remote sensing

MDPI

Article

Interpretability of Transformer Structures Based on Attribution Inference for Polarimetric SAR Image Classification

Lijia Dong [1], Jie Geng [1,*], Yuhang Zhang [1], Yuhan Zhou [2] and Wen Jiang [1]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China; donglijianwpu@mail.nwpu.edu.cn (L.D.); zhangyuhang@mail.nwpu.edu.cn (Y.Z.); jiangwen@nwpu.edu.cn (W.J.)
[2] Queen Mary University of London Engineering School, Northwestern Polytechnical University, Xi'an 710072, China; zhouyvhan@mail.nwpu.edu.cn
* Correspondence: gengjie@nwpu.edu.cn

Highlights

What are the main findings?

- We propose an attribution inference framework that integrates structural dependencies and gradient-based saliency to reveal the internal decision mechanisms of Transformer-based PolSAR classification models.
- The proposed framework quantitatively evaluates attention head contributions, traces key information flows, and identifies salient feature regions, demonstrating its effectiveness across PolSAR datasets.

What are the implications of the main findings?

- The proposed framework provides a systematic and comprehensive approach to interpreting the internal decision mechanisms of Transformer-based PolSAR classification models.
- The attribution-based analysis offers valuable insights into model interpretability and reliability, providing practical guidance for attention head pruning, model compression, and feature visualization.

Abstract

Transformer-based deep networks have gained prominence in Polarimetric Synthetic Aperture Radar (PolSAR) image classification due to their strong capability for modeling long-range dependencies. However, existing Transformer-based methods fail to reveal the polarimetric scattering characteristics and their spatial interactions in classification decisions, which leads to limited interpretability and reduced reliability of the models. To address this challenge, this paper proposes an interpretability analysis framework based on attribution inference for Transformer-based PolSAR image classification. Specifically, an attribution inference module is first constructed by integrating structural dependency relationships with gradient-based saliency information, enabling explicit simulation of polarimetric scattering features and their interactions within the model. To enhance the interpretability of PolSAR classification models, we further conduct a full-chain analysis of the Transformer's internal structure based on the attribution inference matrix, rather than relying on qualitative visualization from a single attention map. This integrated approach enables precise localization of critical attention heads and information flows, while simultaneously identifying discriminative regions that exert a decisive influence on the prediction results. To validate the proposed method, quantitative and qualitative experiments are performed on three PolSAR image classification datasets, which clearly demonstrate the effectiveness of the proposed interpretability framework.

Check for updates

Academic Editor: John Trinder

Received: 28 September 2026
Revised: 6 November 2026
Accepted: 6 November 2026
Published: 8 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: interpretability analysis; attribution inference; Transformer framework; polarimetric synthetic aperture radar

1. Introduction

Synthetic Aperture Radar (SAR) is a powerful high-resolution imaging system with the ability to penetrate clouds, rain, fog, and can complete the observation of ground targets even in harsh weather conditions [1–3]. Compared with conventional SAR systems, fully polarized synthetic aperture radar (PolSAR) transmits both horizontal and vertical polarized waves, enabling the acquisition of target backscatter under diverse polarization combinations. Therefore, PolSAR systems can perform more comprehensive coherent scattering measurements of ground objects, providing richer and more informative data. These advantages give PolSAR systems broad application prospects in environmental monitoring, agriculture, geological disaster detection, urban planning, and military reconnaissance. PolSAR image classification serves as a key technology for the automatic interpretation of PolSAR data [4–7]. Specifically, PolSAR image classification assigns a category label to each pixel, enabling the interpretation of feature target information within the scene. Currently, the rise of deep learning algorithms continues to promote the development of image classification. Among them, the visual Transformer network has also been used in the PolSAR image classification task and achieved excellent performance [8–10]. Although deep learning has made remarkable progress in PolSAR image classification, its inherent "black box" character has raised concerns regarding interpretability. In deep neural networks, the decision-making process is often opaque, where inputs are provided and outputs are returned, but the internal reasoning remains unclear. This lack of interpretability limits their wide application, particularly in complex applications such as pixel-level PolSAR image classification. To improve the transparency and reliability of deep learning models, extensive research efforts have been made. Interpretability methods are generally categorized into ante-hoc and post-hoc approaches, depending on whether the explanation is incorporated.

