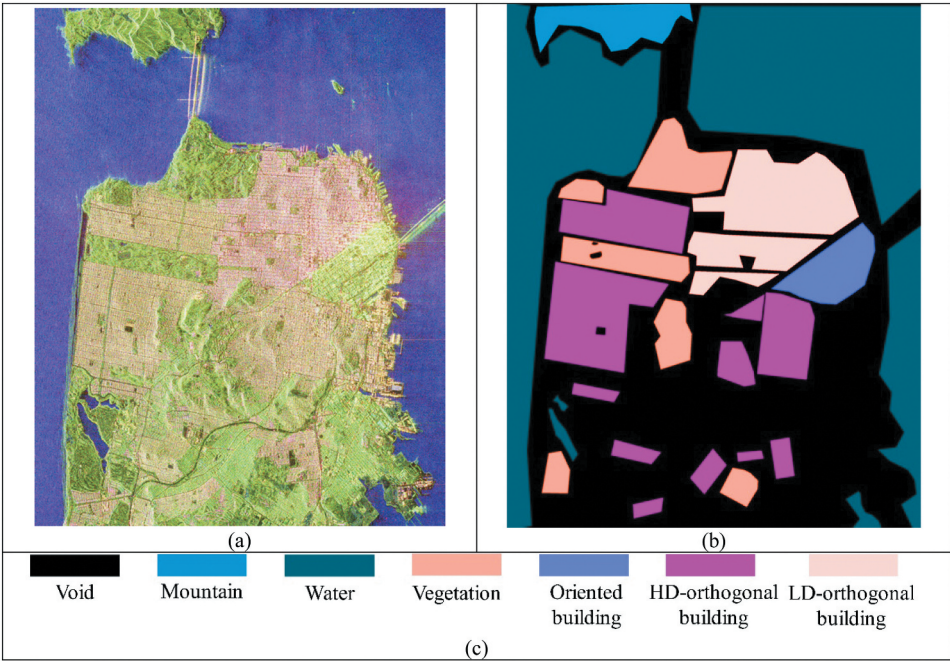


Figure 6. Visual representation of the SF-GF3 dataset. (a) Pseudo-color PolSAR image, (b) ground truth land cover map, and (c) class color legend.

acquisition parameters as well as data granularity, it was possible to obtain a very good agreement. In particular, SF-GF3 defines finer grained building categories like high-density orthogonal (HD-orthogonal) buildings, low-density orthogonal (LD-orthogonal) buildings, and oriented buildings as well as water and vegetation classes.

The third dataset, SF-ALOS2 (Figure 7), also covers the San Francisco region but is acquired using the ALOS-2 satellite operating in the L-band. With a coarser resolution of 18 metres, SF-ALOS2 adds a distinct sensor shift component to the evaluation, offering the same detailed building subclasses as SF-GF3. With C-band and L-band imagery from the same region available, controlled sensor shift experiments are possible. Finally, FleRS2 (Figure 8) comes from the RADARSAT-2 platform but it deals with the Flevoland region in the Netherlands, which is mainly agricultural. The classes in this dataset include water, vegetation, and several types of building structures, which are useful for region-based domain adaptation tasks.

These four datasets are complementary in terms of both spatial and semantic diversity. They facilitate comprehensive analysis of model performance across a spectrum of real-world scenarios, including cross-sensor, cross-region, and combined domain shifts. Table 1 provides a detailed summary of their geographic scope, sensor characteristics, frequency bands, land cover categories, spatial resolutions, and image dimensions.

All datasets used in this study are acquired using PolSAR sensors. These include RADARSAT-2 (C-band), Gaofen-3 (C-band), and ALOS-2 (L-band), which are capable of capturing quad-polarization data. The datasets represent a range of geographic and

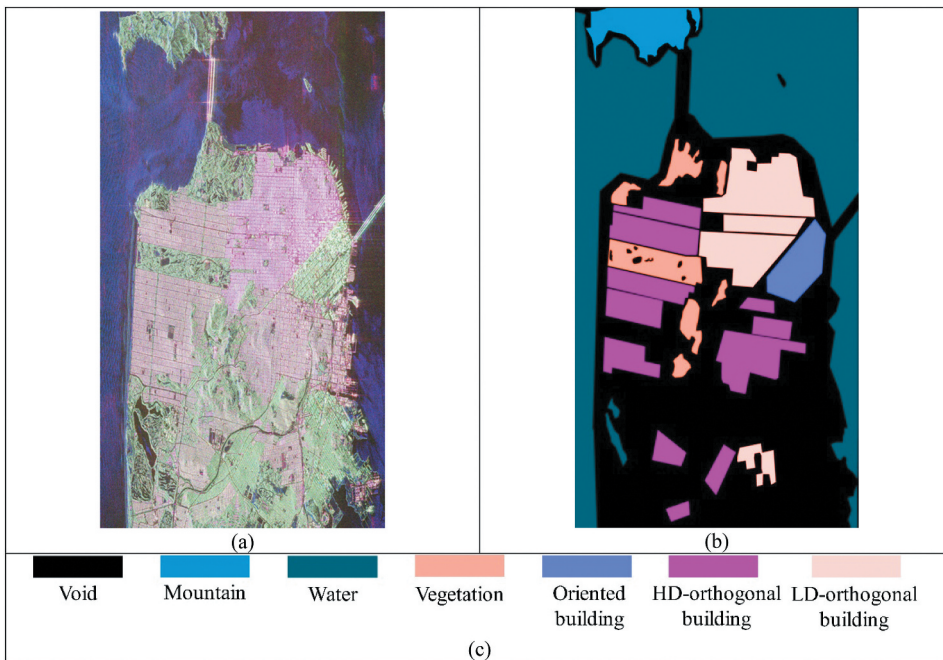


Figure 7. Visual representation of the SF-ALOS2 dataset. (a) Pseudo-color PolSAR image, (b) ground truth land cover map, and (c) class color legend.

acquisition conditions to evaluate robustness across sensor frequency bands, spatial resolutions, and terrain types.

The visual representations of the four datasets (Figure 5–8) and their corresponding ground truth labels used for evaluation in this study were adapted from S. Wang et al. (2025). The land cover categories in the SF-RS2, SF-GF3, and SF-ALOS2 datasets were produced using manual interpretation. For the Fle-RS2 dataset, ground truth was generated using a semi-automated classification process followed by manual refinement. These label sets are widely adopted in the remote sensing community and are considered benchmark-standard for domain adaptation tasks.

4.2. Sample distribution

A critical aspect of training and evaluating terrain classification models, particularly under unsupervised domain adaptation settings, lies in the distribution of labelled samples across terrain categories. Table 2 shows the number of annotated pixels per land cover class in the four PolSAR datasets used in this study that present significant inter-dataset variability in terms of class proportions and which may, or may not, include specific categories. The SF-RS2 dataset, collected over San Francisco using RADARSAT-2, includes annotations for five major classes: water, vegetation, HD-orthogonal buildings, LD-orthogonal buildings, and oriented buildings. The water class is the largest proportion of labelled samples (about 37%), but other building categories are more evenly distributed with HD-orthogonal and LD-orthogonal at about 19% and 16% respectively. On the

