remote sensing

MDPI

Article

A Complex-Valued Multiscale Feature Interaction Attention Vision Transformer for PolSAR Image Classification with Amplitude–Phase Decoupling

Yingying Kong * and Dongyu Yuan

College of Electrical and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; sz2404061@nuaa.edu.cn
* Correspondence: yayako_zy@nuaa.edu.cn

Highlights

What are the main findings?

- The CV-MfIaViT combines a spatial analytic-envelope front end, multiscale complex-valued feature extraction and cross-pooling attention. The attention applies positive real gains that leave the phase of its non-zero input features unchanged. The complete model attains the highest reported mean overall accuracy (OA) and average accuracy (AA) among the compared methods on the three scenes.
- The model retains its accuracy advantage under spatial-block partitioning with shared label-free scene-level normalization. This complementary evaluation tests performance across separated regions without isolating the causal contribution of individual components.

What are the implications of the main findings?

- Cross-pooling positive real attention provides a practical phase-preserving modulation of learned complex features; its classification contribution is evaluated as part of the combined attention design.
- Reporting random-pixel and spatial-block results together makes the dependence of patch-based classification on the sampling protocol more visible, particularly under limited labeled data.

Abstract

Limited labeled data and class imbalance complicate polarimetric synthetic aperture radar (PolSAR) image classification. We propose a complex-valued multiscale feature interaction attention vision transformer (CV-MfIaViT) that combines spatial analytic-envelope preprocessing, complex-valued local feature extraction and global modeling. A two-dimensional Hilbert transform (2D HT) provides envelope features to the shallow pathway. The complex-valued amplitude–phase interaction attention (CV-APIA) module generates positive real gains through cross-modulation of pooled features; the final gating preserves the element-wise phase of non-zero input features. A complex-valued feature complementary mapper (CV-FCM), a complex-valued multiscale extractor (CV-MSE) and a complex-valued vision transformer (CV-ViT) complete the architecture. Under random-pixel sampling, the reported average accuracy (AA) exceeds CV-MsAtViT by 1.83, 3.42 and 2.48 percentage points on Flevoland, San Francisco and Nanjing, respectively. The advantage is also retained under spatial-block partitioning with shared label-free scene-level normalization. Component ablations and comparisons with two simple learned front ends characterize the combined design. The results support the evaluated architecture without isolating phase preservation as the cause of its accuracy gains.

Check for updates

Academic Editor: Fabio Rocca

Received: 2 September 2026
Revised: 22 October 2026
Accepted: 31 October 2026
Published: 2 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: complex-valued neural network; interaction attention; feature fusion; polarimetric synthetic aperture radar (PolSAR); vision transformer; two-dimensional Hilbert transform; amplitude–phase decoupling

1. Introduction

Polarimetric synthetic aperture radar (PolSAR) records complex-valued measurements across transmit and receive polarization states, typically horizontal (H) and vertical (V), and provides information about target scattering. It supports applications such as urban mapping [1] and retrieval of subsurface properties [2]. Related SAR research also addresses small-ship instance segmentation [3], moving-target shadow detection in video SAR [4] and ship-target refocusing [5]; these target-level tasks provide application context but are distinct from the pixel-level PolSAR classification studied here. Convolutional neural networks (CNNs) learn discriminative spatial features [6], while reliable PolSAR annotation remains labor-intensive [7]. Limited labels and uneven class support therefore motivate the present study.

PolSAR classification research combines classifier design with feature extraction. The classical examples include support vector machines [8], decision-tree classification [9] and active-learning/random-forest classification [10]. Their performance depends on the feature representation and classifier design.

Classical feature extraction includes statistical descriptors and polarimetric target decompositions. Phenomenological descriptions and decomposition reviews provide their physical background [11,12]; examples include entropy/alpha descriptors [13] and Krogager, Touzi and Freeman–Durden formulations [14–16]. Selecting features remains application-dependent, which motivates learned representations without replacing the definitions of these physical observables.

Models that combine CNNs with attention mechanisms have shown clear effectiveness in PolSAR image classification [17]. Early deep convolutional networks established the advantage of deep convolution for extracting discriminative spatial features [6,18] and laid the groundwork for CNN-based PolSAR classification [19], producing strong results with 2D CNN [20], 3D CNN [21] and 2D–3D CNN [22] architectures. More recently, the vision transformer (ViT) introduced a global representation paradigm by modeling long-range dependencies [23], and hybrid CNN–ViT architectures extract local and global features in parallel; PolSARFormer [24] is a representative example, balancing local detail against global semantics and improving the recognition of multiscale targets under limited supervision. Most recently, the complex-valued multiscale attention vision transformer (CV-MsAtViT) [25] has further embedded multiscale attention into a complex-valued ViT and achieved strong results on PolSAR benchmarks.

Attention mechanisms reweight feature channels or spatial positions. Examples include squeeze-and-excitation (SE) networks [26], efficient channel attention (ECA) [27], the convolutional block attention module (CBAM) [28] and coordinate attention (CA) [29]. Masked autoencoding and self-supervised pretraining offer additional strategies for learning from limited labels [30]. Their use with complex PolSAR representations requires an explicit definition of how magnitudes, phases and modulation weights interact.

A persistent problem is class imbalance: minority classes, such as small-scale targets and weakly scattering land-cover types, occupy a very small share of the samples yet carry high practical value [31]. Conventional attention favors dominant features and overlooks the evidence that distinguishes weak minority classes, raising both false alarms and missed detections. Processing real and imaginary components with different gains or Cartesian

nonlinearities can change the phase of a learned complex feature. Positive real modulation offers an alternative that leaves the phase of a non-zero input feature unchanged. This algebraic property motivates the attention design but does not establish that any resulting accuracy gain is caused by phase preservation.

We propose a complex-valued framework with two representation choices: a spatial two-dimensional Hilbert transform (2D HT) supplies analytic-envelope features to the shallow pathway, and complex-valued amplitude–phase interaction attention (CV-APIA) applies cross-pooling magnitude modulation while passing through the input feature phase. These components are combined with multiscale feature extraction and a complex-valued vision transformer (CV-ViT).

The main contributions are as follows.

- We introduce a complex-valued multiscale feature interaction attention vision transformer (CV-MfIaViT), combining deterministic spatial envelope preprocessing with a complex-valued feature complementary mapper (CV-FCM), a complex-valued multiscale extractor (CV-MSE) and a transformer classification head.
- We construct cross-pooling CV-APIA attention whose positive real output gate preserves the phase of each non-zero input feature. The pooling interaction and gain construction instantiate an existing magnitude–phase separation principle.
- We evaluate the complete model on three scenes using random-pixel and spatial-block protocols and report component ablations, learned-front-end comparisons and input-window sensitivity. These experiments describe architectural performance and its limits of attribution.

1.1. Related Work

1.1.1. PolSAR Data Processing

Each pixel of polarimetric synthetic aperture radar (PolSAR) data can be represented by a scattering matrix $\mathbf{S}$, which describes the linear transformation between the Jones vectors of the incident and scattered waves,

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix}, \tag{1}$$

where S_{AB} $(A, B \in \{H, V\})$ denotes the backscattering coefficient of the electromagnetic wave transmitted in polarization state A and received in state B. Under the backscattering reciprocity condition, $S_{VH} = S_{HV}$ generally holds, so $\mathbf{S}$ is symmetric and can be reduced to a three-dimensional scattering vector $\vec{k}$ through Pauli decomposition [12]:

$$\vec{k} = \frac{1}{\sqrt{2}}\left[S_{HH} + S_{VV}, \quad S_{HH} - S_{VV}, \quad 2S_{HV}\right]^T. \tag{2}$$

To improve statistical stability, PolSAR data are usually multilook-processed; the coherency matrix $\mathbf{T}$ estimates the second-order moment of the scattering vector by averaging outer products across looks,

$$\mathbf{T} = \frac{1}{n}\sum_{j=1}^{n} \vec{k}_j \vec{k}_j^H = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}, \tag{3}$$

where n is the number of looks and $(\cdot)^H$ denotes the conjugate transpose. Since $\mathbf{T}$ is Hermitian, its diagonal elements are real and its off-diagonal elements complex, yielding six independent components $T_{11}, T_{22}, T_{33}, T_{12}, T_{13}$ and T_{23} as input features.

Each coherency-derived channel is first centered and scaled using scene-level statistics. Spatial patches are then extracted from the normalized channels. The shallow-path 2D HT

and envelope reconstruction operate separately within each patch after this channel-wise normalization and patch extraction. The channel mean T_{mn_avg} and real non-negative standard deviation T_{mn_std} are computed over the whole scene without using class labels. This is numerical feature standardization, not a calibrated polarimetric observable or evidence of Gaussian-distributed inputs. For each real-valued diagonal channel T_{nn}, the statistics are

$$T_{nn_avg} = \frac{\sum_{i=1}^{n_v} T_{nn}(i)}{n_v}, \tag{4}$$

$$T_{nn_std} = \sqrt{\frac{\sum_{i=1}^{n_v} \left(T_{nn}(i) - T_{nn_avg}\right)^2}{n_v}}. \tag{5}$$

For each complex-valued off-diagonal channel T_{mn} $(m \neq n)$, these are computed as

$$T_{mn_avg} = \frac{\sum_{i=1}^{n_v} T_{mn}(i)}{n_v}, \tag{6}$$

$$T_{mn_std} = \sqrt{\frac{\sum_{i=1}^{n_v} \left(T_{mn}(i) - T_{mn_avg}\right)\left(T_{mn}(i) - T_{mn_avg}\right)^*}{n_v}}, \tag{7}$$

Here n_v is the number of spatial samples in the scene used to calculate the channel statistics. For a non-constant channel, normalization uses

$$\widehat{T}_{mn}(i) = \frac{T_{mn}(i) - T_{mn_avg}}{T_{mn_std}}.$$

For off-diagonal channels, the denominator is the square root of the mean squared complex magnitude in Equation (7), so it is a positive real scale shared by the real and imaginary components. The notation $\widehat{T}_{mn}$ denotes standardized coherency-derived feature channels rather than the entries of a new physically calibrated coherency matrix. The network processes these channels as input features; no Wishart likelihood or physical target decomposition is applied to them. Channel-wise complex Z-score standardization of the same six coherency channels is also used by Kuang et al. [32] for PolSAR classification.

Numerical rationale and retained information. Writing $\mu_{mn} = T_{mn_avg}$ and $\sigma_{mn} = T_{mn_std} > 0$, the definitions give

$$\frac{1}{n_v}\sum_i \widehat{T}_{mn}(i) = 0, \qquad \frac{1}{n_v}\sum_i \left|\widehat{T}_{mn}(i)\right|^2 = 1, \qquad T_{mn}(i) = \sigma_{mn}\widehat{T}_{mn}(i) + \mu_{mn}.$$

Thus each non-constant channel has zero empirical mean and unit empirical centered second moment, balancing channel scales before learned filtering without assuming Gaussian inputs. Conditional on the fixed statistics, standardization is an invertible affine change of feature coordinates, not an approximation to the physical coherency matrix. The original values remain recoverable from the standardized features and these statistics in exact arithmetic. Moreover,

$$\widehat{T}_{mn}(i) - \widehat{T}_{mn}(j) = \frac{T_{mn}(i) - T_{mn}(j)}{\sigma_{mn}}.$$

Spatial differences within a channel therefore retain their complex-plane direction and relative magnitudes. This supports using the representation for discriminative learning; it does not prove an improvement in classification accuracy or optimality over other normalizations. The invertibility statement concerns this standardization step only before the subsequent nonlinear envelope construction and network processing.

Relation to PolSAR matrix properties. Real diagonal channels remain real. If a full matrix is reconstructed by setting $\widehat{T}_{nm} = \widehat{T}^*_{mn}$, Hermitian symmetry is retained; equivalently, conjugate-paired channels have $\mu_{nm} = \mu^*_{mn}$ and $\sigma_{nm} = \sigma_{mn}$. Positive semidefiniteness, however, is not generally retained: $T_{nn}(i) < \mu_{nn}$ implies $\widehat{T}_{nn}(i) < 0$, which is incompatible with

a positive-semidefinite coherency matrix. In addition, complex centering can change an entry's phase relative to the origin: for a non-zero centered entry, $\arg \widehat{T}_{mn} = \arg(T_{mn} - \mu_{mn})$, not generally $\arg T_{mn}$. Physical scattering interpretation therefore belongs to $\mathbf{T}$; the standardized tensor is a numerical input representation. Phase preservation by CV-APIA refers only to the learned features entering its final gate.

Under the ideal homogeneous independent-look complex-Gaussian scattering model, multilook covariance/coherency statistics follow a complex-Wishart model [10]; its diagonal intensity marginals are Gamma-distributed. A scene containing different land-cover types need not follow one such marginal distribution globally. We therefore do not assume that the observed $T_{nn}(i)$ values are Gaussian or claim a fitted probability density for these scenes. The distribution changes under centering and scaling; neither the original Wishart/Gamma form nor Gaussianization is claimed for $\widehat{T}_{mn}$. Patches of these standardized features enter the shallow-path analytic-envelope construction in Section 2.2.1. No class labels enter the normalization statistics, but these whole-scene statistics are shared across spatial regions, which limits a strictly inductive interpretation of the block evaluation.

Figure 1 summarizes the representation from the scattering matrix to the Pauli vector and coherency-derived channels. Equations (4)–(7) provide the scene-level statistics for channel-wise normalization. In the shallow pathway, the order is coherency-derived channels, scene-level normalization, patch extraction, and then spatial 2D HT and envelope reconstruction within each patch. The deep pathway uses the corresponding normalized complex-valued patches without this analytic-envelope transformation.

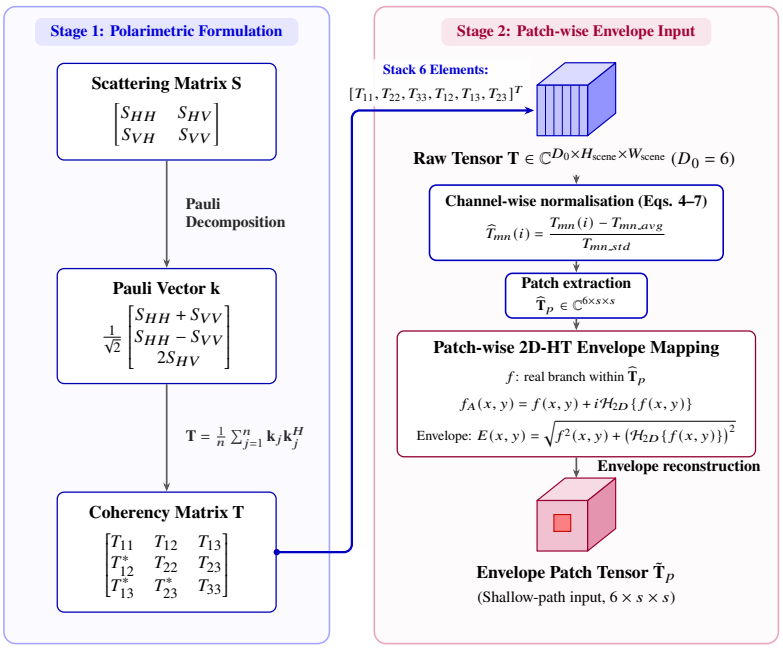


Figure 1. Shallow-path preprocessing from the scattering matrix to network inputs. Channel-wise normalization using Equations (4)–(7) precedes patch extraction. The 2D HT and envelope reconstruction are applied separately within each extracted patch. The deep pathway uses the normalized complex-valued patches without this front end. The standardized and envelope-reconstructed channels are network features, not physical coherency matrices.