Ante-hoc interpretability models are typically designed with simple and transparent structures, making them inherently easy to interpret [11,12]. For example, decision trees [13,14] consist of internal nodes representing attributes and leaf nodes representing classes, with each path from the root to a leaf corresponding to a clear decision rule. Other inherently interpretable models that do not require additional algorithms to explain their decision processes include linear models [15] and naive Bayes classifiers [16]. Yellinek et al. [17] proposed a tree-augmented visual language model that expands the text data in the input image–text pairs into a hierarchical structure and fuses it with the visual representation, thereby enhancing the transparency and interpretability of the network. Although structurally simple ante-hoc interpretability models are inherently interpretable, their limited parameter often constrains their capacity to fit complex patterns. Therefore, additional interpretability structures are often integrated as modules within deep learning models to achieve built-in interpretability. Huang et al. [18] proposed an interpretable model for fine-grained recognition, with the core idea of integrating region-based local exploration and decision attribution into the deep network. The model evaluates the importance of target parts to the final prediction, which enhances the interpretability of the network. Chen et al. [19] proposed a prototype part network to simulate the process of human inference for image classification. This network analyzes the image by identifying prototype parts and aggregates their associated evidence to perform the final classification. Zhao et al. [20] proposed an interpretability framework based on token-level attribution

analysis, which evaluates contextual sensitivity at different embedding positions and identifies the critical locations contributing to model decisions. Wang et al. [21] proposed a region-aware Transformer architecture that improves model interpretability by adaptively modeling network parameters. Although embedding interpretability modules is effective in certain specific scenarios, its generalizability is limited. Such methods often require task-specific and dataset-dependent design of interpretability components, which becomes particularly challenging when dealing with PolSAR image classification.

Post-hoc interpretability methods, known for their greater generalizability, aim to explain already trained models by constructing explanatory frameworks that reveal the underlying mechanisms, decision-making processes, and reasoning basis of deep learning models. Post-hoc interpretability methods can be divided into global and local approaches based on the target of explanation. Global interpretability focuses on understanding the overall behavior of the model and unveiling the internal logic behind its "black-box" nature. Zhang et al. [22] employed decision trees to decompose the deep features of the convolutional neural network (CNN) into fundamental concepts corresponding to target parts. Xu et al. [23] used model distillation to reduce the complexity of the deep learning network, and extracted dark knowledge through the compact student network to intuitively understand the decision-making logic of the model. He et al. [24] introduced the geometric median of the convolution kernel to evaluate their contribution and analyze their influence on the network's decision-making process. Local interpretability methods aim to understand the model's decision-making process by analyzing the specific local feature regions that the model focuses on during prediction. Class Activation Mapping (CAM) [25] is a classic local interpretable method, which uses class activation maps to identify image regions related to specific categories. However, its applicability is limited by the requirement that the original network must include a global average pooling layer. To address this limitation, Selvaraju et al. [26] proposed Grad-CAM to obtain the fusion weight of the target feature map through reverse gradient, so that the class activation map can be extended to each layer of the network. Subsequently, Chattopadhay et al. [27] proposed Grad-CAM++, which refines the localization by considering the individual contributions of each element in the gradient map, producing more precise class activation maps. Khalid et al. [28] improved the capability of existing interpretability methods in handling overlapping objects and conveying linguistic information by combining instance-level thermodynamic maps for spatial localization with a conceptual approach based on non-negative matrix factorization, thereby integrating the advantages of both types of interpretability methods. Although post-hoc interpretability has advanced in the field of computer vision, the decision-making mechanisms of current deep learning models for remote sensing remain poorly understood. This limitation reduces the reliability of model predictions, particularly in high-risk applications.

With the outstanding performance of Transformer networks across a wide range of tasks, research on their interpretability has gained increasing attention [9]. The powerful performance of the Transformer largely benefits from the multi-head attention mechanism, which allows the model to learn different interdependencies on multiple feature subspaces [29]. Chefer et al. [30] believe that each attention head in the multi-head attention mechanism obtains different information when participating in model prediction, and introduce gradients to analyze the contribution of each attention head. In addition, there are a huge number of input tokens in the Transformer block of each layer, and each token has different effects on model prediction. Abnar et al. [31] explained the model decision-making process by analyzing the linear transfer of Token information between Transformer layers. However, existing interpretability methods for Transformer-based frameworks mainly focus on specific internal structures, such as attention heads or tokens, and the tailored algorithms are designed accordingly. There is still a lack of a general

framework to explain the overall decision-making mechanism of Transformer architectures, which limits a comprehensive understanding of model behavior.

To solve the above issues, this paper proposes an interpretability analysis framework for PolSAR image classification based on attribution inference. This framework analyzes multiple interpretability perspectives and proposes three interpretability analysis methods. Specifically, from the perspective of network structure analysis, an attention head contribution measurement method is proposed to score and rank the importance of different heads. From the perspective of feature transfer, an information flow inference for key tokens is introduced to reveal the hierarchical interactions within the Transformer. Additionally, from the perspective of feature attention, a salient feature position method is designed to accurately identify object-level critical decision regions in the features. The key contributions of our study can be summarized as follows:

(1) This paper introduces the concept of attribution inference into the interpretability evaluation of Transformer models. To the best of our knowledge, this is the first deep learning-based interpretability approach specifically designed for PolSAR image classification, which provides practical application significance support for model compression and feature visualization.
(2) An attention causality measurement model is proposed to compute the attribution inference matrix for each layer of the Transformer network. These matrices capture both the saliency of individual inputs and their interaction dependencies, which provide a quantitative basis for evaluating the contributions of different components to the model's decisions.
(3) A unified interpretability analysis method based on the attribution inference matrix is proposed to analyze the internal mechanisms of Transformers from three perspectives, including attention head importance, key Token information flow, and salient feature position, which reduces the computational complexity of designing separate interpretation schemes for different components within the model.