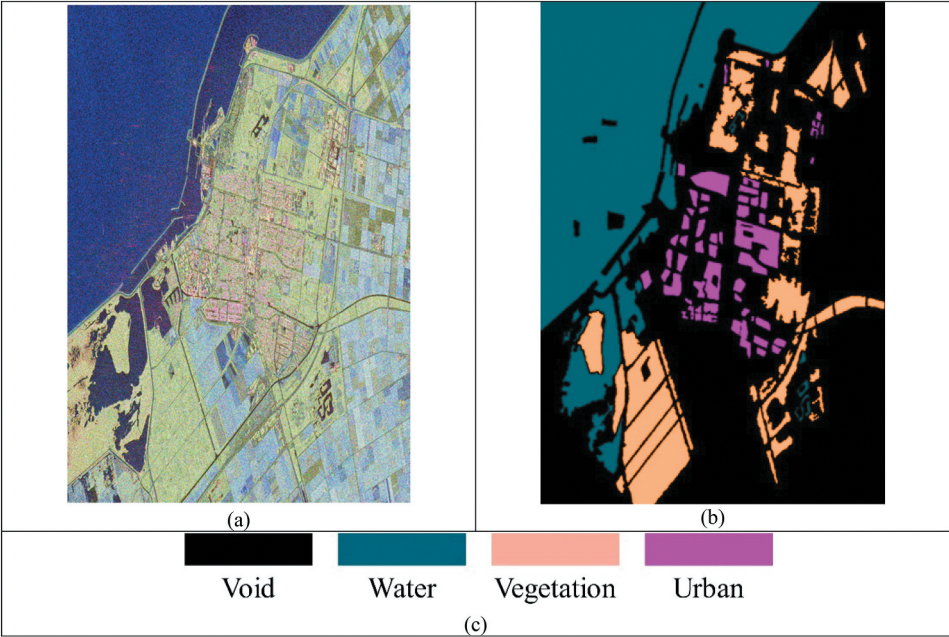


Figure 8. Visual representation of the Fle-RS2 dataset. (a) Pseudo-color PolSAR image, (b) ground truth land cover map, and (c) class color legend.

Table 1. Overview of employed PolSAR datasets for cross-domain terrain classification.

| Dataset | Region | Sensor | Band | Categories | Resolution | Image size (pixels) |
|----------|---------------|------------|------|--|------------|---------------------|
| SF-RS2 | San Francisco | RadarSAT-2 | C | Water, Vegetation, Building | 8 m | 1380 × 1800 |
| SF-GF3 | San Francisco | GaoFen-3 | C | Water, Vegetation, Mountain, HD-orthogonal building, LD-orthogonal building, Oriented building | 8 m | 2304 × 2912 |
| SF-ALOS2 | San Francisco | ALOS-2 | L | Water, Vegetation, Mountain, HD-orthogonal building, LD-orthogonal building, Oriented building | 18 m | 2784 × 5056 |
| Fle-RS2 | Flevoland | RadarSAT-2 | C | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | 8 m | 1635 × 2375 |

Table 2. Class-wise distribution of annotated samples in PolSAR terrain datasets.

| Dataset | Water | Vegetation | HD-orthogonal building | LD-orthogonal building | Oriented building |
|----------|---------------|--------------|------------------------|------------------------|-------------------|
| SF-RS2 | 852078 (37%) | 237237 (13%) | 351181 (19%) | 282975 (16%) | 80616 (4%) |
| SF-GF3 | 2226521 (58%) | 369939 (10%) | 687977 (18%) | 426466 (11%) | 129118 (3%) |
| SF-ALOS2 | 4694554 (61%) | 459701 (7%) | 1314858 (17%) | 947418 (12%) | 250403 (13%) |
| Fle-RS2 | 1050726 (66%) | 403584 (25%) | – | 136561 (9%) | – |

other hand, the skewness of the SF-GF3 dataset is much higher towards the water class (58% of labelled data) and then the HD-orthogonal and LD-orthogonal buildings. The domain-specific challenge posed by this class imbalance is especially significant when SF-GF3 is used as the target domain in UDA experiments.

SF-ALOS2, collected via the L-band ALOS-2 sensor, also includes detailed annotations for the three building subtypes along with water and vegetation. Notably, SF-ALOS2 has the highest total sample count, with over 4.6 million labelled pixels for water alone. The diversity and volume of this dataset make it a good source domain for adaptation tasks with complex urban topologies. However, Fle-RS2, created from the agricultural area of Flevoland, contains water, vegetation, and LD-orthogonal buildings, but does not contain the HD-orthogonal and Oriented classes. The main class in this dataset is vegetation (25% of the labelled samples), and water comprises 66%. The unequal and non-overlapping class distributions across these datasets create a realistic but challenging setting for domain adaptation. An example of such a problem is the label shift problem – when marginal label distribution between source and target domains is different – having a major effect on classifier performance in the absence of proper treatment. To this end, we integrate label distribution aware strategies like DBS in the framework proposed. This further emphasizes the need to evaluate such models not only on the overall accuracy results but also on the class wise and average accuracy of the model across different terrain categories for the sake of fairness and robustness.

4.3. Experimental units

To evaluate the generalization capability of the proposed dual-stream vision transformer under various domain shift scenarios, we construct multiple experimental units based on source – target domain pairings. Each unit is related to one specific configuration where one PolSAR dataset serves as the source domain (with annotated data) and the other as the target domain (unannotated during training). This corresponds to the setting of the unsupervised domain adaptation paradigm, where models have to adapt knowledge learned from the source to perform accurate classification on the target, without using any ground truth labels for the target.

The selection of domain pairs is guided by two primary shift types: sensor shift and region shift. Sensor shift experiments involve datasets from the same geographic region but acquired using different SAR sensors. For example, the SA-SG unit means the adaptation from SF-ALOS2 (L-band) to SF-GF3 (C-band), and SG-SA means the reverse direction. In the same manner, RADARSAT-2 and ALOS-2 exchanges over San Francisco are represented by SR-SA and SA-SR, respectively. These are designed to test the model in different sensor modality, imaging frequency, and resolution while controlling for geographic context.

Region shift experiments, by contrast, explore the impact of spatial variation in terrain characteristics. The SR-FR and FR-SR units represent adaptation between the urban domain of San Francisco (SF-RS2) and the agricultural domain of Flevoland (Fle-RS2), both captured using the same sensor (RADARSAT-2). The model is challenged to learn domain invariant features that generalize across semantically distinct environments without access to the environment information. Combined sensor and region shift units, such as FR-SG and SG-FR, are the most complex and realistic evaluation scenario, where source

Table 3. Source – target domain configurations for PolSAR adaptation experiments.

| Unit name | Class | Different | Source | Target |
|-----------|---|-------------------|----------|----------|
| FR-SG | Water, Vegetation, Building | Sensor and Region | Fle-RS2 | SF-GF3 |
| SG-FR | Water, Vegetation, Building | Sensor and Region | SF-GF3 | Fle-RS2 |
| FR-SR | Water, Vegetation, Building | Region | Fle-RS2 | SF-RS2 |
| SR-FR | Water, Vegetation, Building | Region | SF-RS2 | Fle-RS2 |
| SA-SG | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-ALOS2 | SF-GF3 |
| SG-SA | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-GF3 | SF-ALOS2 |
| SR-SA | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-RS2 | SF-ALOS2 |
| SA-SR | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-ALOS2 | SF-RS2 |
| SG-SR | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-GF3 | SF-RS2 |
| SR-SG | Water, Vegetation, HD-orthogonal building, LD-orthogonal building, Oriented building | Sensor | SF-RS2 | SF-GF3 |

and target differ by both geographic location and sensor type. These experiments replicate real-world applications in which one area and sensor labelled data must be used to classify scenes in a completely different region and acquisition platform. The domain pairs chosen and the type of domain shift for each of the ten experimental units considered in this study are summarized in Table 3. We systematically vary the domain pairings, and consider our model across a broad range of cross-domain classification problems from which we can comprehensively evaluate the adaptability and stability of the model as well as its robustness to domain shifts.