1.1.2. Hybrid-Model-Based PolSAR Classification

PolSAR classification includes real- and complex-valued CNNs and hybrid CNN–ViT architectures. Complex-valued convolution is presented by Zhang et al. [19]; two-dimensional CNN classification is presented by Zhou et al. [20]; three-dimensional CNN

classification is presented by Zhang et al. [21]; and a complex-valued 2D–3D hybrid with attention is presented by Li et al. [22]. These examples have distinct dimensional structures.

Transformer-based approaches include the PolSAR evaluation by Dong et al. [33], hybrid patch interaction for dual-frequency data [34] and temporal–polarimetric–spatial modeling [35]. PolSARFormer [24], HybridCVNet [36] and CV-MsAtViT [25] combine local and global modeling in different ways. Our contribution is the specific envelope-front-end and cross-pooling attention combination evaluated below rather than the first use of complex-valued processing or magnitude–phase separation.

1.1.3. Attention Mechanisms

Class imbalance is a long-standing challenge in PolSAR classification. The existing mitigation strategies, such as cost-sensitive latent space learning (CLSL) [37], correct part of the prior imbalance but act mainly at the loss or data level. Attention mechanisms, in contrast, calibrate the feature distribution directly at the architectural level: through channel, spatial or coordinate attention, such as SE, ECA, and CBAM, the network adaptively weights discriminative channels and key locations, amplifying minority-class signals while suppressing interference from dominant classes. In PolSAR-specific studies, attention-based algorithms such as AFS-CNN [38] and attention-augmented multiscale LSTM [39] improve the response to weak targets. Complex attention may use Cartesian processing, magnitude-derived gates or combinations of both. The distinction examined here concerns the cross-modulation of two pooled feature branches and a positive real output gate. The general phase-preserving rescaling principle is related to existing complex activations and is discussed explicitly in Section 2.3. We evaluate the combination with multiscale complex-valued feature extraction without asserting that all earlier attention mechanisms lack magnitude–phase separation.

2. Materials and Methods

2.1. Architecture of the CV-MfIaViT Model

The proposed model is designed for PolSAR image classification; its architecture appears in Figure 2. The CV-MfIaViT framework consists of a complex-valued feature-mapping mechanism, the CV-APIA attention block tailored for complex-valued minority-class data, a complex-valued ViT (Section 2.4), and fully connected layers for classification.

Before dual-path feature extraction, the coherency-derived channels are normalized using Equations (4)–(7). Patches are extracted from the normalized channels before the pathways diverge. The shallow pathway applies the 2D HT and envelope reconstruction within each patch followed by two 3D CNN layers. The deep pathway uses the corresponding normalized complex-valued patches directly and two 3D CNN layers. The feature complementary mapper and multiscale extractor refine the extracted features, which are passed to CV-APIA, the complex-valued ViT block (Section 2.4) and the classification head. The role and limits of the shallow-path front end are described in Section 2.2.1.

2.2. Complex-Valued Feature Mapping Mechanism

Architectures based on 2D CNNs or 3D CNNs are widely used in PolSAR classification because they capture spatial features effectively, but they have limitations in multi-dimensional feature extraction and fine-grained information capture. To overcome these, the proposed complex-valued feature-mapping mechanism adopts a dual-path parallel convolutional design, which extracts spatial and polarimetric features jointly through complex-valued convolutional filters. Such filters maintain a complex representation but do not generally leave an input phase unchanged. The 3D CNN performs convolution along

the height, width and scattering dimensions so that spatial context and the dependencies along the scattering dimension are extracted simultaneously.

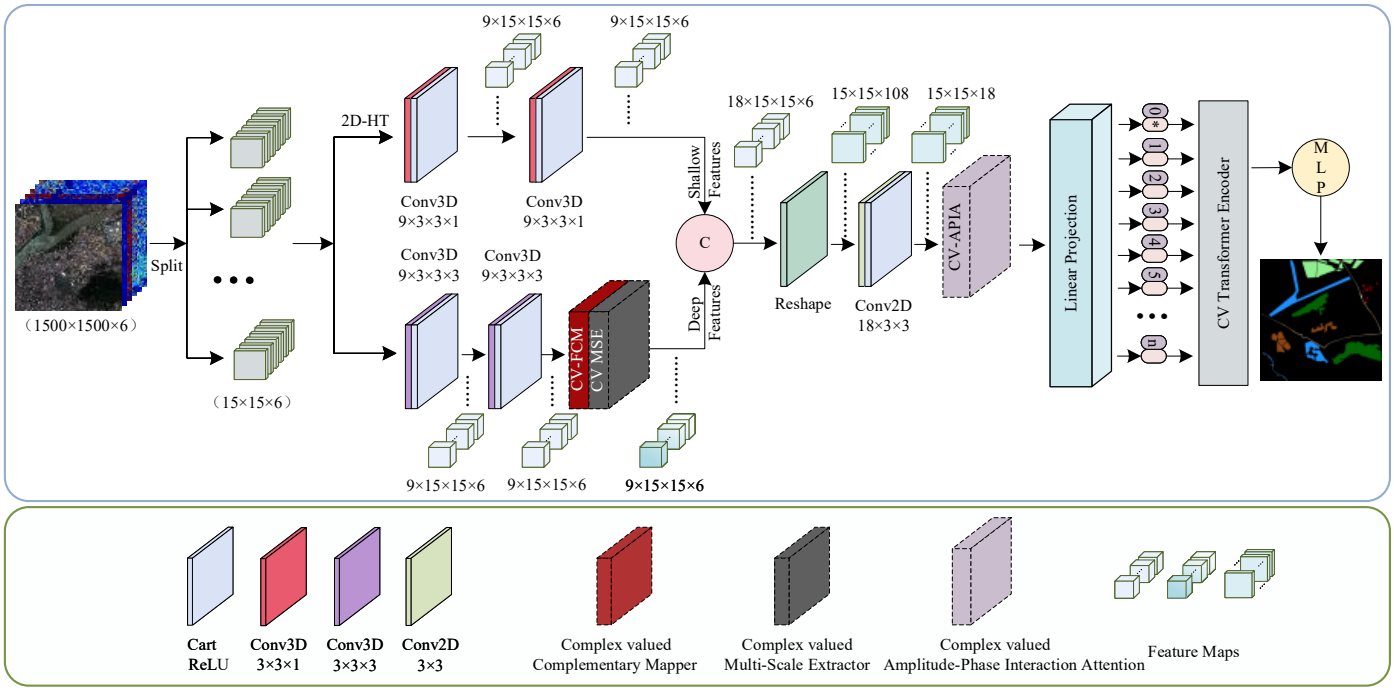


Figure 2. Overall framework of the proposed CV-MfIaViT method. Patch extraction precedes the shallow-path 2D HT; channel normalization is shown in Figure 1.

For a single training sample, the enhanced representation $\tilde{\mathbf{T}}$ forms a complex-valued volume of shape $D_0 \times H \times W$, where the polarimetric depth axis $D_0 = 6$ stacks six envelope-derived channels corresponding to the independent coherency components (the three diagonal envelope channels are stored as complex values with zero imaginary part) and $H \times W$ is the spatial patch size. The 3D CNN stages convolve this volume jointly along the height, width and polarimetric depth axes, producing the five-dimensional feature tensor $X \in \mathbb{C}^{B\times D\times H'\times W'\times C}$, where D is the polarimetric depth after convolution, $H' \times W'$ the spatial size and C the number of filters. We refer to the combination of the two refinement modules introduced below, the complex-valued feature complementary mapper (CV-FCM) and the complex-valued multiscale extractor (CV-MSE), as the feature-mapping mechanism (FM).

Specifically, for an input image X of size $H_1 \times W_1 \times C_1$ and a convolution kernel K of size $H_2 \times W_2 \times C_2$, the output of the 3D convolution at position (x, y, z) is

$$\tilde{Y}(x,y,z) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} \sum_{k=-c}^{c} K(i,j,k)\, X(x+i,\, y+j,\, z+k) + b_0, \tag{8}$$

where the half-sizes are $a = \lfloor H_2/2 \rfloor$, $b = \lfloor W_2/2 \rfloor$ and $c = \lfloor C_2/2 \rfloor$ so that the convolutional kernel is centered at (x, y, z); b_0 is the additive bias and is distinct from the kernel half-width b. The expression specifies the corresponding discretely correlated response rather than a flipped convolution. The symbol $\tilde{Y}$ denotes the real-valued output of the 3D convolution and is deliberately distinguished from the complex convolution output Y of Equation (9). This operation lets the network perceive spatial neighborhoods and inter-channel patterns at the same time.

PolSAR images are stored in complex form, and their phase carries important physical scattering characteristics. Complex-valued convolutional neural networks (CV-CNNs) [40] are therefore introduced to process complex data directly, and they have delivered strong performance in PolSAR classification. Figure 3 shows that, to exploit both the real and

imaginary parts of the complex data, this work adopts a complex-valued 3D CNN (CV-3D CNN). Let the complex image be $X = X_r + iX_i$ and the complex kernel be $K = K_r + iK_i$, where X_r, X_i, K_r, K_i are real tensors and i is the imaginary unit. The complex convolution output Y is

$$Y = X * K = (X_r * K_r - X_i * K_i) + i(X_r * K_i + X_i * K_r), \tag{9}$$

where $*$ denotes the convolution operator. This operation combines the convolution results of the real and imaginary parts to produce complex-valued feature maps. The argument of the output may differ from that of the input because complex convolution combines multiple weighted contributions.

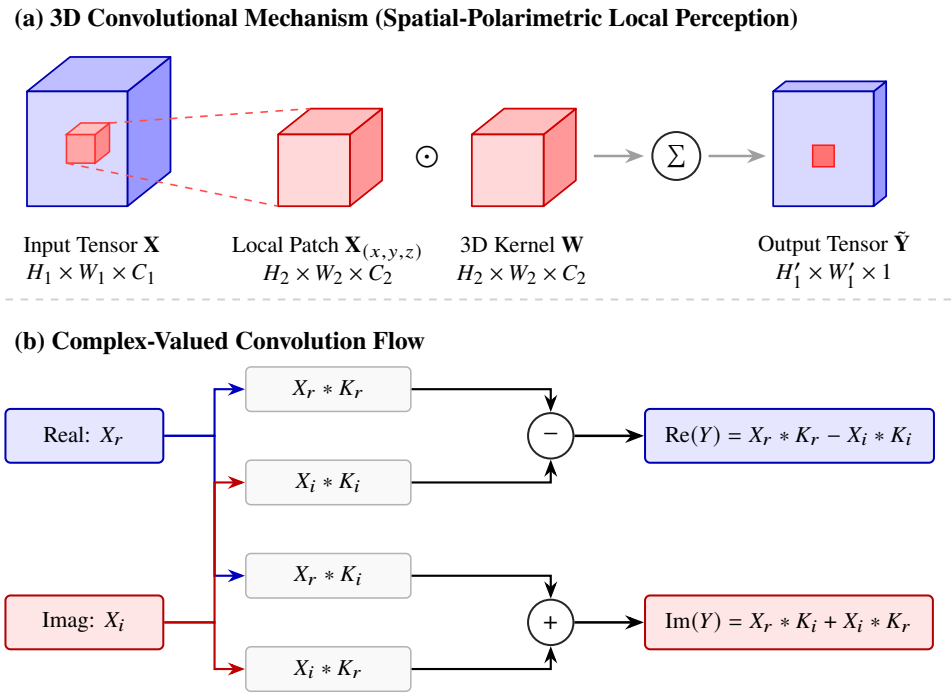


Figure 3. Illustration of 3D convolution and complex-valued convolution.

The shallow feature-extraction pathway uses two 3D convolutional layers with small $(3 \times 3 \times 1)$ filters, which capture basic spatial information efficiently and extract fine-grained local features at low computational cost, laying the foundation for subsequent deep feature extraction [41]. The deep pathway uses two $(3 \times 3 \times 3)$ 3D convolutional filters, extracting spatial and polarimetric features jointly. Combined with the CV-FCM and CV-MSE modules, complex-valued spatial and channel attention ensure that the polarimetric features retain their detail without the information bias caused by channel-count disparities, helping the model to deal with feature-class imbalance—particularly in minority-class classification.

2.2.1. Two-Dimensional Hilbert Transform for Spatial-Envelope Features

The shallow pathway forms spatial analytic-envelope features from normalized coherency-derived channels. A spatial HT supplies the quadrature component; its nonlinear combination with the input yields an amplitude descriptor. The associated spatial oscillation phase is distinct from the polarimetric phase and does not define a new scattering observable.

After scene-level normalization and patch extraction, each original $s \times s$ patch is transformed directly using the two-dimensional discrete Fourier transform (DFT) and its inverse, implemented by the fast Fourier transform (FFT), without additional HT-specific padding or windowing. This is separate from scene-edge padding during patch extraction. Branch-

wise envelopes replace the original polarimetric phase in this pathway. The deep pathway receives normalized complex-valued patches; phase preservation by the attention gate concerns its learned input features only.

Let $f(x,y)$ denote one real-valued branch of an extracted $s \times s$ patch of the normalized coherency-derived channels $\widehat{T}_{mn}$. Nine real branches are obtained: the three real diagonal channels $\widehat{T}_{11}$, $\widehat{T}_{22}$ and $\widehat{T}_{33}$ and the real and imaginary parts of $\widehat{T}_{12}$, $\widehat{T}_{13}$ and $\widehat{T}_{23}$. The same analytic-envelope construction is applied to each branch. The three diagonal envelopes form three real-valued channels, while the six off-diagonal branch envelopes are paired into three complex storage channels. The resulting volume $\tilde{\mathbf{T}}$ retains the six-channel storage format, but it is an envelope-derived feature tensor rather than the original coherency matrix. It does not preserve the original physical polarimetric phase and is not required to satisfy the Hermitian positive-semidefinite constraints of a coherency matrix. The 2D Hilbert transform is defined in the frequency domain as

$$\mathcal{H}_{2D}\{f(x,y)\} = \mathcal{F}^{-1}\Big\{-i \cdot \mathrm{sgn}(\omega_x,\omega_y) \cdot \mathcal{F}\{f(x,y)\}\Big\}, \tag{10}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the 2D DFT and its inverse, and $q(\omega_x,\omega_y) \equiv \mathrm{sgn}(\omega_x,\omega_y)$ is an odd half-plane sign function. On each $s \times s$ frequency grid, $q = +1$ for $\omega_x > 0$ and $\omega_x = 0, \omega_y > 0$; $q = -1$ on the complementary half-plane and $q = 0$ at the origin. Thus $q(-\boldsymbol{\omega}) = -q(\boldsymbol{\omega})$, and the multiplier $M = -iq$ satisfies $M(-\boldsymbol{\omega}) = M(\boldsymbol{\omega})^*$, ensuring a real quadrature output for real input. A constant input has zero quadrature component. All the evaluated sizes are odd, so no Nyquist bins occur. The discrete frequency indexing follows this conjugate-pair convention.

The analytic signal $f_A(x,y)$ is then constructed by combining the original signal with its Hilbert transform as the imaginary part:

$$f_A(x,y) = f(x,y) + i \cdot \mathcal{H}_{2D}\{f(x,y)\}. \tag{11}$$

The envelope (instantaneous amplitude) of the analytic signal is

$$E(x,y) = |f_A(x,y)| = \sqrt{f(x,y)^2 + \big(\mathcal{H}_{2D}\{f(x,y)\}\big)^2}. \tag{12}$$

A discrete example explains the amplitude descriptor. Let $A \geq 0$, let $\mathbf{k}$ be a non-zero angular frequency on the $s \times s$ DFT grid, and define $\theta(\mathbf{r}) = \mathbf{k}^\top \mathbf{r} + \phi$ for integer pixel coordinates $\mathbf{r} = (x,y)$. Then

$$\begin{aligned} f(\mathbf{r}) &= A\cos\theta(\mathbf{r}), \\ \mathcal{H}_{2D} f(\mathbf{r}) &= q(\mathbf{k}) A \sin\theta(\mathbf{r}), \qquad E(\mathbf{r}) = A. \end{aligned}$$

The envelope retains amplitude and removes spatial carrier phase in this single-component model. This motivates a complementary cue that is less sensitive to that phase rather than automatic high-frequency or weak-scatterer enhancement. For a non-negative varying amplitude $A(\mathbf{r})$, a sufficient extension condition is that both carrier-shifted spectra remain in their corresponding half-planes without Nyquist wrapping. Slow amplitude variation alone is insufficient. Related Hilbert product conditions are discussed in [42]; the condition here follows directly from the stated multiplier. Arbitrary broadband, multicomponent and off-grid textures need not obey this interpretation.