The rest of this article is organized as follows: Sections 2.1 and 2.2 review related works on interpretable PolSAR image classification and attribution inference. Sections 2.3–2.7 introduce the proposed AIIA framework and detail its core components. Section 3 designs experiments on three public PolSAR datasets and reports the experimental results. Section 4 discusses the ablation study and the generalizability of the proposed method to PolSAR imagery acquired at other frequency bands. Finally, Section 5 summarizes the conclusions of the proposed method.

2. Materials and Methods

2.1. Interpretable PolSAR Image Classification

Deep learning algorithms have made significant advances in PolSAR image classification. However, the end-to-end training approach often results in limited model interpretability. The interpretability of the model is crucial for tasks such as high-reliability scenario applications, model security, and model lightweight [32]. Huang et al. [33] innovatively regarded the SAR image dataset as a corpus, generated abstract physical prior knowledge through unsupervised learning, and used this interpretable feature for SAR image classification. However, such prior knowledge obtained through unsupervised learning can only serve as auxiliary guidance for model training, and the underlying decision-making mechanism of the entire network still lacks interpretability. Fan et al. [34] introduced an adaptive grad-weight selection class activation explainer for oil spill detection in SAR imagery. This method adaptively selects feature importance weights for the target class based on activation maps, which effectively reduces the impact of background noise. The explainer's output is

used to provide feedback for network adjustment, thereby enhancing the interpretability of the deep learning framework. Feng et al. [35] incorporated the attribute scattering center model from traditional methods into deep learning networks to enhance the interpretability of the generated features. They further conducted interpretability analysis through feature importance visualization and decision maps. The aforementioned methods enhance network interpretability by introducing additional modules. However, such data-specific module designs are often complex and task-dependent. Moreover, these approaches do not establish a systematic interpretability framework for deep learning networks, as they primarily focus on evaluating the local importance of features rather than providing a comprehensive explanation.

2.2. Attribution Inference

Attribution inference aims to reveal the causal relationship between input samples and output decisions [36]. It explains the model's decision-making process by assessing the contribution of input features to the final output. With the development of large models and massive data volume, the richness of semantic information and the increased number of parameters have significantly reduced model interpretability. To improve model reliability, there has been a growing emphasis on the interpretability of networks. Petsiuk et al. [37] generated saliency maps by measuring detection similarity, providing local interpretability of input features in a "black-box" scenario where only the input and output are accessible. In addition, different tasks require specific interpretability methods. In [38], for contrastive language-image pretraining models, a gradient-based interpretability method is proposed. This method evaluates the impact of local image regions or words on the output through decoder decoupling and feature similarity matching. Zhao et al. [39] proposed a gradient-weighted object detector activation map, which generates instance-specific explanations compared to traditional CAM. This work focuses on the characteristics of object detection tasks and performs interpretability analysis from both target specification and object recognition perspectives. Pixel-level PolSAR classification tasks face higher reliability requirements, which motivates us to propose a systematic interpretability framework specifically for PolSAR classification. This framework not only addresses feature-level interpretability but also considers the aspects of network architecture and information flow.

2.3. Overview of the AIIA Framework

The framework of the attribution inference interpretability analysis method (AIIA) of Transformer models for PolSAR image classification is shown in Figure 1. The proposed method first derives the attribution inference matrix for each layer of the Transformer network using the attention causality measurement model. This matrix captures the contribution of token-to-token interactions within the attention weight matrix to the network's decision-making process. Completely based on the attribution inference matrix, the proposed interpretability analysis framework explains the model's mechanism from three perspectives, thereby enhancing the interpretability of Transformer-based PolSAR image classification networks: (1) *Attention head contribution measurement*. By identifying high-contribution connections within the attention mechanism, the framework analyzes components of the network structure that significantly influence prediction outcomes, which is a perspective of network structure analysis. (2) *Information flow inference of key tokens*. Constructing the critical path by quantifying the contribution of input tokens in each layer to network decision-making is an interpretable perspective of network feature transfer. (3) *Salient feature position*. By generating attribution masks of each layer to obtain the feature areas that model decision-making focuses on, it is an interpretable perspective of feature attention. Based on attribution inference in Transformer networks, we construct a

comprehensive interpretability analysis framework that integrates multiple perspectives, which offers support for network structure pruning, decision path inference, and feature visualization techniques.