4.4. Evaluation protocol

To ensure a rigorous and fair comparison across models and domain adaptation setups, we adopt a standardized evaluation protocol that aligns with established practices in unsupervised domain adaptation for remote sensing. Experiments are conducted in a source to target setting in which we have labelled data in the source domain and perform supervised training using it and the target domain remains entirely unlabelled throughout the training process. This is similar to the real-world where annotation is expensive or not available in a target area. Each model is trained using only the labelled samples from the source dataset, following a fixed partition that includes both training and validation splits. The target dataset is only used for inference and performance evaluation. Unless explicitly used as part of the model adaptation mechanism, no pseudo labelling or self-training is applied. We use the same training and testing splits for all domain pairs as defined in previous studies for reproducibility and fair benchmarking.

Performance is assessed using three standard metrics. Overall Accuracy (OA) measures the proportion of correctly classified pixels across the entire test set, providing a global performance estimate. The Average Accuracy (AA) is computed as mean of classification accuracy for individual classes and is very effective in imbalanced datasets to avoid result skewing because of dominant classes (e.g. water). Last, the Kappa coefficient is reported to measure the agreement between predicted and true labels, taking into account random chance. The proposed model is evaluated in terms of robustness and statistical significance by repeating each experiment three times with different random seeds and

averaging the final results. Any variability due to stochastic optimization, batch sampling is accounted for by this repetition.

We also report class wise accuracies across all the three terrain categories (water, vegetation, and various building classes) to get a better idea of how the model performs across heterogeneous land cover types. The breakdown of this approach is especially important for generalization to minority classes under domain shift conditions. All models are trained under the same training conditions, having the same hyperparameters, the same input patch sizes, and the same optimization strategies as specified in the experimental unit settings. Such uniformity guarantees that any performance differences observed are due to architectural or adaptation capabilities rather than due to differences in the implementation or the particular training regime employed.

5. Results and evaluation

5.1. Ablation study

To validate the effectiveness of the proposed dual-stream vision transformer architecture, we conduct a comprehensive ablation study by selectively disabling key components of the model. In particular, we evaluate the individual and joint effects of the SimPool+ transformer stream and the ResMLP+ stream, which constitute the basis of our architecture. This analysis provides empirical insights into the importance of each module regarding learning spatial – spectral representations and improving generalization under shift in domains.

The ablation experiments are conducted using different units (SA-SG, SG-SA, SR-SA, SA-SR, SG-SR, and SR-SG), which present a significant challenge due to the difference in sensor modality (L-band vs. C-band) and frequency-dependent scattering characteristics. In comparison, we investigate five configurations: (1) SimPool only: This is the model where attention stream is active and ResMLP+ branch is removed; (2) SimPool+ only: This is the model where attention stream is active and ResMLP+ branch is removed; (3) ResMLP only: This is the baseline model without SimPool+ integration using MLP stream; (4) ResMLP+ only: This is the ResMLP+ model without SimPool+ integration using MLP stream; and (5) Full Model: This is the composite model using both SimPool+ and ResMLP+ streams with element-wise fusion.

As shown in [Table 4](#), the SimPool+ only variant outperforms the ResMLP+ only configuration, achieving higher overall accuracy. This result indicates that the SimPool+ mechanism can capture global contextual dependencies well via simplified attention. Nevertheless, the ResMLP+ stream is still beneficial in local feature modelling, especially on spatially coherent terrain classes such as vegetation and low-density buildings.

Table 4. Ablation study on the effectiveness of SimPool+ and ResMLP+ streams.

| Module | SA-SG | SG-SA | SR-SA | SA-SR | SG-SR | SR-SG |
|---------------|-------|-------|-------|-------|-------|-------|
| SimPool only | 93.15 | 94.77 | 94.28 | 96.01 | 96.40 | 94.49 |
| SimPool+ only | 96.85 | 95.06 | 96.96 | 98.78 | 97.77 | 96.58 |
| ResMLP only | 94.73 | 94.38 | 94.98 | 96.61 | 96.12 | 94.69 |
| ResMLP+ only | 95.54 | 94.83 | 95.44 | 97.34 | 96.84 | 95.64 |
| Full model | 99.31 | 99.07 | 99.26 | 99.84 | 99.64 | 99.85 |

The full dual-stream model achieves the highest performance across all units, demonstrating that the fusion of global and local information pathways leads to superior generalization. The performance gain in support of our hypothesis that combining spectral – spatial sensitivity and global structural awareness is necessary for robust PolSAR terrain classification, especially when features are distorted by the sensor and labels shift. They confirm the architectural synergy of SimPool+ and ResMLP+ (our design choice) and substantiate our proposition that dual stream processing is valuable for domain adaptive transformer models.

5.2. Cross-sensor domain adaptation

To evaluate the robustness of the proposed model in handling sensor-induced domain shifts, we conduct experiments involving datasets captured over the same geographic region (San Francisco) but acquired using different SAR sensors operating in distinct frequency bands. This type of setup is a practical simulation of remote sensing scenarios where training data and deployment data come from different platforms having different sensor geometries, resolutions, and polarimetric attributes.

We consider four cross-sensor adaptation units: SA-SG, SG-SA, SR-SA, and SA-SR, where SF-ALOS2 (L-band), SF-GF3 (C-band), and SF-RS2 (C-band) are alternately used as source and target domains. Such configurations push models to learn representations of a scene that are invariant to the imaging system and the spectral band in effect, thereby facilitating easier transfer across spectral bands and different systems. Figure 9 shows the predicted terrain classification maps for target domains across experimental units under cross-sensor domain adaptation scenarios (S. Wang et al. 2025). As reported in Table 5, methods including CV-CNN, DAAN, MCC, and PSCAN struggle to maintain high classification performance under cross-sensor shifts, particularly in the L-band to C-band direction. For example, PSCAN attains an OA of 89.84% in the SA-SG unit compared to mild improvements provided by DAAN and CV-CNN with OA scores of 87.95% and 86.63%, respectively. These models are outperformed by MCC (92.26%), which, nevertheless, is not sufficiently adaptive.

In contrast, the proposed dual-stream vision transformer achieves a substantial improvement in performance across all sensor-shift configurations. Finally, in the SA-SG unit, it achieves 99.31% OA, 94.46% Kappa, and 96.42% AA, surpassing the state-of-the-art CDFNet (S. Wang et al. 2025) by a considerable margin. The SG-SA and SR-SA units show similar gains, indicating that the model can properly align heterogeneous polarimetric features from different SAR sensors. This confirms the architectural design of the model, in particular its capacity to fuse global and local features in a way that is robust against sensor variability. Collectively, the SimPool+ attention mechanism maintains structural context and the ResMLP+ stream retains spatial consistency to form a unified representation space that generalizes across sensor modalities. The results under the cross-sensor settings are very strong, indicative of the model's practicality and applicability for real-world deployment where training data from one sensor needs to be used to classify scenes captured by another without target labels. It equally stresses the significance of architectural efficiency and semantic consistency to domain adaptive PolSAR classification.



Figure 9. Predicted terrain classification maps by the proposed model for target domains across experimental units under cross-sensor domain adaptation scenarios.

5.3. Cross-region domain adaptation

To further evaluate the generalizability of the proposed model under domain shift conditions, we conduct experiments involving cross-region domain adaptation, where training and testing datasets are acquired from different geographic locations using the same sensor. The challenge this setting poses is different from that of cross-sensor adaptation: although the imaging modality is the same, the distributions of terrain categories in terms of both semantics and spatial distributions may vary substantially, because of landscape variability, cultural infrastructure, and ecological patterns.