The original-size DFT imposes periodic boundary conditions; mismatched opposite edges can cause circular wrap-around and boundary-dependent responses. The fixed half-plane convention is orientation-dependent and does not guarantee rotation invariance. The non-zero-frequency unit-magnitude multiplier applies only to the quadrature branch and does not establish artifact-free envelope processing. No artifact suppression or signal-to-noise ratio (SNR) improvement has been measured.

Shallow-only placement is an architectural choice: input-derived channels are transformed while deeper task-dependent activations are left to the learned pathway. No shallow/deep/both-path placement comparison or feature-noise statistics establish this choice as a measured optimum.

Within each patch, the nine real-branch envelopes are re-packed into the six-slot feature tensor $\tilde{\mathbf{T}}$ and passed to the shallow pathway's two $(3 \times 3 \times 1)$ 3D CNN layers.

Multiscale feature fusion networks are widely used in PolSAR classification, but the choice of filter size remains a matter of empirical optimization: PolSARFormer, for example, uses 3D convolutional layers of size $12 \times 12 \times 12$, whereas HybridCVNet uses filters of size $11 \times 7 \times 3$, $3 \times 5 \times 5$ and 5×7. In our scheme, small kernels are used and the number of filters is set to nine following the empirical design guidelines of the complex-valued hybrid networks in [36]; this small-kernel low-filter-count setting balances computational efficiency against model performance, mining the deep information in the polarimetric features adequately while avoiding the performance fluctuations caused by too many or too few filters.

2.2.2. Complex-Valued Feature Complementary Mapper (CV-FCM)

After the coherency matrix of PolSAR data is unfolded, the spatial texture and polarimetric phase information of different channels tend to be intermingled, and the discriminative features of weak channels are masked by dominant ones [43]. To solve this problem, we design the complex-valued feature complementary mapper (CV-FCM): as shown in Figure 4, the module splits the input features, weights the shallow positional responses through complex-valued spatial attention and balances the channel magnitudes through complex-valued channel attention, fusing the features implicitly for the subsequent network.

The module first performs channel splitting. The model input is $X \in \mathbb{C}^{B \times D \times H \times W \times C}$. The input channels are split into αC and $(1-\alpha)C$, giving X_{main} and X_{sub},

$$(X_{\text{main}}, X_{\text{sub}}) = \text{Split}(X; \alpha), \tag{13}$$

where X_{main} denotes the main channels, X_{sub} the subordinate channels, and α the proportion coefficient of the main channels. The main branch therefore has shape $\mathbb{C}^{B \times D \times H \times W \times \lfloor \alpha C \rfloor}$ and the subordinate branch shape $\mathbb{C}^{B \times D \times H \times W \times (C - \lfloor \alpha C \rfloor)}$. We set $\alpha = 2/3$, giving six main-branch and three subordinate-branch channels when $C = 9$. This assigns more channels to the branch with $3 \times 3 \times 3$ kernels than to the pointwise branch. It is a capacity-allocation choice, not a measured separation into semantic and purely positional feature types. The reported sensitivity check considers integer main-branch counts 3 through 8 (ratios approximately 0.33–0.89) and states that average accuracy (AA) varies by less than 0.14 percentage points (pp) in the 0.5–0.75 ratio range. This descriptive result does not establish a validation-selected optimum.

After the split, X_{main} is fed into a branch of $(3 \times 3 \times 3)$ convolutions to extract richer information, yielding X_1. Meanwhile, X_{sub} is placed in a branch of pointwise convolutions; because pointwise convolution extracts more weakly and the branch is shallow by design, it retains a large amount of low-level spatial positional information. We denote its output as X_2:

$$X_1 = \text{Conv}_{1\times1\times1}\Big(\text{Conv}_{3\times3\times3}\big(\text{Conv}_{3\times3\times3}(X_{\text{main}})\big)\Big), \tag{14}$$

Throughout this feature mapper, Conv denotes a convolutional layer and BN denotes batch normalization. The complex-valued variants act on real and imaginary components through complex arithmetic. Kernel sizes are stated as spatial height × spatial width × polarimetric depth; this kernel convention is distinct from the written ordering of tensor axes,

$$X_2 = \text{Conv}_{1\times1\times1}\big(X_{\text{sub}}\big), \tag{15}$$

where X_1 represents the dominant features and X_2 the weak features.

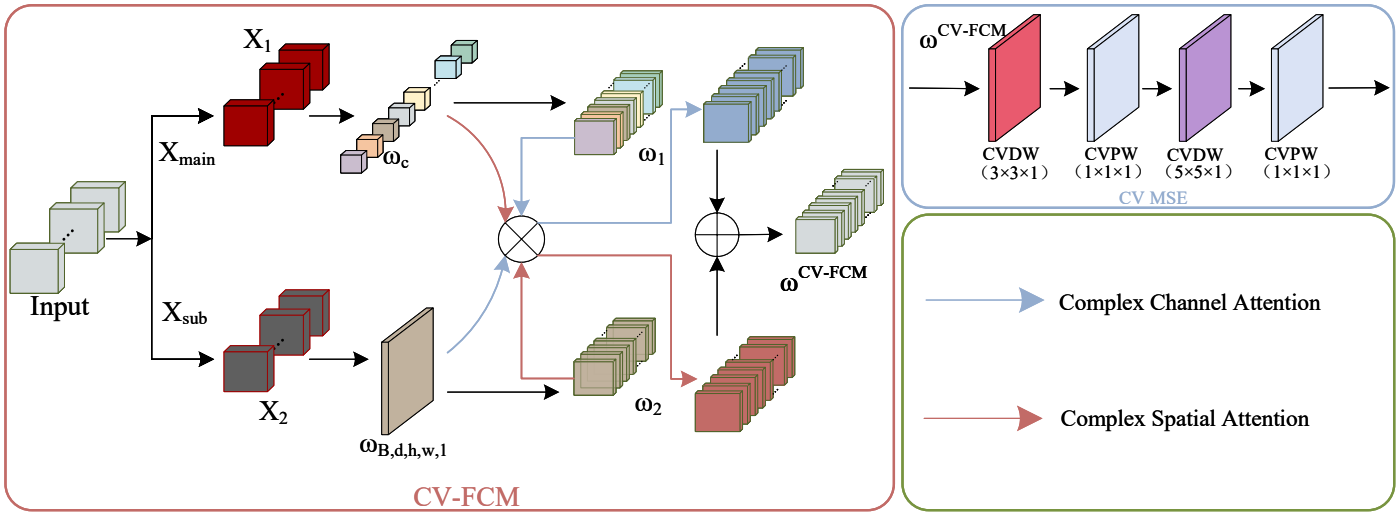


Figure 4. Flowchart of the CV-FCM and CV-MSE modules.

The features X_1 and X_2 are effective but discrete; we therefore perform complementary mapping between them to compensate for their respective missing feature maps and achieve efficient feature matching.

The complex-valued channel attention is implemented as follows. A depthwise separable complex-valued convolution is first applied to each channel followed by normalization, which cuts off the information exchange between channels,

$$Z = \mathrm{BN}\Big(\mathrm{ConvGroup}(X_1)\Big), \tag{16}$$

where Z is the feature after grouped convolution.

Global average pooling is then applied to obtain the global information of each channel. The key-information weights are obtained as the channel-averaged intensity,

$$\omega_c = \frac{1}{DHW}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W}\left|Z_{b,d,h,w,c}\right|, \quad c = 1,\ldots,C, \tag{17}$$

where ω_c holds the intensity value of the c channels that is broadcast to ω_1 and has the same number of channels as the original input.

The complex-valued spatial attention first passes through a $(1\times1\times1)$ spatial convolution layer; after normalization, the weights are obtained through sigmoid activation,

$$\omega_{b,d,h,w,1} = \sigma\Big(\left|\mathrm{BN}\big(\mathrm{Conv}_{1\times1\times1}(X_2)\big)_{b,d,h,w,1}\right|\Big) \in (0,1), \tag{18}$$

where $\omega_{b,d,h,w,1}$ is a single-channel feature carrying the spatial feature intensity, broadcast to ω_2, which has the same number of channels as the original input.

Next, X_1, which carries rich channel information, is weighted by ω_2 so that each position receives its own weight and X_2 by ω_1 so that each channel receives its own weight, achieving complementary feature fusion that exchanges higher-level information while retaining low-level spatial information,

$$X^{\text{CV-FCM}} = (X_1 \otimes \omega_2) \oplus (X_2 \otimes \omega_1), \tag{19}$$

where $X^{\text{CV-FCM}}$ is the fused output feature. The process uses dominant features to guide weak features, alleviating the information imbalance.

2.2.3. Complex-Valued Multiscale Extractor (CV-MSE)

In PolSAR data, minority classes are often masked by background noise, leading to a loss of effective information [44,45]. To exploit the available features fully, we employ multi-kernel perception units to detect targets at different scales and establish spatial relationships between them, strengthening the feature representation of contextual and small-target information. Figure 4 shows the complex-valued multiscale extractor, which connects convolution kernels of different sizes sequentially and interleaves pointwise convolutions between kernels of different scales. The process is expressed mathematically as

$$X' = T_{2k+1}\big(A(\ldots A\big(T_k(X)\big)\ldots)\big), \tag{20}$$

where T_k is a convolution with kernel size k and $A(\cdot)$ the interleaved pointwise convolution and activation.

2.2.4. Dimensionality Reduction Before the Attention Stage

CV-FCM and CV-MSE operate on the five-dimensional tensor $X \in \mathbb{C}^{B\times D\times H'\times W'\times C}$ in which the polarimetric depth axis D and the channel axis C jointly encode the polarimetric and spectral information. The subsequent attention and transformer stages, however, are defined on two-dimensional spatial feature maps because both the interaction attention and the ViT model spatial relationships within each spatial slice. The transition is realized by a purely structural reshape that folds the polarimetric depth axis into the channel axis:

$$X_{2\mathrm{D}} = \mathrm{Reshape}(X) \in \mathbb{C}^{B\times H'\times W'\times C''}, \qquad C'' = D\cdot C, \tag{21}$$

This notation preserves spatial coordinates: the feature at (b,d,h,w,c) is mapped to $(b,h,w,dC+c)$ with the corresponding zero-based depth/channel indices. Equivalently, axes are arranged as (B,H',W',D,C) before merging the last two. The written tensor ordering and the framework's storage convention must be distinguished.

This folding introduces no learnable parameters and no information loss: every spatial position retains all its polarimetric depth–channel features, re-arranged along a single channel axis. After this folding step, the tensor has exactly the form $\mathbb{C}^{B\times H'\times W'\times C''}$ consumed by the attention module below, and all the subsequent operations (CV-APIA, CV-ViT and the classification head) share the same four-dimensional complex tensor convention, which closes the data-flow loop of the whole architecture.

2.3. Complex-Valued Amplitude–Phase Interaction Attention (CV-APIA)

CV-APIA exchanges magnitude-derived gains between two pooled descriptors of the same complex feature map. The intended purpose is to modulate relative feature responses under limited labels. Here interaction denotes cross-modulation of the pooling branches rather than learned deformation offsets. Figure 5 introduces the complete structure before its operations are defined below.

Given $X \in \mathbb{C}^{B\times H\times W\times C}$, CV-APIA aggregates the channel axis separately at each spatial position. The average and maximum pooling branches produce $A, M \in \mathbb{C}^{B\times H\times W\times 1}$:

$$A = \mathcal{P}_{\mathrm{avg}}(X), \qquad M = \mathcal{P}_{\mathrm{max}}(X). \tag{22}$$

Here $\mathcal{P}$ denotes channel aggregation. It retains H and W rather than collapsing the spatial map, and the resulting gains form a spatial attention mask. The complex maximum-pooling implementation is a model choice; complex numbers have no intrinsic scalar ordering.

The aggregated features are then spatially downsampled (stride 2) to reduce computational cost while preserving correlations between neighboring pixels,

$$A_d = \mathcal{D}(A), \quad M_d = \mathcal{D}(M), \tag{23}$$

where $\mathcal{D}$ is the downsampling operation with stride 2, producing an output of size $\lceil H/2 \rceil \times \lceil W/2 \rceil$.

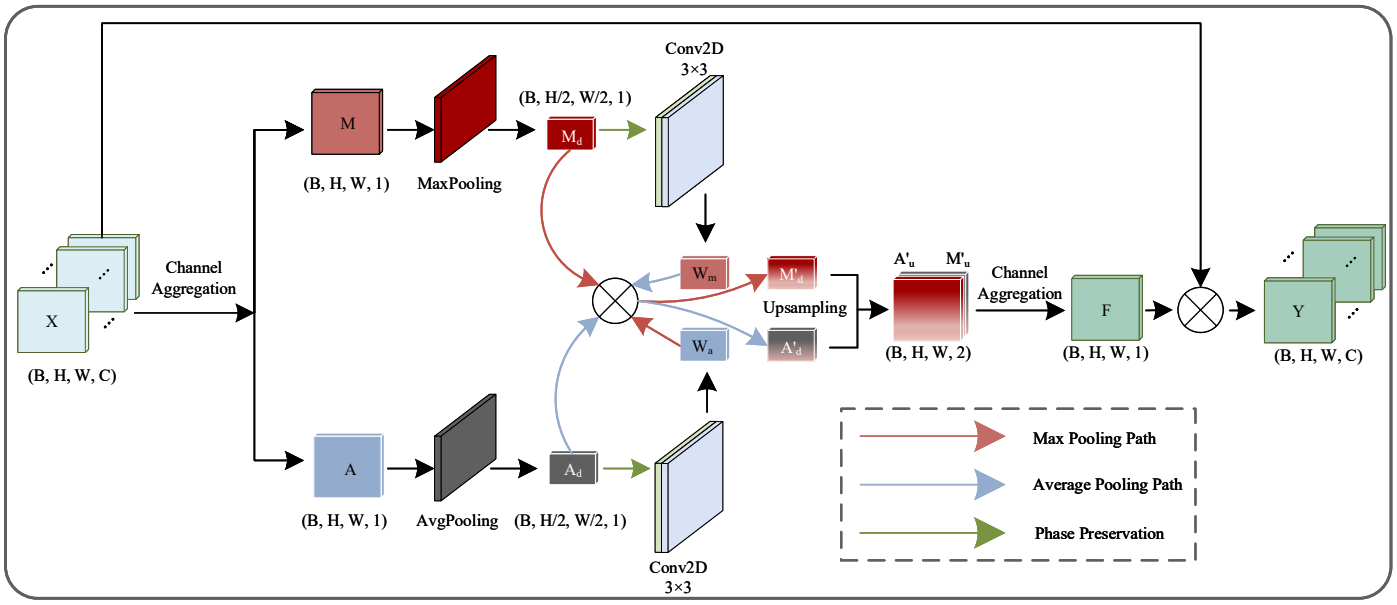


Figure 5. Cross-pooling CV-APIA structure. The final positive real output gate preserves the phase of non-zero input features; internal complex feature mixing is distinct from this property. Half-resolution labels abbreviate $\lceil H/2 \rceil \times \lceil W/2 \rceil$; interpolation returns $H \times W$.