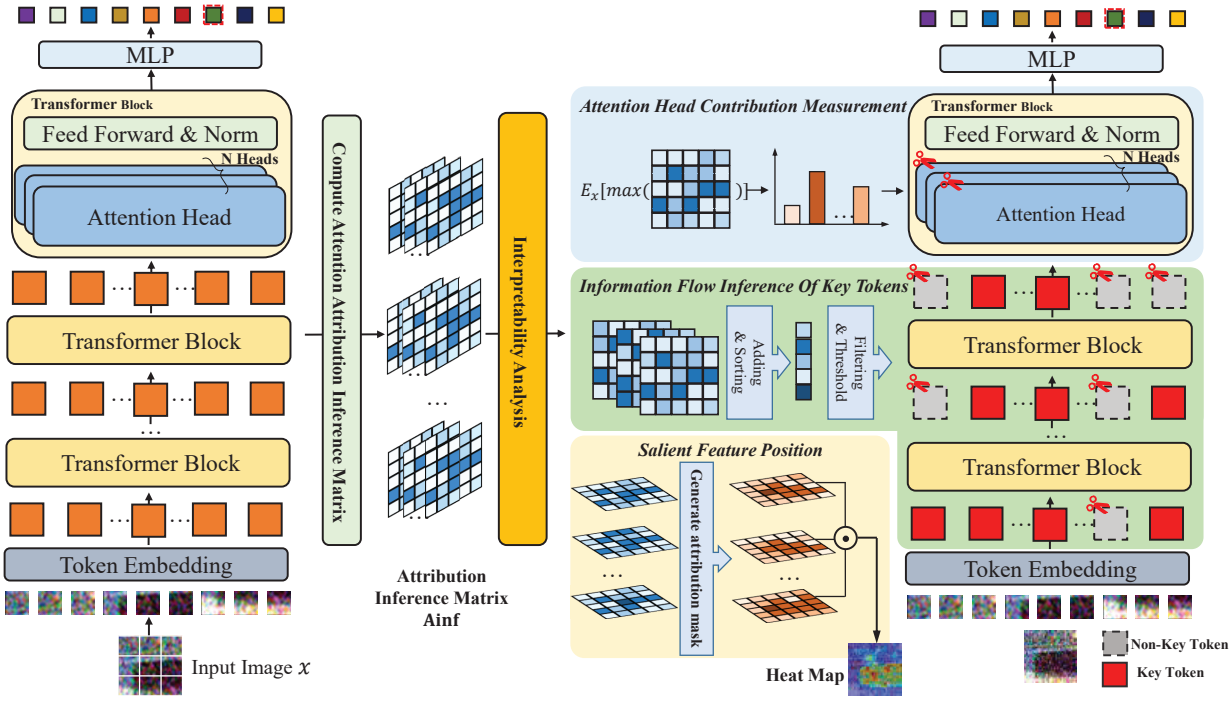

Figure 1. The overall pipeline of the interpretability analysis framework based on attribution inference (AIIA) for Transformer-based PolSAR image classification. The framework includes three interpretability perspectives, namely attention head contribution measurement, key Token information flow inference, and salient feature position. In the figure, ⊙ denotes the Hadamard product.

2.4. Attention Causality Measurement Model

Similar to the CNN network, the Transformer network model is also composed of multiple-layer stacks, which largely rely on the multi-head self-attention mechanism to capture the relationship between feature tokens. The output token of a Transformer layer is obtained by computing a weighted sum of the input tokens, where the weights are derived from the self-attention matrix. Given that the input of layer l is x_l, the working mechanism of the self-attention mechanism of head h can be described by the following formula:

$$\begin{aligned} Q_h = X_l W_h^Q, K_h = X_l W_h^K, V_h = X_l W_h^V, \\ A_h = softmax(\frac{Q_h, K_h^T}{\sqrt{d_k}}), \\ O_h = AttentionHead(X) = A_h V_h, \end{aligned} \tag{1}$$

where $Q, K \in R^{n \times d_k}, V \in R^{n \times d_v}$ is the linear mapping projection matrix, A_h is the self-attention weight matrix of the h-th head, and the score $A_h^{i,j}$ is the i,j-th element of the self-attention weight matrix. Intuitively, $A_h^{i,j}$ represents the degree of attention of the i-th token to the j-th token. Generally, there are multiple self-attention heads in a Transformer layer, and the input of the multi-head attention mechanism can be expressed as follows:

$$MultiH(X) = [O_1, O_2, ..., O_H]W^o, \tag{2}$$

where H is the number of attention heads of the multi-head attention mechanism, and O_h is the output of the h-th head. The multi-head attention mechanism is a fundamental

component of Transformers. Thus, analyzing and understanding its functioning is crucial to deciphering the black-box model of Transformer networks. In previous work, researchers usually focus on the self-attention weight matrix to analyze the information interaction between input tokens [40]. However, attention weight matrices typically contain dense Token connections, and the presence of multiple attention heads further complicates the analysis of information interactions within the Transformer. The attention weights in Equations (1) and (2) characterize the interaction strengths between different Tokens, but do not directly reflect the contributions of these interactions to the final classification decision [41]. A connection with a high attention weight may have a limited influence on model predictions, as its actual contribution also depends on subsequent feature transformations and classification operations. Therefore, attention weights alone are insufficient for evaluating the importance of individual Token interactions in model decision-making. To overcome this limitation, we introduce gradient information that captures the sensitivity of model predictions to variations in attention connections and integrate it with the original attention weights through the proposed attention causality measurement model (ACMM). This integration produces an attribution inference matrix in which each element quantifies the contribution of a specific Token interaction to model predictions, thereby establishing a quantitative relationship between attention connections and classification decisions. The overall architecture of ACMM is illustrated in Figure 2.