In this study, we utilize the SR-FR and FR-SR experimental units, which pair the urban environment of San Francisco (SF-RS2) with the predominantly agricultural region of Flevoland (Fle-RS2). The two datasets are collected with the RADARSAT-2 C-band sensor to maintain consistency in radar configuration but to also induce meaningful changes in scene content and class distribution. Fle-RS2 has a more homogeneous class structure with fewer building subclasses than SF-RS2, which is more complex and structurally varied.

Table 5. Classification performance under cross-sensor domain adaptation scenarios.

| Unit | Model | Classification accuracy | | | | | | | |
|-------|-------------------------------|-------------------------|------------|------------------------|------------------------|-------------------|--------|-----------|--------|
| | | Water | Vegetation | HD-orthogonal building | LD-orthogonal building | Oriented building | OA (%) | Kappa (%) | AA (%) |
| SA-SG | CV-CNN (Alkhatib et al. 2023) | 95.32 | 83.04 | 89.74 | 78.55 | 98.41 | 86.63 | 76.34 | 89.01 |
| | DAAN (Ganin et al. 2016) | 98.71 | 90.12 | 84.71 | 48.24 | 44.87 | 87.95 | 80.22 | 73.13 |
| | MCC (Jin et al. 2020) | 99.80 | 94.34 | 75.93 | 81.09 | 80.13 | 92.26 | 87.38 | 86.86 |
| | PSCAN (Dong et al. 2023) | 99.85 | 87.79 | 54.04 | 96.35 | 92.47 | 89.84 | 83.49 | 86.90 |
| | CDFNet (S. Wang et al. 2025) | 99.96 | 85.59 | 87.82 | 89.75 | 82.81 | 94.37 | 91.19 | 89.18 |
| SG-SA | Proposed model | 99.98 | 96.41 | 90.49 | 97.77 | 97.43 | 99.31 | 94.46 | 96.42 |
| | CV-CNN (Alkhatib et al. 2023) | 87.55 | 93.27 | 92.98 | 81.70 | 65.22 | 83.47 | 71.75 | 84.14 |
| | DAAN (Ganin et al. 2016) | 91.35 | 87.65 | 88.56 | 57.92 | 62.02 | 85.56 | 75.99 | 77.50 |
| | MCC (Jin et al. 2020) | 99.82 | 74.62 | 75.15 | 85.16 | 87.61 | 91.87 | 86.00 | 84.91 |
| | PSCAN (Dong et al. 2023) | 87.17 | 92.03 | 84.35 | 88.60 | 65.20 | 86.43 | 78.14 | 83.47 |
| SR-SA | CDFNet (S. Wang et al. 2025) | 99.93 | 74.83 | 87.73 | 97.97 | 77.07 | 94.65 | 91.52 | 87.82 |
| | Proposed model | 99.97 | 94.41 | 95.94 | 98.89 | 91.87 | 99.07 | 93.69 | 96.22 |
| | CV-CNN (Alkhatib et al. 2023) | 86.44 | 71.43 | 94.47 | 87.76 | 96.47 | 79.94 | 75.84 | 87.31 |
| | DAAN (Ganin et al. 2016) | 89.70 | 65.61 | 60.10 | 84.92 | 87.70 | 82.52 | 71.57 | 77.21 |
| | MCC (Jin et al. 2020) | 99.92 | 75.27 | 89.59 | 64.02 | 71.43 | 91.30 | 84.85 | 80.46 |
| SA-SR | PSCAN (Dong et al. 2023) | 91.07 | 91.88 | 91.36 | 82.62 | 29.73 | 88.09 | 80.26 | 65.43 |
| | CDFNet (S. Wang et al. 2025) | 99.97 | 74.69 | 87.08 | 97.39 | 81.67 | 94.56 | 91.48 | 88.17 |
| | Proposed model | 99.98 | 93.91 | 94.85 | 98.93 | 97.79 | 99.26 | 94.74 | 97.09 |
| | CV-CNN (Alkhatib et al. 2023) | 95.43 | 62.17 | 91.47 | 98.65 | 99.94 | 88.67 | 88.75 | 89.53 |
| | DAAN (Ganin et al. 2016) | 99.99 | 65.17 | 94.55 | 74.95 | 97.00 | 90.29 | 85.96 | 86.23 |
| SG-SR | MCC (Jin et al. 2020) | 96.84 | 95.84 | 85.68 | 78.14 | 86.14 | 91.13 | 87.33 | 88.12 |
| | PSCAN (Dong et al. 2023) | 99.95 | 92.48 | 96.28 | 52.57 | 94.04 | 90.55 | 86.38 | 87.46 |
| | CDFNet (S. Wang et al. 2025) | 99.96 | 95.81 | 92.81 | 88.99 | 90.31 | 96.16 | 94.23 | 93.78 |
| | Proposed model | 99.99 | 97.26 | 98.63 | 99.66 | 99.37 | 99.84 | 97.83 | 98.98 |
| | CV-CNN (Alkhatib et al. 2023) | 99.92 | 93.47 | 94.82 | 81.65 | 96.27 | 93.19 | 90.83 | 93.23 |
| SG-SR | DAAN (Ganin et al. 2016) | 99.91 | 76.30 | 91.35 | 95.28 | 86.86 | 93.67 | 90.89 | 89.14 |
| | MCC (Jin et al. 2020) | 99.98 | 94.95 | 96.13 | 84.87 | 79.02 | 95.27 | 93.18 | 88.19 |
| | PSCAN (Dong et al. 2023) | 99.64 | 95.36 | 84.83 | 96.55 | 84.23 | 95.02 | 92.85 | 92.34 |
| | CDFNet (S. Wang et al. 2025) | 99.95 | 93.74 | 93.51 | 90.56 | 92.70 | 96.21 | 95.00 | 95.07 |
| | Proposed model | 99.98 | 97.22 | 98.46 | 98.71 | 98.72 | 99.64 | 98.56 | 98.62 |

(Continued)

Table 5. (Continued).

| Unit | Model | Classification accuracy | | | | | | | |
|-------|-------------------------------|-------------------------|------------|------------------------|------------------------|-------------------|--------|-----------|--------|
| | | Water | Vegetation | HD-orthogonal building | LD-orthogonal building | Oriented building | OA (%) | Kappa (%) | AA (%) |
| SR-SG | CV-CNN (Alkhatib et al. 2023) | 99.91 | 86.76 | 83.65 | 78.57 | 99.85 | 90.67 | 87.46 | 89.75 |
| | DAAN (Ganin et al. 2016) | 99.77 | 79.24 | 78.03 | 91.95 | 85.58 | 92.55 | 87.82 | 86.91 |
| | MCC (Jin et al. 2020) | 99.81 | 93.38 | 70.64 | 83.93 | 73.98 | 91.33 | 85.82 | 84.18 |
| | PSCAN (Dong et al. 2023) | 99.79 | 88.55 | 92.46 | 54.63 | 82.35 | 91.79 | 86.48 | 83.82 |
| | CDFNet (S. Wang et al. 2025) | 99.94 | 87.95 | 81.59 | 98.01 | 79.56 | 94.92 | 90.29 | 89.42 |
| | Proposed model | 99.98 | 95.33 | 94.87 | 99.54 | 99.38 | 99.85 | 94.84 | 97.82 |

Figure 10 shows the predicted terrain classification maps for target domains across experimental units under cross-region domain adaptation (S. Wang et al. 2025). The results, presented in Table 6, indicate that CV-CNN and early domain adaptation methods such as DAAN and MCC experience a notable performance drop when transferring knowledge across these regions. For example, in the SR-FR unit, CV-CNN achieves the OA of 95.94%, DAAN and MCC achieve 92.92% and 95.41%, respectively. Despite having limited ability to align semantic representations in urban and agricultural landscapes, PSCAN has moderate improvement as well.