The interaction branch decomposes its pooled complex features as $z = |z|e^{i\theta_z}$ and generates real modulation weights from their magnitudes. This operates on learned feature magnitudes and phases; these quantities are not treated as calibrated backscattering power or polarimetric scattering observables.

For independently weighted components, $y = w_r x_r + i w_i x_i$, the phase is

$$\theta_y = \operatorname{atan2}(w_i x_i, w_r x_r).$$

Unequal gains can change the phase but do not do so for every input. In contrast, let $g = \sigma(|F|) > 0$ and $Y = gX$. For each non-zero input element, $\arg Y = \arg X$; if $X = 0$, then $Y = 0$ and its phase is undefined. Thus the output gate provides an algebraic phase-preservation property without establishing its separate causal contribution to classification accuracy.

Concretely, the downsampled features are first expressed in exponential form,

$$A_d = |A_d| \cdot e^{i\Theta_A}, \quad M_d = |M_d| \cdot e^{i\Theta_M}, \tag{24}$$

where $|\cdot|$ denotes the element-wise magnitude and $\Theta_A, \Theta_M \in \mathbb{R}^{B\times\lceil H/2\rceil\times\lceil W/2\rceil\times 1}$ are the corresponding phase tensors.

Two sets of modulation coefficients are then generated from the amplitude components alone through real-valued convolution and sigmoid activation, while the phase tensors are left unchanged,

$$W_a = \sigma\big(C_{\text{mag}}(|A_d|)\big), \quad W_m = \sigma\big(C_{\text{mag}}(|M_d|)\big), \tag{25}$$

where C_{mag} denotes a real-valued convolutional layer operating only on the amplitude domain and σ is the standard (real-valued) sigmoid function. By confining the nonlinear activation to the amplitude path, the phase information Θ_A and Θ_M is preserved from activation-induced distortion.

The cross-weighting mechanism then modulates the amplitudes while re-attaching the preserved phases—an operation we call amplitude modulation with phase preservation:

$$\begin{aligned} A'_d &= \left(W_m \odot |A_d|\right) \cdot e^{i\Theta_A}, \\ M'_d &= \left(W_a \odot |M_d|\right) \cdot e^{i\Theta_M}, \end{aligned} \tag{26}$$

The gain W_m weights the magnitude of the average-pooled branch, while W_a weights that of the maximum-pooled branch. The symbol $\odot$ denotes element-wise multiplication. This cross-modulation changes the relative feature responses; it does not algebraically guarantee enhancement of a particular class or protection of every weak scatterer.

The modulated dual-path features are then restored to their original spatial dimensions by nearest-neighbor upsampling,

$$A'_u = \mathcal{R}(A'_d), \quad M'_u = \mathcal{R}(M'_d), \tag{27}$$

where $\mathcal{R}$ is nearest-neighbor interpolation to the target spatial size $H \times W$.

The upsampled features are then concatenated along the channel dimension, and a fused feature that encodes both spatial coordinate information and cross-branch interaction gains is reconstructed through a complex convolution, complex batch normalization and Cartesian ReLU (Cart-ReLU, also denoted CReLU) activation,

$$F = \phi(\mathcal{BN}(\mathcal{C}(\text{Concat}[A'_u, M'_u]))), \tag{28}$$

where ϕ is the Cart-ReLU activation, $\mathcal{BN}$ denotes complex batch normalization, and $F \in \mathbb{C}^{B\times H\times W\times 1}$ is the fused feature.

The fusion feature F produces a positive real spatial mask, which is applied to the original module input:

$$Y = \left(|X| \odot \sigma(|F|)\right) e^{i\Theta_X}. \tag{29}$$

Here $X = |X|e^{i\Theta_X}$ is the fused learned feature entering CV-APIA, and F is the internal feature of Equation (28). The output gate preserves the phase of each non-zero element of X. No Cartesian activation follows this gate; Cart-ReLU is confined to the internal fusion branch. The internal convolutions may depend on relative phase, so this is not a claim that the entire mask-generation branch is phase-independent. We call this positive real magnitude modulation amplitude–phase decoupling (APD).

Decoupling amplitude from phase is not in itself new in complex-valued networks: magnitude-threshold activations such as modReLU, introduced by Arjovsky et al. [46] and analyzed further by Trabelsi et al. [40], rescale a complex feature by a real factor derived from its magnitude, and complex-valued convolutional architectures have been adopted widely for PolSAR classification [19,22,36]. We therefore do not present the decoupling of amplitude and phase as a new principle; what is specific here is how the modulating gain is produced and where it is applied. First, the gain is not a pointwise function of the feature's own magnitude, as in modReLU, but a spatial attention mask that was obtained from the magnitudes of average- and maximum-pooled features over the channel dimension (Equation (22)). Therefore, the gain depends on the aggregated amplitude of all the channels rather than the magnitude of a single feature alone. Second, the mask is a bounded real field in $(0, 1)$ produced by a sigmoid and is applied in the exponential domain so that the final gate leaves the phase of every non-zero input element of X unchanged; Cartesian-coordinate activations such as CReLU can change phase by truncating real or imaginary components. Third, one and the same real mask multiplies the real and imaginary parts, so the two components are modulated jointly rather than as independent real channels. These points concern the construction of the modulation, not the decoupling principle itself. Two further distinctions should be stated for precision: the mask is derived from the magnitudes

of the pooled features and is positive real, and the output of the module preserves the phase of its input feature X (Equation (29)); we do not claim that every quantity inside the module is independent of phase since the internal complex convolution combines contributions whose sum depends on their relative phases. Finally, the ablation in Section 3.3 quantifies this design relative to the Cartesian-coordinate variant; as discussed there, that comparison evaluates the amplitude–phase-decoupled formulation as a whole and does not isolate phase preservation from the remaining design choices, and it is not a comparison against previously published decoupling methods, which differ in gating source, pooling scheme and task and would require a separate re-implementation.

2.4. Complex-Valued Vision Transformer (CV-ViT)

The attention stage is followed by a complex-valued vision transformer (CV-ViT) that captures long-range dependencies over the folded feature map. All the linear operations in this block are complex-valued and follow the deep complex network framework of Trabelsi et al. [40]: the real and imaginary parts are processed jointly through complex arithmetic instead of being treated as independent real-valued channels.

Complex patch embedding. Let $X \in \mathbb{C}^{B\times H'\times W'\times C''}$ denote the folded feature tensor. The spatial map is partitioned into $N = (H'/P)\,(W'/P)$ non-overlapping patches of size $P \times P$, and each patch is flattened into a complex vector $x_p \in \mathbb{C}^{P^2C''}$. A trainable complex embedding matrix $E \in \mathbb{C}^{P^2C''\times d_{\text{emb}}}$ projects each patch into the transformer dimension d_{emb}, and a learnable complex positional embedding $e_p \in \mathbb{C}^{d_{\text{emb}}}$ is added. A complex class token $z_{\text{cls}} \in \mathbb{C}^{d_{\text{emb}}}$ is prepended, giving the input sequence

$$z_0 = [\, z_{\text{cls}};\ x_1E + e_1;\ \ldots;\ x_NE + e_N \,] \in \mathbb{C}^{(N+1)\times d_{\text{emb}}}. \tag{30}$$

Component-wise layer normalization. Complex layer normalization (CLN) is implemented by applying layer normalization separately to the real and imaginary parts. For $z \in \mathbb{C}^{B\times(N+1)\times d_{\text{emb}}}$, the statistics are calculated over the final feature axis for each sample b and token n without averaging across the batch or token axes:

$$\begin{aligned}
\mu_{r,b,n} &= \frac{1}{d_{\text{emb}}}\sum_{d=1}^{d_{\text{emb}}} z_{r,b,n,d},\\
v_{r,b,n} &= \frac{1}{d_{\text{emb}}}\sum_{d=1}^{d_{\text{emb}}} (z_{r,b,n,d} - \mu_{r,b,n})^2,\\
\hat{z}_{b,n,d} &= \gamma_{r,d}\frac{z_{r,b,n,d} - \mu_{r,b,n}}{\sqrt{v_{r,b,n} + \epsilon}} + \beta_{r,d}\\
&\quad + i\left(\gamma_{i,d}\frac{z_{i,b,n,d} - \mu_{i,b,n}}{\sqrt{v_{i,b,n} + \epsilon}} + \beta_{i,d}\right).
\end{aligned} \tag{31}$$

The imaginary-part statistics are defined analogously. The real scale and shift parameters are learned separately for the two parts, and $\epsilon = 10^{-6}$. This is component-wise layer normalization rather than batch normalization or a full complex-covariance whitening operation.

Hermitian attention and head dimensions. The embedding width is $d_{\text{emb}} = 32$ and the number of heads is $h = 4$. The configured key and value width is $d_k = d_v = 32$ *per head*; it is not d_{emb}/h. Suppressing the batch axis, each head $m \in \{1, \ldots, h\}$ uses projections $W_{Q,m}, W_{K,m}, W_{V,m} \in \mathbb{C}^{d_{\text{emb}}\times d_k}$:

$$Q_m = zW_{Q,m}, \qquad K_m = zW_{K,m}, \qquad V_m = zW_{V,m} \in \mathbb{C}^{(N+1)\times d_k}. \tag{32}$$

Separately projected queries and keys need not have real diagonal entries in $Q_m K_m^H$. Each head takes the real part of the conjugate-transposed product before the row-wise softmax:

$$A_m = \text{softmax}\left(\frac{\text{Re}(Q_m K_m^H)}{\sqrt{d_k}}\right) \in \mathbb{R}^{(N+1)\times(N+1)},$$
$$Z_m = A_m V_m \in \mathbb{C}^{(N+1)\times d_k}. \tag{33}$$

The four head outputs concatenate to width $hd_k = 128$ and are projected back to the embedding width for the residual connection, defining multi-head attention (MHA):

$$\text{MHA}(z) = \text{Concat}[Z_1, \ldots, Z_h]W_O, \qquad W_O \in \mathbb{C}^{128\times 32}.$$

The attention weights are real and non-negative, but weighted sums of complex values can alter phase. The element-wise phase-preservation claim remains limited to the final positive real CV-APIA gate.

Complex multilayer perceptron and activation. Each transformer block ends with a two-layer complex multilayer perceptron (MLP)

$$\text{MLP}(z) = \sigma_c\big(zW_1 + b_1\big)W_2 + b_2, \tag{34}$$

with complex weights $W_1 \in \mathbb{C}^{d_{\text{emb}}\times d_{\text{ff}}}$ and $W_2 \in \mathbb{C}^{d_{\text{ff}}\times d_{\text{emb}}}$, complex biases b_1, b_2, and the Cartesian ReLU (CReLU) activation applied to the real and imaginary parts separately following Trabelsi et al. [40]:

$$\sigma_c(z) = \text{ReLU}\big(\text{Re}(z)\big) + i\ \text{ReLU}\big(\text{Im}(z)\big). \tag{35}$$

Each block is wrapped by residual connections and CLN in the standard ViT arrangement,

$$\hat{z}_\ell = \text{MHA}\big(\text{CLN}(z_{\ell-1})\big) + z_{\ell-1}, \qquad z_\ell = \text{MLP}\big(\text{CLN}(\hat{z}_\ell)\big) + \hat{z}_\ell, \tag{36}$$

for $\ell = 1, \ldots, L$ blocks, where MHA denotes the multi-head Hermitian attention.

Classification head. After the last block, the class token z_{cls}^L is projected by a complex dense layer to the number of classes; the classification logits are taken as the real part of this projection, and a softmax produces the final probabilities. The CV-ViT supplies global complex-valued modeling after the local feature extraction and CV-APIA gating. Complex representation and phase-preserving gating are distinct: the transformer mixes features, while only the specified positive real gate is asserted to preserve its non-zero input phase.

2.5. PolSAR Datasets

Table 1 introduces two public PolSAR scenes, Flevoland and San Francisco, and an author-curated Nanjing scene. It lists the sensor, reported acquisition information, image dimensions and number of labeled classes.

Table 1. Reported scene characteristics. The Nanjing scene is author-curated rather than a public benchmark.

Name	Sensor	Time	Image Band	Dimensions	Class
Flevoland 1989	AIRSAR	1989	L-band	750 × 1024	15
San Francisco	AIRSAR	2008	L-band	900 × 1024	5
Nanjing	RADARSAT-2	2017	C-band	1500 × 1500	6

The Flevoland image in Figure 6 was acquired by NASA/JPL AIRSAR over the Netherlands on 16 August 1989. It contains four-look L-band data of 750 × 1024 pixels at the reported 12 m resolution with 15 classes [47].

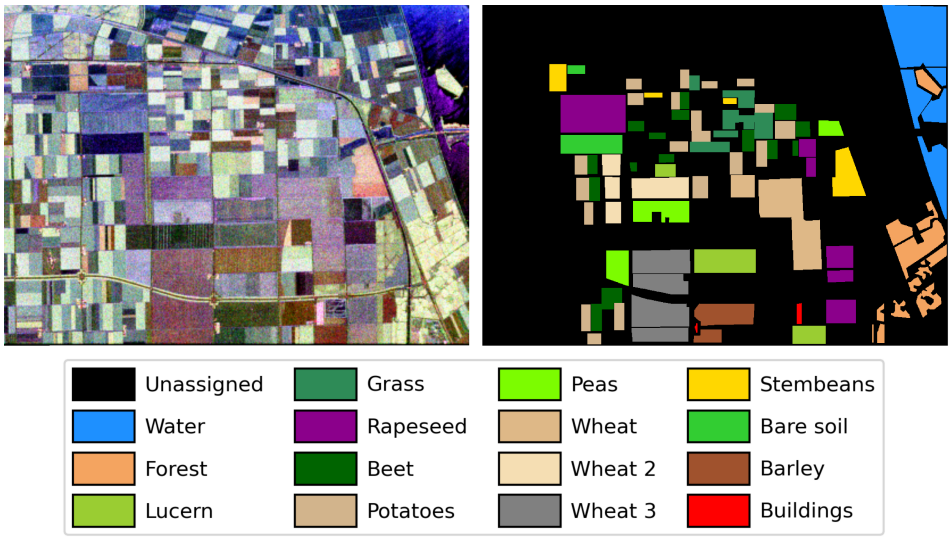


Figure 6. Flevoland scene. (**a**) Pauli RGB composite; (**b**) reference-label map. Unassigned pixels are excluded from the scored classes.

The San Francisco image in Figure 7 is an L-band AIRSAR scene of 900 × 1024 pixels at the reported 10 m resolution with five labeled classes. Xing et al. identify the San Francisco AIRSAR data as L-band in their experimental description and Figure 4 caption [48].

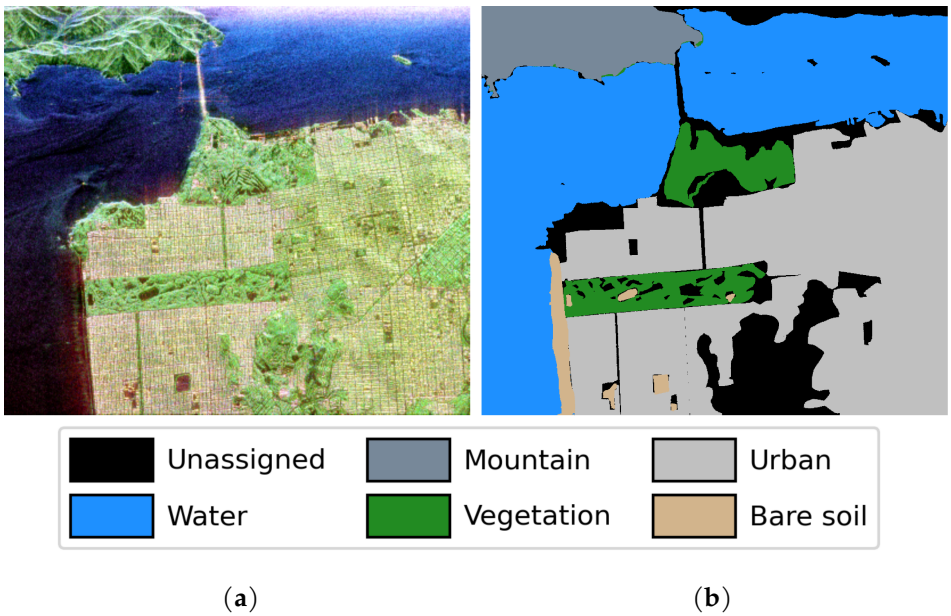


Figure 7. San Francisco scene. (**a**) Pauli RGB composite; (**b**) reference-label map. Unassigned pixels are excluded from the scored classes.