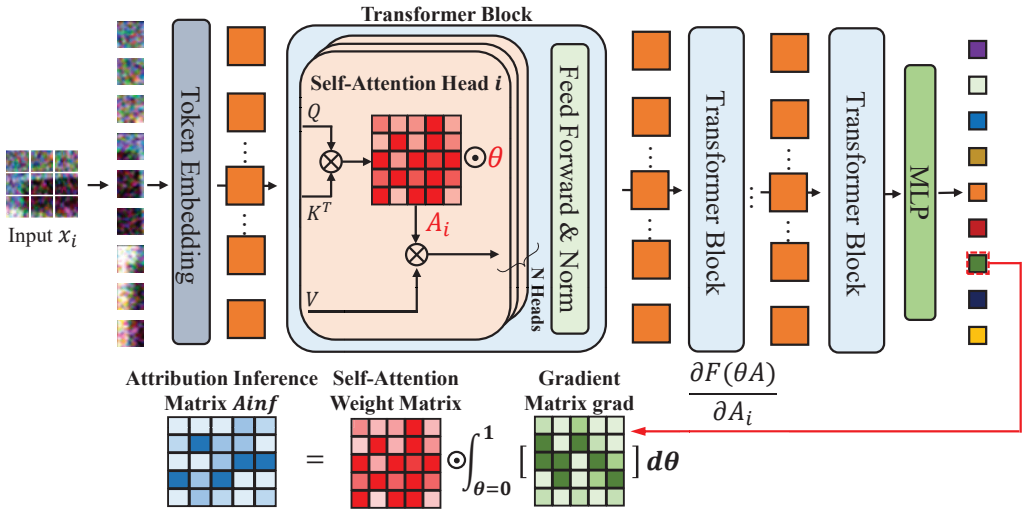

Figure 2. Illustration of the attention causality measurement model (ACMM). The attribution inference matrix is constructed by integrating the self-attention matrix with the gradient matrix (grad), reflecting the contributions of different Token interactions to model decisions. The symbols ⊙ and ⊗ denote the Hadamard product and Kronecker product, respectively.

Specifically, given the input sequence x and the Transformer network $F_x(\cdot)$ to be analyzed, we take the attention weight matrix A as the input of $F_x(\cdot)$. In addition, a scaling factor θ is introduced to manipulate the intensity of the internal attention weights within the Transformer network. It allows for the examination of how variations in the strength of token-to-token interactions, represented by $F_x(\theta A)$, influence the model's decision-making process. The function of ACMM is to calculate the attribution score of the attention connection between each group of tokens. Taking one layer of the Transformer model as an example, the calculation formula for the attribution inference matrix of head h is as follows:

$$Ainf_h(A) = A_h \odot \int_{\theta=0}^{1} \frac{\partial F_x(\theta A)}{\partial A_h} d\theta, \tag{3}$$

where $A = [A_1, A_2, ..., A_H]$ represents the self-attention matrix of the multi-head attention mechanism at that layer, and symbol $\odot$ denotes element-wise matrix multiplication. $Ainf_h \in R^{n \times n}$ represents the attribution inference matrix corresponding to the h-th head attention weight matrix A_h, and $\partial F_x(\theta A)/\partial A_h$ represents the gradient of model $F_x(\cdot)$ along A_h. The (i, j)-th element $Ainf_h^{i,j}$ of $Ainf_h$ represents the contribution of the connection between input token x_i and input token x_j in the h-th attention head to the model decision-making. In Equation (3), the scaling factor $\theta \in [0, 1]$ controls the strengths of attention connections along a continuous path from the zero matrix to the original attention weight matrix. When $\theta = 0$, the attention connections under analysis are suppressed, whereas $\theta = 1$ corresponds to their original strengths. The gradient information reflects the sensitivity of model predictions to individual Token connections at different points along this path, while the integral accumulates these sensitivities across varying connection strengths. Furthermore, the elementwise multiplication with the original attention weights incorporates the strengths of the corresponding Token connections, thereby quantifying their attribution contributions to model predictions. Consequently, the proposed method considers both the attention connection strengths and their influence on classification decisions, providing a quantitative basis for identifying important Token interactions within the Transformer.