In contrast, the proposed dual-stream transformer consistently outperforms all models across both SR-FR and FR-SR directions. In the SR-FR unit, it achieves 99.86% OA, 98.73% Kappa, and 98.71% AA and similar results in the reverse direction. The results of these outcomes show that the model can maintain discriminative features across different terrains by depending on the complementary capabilities of the SimPool+ and ResMLP + stream. The superior results in these region-shift experiments confirm the importance of

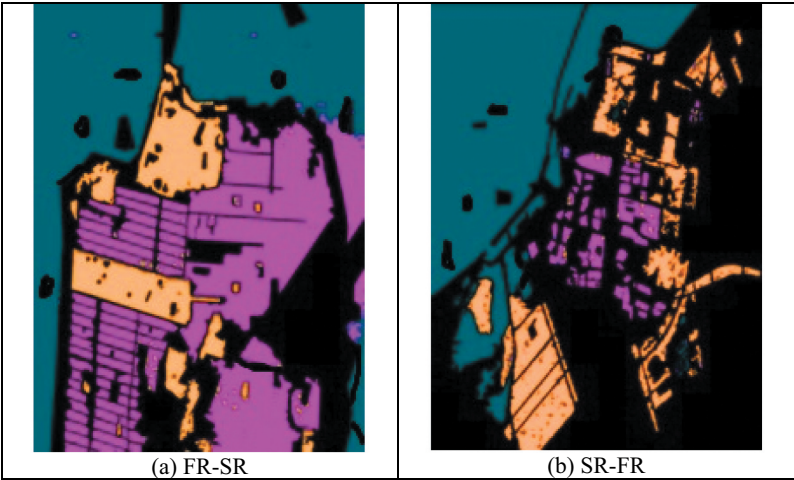


Figure 10. Predicted terrain classification maps by the proposed model for target domains across experimental units under cross-region domain adaptation.

Table 6. Classification results under cross-region domain adaptation.

| Unit | Model | Classification accuracy | | | | | |
|-------|-------------------------------|-------------------------|------------|----------|--------|-----------|--------|
| | | Water | Vegetation | Building | OA (%) | Kappa (%) | AA (%) |
| FR-SR | CV-CNN (Alkhatib et al. 2023) | 93.47 | 96.77 | 95.97 | 91.87 | 89.76 | 95.40 |
| | DAAN (Ganin et al. 2016) | 99.11 | 93.96 | 95.91 | 92.92 | 94.05 | 96.33 |
| | MCC (Jin et al. 2020) | 93.10 | 93.78 | 99.09 | 95.41 | 92.39 | 95.32 |
| | PSCAN (Dong et al. 2023) | 94.62 | 92.71 | 98.45 | 97.51 | 95.86 | 95.26 |
| | CDFNet (S. Wang et al. 2025) | 99.96 | 97.89 | 97.98 | 98.95 | 98.24 | 98.61 |
| | Proposed model | 99.98 | 98.34 | 99.66 | 99.71 | 99.24 | 99.32 |
| SR-FR | CV-CNN (Alkhatib et al. 2023) | 98.89 | 94.46 | 89.76 | 95.94 | 91.66 | 94.37 |
| | DAAN (Ganin et al. 2016) | 99.62 | 95.00 | 95.00 | 98.19 | 96.37 | 96.54 |
| | MCC (Jin et al. 2020) | 99.64 | 90.16 | 96.06 | 96.93 | 93.81 | 95.62 |
| | PSCAN (Dong et al. 2023) | 99.93 | 92.24 | 95.16 | 97.57 | 95.08 | 95.78 |
| | CDFNet (S. Wang et al. 2025) | 99.96 | 96.87 | 97.17 | 98.92 | 97.82 | 98.00 |
| | Proposed model | 99.98 | 97.66 | 98.51 | 99.86 | 98.73 | 98.71 |

combining global contextual modelling and localized spatial sensitivity, especially when dealing with shifts in land cover semantics. In contrast, instead of feature statistics being changed due to acquisition mechanisms (in case of sensor-induced shifts), the spatial structure, and the class balance are affected in case of region-induced shifts. The proposed model provides a good solution to both challenges and demonstrates its usefulness for real-world applications where training is done on one region and deployment on another.

5.4. Combined region and sensor shift

The most challenging evaluation scenario explored in this study involves simultaneous shifts in both geographic region and sensor modality. This dual domain shift is a practical and very relevant setting in remote sensing applications, where models trained on data labelled in one region and sensor need to generalize to a completely new environment and acquisition platform, without access to target domain labels. To simulate this real-world condition, we define two experimental units: FR-SG and SG-FR. The FR-SG setup trains the model on Fle-RS2 dataset, which was collected over an agricultural region of Flevoland with RADARSAT-2 C-band sensor and tests it on SF-GF3, an urban scene in San Francisco imaged by Gaofen-3 satellite. On the opposite, the direction is inverted in SG-FR. These experiments induce both semantic shift due to landscape difference between rural and urban environments and sensor shift due to different radar system specifications and imaging geometry between RADARSAT-2 and GF-3.

Figure 11 shows the predicted terrain classification maps for target domains across experimental units under combined region and sensor domain shifts (S. Wang et al. 2025). The results in Table 7 underscore the difficulty of these combined shifts for conventional and even recent UDA models. We observe OA scores below 94% for CV-CNN, weak improvements to 96.79% OA for DAAN and 97.03% for MCC, and a modest improvement but no assurance on spatial and spectral alignment for PSCAN (97.41%). Despite the fact that CDFNet (98.11%) is specifically developed for PolSAR UDA, it is unable to close the gap completely, especially in terms of class-wise average accuracy.

In stark contrast, the proposed dual-stream vision transformer achieves exceptionally high accuracy, recording 99.95% OA, 98.33% Kappa, and 99.19% AA in FR-SG unit. These



Figure 11. Predicted terrain classification maps by the proposed model for target domains across experimental units under combined region and sensor domain shifts.

Table 7. Classification results under combined region and sensor domain shifts.

| Unit | Model | Classification accuracy | | | OA (%) | Kappa (%) | AA (%) |
|-------|-------------------------------|-------------------------|------------|----------|--------|-----------|--------|
| | | Water | Vegetation | Building | | | |
| FR-SG | CV-CNN (Alkhatib et al. 2023) | 99.93 | 96.81 | 91.44 | 93.84 | 91.73 | 96.06 |
| | DAAN (Ganin et al. 2016) | 99.67 | 97.61 | 90.77 | 96.79 | 94.12 | 96.02 |
| | MCC (Jin et al. 2020) | 97.97 | 90.51 | 96.61 | 97.03 | 94.47 | 95.03 |
| | PSCAN (Dong et al. 2023) | 99.77 | 97.19 | 92.78 | 97.41 | 95.25 | 96.25 |
| | CDFNet (S. Wang et al. 2025) | 99.96 | 92.12 | 94.53 | 98.11 | 96.51 | 95.53 |
| | Proposed model | 99.98 | 98.77 | 98.81 | 99.95 | 98.33 | 99.19 |
| SG-FR | CV-CNN (Alkhatib et al. 2023) | 93.48 | 98.71 | 89.47 | 93.96 | 91.27 | 93.89 |
| | DAAN (Ganin et al. 2016) | 98.93 | 93.64 | 94.17 | 97.16 | 94.32 | 95.58 |
| | MCC (Jin et al. 2020) | 99.60 | 99.79 | 73.87 | 97.43 | 94.76 | 91.08 |
| | PSCAN (Dong et al. 2023) | 97.12 | 95.65 | 93.98 | 96.47 | 93.00 | 95.58 |
| | CDFNet (S. Wang et al. 2025) | 99.98 | 94.53 | 99.22 | 98.46 | 96.88 | 97.91 |
| | Proposed model | 99.98 | 99.96 | 99.90 | 99.93 | 98.55 | 99.94 |

results show that the model is unmatched at generalizing to the most severe forms of domain shift. This architecture succeeds in merging the spectral distortion resilience of ResMLP+ with the topological variability resilience of a global model like SimPool+ into a single unified representation.