The Nanjing SAR imagery was provided by the authors' research group and acquired by RADARSAT-2 on 11 April 2017. The experimental scene in Figure 8 is a 1500 × 1500 crop at the reported 5 m resolution that covers Nanjing and its surroundings in Jiangsu, China. Its six-class reference mask was manually annotated for this study. RapidEye optical imagery was acquired in April 2017 at 5 m resolution, providing a visual reference. The optical images were mosaicked for the reference coverage; the model input is the SAR crop. This is an author-curated experimental scene rather than a publicly released classification benchmark.

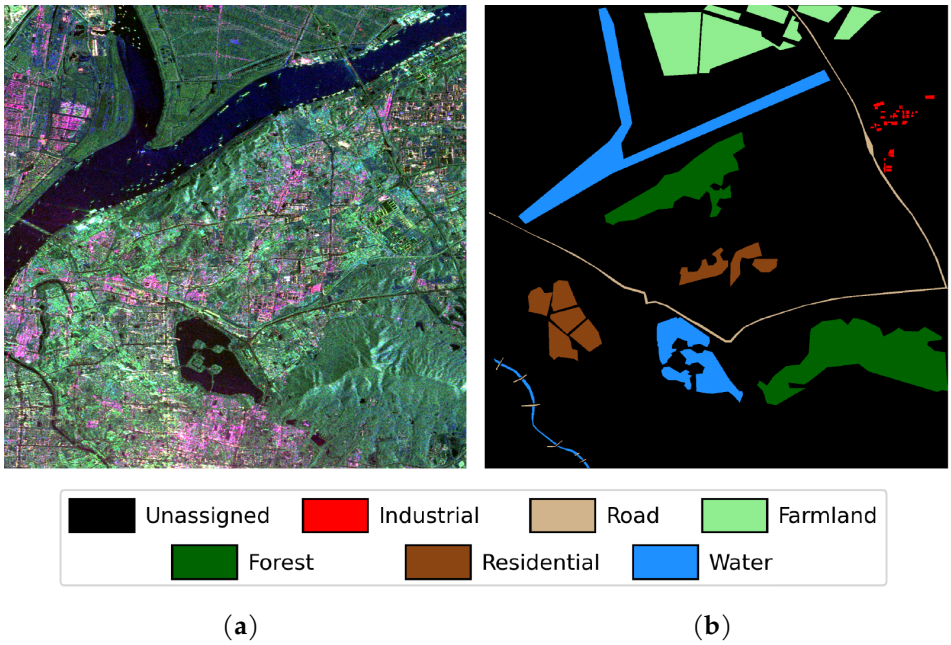


Figure 8. Author-curated Nanjing scene. (**a**) Pauli RGB composite of the SAR crop; (**b**) manually prepared partial reference mask. Water labels cover selected interior portions of the water surface rather than its full extent. Black denotes unassigned pixels, which are excluded from supervised label sampling and accuracy assessment.

The reference mask deliberately covers selected regions, not the complete geographical extent of every class. For water, annotation focused on central or interior portions of the water surface where the annotator judged the class membership to be unambiguous rather than tracing the full shoreline. Partial coverage was also intended to keep the labeled pool and, under the fixed-proportion sampling protocol, the training-sample budget manageable. The remaining unannotated water pixels are left unassigned. A smaller water annotation area therefore describes incomplete annotation coverage, not a narrower physical river boundary. Black unassigned pixels are not an additional scored class and are excluded from supervised label sampling and the overall accuracy (OA), AA and kappa coefficient (κ) calculations. The class totals reported for the Nanjing dataset (Section 3.4) sum to 342,689 labeled pixels, approximately 15.23% of the 1500 × 1500 crop. Accuracy is evaluated on the labeled test pixels under the stated sampling protocols; predictions in unannotated areas are not quantitatively validated by this reference mask.

2.6. Experimental Configuration

2.6.1. Input Patches and Labeled-Sample Budgets

Each sample is a square spatial patch centered on a labeled pixel, and its supervised target is the class of that central pixel. For the random-pixel protocol, patches are enumerated in row-major order with zero padding at scene boundaries; centers with the unassigned label are removed from the labeled sampling pool. The reported random-pixel experiments use class-stratified sampling with nominal training budgets of 1% of labeled pixels on Flevoland and San Francisco and 5% on Nanjing. The split uses `train_test_split` with `stratify=y` and a fixed split `random_state=345`; the corresponding test fractions are 0.99, 0.99 and 0.95. Thus the percentage denominator is the labeled-pixel pool, not the area of the full image. If N_L is the number of labeled pixels and N_{tr} the selected training count, the realized ratio is $100N_{\text{tr}}/N_L$ percent. Integer class allocations are assigned by the stratified splitter rather than independently rounding the nominal fraction within every class.

The training counts tabulated for Flevoland and San Francisco sum to 2078 and 8023, respectively. The Nanjing table instead lists the annotated class totals, not the class training counts. In the random-pixel protocol, labeled pixels that were not selected for training form the test pool. The spatial-block protocol in Section 3.6 restricts training and test samples to their assigned regions.

Input windows are 15 × 15 on Flevoland and Nanjing and 13 × 13 on San Francisco. The same dataset-specific window and reported training split are used by the compared methods. Window sizes follow the cited baseline settings; the separate sensitivity curves are descriptive evaluations on test data, not independent validation-set selection. We do not present these settings as the result of a held-out grid-search procedure.

2.6.2. Optimization and Repeated Evaluation

The experiments use Python (v3.9) and TensorFlow (v2.10.0) on a Windows 11 machine with 32 GB RAM and an NVIDIA GeForce RTX 4070 Ti GPU (NVIDIA Corporation, Santa Clara, CA, USA) with 12 GB VRAM. The CV-MfIaViT is trained with Adam, a learning rate of 10^{-3}, categorical cross-entropy loss, batch size 128 and a maximum of 100 epochs. Training samples are shuffled between epochs. Early stopping monitors training accuracy with patience 10 and restoration of the weights selected by this criterion. No separate validation subset is reserved.

Configuration development and subsequent evaluation. The proposed configuration was established through manual exploratory trials starting from reference-method or code defaults rather than a formal grid search. Training loss and accuracy, together with random-pixel test OA and AA, informed this development stage. The model and hyperparameter configuration was fixed before the spatial-block experiments in Section 3.6; their results were not used for further model revision or hyperparameter selection. The reference methods retain their original architecture and optimization settings from the publications in Section 3.4. The common labeled splits, sampling budgets and input windows are described here. The Adam and stopping settings above apply to the proposed model. Section 4 distinguishes the development-stage comparisons from the subsequent fixed-configuration assessment.

After selecting the configuration, three random-pixel main runs separately initialize and train the model on the same labeled training/test center-pixel split. Aggregate OA, AA and κ are reported as mean ± standard deviation over these three runs. This standard deviation measures initialization and training variation that are conditional on the fixed split. The subsequent spatial-block runs retain the fixed configuration and use different block assignments, as specified in Section 3.6. Per-class entries in the Flevoland comparison are from one representative run of each method. The CV-MfIaViT vegetation entry of 93.67% reported for San Francisco (Section 3.4) is also a single-run class accuracy that is separately reported from the three-run aggregate metrics. Other class summaries follow their stated captions. Figure 9 follows the single-run protocol in Section 3.2.

The comparison controls the labeled-sample budget, dataset-specific input window and reported split at a given configuration. Model architectures are identified by the cited reference methods. A common window controls the supplied spatial context but does not imply identical effective receptive fields or independently optimal settings for every architecture.

2.6.3. Reported Parameter and Forward-Cost Accounting

For the default Flevoland configuration (15 × 15 input and 15 classes), the CV-MfIaViT has a reported total count of 2,933,146 entries in the Keras model summary: 2,933,046 trainable parameters and 100 non-trainable batch normalization (BN) state entries. The re-

ported network forward cost is 146.5 million floating-point operations (FLOPs) per patch. The Cartesian-coordinate attention variant has a reported cost of 148.2 M FLOPs. The distinction concerns the modulation step: complex multiplication combines four real products and two real additions, whereas multiplication of a complex feature by one real gain uses two real products. Complex fusion convolution, batch normalization and Cart-ReLU remain in the attention module, as shown in Equation (28); APD is not a wholly real-valued network.

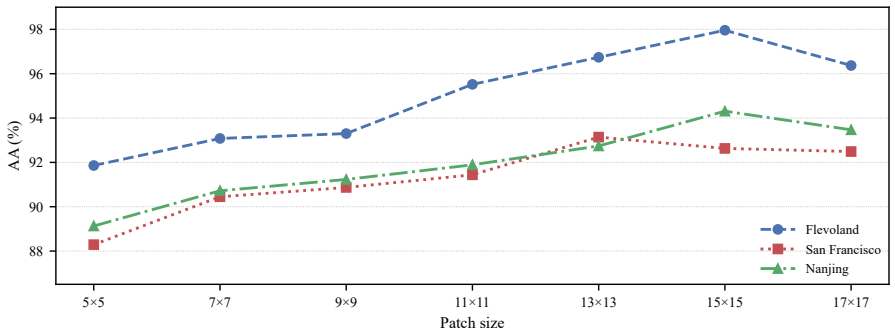


Figure 9. Single-run test-set AA versus input-window size. Each size is separately trained using the same center-pixel split, sampling proportion and other hyperparameters within a dataset. The curves describe window sensitivity and differ from the three-run aggregate results of the main comparisons.

Counts sum stored tensor elements: each learned complex coefficient contributes its separate real and imaginary entries, while BN moving statistics follow their stored tensor shapes. The total therefore includes non-trainable state. The attention module contributes 82 trainable and 5 non-trainable entries, totaling 87; FM contributes 5626 trainable and 95 non-trainable entries, totaling 5721. Parameter increments for the learned front ends are stated separately in the front-end ablation summary (Section 3.3).

Scene-level standardization precedes patch extraction, whereas the shallow-path 2D HT is computed separately within each extracted patch. The reported network FLOPs exclude standardization and this patch-wise front end. Parameter-free HT is therefore not zero-cost end-to-end processing; its computation must be included when measuring the complete pipeline. No measured GPU latency, peak memory or matched-hardware baseline training time is supplied here. The conditional area-scaling analysis in Section 3.2 is labeled separately from the reported default-configuration cost.

2.7. Evaluation Metrics

The predicted classification map is compared with the reference (ground-truth) data. Visual inspection alone is subjective and incomplete, so quantitative metrics are used; they are defined below.

Overall accuracy (OA) is the proportion of correctly classified pixels in the dataset,

$$\mathrm{OA} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}}, \tag{37}$$

where N_{correct} is the number of correctly classified pixels and N_{total} the number of scored labeled test pixels.

Average accuracy (AA) is the arithmetic mean of the per-class accuracies,

$$\mathrm{AA} = \frac{1}{c}\sum_{i=1}^{c} x_i, \tag{38}$$

where c is the number of scored classes and x_i is the fraction of correctly predicted test pixels among reference pixels of class i.

Class accuracy measures the proportion of correctly classified samples within a given class.

The Kappa coefficient (κ) evaluates the agreement between predictions and ground truth while discounting chance agreement,

$$\kappa = \frac{\mathrm{OA} - P_e}{1 - P_e}, \tag{39}$$

where P_e is the expected agreement under random chance; Kappa equals 1 for perfect agreement, 0 for agreement at the level of chance, and can be negative when the agreement is worse than chance. Its values are reported multiplied by 100, as in the tables below.

3. Results

3.1. Evaluation Overview

The experiments evaluate two public scenes, Flevoland and San Francisco, and the author-curated Nanjing scene. We first describe input-window sensitivity, then component ablations, reference-method comparisons, training-budget sensitivity and spatial-block evaluation. Sampling and optimization are defined in Section 2.6; each table or figure specifies its configuration and statistical scope.

3.2. Sensitivity Analysis of the Patch Receptive Field

A patch supplies spatial context for classifying its labeled central pixel. Larger windows provide more neighborhood information but can also include pixels from other classes, so the balance depends on scene layout and spatial sampling. Figure 9 evaluates 5×5, 7×7, 9×9, 11 × 11, 13 × 13, 15 × 15 and 17 × 17 windows. Within each dataset, the labeled training/test center-pixel split, sampling proportion and other hyperparameters are held fixed according to Section 2.6. A model is separately initialized and trained for each size. Each plotted AA is a single-run test result; this experiment does not estimate run-wise uncertainty.

On Flevoland, 13 × 13 → 15 × 15 increases AA by 1.22 percentage points (pp), whereas 15 × 15 → 17 × 17 decreases it by 1.59 pp. The corresponding changes on Nanjing are +1.57 and −0.84 pp. On San Francisco, the AA values at 15 × 15 and 17 × 17 are 0.51 and 0.65 pp lower than at 13 × 13. These observations identify the peak within the tested range; they do not establish a universally optimal size or quantitatively measure the underlying land-cover scales.

The main comparisons use 15×15 on Flevoland and Nanjing and 13×13 on San Francisco in accordance with the baseline settings [25]. A common window within each dataset controls the supplied input context, although effective receptive fields differ across architectures. Figure 9 characterizes sensitivity of the evaluated model to spatial context; it is a descriptive test-set analysis rather than an independent validation-based selection procedure.

Conditional computation-scale analysis. The 13 × 13, 15 × 15 and 17 × 17 inputs contain 169, 225 and 289 spatial positions: increases of 33.1% and 28.4% for the two consecutive size changes. Convolutional operation counts depend on the corresponding feature-map areas, while transformer cost also depends on the token count and includes quadratic attention terms. Thus input area alone is not an exact total-network FLOP formula.

As an illustrative area-only scenario, applying $C_{\mathrm{area}}(s) = 146.5(s/15)^2$ M operations to the reported 15×15 reference cost gives approximately 110.0 and 188.2 M at 13 and 17. These values are conditional extrapolations, not separately profiled model FLOPs or measured timings, and do not prove that convolution dominates the actual network. Within this scenario, Flevoland and Nanjing gain AA up to 15 while the move to 17 increases scale and lowers AA; on San Francisco, 13 is more accurate and has a smaller input than 15 and 17. It is not the smallest or cheapest input among all seven tested sizes. Actual latency, memory and model-specific operation counts across window sizes remain to be measured.

3.3. Ablation Experiments

The Flevoland ablations examine module inclusion and front-end/attention variants under a common reported training budget. Table 2 first compares the four combinations of FM and the attention module while the CNN and transformer structure are fixed. Its columns report OA, AA and $\kappa \times 100$. The run-wise standard deviations are defined in Section 2.6.

Table 2. FM and attention-module combinations on Flevoland with 1% stratified training samples. Aggregate scores are mean ± standard deviation over three runs.