In a deep learning network, the integral cannot be calculated directly, so the Riemann sum formula [42] is used to approximate the above integral equation. Specifically, we sample a certain number of points at small intervals on the path from the zero matrix to the original attention weight matrix and sum the gradients of these points. The calculation formula is defined as follows:

$$Ainf_h{}'(A) = \frac{A_h}{N} \odot \sum_{i=1}^{n} \frac{\partial F_x(\frac{i}{N} A)}{\partial A_h}, \tag{4}$$

where N is the number of sampling points. Based on the attribution inference matrix derived from Equations (3) and (4), the proposed AIIA framework establishes a unified mathematical foundation for subsequent interpretability analyses of Transformer models. Specifically, the attribution contributions of Token connections are aggregated within individual attention heads to evaluate their relative importance, providing a quantitative basis for network structure analysis. Furthermore, Token-level attribution information is filtered and accumulated across network layers to infer key information flows during hierarchical feature propagation. Finally, the resulting Token contribution scores are mapped onto their corresponding image patches, enabling the localization of salient regions associated with classification decisions. Through these operations, the attribution inference matrix establishes explicit functional connections between attention interactions and three interpretability perspectives, including attention head contribution measurement, key Token information flow inference, and salient feature position.

2.5. Attention Head Contribution Measurement

In a Transformer model with a large number of parameters, the impact of different attention heads on the final decision-making of the model is inconsistent. Analyzing the contribution of network structure to prediction results is of great significance to enhance the interpretability of the model. To this end, we propose a measurement method of attention head contribution based on the attribution inference matrix to provide support for Transformer model pruning.

According to the ACMM introduced in Section 2.4, the attribution inference matrix of each attention head in all layers of the Transformer network can be obtained. This allows us to understand how input tokens interact within each attention head of every Transformer layer, as well as the contribution of their attention connections to the model's decision-making process. Important attention heads often contain token connections with higher

contribution values, which appear as larger numerical scores in the attribution inference matrix. Specifically, for sample X_i, the attribution inference matrix corresponding to each Transformer layer can be calculated through ACMM:

$$Ainf_i^l = [At^i{}_{l,1}, At^i{}_{l,2}, ..., At^i{}_{l,H}]. \tag{5}$$

Then, we traverse all the samples in the test set in turn and then accumulate the obtained contribution scores. The formula of this process can be described as follows:

$$G_{l,h} = \sum_i max(At^i_{l,h}), \tag{6}$$

where $max(At^i_{l,h})$ is the maximum value of the element in the attribution inference matrix of the h-th head in the l-th layer. The contribution of the attention heads in each layer is then normalized, resulting in the contribution distribution of each attention head in the Transformer network on the PolSAR dataset:

$$\begin{aligned} I_l &= [I_1, I_2, ..., I_H], \\ I_h &= \frac{1}{\max\limits_i G_{l,i}} G_{l,h}, i \in [1, H]. \end{aligned} \tag{7}$$

Here, I_l represents the contribution distribution of all attention heads in the l-th Transformer layer, where I_h denotes the normalized contribution score of the h-th attention head. Specifically, I_h is calculated by dividing its accumulated contribution score by the maximum score among all attention heads in the same layer. A higher I_h indicates a greater contribution to the model's classification decisions, suggesting that the corresponding attention head is more important.

Finally, we set a pruning rate r% to prune the attention head of r% in every Transformer layer, thus retaining the top $(1 - r)$% most important attention heads in each Transformer layer. After sorting and filtering, the set of attention head index numbers retained by each layer can be obtained:

$$S^l = \text{top}(k, \text{argsort}(I_h)), k = r * H, \tag{8}$$

where S^l represents the set of attention head indexes retained by layer l. $\text{argsort}(\cdot)$ is a sorting function that outputs the sorted index sequence, and $\text{top}(k, \cdot)$ means taking the first k input indexes. Specifically, each layer of the Transformer network needs to retain the same number of attention heads to ensure the model can perform forward propagation properly. The pseudo-code of the attention head pruning algorithm based on the attribution inference matrix is shown in Algorithm 1.

2.6. Information Flow Inference of Key Tokens

As the basic unit of forward propagation of information flow in the Transformer network, the token is an indispensable component of model inference. Thus, analyzing the transmission patterns of feature tokens within Transformer networks provides a crucial entry point for understanding the underlying operational mechanisms of these networks. Tokens are input to the Transformer model as sequences, where they capture the relationships between tokens through multi-head attention mechanisms. Due to the absence of feature compression operations like pooling and downsampling in convolutional neural networks, Transformer networks do not simplify the information complexity at each layer. As a result, redundant information continually accumulates within the network, further complicating the understanding of the operating mechanisms of these inherently complex deep networks. Therefore, to address the issue of a vast number of input tokens and unclear information flow evolution in the layers of the Transformer model, a method based on attribution inference matrices is proposed. This method aims to quantify the contribution of

each input token to the network's decision-making process, thereby constructing the critical pathways in the forward propagation of the Transformer network. By pruning redundant tokens and retaining those with significant contributions, we can simplify the network's depth and improve the interpretability of the Transformer network's underlying mechanisms.