The performance gains observed in this setting highlight the real-world applicability of the proposed model. It proves capable of classifying terrain accurately even when faced with unfamiliar regions and previously unseen sensor configurations, making it highly suitable for large-scale, deployment-ready PolSAR classification systems.

5.5. Per-class accuracy analysis

Beyond global metrics such as overall accuracy and Kappa coefficient, evaluating per-class accuracy provides critical insights into a model's ability to generalize across different terrain types – particularly in the presence of class imbalance and label distribution shifts.

In PolSAR terrain classification, this analysis is important as some of the categories (e.g. water) are typically overrepresented, whereas others (e.g. oriented or low-density buildings) are underrepresented and subject to misclassification.

Table 5 through 7 show consistent trends across all experimental units of region, sensor, and combined shifts. Existing CV-CNN models show a significant bias towards the dominant classes, which are water classes, and the performance on vegetation and building subtypes is not well generalized. The performance of DAAN, MCC, and PSCAN on minority classes is improved to some extent but also suffers on maintaining a uniform classification profile in the presence of severe domain shifts.

On the other hand, the proposed dual stream vision transformer achieves very balanced performance across all terrain types. For example, in SG-FR unit (Table 7), it shows over 99% accuracy in each of the three main classes (Water, Vegetation, and Building) and existing models have significant drop for building class. In the same way, the proposed model also achieves high accuracy on more complex categories like HD-orthogonal, LD-orthogonal, and Oriented buildings, which are often confused by other methods in cross-sensor units like SA-SG. From the dual stream design of the proposed architecture, it is able to provide resilience to intra class variability and under represented categories. The spatially similar regions are well brought together by the SimPool+ stream and still manage to be distinguished because of global scene context information that the stream captures, whereas the role of the ResMLP+ stream is to reinforce the local texture and structural patterns that are important for distinguishing closely related classes like building types.

These results affirm the model's robustness not only in general domain adaptation but also in achieving semantic consistency across fine-grained terrain labels. The importance is accentuated for applications where class-wise reliability matters as in disaster mapping, urban planning, environmental monitoring, etc.

5.6. Summary of results

The comprehensive experimental results presented across the various domain adaptation scenarios affirm the superior performance and robustness of the proposed dual-stream vision transformer for PolSAR terrain classification. The model is consistently superior to state-of-the-art domain adaptation techniques CV-CNN, DAAN, MCC, PSCAN, and CDFNet across all adaptation units (cross sensor, cross region, and combined domain). In cross-sensor experiments, the model demonstrates its ability to handle spectral discrepancies arising from different radar bands and sensor geometries, achieving over 99% OA and maintaining high accuracy across all land cover classes. The proposed model shows exceptional generalization in cross-region tasks where there was a significant difference in semantic and spatial distribution in the urban and the agricultural landscapes and preserves class wise accuracy and structure consistently. For example, in perhaps the most challenging real-world condition of simultaneous region and sensor adaptation, the model achieves near ideal classification performance and thus sets a new benchmark for real-world unsupervised PolSAR domain adaptation.

Ablation study proves the individual and synergistic contribution of SimPool+ and ResMLP+ streams. SimPool+ enables global context modelling, ResMLP+ improves local spectral – spatial sensitivity, and their fusion achieves the best performance in

all metrics. Additionally, the per class analysis shows that the model is able to keep semantic consistency in classes having fewer training samples or higher intra class variability. Overall, the results underscore the efficacy of combining lightweight attention mechanisms with MLP-based spatial modelling in a dual-stream framework. The proposed model is thus shown to be highly capable and generalizable for PolSAR terrain classification under domain shift across heterogeneous sensing environments, diverse geographic terrains, and complex land cover structures via this architectural strategy.

5.7. Computational efficiency analysis

To evaluate the practicality of the proposed dual-stream vision transformer framework for real-world PolSAR applications, we conducted an in-depth analysis of its computational complexity, memory usage, and inference time. The dual-stream model avoids the quadratic complexity of traditional multi-head self-attention by employing SimPool+, which reduces the attention calculation to a linear operation via adaptive pooling. SimPool+ has complexity $\mathcal{O}(N \cdot d)$, where N is the number of tokens and d is the embedding dimension. ResMLP+ employs affine and channel-wise MLPs with SE blocks, maintaining near-linear scaling with respect to input size.

In addition, the model maintains a compact memory footprint due to the use of lightweight modules (SimPool+, ResMLP+) and the absence of deep attention stacks. On a standard RTX 3090 GPU with 24 GB memory, training with batch size 8 and patch size 16×16 utilized only ~ 5.4 GB of GPU memory. Moreover, the model converges within ~ 40 epochs for each domain adaptation unit. On average, training took approximately 2.3 hours per domain pair. Inference on a full-size PolSAR scene takes ~ 3.7 seconds, making it suitable for large-scale terrain analysis. These findings demonstrate that the proposed architecture offers a strong trade-off between accuracy and computational cost, making it well-suited for scalable and deployable PolSAR classification systems.

5.8. Limitations

Despite its strong performance across various domain shift scenarios, the proposed dual-stream vision transformer framework has some notable limitations:

- Sensitivity to spatial resolution disparities: The model's effectiveness may decline when there are significant mismatches in spatial resolution between the source and target domains, which remains unaddressed in the current study.
- Reliance on decomposition-based features: The framework uses polarimetric decomposition as input, which requires prior processing and may introduce inconsistencies under noisy or incomplete data conditions.
- Limited interpretability of learned features: While dual-stream fusion improves accuracy, the interpretability of intermediate features (especially in domain shift scenarios) remains limited and could benefit from explainable AI techniques.
- Computational limitations for large-scale scenes: Although the model is lightweight, inference on full-size high-resolution satellite images may be constrained by memory and computation, especially when deployed in real-time systems.

6. Conclusion

In this paper, we proposed a novel dual-stream vision transformer framework for unsupervised PolSAR terrain classification under domain shift conditions. Our model integrates a SimPool+ transformer stream for capturing global contextual dependencies and a ResMLP+ stream for modelling local spectral – spatial relationships. This architecture is specifically tailored to address the challenges of domain adaptation in PolSAR data, such as sensor heterogeneity, regional variability, and class imbalance. We conducted extensive experiments across ten sources – target domain adaptation units using four benchmark PolSAR datasets. These included cross-sensor, cross-region, and combined shift scenarios, enabling a comprehensive evaluation of model generalization. On all metrics (overall accuracy, average accuracy, and Kappa coefficient), the proposed method was consistently superior to state-of-the-art UDA models such as CV-CNN, DAAN, MCC, PSCAN, and CDFNet. Our ablation studies confirmed the complementary strengths of the two streams and their synergistic fusion. In addition, per class analysis showed the model's robustness for minority and spatially ambiguous classes like low density and oriented buildings. Overall, this work contributes a scalable, data-efficient, and semantically consistent solution for PolSAR terrain classification in real-world scenarios where target domain labels are unavailable.