FM	APIA	OA (%)	AA (%)	$\kappa \times 100$
×	×	96.43 ± 0.38	95.56 ± 0.31	96.21 ± 0.42
✓	×	97.46 ± 0.51	97.21 ± 0.64	97.44 ± 0.62
×	✓	97.23 ± 0.45	96.63 ± 0.42	96.89 ± 0.58
✓	✓	97.78 ± 0.52	97.35 ± 0.68	97.61 ± 0.54

Without either module, OA is 96.43% and AA is 95.56%. FM alone gives 97.46% OA and 97.21% AA; attention alone gives 97.23% OA and 96.63% AA; the combination gives 97.78% OA and 97.35% AA. The marginal OA changes are 0.55 pp for FM with attention present and 0.32 pp for attention with FM present. These are descriptive differences between the compared architectures.

Using the total-count convention in Section 2.6, the attention module adds $82 + 5 = 87$ entries and FM adds $5626 + 95 = 5721$, giving $87 + 5721 = 5808$. Of these, 5708 are trainable parameters and 100 are non-trainable BN state entries. Relative to the full-model total, $100 \times 5808/2933146 = 0.198\%$ after rounding. This is total-count overhead, not trainable overhead or measured execution cost.

Table 3 then compares identity and 2D HT front ends with and without APD plus learned front ends G1 and G3 with APD enabled. The reference is the FM-plus-Cartesian-attention configuration from Table 2. G1 applies $x' = 2\sigma(W|x| + b) \odot x$ to the six-component vector at each spatial location; $W \in \mathbb{R}^{6\times6}$ and $b \in \mathbb{R}^6$ are real-valued and correspond to 42 reported parameters. The reported G3 count corresponds to one complex input and output channel, a $3 \times 3 \times 1$ kernel and complex bias: $2(9 + 1) = 20$ independent real scalars. These are distinct from the six polarimetric-depth slots.

Table 3. Front-end and APD ablation on Flevoland with a 15×15 input and 1% training samples. Mean ± standard deviation is reported over three runs. Learned G1/G3 replace the front end with APD enabled; their reported parameter increments and network forward costs exclude channel standardization and patch-wise HT preprocessing.

Front End	APD	Extra Params	FLOPs (M)	OA (%)	AA (%)	$\kappa \times 100$
Identity	×	0	148.2	97.78 ± 0.52	97.35 ± 0.68	97.61 ± 0.54
2D HT	×	0	148.2	97.95 ± 0.50	97.65 ± 0.64	97.81 ± 0.52
Identity	✓	0	146.5	97.91 ± 0.48	97.53 ± 0.61	97.74 ± 0.51
G1 (learned gate)	✓	+42	146.6	97.98 ± 0.46	97.70 ± 0.59	97.81 ± 0.49
G3 (local conv)	✓	+20	147.1	98.04 ± 0.47	97.78 ± 0.60	97.87 ± 0.49
2D HT (proposed)	✓	0	146.5	98.15 ± 0.48	97.96 ± 0.71	97.98 ± 0.45

Adding 2D HT to the reference changes OA by +0.17 pp and AA by +0.30 pp. APD alone changes them by +0.13 and +0.18 pp. The combined configuration reaches 98.15% OA and 97.96% AA, changes of +0.37 and +0.61 pp. The individual changes sum to $0.17 + 0.13 = 0.30$ pp for OA and $0.30 + 0.18 = 0.48$ pp for AA. These point differences

describe the joint design; they are not a significance test of interaction or proof of phase preservation as the cause of the gains.

G1 gives 97.98 ± 0.46% OA and 97.70 ± 0.59% AA, and G3 gives 98.04 ± 0.47% OA and 97.78 ± 0.60% AA. Their OA means are 0.17 and 0.11 pp below the fixed-front-end mean, respectively. Three-run dispersion is larger than these point differences, and no formal superiority or equivalence test is reported. The results show similar reported means with different front-end parameter requirements. The fixed transform requires no learned front-end parameters, while its patch-wise computation still has a cost.

Statistical Interpretation

OA weights classes according to test-pixel frequency, whereas AA gives each scored class equal weight. The joint front-end/attention configuration has an AA point gain of 0.61 pp and an OA point gain of 0.37 pp over its reference. This is compatible with the class-balance motivation but does not identify a particular scattering mechanism.

The OA standard deviations in Table 3 range from 0.46 to 0.52 pp and the AA standard deviations from 0.59 to 0.71 pp. Accuracy differences, run-wise variation and individual class point values have distinct interpretations. In particular, large class differences between complete models are not isolated component effects, and a smaller accuracy standard deviation does not by itself demonstrate faster or more stable training convergence. The representative-run class values on Flevoland also do not provide per-class uncertainty estimates.

3.4. Comparison with Reference Methods

The seven reference configurations cover convolutional and transformer families: CV-2D CNN [20], CV-3D CNN [21], CV-2D–3D [22], ViT [23], Swin Transformer [49], PolSAR-Former [24] and CV-MsAtViT [25]. The cited works identify their architectural and original configuration sources. Each reference method uses its original architecture and optimization settings with the common sampling and input-window settings in Section 2.6. The results compare these stated configurations; development and subsequent fixed-configuration evaluation are distinguished in Sections 2.6 and 3.6.

3.4.1. Classification Results on the Flevoland Dataset

Table 4 presents the 15 class accuracies, training counts and aggregate metrics for the 1% training configuration. It distinguishes representative-run class values from three-run aggregate summaries.

The CV-MfIaViT has the highest reported aggregate means among these configurations: 98.15 ± 0.48% OA, 97.96 ± 0.71% AA and 97.98 ± 0.45 for $\kappa \times 100$. Its AA mean is 1.83 pp above CV-MsAtViT. The CNN and transformer references differ in architecture and performance; these comparisons do not isolate dimensionality, parameter count or a single physical effect.

The proposed model attains the best or tied-best class value in nine of 15 classes. The representative buildings value is 97.55% compared with 81.12% for CV-MsAtViT, a difference of 16.43 pp with six tabulated training samples. Grass and peas differ by 4.51 and 4.92 pp. Stem beans and beet are lower by 1.42 and 1.74 pp than CV-MsAtViT and PolSARFormer, respectively. These class differences characterize complete models and do not identify the contribution of phase preservation alone.

Figure 10 provides the classification-map comparison as visual context; numerical accuracy is determined on the scored reference pixels.

3.4.2. Classification Results on the San Francisco Dataset

Table 5 reports the five class summaries, training counts and aggregate metrics under 1% training and a 13 × 13 input. This separates sample support from prediction accuracy.

Table 4. Flevoland reference-class accuracies and reported training counts. Class accuracies are from one representative run per method; aggregate OA, AA and $\kappa \times 100$ are mean ± standard deviation across three runs. Accuracy units are percent.

Class	Training Samples	CV-2D CNN	CV-3D CNN	CV-2D–3D	ViT	Swin T	PolSAR-Former	CV-MsAtViT	CV-MfIaViT
Water	292	98.13	99.19	97.12	99.34	99.59	99.84	99.36	99.96
Forest	159	94.38	**99.43**	94.40	94.04	96.15	99.36	98.34	97.95
Lucerne	112	95.41	**98.64**	96.66	97.43	97.65	96.38	97.39	96.51
Grass	102	3.66	89.89	84.50	90.22	85.04	93.38	89.74	**94.25**
Rapeseed	219	88.13	97.08	87.65	95.01	90.44	96.58	98.84	**98.95**
Beet	147	87.17	94.10	90.28	95.63	95.86	**98.08**	92.46	96.34
Potatoes	213	79.63	96.50	95.83	**98.00**	92.65	90.52	97.83	97.43
Peas	104	86.19	95.31	98.61	94.48	92.92	96.41	93.74	**98.66**
Stem Beans	85	89.96	96.27	94.32	86.77	94.28	96.21	**99.49**	98.07
Bare Soil	**63**	2.94	83.27	90.37	97.19	94.84	84.56	97.04	**98.99**
Wheat	176	82.53	**98.63**	97.04	85.03	96.49	97.64	98.08	96.38
Wheat 2	106	0.10	62.88	87.97	95.58	84.41	81.24	98.87	**98.89**
Wheat 3	220	94.70	98.38	96.06	98.61	96.59	95.57	99.75	**99.85**
Barley	74	90.58	95.97	94.16	94.15	96.77	96.34	99.82	**99.82**
Buildings	**6**	83.30	92.05	84.53	12.96	68.00	87.68	81.12	**97.55**
OA (%)		78.50 ± 2.42	94.94 ± 2.55	93.68 ± 2.23	94.80 ± 3.28	94.33 ± 2.26	95.34 ± 1.08	97.53 ± 0.78	**98.15 ± 0.48**
AA (%)		71.79 ± 2.71	93.17 ± 2.81	92.63 ± 2.51	88.96 ± 3.53	92.11 ± 2.41	93.99 ± 1.31	96.13 ± 0.81	**97.96 ± 0.71**
Kappa ×100		76.33 ± 2.31	94.47 ± 2.45	93.10 ± 2.17	94.32 ± 3.10	93.80 ± 2.13	94.91 ± 1.05	97.30 ± 0.75	**97.98 ± 0.45**

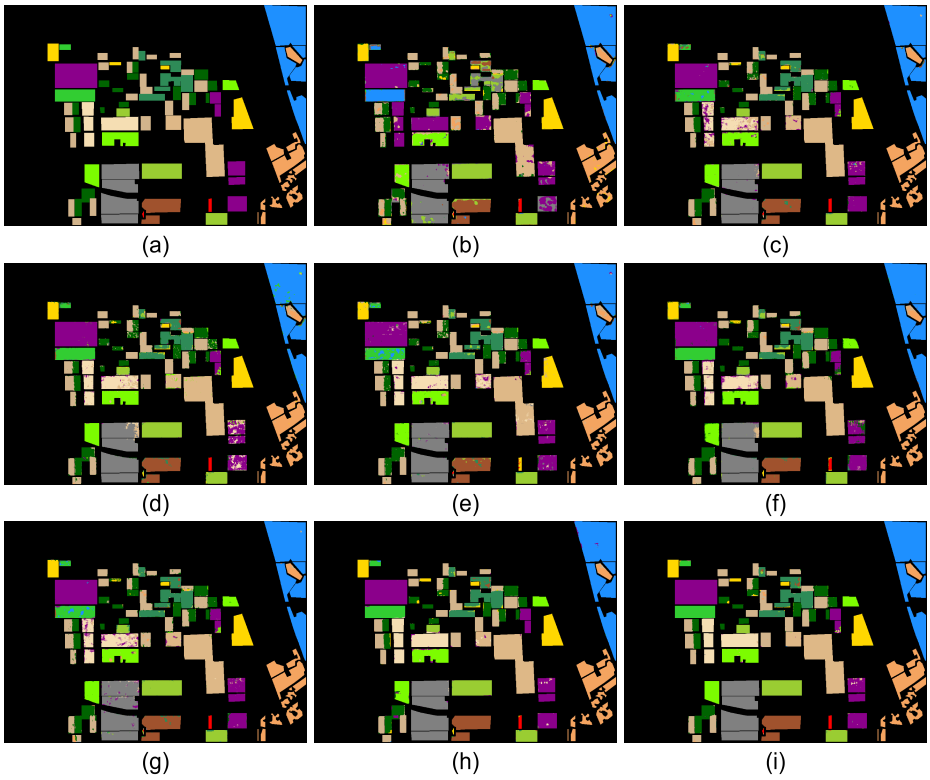


Figure 10. Classification maps on the Flevoland dataset. (**a**) Ground truth; (**b**) CV-2D CNN; (**c**) CV-3D CNN; (**d**) CV-2D–3D; (**e**) ViT; (**f**) Swin T; (**g**) PolSARFormer; (**h**) CV-MsAtViT; (**i**) CV-MfIaViT.

The proposed model gives 97.46 ± 0.87% OA, 93.14 ± 0.96% AA and 96.03 ± 1.03 for $\kappa \times 100$. Its OA mean is 0.11 pp above CV-MsAtViT and its AA mean is 3.42 pp higher. The small OA difference should be read together with the run-wise dispersion.

Vegetation has 535 training samples within the 8023-sample training set. The CV-MfIaViT gives a single-run test accuracy of 93.67% for this class; this value is not a three-run class mean. The tabulated PolSARFormer class value is 71.17%. Training and evaluation follow Section 2.6. Water reaches 99.71%, while other methods lead on mountain, urban and bare soil. These results show class-dependent trade-offs rather than improvement for

every class. Figure 11 displays the corresponding maps; it does not replace the numerical evaluation or isolate an attention mechanism.

Table 5. San Francisco class accuracies and reported training counts. The CV-MfIaViT vegetation entry of 93.67% is a single-run class accuracy. Aggregate OA, AA and $\kappa \times 100$ are mean ± standard deviation over three runs. Accuracy units are percent.

Class	Training Samples	CV-2D CNN	CV-3D CNN	CV-2D–3D	ViT	Swin T	PolSAR-Former	CV-MsAtViT	CV-MfIaViT
Water	3296	98.39	99.37	99.01	97.49	98.53	97.72	99.55	99.71
Mountain	627	90.92	**98.34**	92.99	90.03	90.65	92.54	97.89	92.16
Vegetation	**535**	78.64	87.84	82.65	62.31	55.85	71.17	73.02	**93.67**
Urban	3428	98.20	95.99	95.51	98.91	98.48	99.00	**99.70**	97.45
Bare soil	**137**	69.57	74.82	50.81	70.51	70.12	**83.38**	78.43	82.70
OA (%)		95.92 ± 2.78	96.66 ± 2.46	95.13 ± 2.12	94.71 ± 2.54	94.56 ± 2.41	95.85 ± 1.57	97.35 ± 1.28	**97.46 ± 0.87**
AA (%)		87.14 ± 2.63	91.27 ± 2.33	84.19 ± 1.97	83.85 ± 2.12	82.73 ± 1.96	88.76 ± 1.39	89.72 ± 1.37	**93.14 ± 0.96**
Kappa ×100		93.58 ± 2.93	94.79 ± 2.78	92.38 ± 2.26	91.70 ± 2.61	91.47 ± 2.33	93.49 ± 1.53	95.81 ± 1.11	**96.03 ± 1.03**

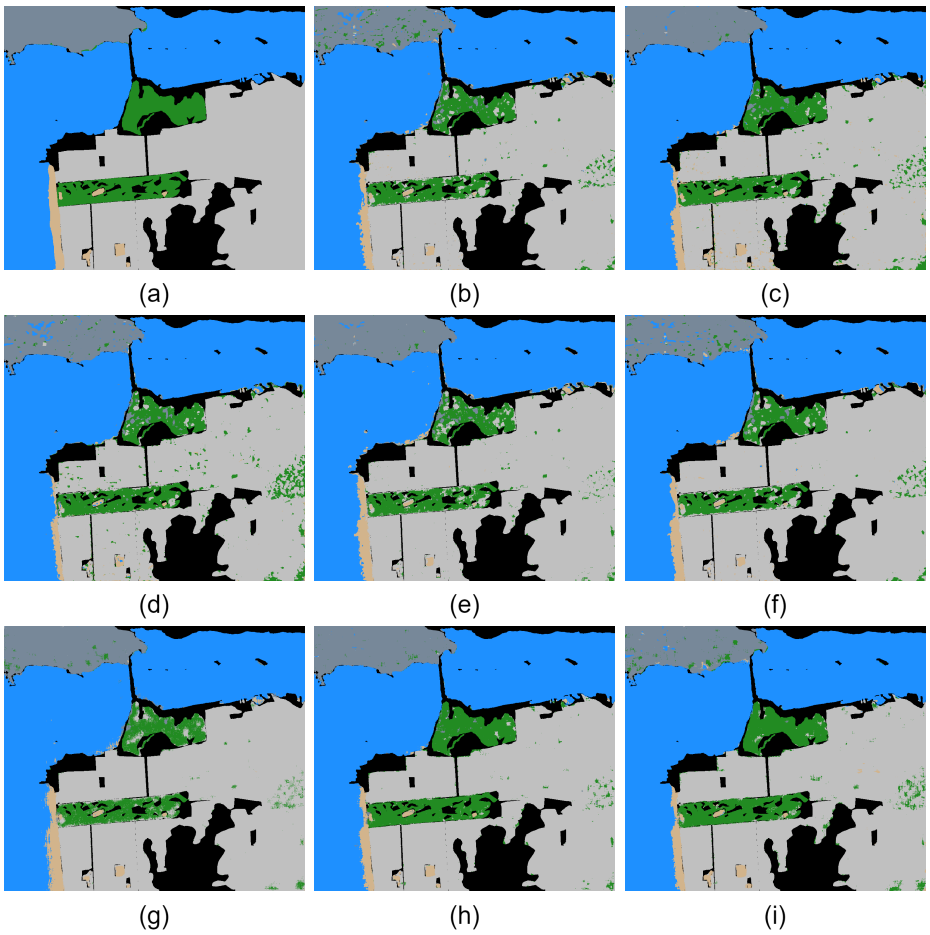


Figure 11. Classification maps on the San Francisco dataset. (**a**) Ground truth; (**b**) CV-2D CNN; (**c**) CV-3D CNN; (**d**) CV-2D–3D; (**e**) ViT; (**f**) Swin T; (**g**) PolSARFormer; (**h**) CV-MsAtViT; (**i**) CV-MfIaViT.