Algorithm 1: Algorithm of Attention Head Pruning Based on Attribution Inference Matrix

Input: Test sample X_i in set $\mathcal{D}$, Transformer model $F_x(\cdot)$, pruning rate r.
Output: The attention head index set $[S^1, S^2, ..., S^l]$ retained by the Transformer model.

Load the trained Transformer model parameters.
▷ Calculate the attribution matrix of each layer according to ACMM:
for X_i ***in*** $\mathcal{D}$ **do**
    Calculate the attribution matrix $Ainf_l^i$ of all layers in $F_x(\cdot)$ corresponding to X_i through Equation (4);
    Calculate the attribution scores in each attention head $G_{l,h} \leftarrow G_{l,h} + G^i{}_{l,h}$ through Equation (6);
▷ Prune attention heads according to the contribution distribution, which is defined as the normalized contribution scores of all attention heads:
Calculate the normalized contribution score of each attention head as $I_h = G_{l,h} / \max_i G_{l,i}$, and obtain the contribution distribution I_l through Equation (7);
Calculate the attention head index $[S^1, S^2, ..., S^l]$ through Equation (8);
Prune the attention head according to $[S^1, S^2, ..., S^l]$.

We focus on a more microscopic perspective, analyzing the interconnection between tokens to examine the impact of each input token on the model's decision at each layer. In the self-attention mechanism, the output token of a layer is obtained by the weighted sum of the input tokens, and its weight is determined by the attention matrix. Accordingly, Figure 3 illustrates how the attribution matrix quantifies the contributions of individual Tokens. The formula is defined as follows:

$$y_i = \sum_j a_{i,j} v_j. \tag{9}$$

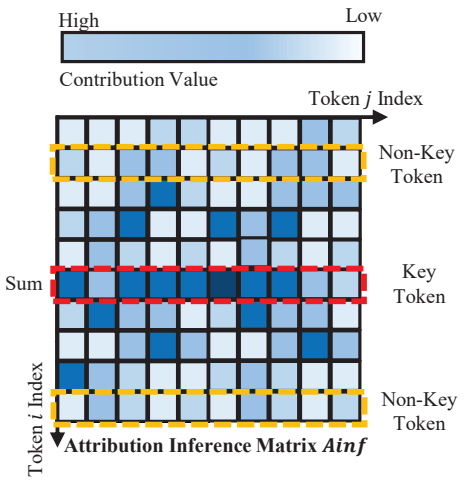

Figure 3. Illustration of Token contribution measurement based on the attribution inference matrix. The colored elements represent the contribution scores of different Token interactions, with darker colors indicating higher contributions. The contribution value of each Token is obtained by aggregating

the attribution scores along the corresponding row. The red dashed box highlights the Key Token, whereas the yellow dashed boxes indicate Non-Key Tokens.

Removing these redundant pieces of information is unlikely to have a significant impact on the model's decision-making process. The pruning process of non-key tokens is illustrated in Figure 4. Given the attribution inference matrix $Ainf_i^l = [At^i{}_{l,1}, At^i{}_{l,2}, ..., At^i{}_{l,H}]$ of H heads in the l-th layer corresponding to sample X_i, it's necessary to filter out certain connections in each attention head's $Ainf_i^l$, including (1) Values on the diagonal of the $Ainf_i^l$ and (2) Values below the threshold τ. The specific expression is defined as follows:

$$\widetilde{At}_{l,h}^i = At_{l,h}^i \odot [ones^{n\times n} - I^{n\times n}] \odot M_{[At_{l,h}^i > \tau]}^{n\times n}, \tag{10}$$

where $\widetilde{At}_{l,h}^i$ is the filtered attribution inference matrix, $ones^{n\times n}$ is a matrix with all values 1, $I^{n\times n}$ is the identity matrix, and $M_{[At_{l,h}^i>\tau]}^{n\times n}$ is a binary matrix that satisfies condition $At_{l,h}^i > \tau$. Then, for the l-th layer of the Transformer network, the attribution inference matrix $\widetilde{At}_{l,h}^i$ of each attention head is accumulated to obtain the attribution score matrix $\widetilde{At}_l^i \in \mathbb{R}^{n\times n}$ of the l-th layer:

$$\widetilde{At}_l^i = \sum_{h=1}^{H} \widetilde{At}_{l,h}^i. \tag{11}$$

The attribution score matrix $\widetilde{At}_l^i$ measures the contribution of token connection relationships at the l-th layer to the model's decision-making process. However, relying solely on it is insufficient to infer the key information flow throughout the entire network. Due to differences in receptive fields and the semantic relevance of features extracted at each layer, tokens considered non-key in the l-th layer may be regarded as key tokens in subsequent layers. Therefore, when measuring the contribution of tokens at the l-th layer, it is necessary to take into account all layers beyond l. Thus, the attribution score matrix As_l^i for the l-th layer is further calculated using the following equation:

$$As_l^i = \sum_{k=l}^{L} \widetilde{At}_k^i. \tag{12}$$

Then, use As_l^i to calculate the contribution value $C_{l,t}^i$ of the t-th token in the l-th layer. The formula is as follows:

$$C_{l,t}^i = \sum_{k=1}^{n} as_{t,k}, as_{t,k} \in [As_l^i]^{n\times n}, \tag{13}$$

where $C_{l,t}^i \in \mathbb{R}^n$, $as_{t,k}$ is the (t,k)-th element of the attribution score matrix of the l-th layer.