In future work, we plan to enhance our dual-stream architecture by incorporating advanced segmentation parameter estimation techniques such as Kurtosis wavelet energy (Akbarizadeh 2012) and skewness wavelet energy (Tirandaz and Akbarizadeh 2015). These wavelet-based descriptors capture higher-order statistical moments, enabling compact yet discriminative modelling of spatial – textural structures. Integrating them could significantly improve the model's sensitivity to nuanced scattering and terrain variations, especially under cross-domain settings involving complex land cover transitions. Furthermore, while this work focuses on PolSAR imagery, the modular architecture of the dual-stream transformer – particularly its reliance on global attention and local spectral – spatial modelling – is potentially adaptable to other remote sensing modalities. For instance, non-polarimetric SAR data may benefit from modified input feature extraction, and hyperspectral imagery could be incorporated by adapting the tokenization and local stream to handle high spectral dimensionality. Future work will explore these extensions and validate the model's versatility beyond PolSAR. Another promising direction involves self-supervised pretraining tailored for PolSAR data. By leveraging unlabelled source and target data to learn invariant representations through contrastive or masked modelling objectives, we can reduce reliance on annotated datasets and improve downstream classification accuracy in label-scarce domains.

Abbreviations

| Abbreviation | Explanation |
|--------------|--|
| AA | Average accuracy |
| CDFNet | Cross-domain feature fusion network |
| CV-CNN | Complex-valued convolutional neural networks |
| DAAN | Domain adaptive attention network |
| DANN | Domain-adversarial neural network |
| DBS | Domain balanced sampling |

| | |
|----------|--|
| Fle-RS2 | Flevoland – RADARSAT-2 |
| GELU | Gaussian error linear unit |
| HD | High-density |
| LD | Low-density |
| LiDAR | Light Detection and Ranging |
| MCC | Minimum class confusion |
| MHSA | Multi-head self-attention |
| OA | Overall accuracy |
| PolSAR | Polarimetric synthetic aperture radar |
| PSCAN | Progressive semantic context-aware network |
| RADARSAT | Radar satellite platform |
| ResMLP | Residual multi-layer perceptron |
| SAR | Synthetic aperture radar |
| SE | Squeeze-and-excitation |
| SF-ALOS2 | San Francisco – ALOS-2 |
| SF-GF3 | San Francisco – Gaofen-3 |
| SF-RS2 | San Francisco – RADARSAT-2 |
| UDA | Unsupervised domain adaptation |
| ViT | Vision transformer |

Acknowledgments

The authors thank the management of Jayaraj Annapackiam CSI College of Engineering, Nazareth, Tamil Nadu for providing infrastructural support.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

A. Rega  <http://orcid.org/0009-0003-4252-8970>

E. Fantin Irudaya Raj  <http://orcid.org/0000-0003-2051-3383>

Data availability statement

Data will be made available on request.

References

- Aghaei, N., G. Akbarizadeh, and A. Kosarian. 2022a. "Osdes_Net: Oil Spill Detection Based on Efficient_Shuffle Network Using Synthetic Aperture Radar Imagery." *Geocarto International* 37 (26): 13539–13560. <https://doi.org/10.1080/10106049.2022.2082545>.
- Aghaei, N., G. Akbarizadeh, and A. Kosarian. 2022b. "Greywolflsm: An Accurate Oil Spill Detection Method Based on Level Set Method from Synthetic Aperture Radar Imagery." *European Journal of Remote Sensing* 55 (1): 181–198. <https://doi.org/10.1080/22797254.2022.2037468>.
- Ahmad, M., M. H. F. Butt, A. M. Khan, M. Mazzara, S. Distefano, M. Usama, and D. Hong. 2025. "Spatial–Spectral Morphological Mamba for Hyperspectral Image Classification." *Neurocomputing* 636:C 129995.

- Akbarizadeh, G. 2012. "A New Statistical-Based Kurtosis Wavelet Energy Feature for Texture Recognition of SAR Images." *IEEE Transactions on Geoscience & Remote Sensing* 50 (11): 4358–4368. <https://doi.org/10.1109/TGRS.2012.2194787>.
- Akbarizadeh, G., and M. Rahmani. 2015. "A New Ensemble Clustering Method for PolSAR Image Segmentation." In *2015 7th Conference on Information and Knowledge Technology (IKT) Urmia, Iran* Suresh Chandra Satapathy Amit Joshi Nilesh Modi Nisarg Pathak, 1–4. IEEE. May.
- Akbarizadeh, G., and M. Rahmani. 2017. "Efficient Combination of Texture and Color Features in a New Spectral Clustering Method for PolSAR Image Segmentation." *National Academy Science Letters* 40 (2): 117–120. <https://doi.org/10.1007/s40009-016-0513-6>.
- Alkhatib, M. Q., M. Al-Saad, N. Aburaed, M. S. Zitouni, and H. Al-Ahmad. 2023. "POLSAR Image Classification Using Attention Based Shallow to Deep Convolutional Neural Network." In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* 16/07/23 - 21/07/23 Pasadena, United States Gerardo Di Martino, 8034–8037. IEEE.
- Chen, P., Y. Ren, B. Zhang, and Y. Zhao. 2025. "Class Imbalance in the Automatic Interpretation of Remote Sensing Images: A Review." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 18:9483–9508. <https://doi.org/10.1109/JSTARS.2025.3555567>.
- Davari, N., G. Akbarizadeh, and E. Mashhour. 2021. "Corona Detection and Power Equipment Classification Based on Googlenet-Alexnet: An Accurate and Intelligent Defect Detection Model Based on Deep Learning for Power Distribution Lines." *IEEE Transactions on Power Delivery* 37 (4): 2766–2774. <https://doi.org/10.1109/TPWRD.2021.3116489>.
- Dong, H., L. Si, W. Qiang, W. Miao, C. Zheng, Y. Wu, and L. Zhang. 2023. "A Polarimetric Scattering Characteristics-Guided Adversarial Learning Approach for Unsupervised PolSAR Image Classification." *Remote Sensing* 15 (7): 1782. <https://doi.org/10.3390/rs15071782>.
- Fang, X., C. He, Q. Zhang, and M. Tong. 2024. "POLSAR Image Classification Framework with POA Align and Cyclic Channel Attention." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 17:10203–10220. <https://doi.org/10.1109/JSTARS.2024.3400409>.
- Fu, Y., Z. Zhu, L. Liu, W. Zhan, T. He, H. Shen, and Z. Ao. 2024. "Remote Sensing Time Series Analysis: A Review of Data and Applications." *Journal of Remote Sensing* 4:1 0285.
- Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and V. Lempitsky. 2016. "Domain-Adversarial Training of Neural Networks." *Journal of Machine Learning Research* 17 (59): 1–35.
- Ghara, F. M., S. B. Shokouhi, and G. Akbarizadeh. 2022. "A New Technique for Segmentation of the Oil Spills from Synthetic-Aperture Radar Images Using Convolutional Neural Network." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 15:8834–8844. <https://doi.org/10.1109/JSTARS.2022.3213768>.
- Huang, Z., H. Yan, Q. Zhan, S. Yang, M. Zhang, C. Zhang, and Y. Wang. 2025. "A Survey on Remote Sensing Foundation Models: From Vision to Multimodality 2503 3 1–29 doi:<https://doi.org/10.48550/arXiv.2503.22081>." *arXiv preprint arXiv:2503.22081*.
- Huo, C., K. Chen, S. Zhang, Z. Wang, H. Yan, J. Shen, and Z. Wang. 2025. "When Remote Sensing Meets Foundation Model: A Survey and Beyond." *Remote Sensing* 17 (2): 179. <https://doi.org/10.3390/rs17020179>.
- Jamali, A., S. K. Roy, B. Lu, L. H. Beni, N. Kakhani, J. Chanussot, and P. Ghamisi. 2025. "MSHCT: A Multi-Scale Compact Convolutional Network for High Resolution Aerial Scene Classification IEEE Geoscience and Remote Sensing Letters 22 1 5001205 doi:[10.1109/LGRS.2025.3556373](https://doi.org/10.1109/LGRS.2025.3556373)."
- Jin, Y., X. Wang, M. Long, and J. Wang. 2020. "Minimum Class Confusion for Versatile Domain Adaptation Andrea Vedaldi Horst Bischof Thomas Brox Jan-Michael Frahm." In *Computer Vision–ECCV 2020: 16th European Conference*, 464–480. Glasgow, UK: Springer International Publishing. August 23–28, 2020, Proceedings, Proceedings.
- Lang, C., G. Cheng, J. Wu, Z. Li, X. Xie, J. Li, and J. Han. 2024. "Toward Open-World Remote Sensing Imagery Interpretation: Past, Present, and Future." *IEEE Geoscience and Remote Sensing Magazine*.
- Liu, X., L. Jiao, F. Liu, D. Zhang, and X. Tang. 2022. "POLSF: Polsar Image Datasets on San Francisco." In *International Conference on Intelligence Science*, 214–219. Cham: Z Shi Y Jin X Zhang Springer International Publishing. October.