3.4.3. Classification Results on the Nanjing Dataset

Table 6 reports the six annotated class totals and model accuracies under the nominal 5% training configuration. The annotated counts sum to 342,689; they must not be interpreted as the training-set size. Figure 12 supplies the classification-map comparison. These results are evaluated against the supplied author-curated labels, not an independently verified complete land-cover map.

Table 6. Nanjing reported class accuracies and annotated reference-pixel counts. The second column is not a training-count column. Class accuracies are averaged over three reported runs; aggregate metrics include run-wise standard deviations. Accuracy units are percent. Results are conditional on the author-curated reference labels.

Class	Labeled Pixels	CV-2D CNN	CV-3D CNN	CV-2D–3D	ViT	Swin T	PolSAR-Former	CV-MsAtViT	CV-MfIaViT
Forest	110,901	87.12	93.57	97.68	92.12	92.40	93.07	**98.73**	97.51
Industrial	**3444**	12.22	63.44	43.50	56.06	75.05	62.99	86.32	**90.02**
Residential	43,018	83.53	92.51	91.79	90.23	90.05	93.52	**94.95**	94.64
Road	14,673	19.18	64.20	51.69	59.70	78.67	81.49	76.06	**86.49**
Water	95,418	96.33	98.86	97.84	96.64	99.19	98.66	**99.55**	98.51
Farmland	75,235	84.61	96.56	87.72	89.95	87.41	94.55	95.35	**98.67**
OA (%)		85.08 ± 3.32	93.99 ± 2.13	92.44 ± 2.34	90.95 ± 3.47	92.15 ± 2.31	94.18 ± 1.86	96.71 ± 1.36	**97.15 ± 0.98**
AA (%)		63.83 ± 3.89	84.86 ± 2.43	78.37 ± 2.16	80.78 ± 3.12	87.13 ± 2.21	87.38 ± 2.32	91.83 ± 1.24	**94.31 ± 0.72**
Kappa ×100		79.86 ± 3.56	91.96 ± 2.36	89.79 ± 2.47	87.87 ± 3.32	89.52 ± 2.43	92.24 ± 1.97	95.57 ± 1.44	**96.17 ± 1.04**

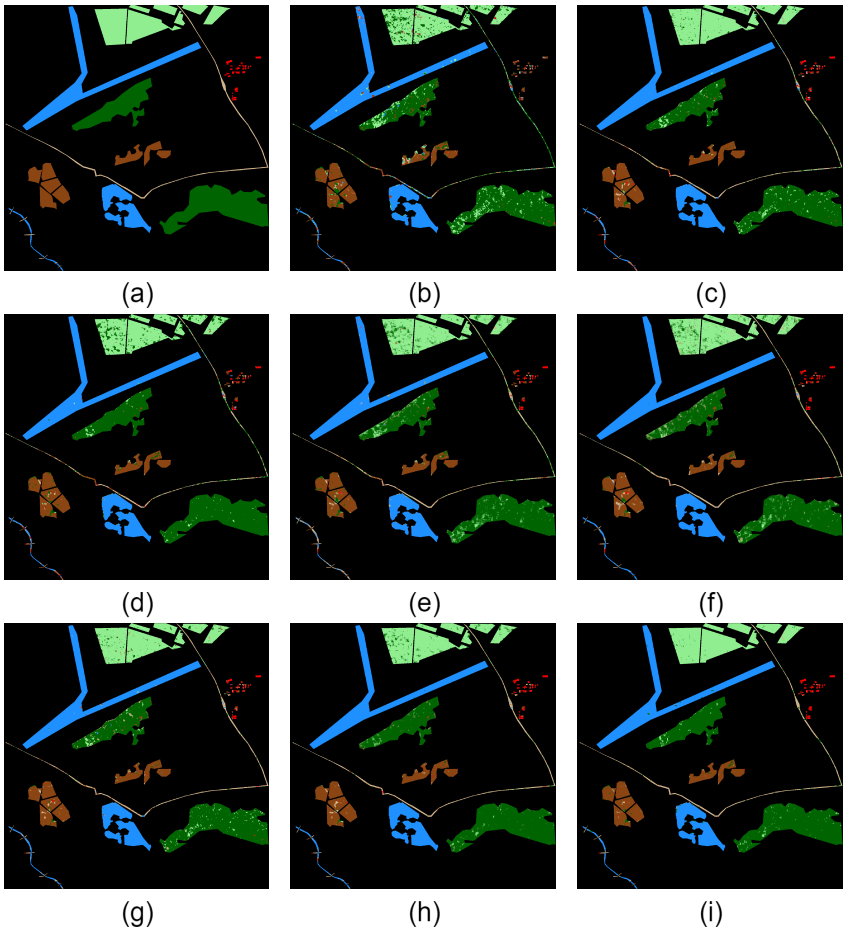


Figure 12. Classification maps on the Nanjing dataset. (**a**) Partial reference mask; (**b**) CV-2D CNN; (**c**) CV-3D CNN; (**d**) CV-2D–3D; (**e**) ViT; (**f**) Swin T; (**g**) PolSARFormer; (**h**) CV-MsAtViT; (**i**) CV-MfIaViT.

The CV-MfIaViT has reported means of 97.15 ± 0.98% OA, 94.31 ± 0.72% AA and 96.17 ± 1.04 for $\kappa \times 100$. The AA mean is 2.48 pp higher than CV-MsAtViT. Industrial and road differ by 3.70 and 10.43 pp in the class summaries, while forest and water are lower than the corresponding CV-MsAtViT values.

The images provide qualitative context for the labeled regions. Their apparent spatial agreement does not independently establish registration or annotation accuracy. The private-scene comparison therefore supports performance on the reported reference mask. The annotation-quality limitation is stated in Sections 2.5 and 4.

3.5. Results Under Different Sampling Rates

Figure 13 compares CV-3D CNN, Swin Transformer, CV-MsAtViT and CV-MfIaViT on Nanjing at nominal training proportions of 3%, 4%, 5%, 6% and 7%. At a given rate, the training budget is defined relative to the 342,689 labeled reference pixels with class stratification as described in Section 2.6; the realized percentage is $100N_{tr}/342{,}689$, where N_{tr} is the selected training count. Unassigned pixels are excluded from this denominator. Each plotted OA, AA and $\kappa \times 100$ value is the mean over three training runs at that rate. The input window and remaining configuration are kept at their reported settings. The curves are training-budget sensitivity evaluations, not an independent hyperparameter-selection procedure.

Larger labeled budgets are generally associated with higher plotted scores, although individual point sequences need not be monotonic. The four configurations show rate-dependent differences in their reported aggregate metrics. The plot does not establish that one architecture intrinsically requires a particular quantity of data, and it provides no isolated measurement of attention-learning stability.

Figure 14 provides enlarged map comparisons for selected regions on the three scenes. These illustrations complement the class summaries without serving as measurements of attention weights or the physical effect of the front end.

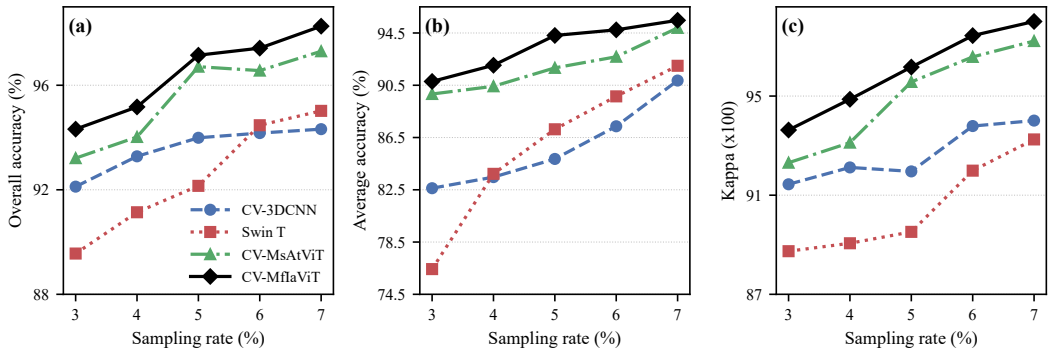


Figure 13. Nanjing aggregate metrics at nominal 3–7% labeled-pixel training rates. The panels (**a**–**c**) show three-run means of OA, AA and $\kappa \times 100$ for the reported configurations; no run-wise uncertainty bars are shown.

3.6. Generalization Under Spatial-Block Partition

After the configuration was fixed, spatial-block experiments were conducted without using their outcomes to modify the model or hyperparameters. Random-pixel sampling can place training and test centers close together, so sliding patches may share spatial context. This subsequent evaluation partitions each scene into a 3×3 grid: six blocks supply training samples and the remaining three supply test samples. The nominal training budget is kept at 1% of labeled pixels for Flevoland and San Francisco and 5% for Nanjing. Samples are drawn only from the assigned training region. The reported class-preserving constraint keeps each class represented in the training selection.

The spatial separation requirement concerns the complete patch footprint, not merely disjoint center coordinates. Three reported runs use different block assignments; their mean $\pm$ standard deviation reflects the combined regional configuration and training variation. Changing the six training blocks also changes the three complementary test blocks, so the variation cannot be attributed only to training-region selection.

Table 7 places these block results alongside the random-pixel metrics. Scene-level standardization is shared label-free preprocessing, while the 2D HT operates separately within each patch. The shared normalization statistics limit a strictly inductive interpretation even when the patch regions are separated. All the methods have lower means under the

reported block evaluation. The CV-MfIaViT's OA decreases by 3.87, 3.61 and 4.55 pp on Flevoland, San Francisco and Nanjing, respectively, and it retains the highest reported OA and AA means among the compared configurations. Relative to CV-MsAtViT, the block-protocol OA means are higher by 1.93, 1.75 and 1.78 pp on Flevoland, San Francisco and Nanjing, respectively; the corresponding AA differences are 3.31, 4.77 and 4.15 pp. These are descriptive differences between the reported means. This supports its cross-region performance under the stated protocol. It does not isolate a physical mechanism, establish an artifact-free HT operator or exclude the influence of shared whole-scene normalization.

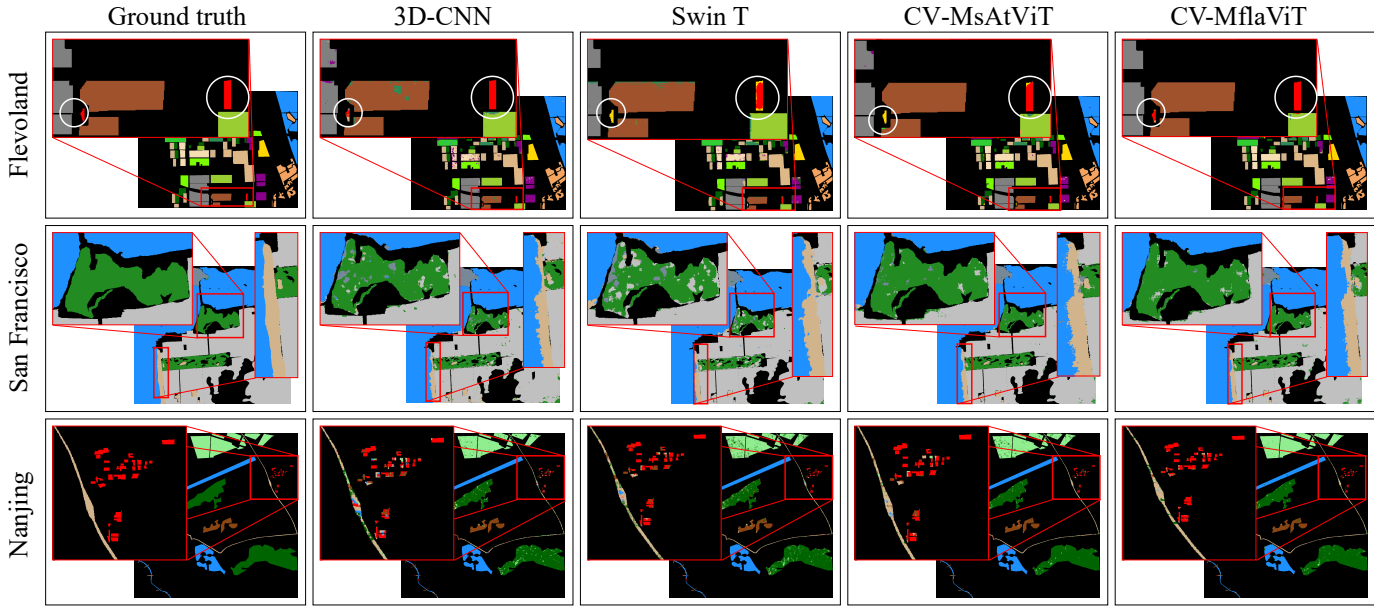


Figure 14. Minority-class results of CV-3D CNN, Swin T, CV-MsAtViT, and CV-MfIaViT across different datasets.

Table 7. Reported aggregate accuracy under random-pixel and spatial-block sampling. Values are mean ± standard deviation over three runs. Block runs retain the configuration fixed before this evaluation and use different complementary region assignments with shared scene-level normalization.

Dataset	Partition	Metric	CV-3D CNN	Swin T	PolSARFormer	CV-MsAtViT	CV-MfIaViT
Flevoland	Random (1%)	OA (%)	94.94±2.55	94.33±2.26	95.34±1.08	97.53±0.78	**98.15 ± 0.48**
		AA (%)	93.17±2.81	92.11±2.41	93.99±1.31	96.13±0.81	**97.96 ± 0.71**
	Block	OA (%)	88.62±1.25	87.15±1.38	89.47±1.12	92.35±0.95	**94.28 ± 0.76**
		AA (%)	85.40±1.42	84.02±1.55	87.12±1.28	89.84±1.10	**93.15 ± 0.88**
San Francisco	Random (1%)	OA (%)	96.66±2.46	94.56±2.41	95.85±1.57	97.35±1.28	**97.46 ± 0.87**
		AA (%)	91.27±2.33	82.73±1.96	88.76±1.39	89.72±1.37	**93.14 ± 0.96**
	Block	OA (%)	89.14±1.35	86.80±1.42	89.72±1.20	92.10±0.98	**93.85 ± 0.82**
		AA (%)	83.25±1.48	77.10±1.62	82.41±1.35	84.65±1.15	**89.42 ± 0.94**
Nanjing	Random (5%)	OA (%)	93.99±2.13	92.15±2.31	94.18±1.86	96.71±1.36	**97.15 ± 0.98**
		AA (%)	84.86±2.43	87.13±2.21	87.38±2.32	91.83±1.24	**94.31 ± 0.72**
	Block	OA (%)	86.74±1.45	85.32±1.50	87.65±1.30	90.82±1.10	**92.60 ± 0.85**
		AA (%)	76.20±1.65	78.90±1.58	80.45±1.42	84.60±1.25	**88.75 ± 0.98**

4. Discussion

Representation and attribution. As shown in Section 1.1.1, channel standardization is an invertible feature reparameterization for fixed statistics and positive channel scales; it does not preserve all the physical coherency-matrix properties. The reported results evaluate the complete pipeline and do not isolate the benefit of normalization against alternative scaling schemes. The spatial front end supplies analytic-envelope features to one pathway; it is not a polarimetric decomposition. Positive real attention preserves the phase of a non-zero learned input feature. The combined classification results and Cartesian/APD

ablation do not separately identify phase preservation, non-negativity, pooling or parameter changes as the cause of the gains. Internal complex mixing can alter phase even when the final gate preserves it.

Fixed and learned front ends. With APD enabled in Table 3, replacing the identity front end with 2D HT changes the reported OA mean from 97.91% to 98.15% and AA from 97.53% to 97.96%, differences of 0.24 and 0.43 pp. The fixed-front-end mean is also numerically higher than those of G1 and G3, demonstrating OA differences of 0.17 and 0.11 pp. These observations support the evaluated representation at the classification level; they do not distinguish its amplitude-description effect from possible boundary-dependent cues or establish physical scattering enhancement. Three runs do not establish superiority or equivalence. A fixed transform avoids learned front-end parameters and optimization but still requires computation within every processed patch. The input-area extrapolation describes a conditional cost scenario; matched operation accounting, memory and timing across sizes and reference models are not available.

Evaluation limits. The ablations are restricted to Flevoland, and the limited support of rare classes constrains class-specific uncertainty estimates. The random-pixel results informed configuration development and therefore describe development-stage comparisons rather than untouched-test estimates. In contrast, the subsequent spatial-block experiments used the previously fixed configuration, and no model or hyperparameter changes were made from their results. They provide a further assessment under separated-region sampling rather than another tuning stage. These experiments still use the same scenes and shared label-free scene-level normalization; they do not constitute evaluation on entirely new scenes or a strictly inductive protocol. The reference methods retain their original settings, so the comparison does not equalize per-model tuning budgets. The single-run window curves do not estimate training variability; the random-pixel standard deviations describe variation that is conditional on a fixed labeled split. The patch-wise FFT remains local to each input patch. Exact split indices, block-boundary handling, initialization seeds and final settings are needed for exact numerical replication.

Nanjing is an author-curated SAR crop with a manually prepared partial reference mask. Water annotation favors interior regions where the annotator judged the class membership to be unambiguous. This choice may favor easier water pixels and does not establish performance at shorelines or over the full water extent. The reported accuracies concern the selected labeled test population and do not measure prediction accuracy in unassigned regions. Partial coverage was intended to limit the training-sample budget; it does not itself establish stronger prediction accuracy or a measured training-speed benefit. Selective annotation limits the representativeness of the assessment for the entire scene; broader validation would require additional reference observations.

Further evaluation. Useful extensions include controlled HT placement comparisons, direct envelope and attention-weight visualization, a matched-mask phase experiment and a common-hardware accuracy–cost study. For large scenes, pruning, distillation, mixed precision and parallel inference may reduce the deployment burden; their benefits require measurement in the target implementation.

5. Conclusions

The CV-MfIaViT combines a spatial analytic-envelope front end, complex-valued multiscale feature extraction, cross-pooling attention and transformer modeling. Its final positive real attention gate preserves the phase of non-zero input features; the empirical evaluation concerns the complete design rather than phase preservation alone.

On the supplied reference labels, the reported random-pixel OA means are 98.15%, 97.46% and 97.15% for Flevoland, San Francisco and Nanjing, respectively. The proposed

model also retains the highest reported OA and AA means in the spatial-block comparison of Table 7 under shared scene-level normalization. The class gains and declines show that the effects depend on the evaluated scene and class.

Figure 9 identifies AA peaks at 15×15, 13×13 and 15×15 within the tested ranges. Larger windows do not improve accuracy monotonically, and their input scale increases; the area-only extrapolation is a conditional illustration, not a measured model cost. Figure 13 further shows the dependence of the reported Nanjing metrics on the labeled training budget.

The spatial-block results provide a subsequent assessment of the fixed configuration without further model revision. Future work should extend the evaluation to new scenes and strengthen annotation validation, reproducible split/run reporting, direct feature interpretation and matched-hardware cost measurement. The current conclusions remain specific to the three scenes, the stated configuration-development and sampling protocols, and the supplied reference labels.

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

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers 61501228 and 62171220), the Natural Science Foundation of Jiangsu Province (grant number BK20140825), the Aeronautical Science Foundation of China (grant numbers 20152052029 and 20182052012), the Basic Research Foundation (grant numbers NS2015040 and NS2021030), the National Science and Technology Major Project (grant number 2017-11-0001-0017), and the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education (grant number NJ20240002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Flevoland and San Francisco datasets are publicly available from the NASA/JPL AIRSAR archive (https://airsar.jpl.nasa.gov/ (accessed on 30 October 2026)). The Nanjing SAR imagery was supplied by the authors' research group, and its partial reference labels were manually prepared for this study. The experimental crop and reference mask are available from the corresponding author upon reasonable request.

Acknowledgments: During preparation and revision, the authors used Claude Opus 5.5 (Anthropic) and GPT-6.1 Sol (OpenAI) to assist with language editing, manuscript organization, technical exposition and LaTeX formatting. The authors have reviewed and edited the content and take full responsibility for the published version.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PolSAR	polarimetric synthetic aperture radar
ViT	vision transformer
CV-MfIaViT	complex-valued multiscale feature interaction attention vision transformer
CV-APIA	complex-valued amplitude–phase interaction attention
CV-FCM	complex-valued feature complementary mapper
CV-MSE	complex-valued multiscale extractor
CV-ViT	complex-valued vision transformer
CV-MsAtViT	complex-valued multiscale attention vision transformer

FM	feature-mapping mechanism
APD	amplitude–phase decoupling
2D HT	two-dimensional Hilbert transform
CLN	complex layer normalization
CReLU	Cartesian ReLU
OA	overall accuracy
AA	average accuracy
MHA	multi-head attention
MLP	multilayer perceptron
SE	squeeze-and-excitation
ECA	efficient channel attention
CBAM	convolutional block attention module
CA	coordinate attention
Conv	convolutional layer
BN	batch normalization
SNR	signal-to-noise ratio
DFT	discrete Fourier transform
FFT	fast Fourier transform
FLOPs	floating-point operations
GT	reference (ground-truth) labels
C-APIA	Cartesian-coordinate attention variant

References

1. Ban, Y.; Jacob, A.; Gamba, P. Spaceborne SAR Data for Global Urban Mapping at 30 M Resolution Using a Robust Urban Extractor. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 28–37. [CrossRef]
2. Whitcomb, J.; Chen, R.; Clewley, D.; Kimball, J.; Pastick, N.; Yi, Y.; Moghaddam, M. Active Layer Thickness Throughout Northern Alaska by Upscaling from P-Band Polarimetric Sar Retrievals. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022 ; pp. 3660–3663. [CrossRef]
3. Zhang, Y.; Sun, Z.; Chang, S.; Tang, B.; Hou, B. SegS-YOLO: A YOLO-based instance segmentation approach for small-scale ship targets in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 2658–2679. [CrossRef]
4. Wu, X.; Sun, Z.; Wu, H.; Ji, K. DSENet: A Detail and Semantic Enhanced Network for Video SAR Moving Target Shadow Detection. *Remote Sens.* **2026**, *18*, 1623. [CrossRef]
5. Xu, X.; Leng, X.; Sun, Z.; Tan, X.; Ji, K.; Kuang, G. Adaptive fast refocusing for ship targets with complex motion in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 8559–8572. [CrossRef]
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
7. Hu, Y.; Fan, J.; Wang, J. Classification of PolSAR Images Based on Adaptive Nonlocal Stacked Sparse Autoencoder. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1050–1054. [CrossRef]
8. Lardeux, C.; Frison, P.L.; Tison, C.; Souyris, J.C.; Stoll, B.; Fruneau, B.; Rudant, J.P. Support Vector Machine for Multifrequency SAR Polarimetric Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 4143–4152. [CrossRef]
9. Yin, Q.; Cheng, J.; Zhang, F.; Zhou, Y.; Shao, L.; Hong, W. Interpretable POLSAR Image Classification Based on Adaptive-Dimension Feature Space Decision Tree. *IEEE Access* **2020**, *8*, 173826–173837. [CrossRef]
10. Liu, W.; Yang, J.; Li, P.; Han, Y.; Zhao, J.; Shi, H. A Novel Object-Based Supervised Classification Method with Active Learning and Random Forest for PolSAR Imagery. *Remote Sens.* **2018**, *10*, 1092. [CrossRef]
11. Pottier, E. Dr. J.R. Huynen's Main Contributions in the Development of Polarimetric Radar Techniques and How the 'Radar Targets Phenomenological Concept' Becomes a Theory. In Proceedings of the SPIE 1748, Radar Polarimetry, San Diego, CA, USA, 23–24 July 1992; pp. 72–85. [CrossRef]
12. Cloude, S.R.; Pottier, E. A Review of Target Decomposition Theorems in Radar Polarimetry. *IEEE Trans. Geosci. Remote Sens.* **1996**, *34*, 498–518. [CrossRef]
13. Cloude, S.R.; Pottier, E. An Entropy Based Classification Scheme for Land Applications of Polarimetric SAR. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 68–78. [CrossRef]
14. Krogager, E.; Boerner, W.M.; Madsen, S.N. Feature-Motivated Sinclair Matrix (sphere/diplane/helix) Decomposition and Its Application to Target Sorting for Land Feature Classification. In Proceedings of the SPIE 3120, Wideband Interferometric Sensing and Imaging Polarimetry, San Diego, CA, USA, 28–29 July 1997; pp. 144–154. [CrossRef]

15. Touzi, R. Target Scattering Decomposition in Terms of Roll-Invariant Target Parameters. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 73–84. [CrossRef]
16. Freeman, A.; Durden, S.L. A Three-Component Scattering Model for Polarimetric SAR Data. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 963–973. [CrossRef]
17. Wang, W.; Wang, J.; Quan, D.; Yang, M.; Sun, J.; Lu, B. PolSAR Image Classification via a Multigranularity Hybrid CNN-ViT Model with External Tokens and Cross-Attention. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8003–8019. [CrossRef]
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NeurIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.
19. Zhang, Z.; Wang, H.; Xu, F.; Jin, Y.Q. Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7177–7188. [CrossRef]
20. Zhou, Y.; Wang, H.; Xu, F.; Jin, Y.Q. Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1935–1939. [CrossRef]
21. Zhang, L.; Chen, Z.; Zou, B.; Gao, Y. Polarimetric SAR Terrain Classification Using 3D Convolutional Neural Network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 4551–4554. [CrossRef]
22. Li, W.; Xia, H.; Zhang, J.; Wang, Y.; Jia, Y.; He, Y. Complex-Valued 2D-3D Hybrid Convolutional Neural Network with Attention Mechanism for POLSAR Image Classification. *Remote Sens.* **2024**, *16*, 2908. [CrossRef]
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
24. Jamali, A.; Roy, S.K.; Bhattacharya, A.; Ghamisi, P. Local Window Attention Transformer for Polarimetric SAR Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4004205. [CrossRef]
25. Alkhatib, M.Q. PolSAR Image Classification Using Complex-Valued Multiscale Attention Vision Transformer (CV-MsAtViT). *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *137*, 104412. [CrossRef]
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542. [CrossRef]
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722. [CrossRef]
30. Wang, J.; Li, Y.; Quan, D.; Hou, B.; Wang, Z.; Sima, H.; Sun, J. MAPM: PolSAR Image Classification with Masked Autoencoder Based on Position Prediction and Memory Tokens. *Remote Sens.* **2024**, *16*, 4280. [CrossRef]
31. Kuang, Z.; Bi, H.; Li, F.; Xu, C.; Sun, J. Polarimetry-Inspired Contrastive Learning for Class-Imbalanced PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5212819. [CrossRef]
32. Kuang, Z.; Liu, S.; Bi, H.; He, L.; Li, F. CV-CPKAN: Complex-Valued Convolutional Kolmogorov–Arnold Framework for PolSAR Image Classification. *Remote Sens.* **2026**, *18*, 330. [CrossRef]
33. Dong, H.; Zhang, L.; Zou, B. Exploring Vision Transformers for Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5219715. [CrossRef]
34. Wang, H.; Xing, C.; Yin, J.; Yang, J. Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer. *Remote Sens.* **2022**, *14*, 4656. [CrossRef]
35. Yin, Q.; Lin, Z.; Hu, W.; López-Martínez, C.; Ni, J.; Zhang, F. Crop Classification of Multitemporal PolSAR Based on 3-D Attention Module with ViT. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4005405. [CrossRef]
36. Alkhatib, M.Q. PolSAR Image Classification Using a Hybrid Complex-Valued Network (HybridCVNet). *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4017705. [CrossRef]
37. Wu, Q.; Hou, B.; Wen, Z.; Ren, Z.; Jiao, L. Cost-Sensitive Latent Space Learning for Imbalanced PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4802–4817. [CrossRef]
38. Dong, H.; Zhang, L.; Lu, D.; Zou, B. Attention-Based Polarimetric Feature Selection Convolutional Network for PolSAR Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4001705. [CrossRef]
39. Hua, W.; Wang, X.; Zhang, C.; Jin, X. Attention-Based Multiscale Sequential Network for PolSAR Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4506505. [CrossRef]
40. Trabelsi, C.; Bilaniuk, O.; Zhang, Y.; Serdyuk, D.; Subramanian, S.; Santos, J.F.; Mehri, S.; Rostamzadeh, N.; Bengio, Y.; Pal, C.J. Deep Complex Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

41. Pei, S.; Song, H.; Lu, Y. Small Sample Hyperspectral Image Classification Method Based on Dual-Channel Spectral Enhancement Network. *Electronics* **2022**, *11*, 2540. [CrossRef]
42. Zhang, H. Multidimensional Analytic Signals and the Bedrosian Identity. *Integral Equ. Oper. Theory* **2014**, *78*, 301–321. [CrossRef]
43. Xiao, D.; Liu, C.; Wang, Q.; Wang, C.; Zhang, X. PolSAR Image Classification Based on Dilated Convolution and Pixel-Refining Parallel Mapping Network in the Complex Domain. *arXiv* **2019**, arXiv:1909.10783.
44. Zhao, W.; Yu, X.; Yang, Z.; Yin, Z.; Zhang, Q. Polarimetric SAR Deep Learning Classification Using Multi-Scale Features. *Mod. Radar* **2023**, *45*, 48–54.
45. Gao, G.; Li, C.; Zhang, X.; Yao, B.; Chen, Z. Mamba-CNN Hybrid Multi-Scale Ship Detection Network Driven by a Dual-Perception Feature of Doppler and Scattering. *ISPRS J. Photogramm. Remote Sens.* **2026**, *232*, 656–674. [CrossRef]
46. Arjovsky, M.; Shah, A.; Bengio, Y. Unitary Evolution Recurrent Neural Networks. In Proceedings of the Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1120–1128.
47. Cao, Y.; Wu, Y.; Li, M.; Liang, W.; Zhang, P. PolSAR Image Classification Using a Superpixel-Based Composite Kernel and Elastic Net. *Remote Sens.* **2021**, *13*, 380. [CrossRef]
48. Xing, X.; Chen, Q.; Yang, S.; Liu, X. Feature-Based Nonlocal Polarimetric SAR Filtering. *Remote Sens.* **2017**, *9*, 1043. [CrossRef]
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]

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