- Parida, B. R., and S. P. Mandal. 2020. "Polarimetric Decomposition Methods for LULC Mapping Using ALOS L-Band PolSAR Data in Western Parts of Mizoram, Northeast India." *SN Applied Sciences* 2 (6): 1049. <https://doi.org/10.1007/s42452-020-2866-1>.
- Parikh, H., S. Patel, and V. Patel. 2020. "Classification of SAR and PolSAR Images Using Deep Learning: A Review." *International Journal of Image and Data Fusion* 11 (1): 1–32. <https://doi.org/10.1080/19479832.2019.1655489>.
- Psomas, B., I. Kakogeorgiou, K. Karantzas, and Y. Avrithis. 2023. "Keep It Simpool: Who Said Supervised Transformers Suffer from Attention Deficit?" In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 01.10.2023 - 06.10.2023 Paris, France Frédéric Jurie Gaurav Sharma (IEEE), 5350–5360 doi:10.1109/ICCV51070.2023.00493.
- Ren, Z., Z. Du, Y. Zhang, F. Sha, W. Li, and B. Hou. 2024. "Multi-Step Unsupervised Domain Adaptation in Image and Feature Space for Synthetic Aperture Radar Image Terrain Classification." *Remote Sensing* 16 (11): 1901. <https://doi.org/10.3390/rs16111901>.
- Roy, S. K., A. Jamali, J. Chanussot, P. Ghamisi, E. Ghaderpour, and H. Shahabi. 2025. "Simpoolformer: A Two-Stream Vision Transformer for Hyperspectral Image Classification." *Remote Sensing Applications: Society & Environment* 37:101478. <https://doi.org/10.1016/j.rsase.2025.101478>.
- Samadi, F., G. Akbarizadeh, and H. Kaabi. 2019. "Change Detection in SAR Images Using Deep Belief Network: A New Training Approach Based on Morphological Images." *IET Image Processing* 13 (12): 2255–2264. <https://doi.org/10.1049/iet-ipr.2018.6248>.
- Shang, R., J. Wang, L. Jiao, X. Yang, and Y. Li. 2022. "Spatial Feature-Based Convolutional Neural Network for PolSAR Image Classification." *Applied Soft Computing* 123:108922. <https://doi.org/10.1016/j.asoc.2022.108922>.
- Sharifzadeh, F., G. Akbarizadeh, and Y. Seifi Kavian. 2019. "Ship Classification in SAR Images Using a New Hybrid CNN–MLP Classifier." *The Journal of the Indian Society of Remote Sensing* 47 (4): 551–562. <https://doi.org/10.1007/s12524-018-0891-y>.
- Takahashi, S., Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, and R. Hamamoto. 2024. "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review." *Journal of Medical Systems* 48 (1): 84.
- Tao, L., H. Zhang, H. Jing, Y. Liu, D. Yan, G. Wei, and X. Xue. 2025. "Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques." *Remote Sensing* 17 (1): 162. <https://doi.org/10.3390/rs17010162>.
- Tirandaz, Z., and G. Akbarizadeh. 2015. "A Two-Phase Algorithm Based on Kurtosis Curvelet Energy and Unsupervised Spectral Regression for Segmentation of SAR Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 9 (3): 1244–1264. <https://doi.org/10.1109/JSTARS.2015.2492552>.
- Tirandaz, Z., G. Akbarizadeh, and H. Kaabi. 2020. "POLARS Image Segmentation Based on Feature Extraction and Data Compression Using Weighted Neighborhood Filter Bank and Hidden Markov Random Field-Expectation Maximization." *Measurement* 153:107432. <https://doi.org/10.1016/j.measurement.2019.107432>.
- Touvron, H., P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, and H. Jégou. 2022. "ResMLP: Feedforward Networks for Image Classification with Data-Efficient Training." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45 (4): 5314–5321.
- Wang, N., W. Jin, H. Bi, C. Xu, and J. Gao. 2024. "A Survey on Deep Learning for Few-Shot PolSAR Image Classification." *Remote Sensing* 16 (24): 4632. <https://doi.org/10.3390/rs16244632>.
- Wang, S., Z. Sun, T. Bian, Y. Guo, L. Dai, Y. Guo, and L. Jiao. 2025. "CDFNet: Cross-Domain Feature Fusion Network for PolSAR Terrain Classification IEEE Transactions on Geoscience and Remote Sensing 63 1 5200215 doi:10.1109/TGRS.2024.3506927."
- Wang, X., X. Jin, Z. Dai, Y. Wu, and A. Chehri. 2025. "Deep Learning-Based Methods for Road Extraction from Remote Sensing Images: A Vision, Survey, and Future Directions." *IEEE Geoscience and Remote Sensing Magazine*.
- Wang, Y., Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang. 2025. "Vision Transformers for Image Classification: A Comparative Survey." *Technologies* 13 (1): 32. <https://doi.org/10.3390/technologies13010032>.

- Yan, K., C. Han, and X. Hou. 2025. "Research on the Current Status of Remote Sensing Image Land Classification Information Retrieval Technology Based on Deep Learning." *The International Conference Optoelectronic Information and Optical Engineering (OIOE2024)* (Vol. 13513) 28.02.2025 - 02.03.2025 Dali, China Yang Yue Ming Jiang Qingyang Wei, 137–141 doi:<https://doi.org/10.1117/12.3045361>. SPIE. January.
- Yang, D., X. Gao, Y. Yang, K. Guo, K. Han, and L. Xu. 2025. "Advances and Future Prospects in Building Extraction from High-Resolution Remote Sensing Images." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 18:6994–7016. <https://doi.org/10.1109/JSTARS.2025.3538662>.
- Zhang, F., X. Sun, F. Ma, and Q. Yin. 2024. "Superpixelwise Likelihood Ratio Test Statistic for PolSAR Data and Its Application to Built-Up Area Extraction." *ISPRS Journal of Photogrammetry & Remote Sensing* 209:233–248. <https://doi.org/10.1016/j.isprsjprs.2024.02.009>.