Furthermore, considering the flow of information in the forward propagation process of the Transformer network, pruning methods for non-key tokens should adhere to the following basic principles: (1) Tokens pruned at the current layer should not appear in any subsequent layers. (2) The total number of tokens pruned at the current layer should be greater than or equal to the total number pruned at the previous layer. (3) For each sample, the key information flow within the network should be different. Given the number of tokens cn_l that need to be cut off in the l-th layer, then the number of tokens $k_l^{'}$ that need to be retained in the l-th layer is

$$k_l^{'} = n - \sum_{j=1}^{l} cn_j. \tag{14}$$

Based on the above principles, the key token index number set R_l^i retained by sample X_i in the l-th layer can be obtained. The formula of this process is described as follows:

$$\begin{aligned} R_l^i &= \begin{cases} top(k_l^{'}, C_l^{'} \odot Mask_l^{'}), & \text{if } l \geq 2, \\ top(k_1^{'}, C_1^{'}), & \text{otherwise,} \end{cases} \\ m_k &= \begin{cases} 0, \text{ if } k \in R_{l-2} \\ 1, \text{ otherwise} \end{cases}, m_k \in Mask_l^i, l \geq 2. \end{aligned} \tag{15}$$

Here, $Mask_l^i$ is the index mask. According to principle (1), it is used to block the impact of the pruned tokens on subsequent information flow. Finally, determine the key token information flow of sample X_i in the Transformer network based on the index numbers $[R_1^i, R_2^i, ..., R_L^i]$ retained for each layer.

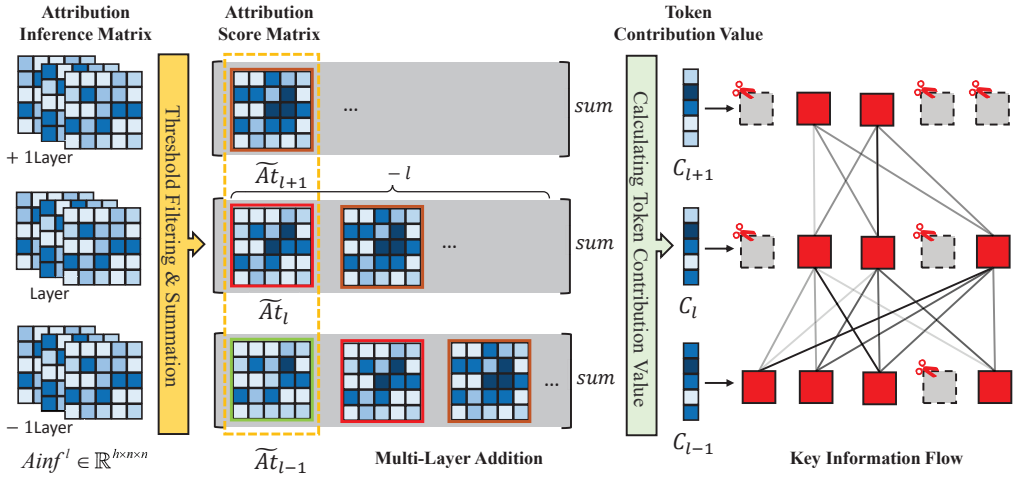

Figure 4. The pruning process of non-key tokens. Self-connections and low-contribution Token interactions are removed through threshold filtering, with the retained attribution scores aggregated across attention heads and accumulated across network layers. Token contribution values are calculated from the resulting matrices, enabling the inference of key information flow.

2.7. Salient Feature Position

Salient feature Position is a technology that uses a visual method to image the features of concern in deep networks, which is of great significance for model interpretability research. Due to the particularity of the feature transfer of the Transformer network, the feature visualization methods currently widely used in CNN models do not work well on the Transformer model, and there are relatively few feature visualization technologies for Transformer networks. Therefore, a salient feature positioning method for the Transformer network is proposed based on the attribution inference matrix to locate the feature areas in the input image that affect the model's decision-making, thereby improving the interpretability of the model. The proposed method can be used in the Transformer-based PolSAR image classification network to analyze the model's attention to different land features.

Unlike 2D feature maps in CNN models, in Transformer models, the input image is first divided into n^2 image patches of a certain size. Then, each image patch is mapped to a vector and sequentially arranged into a token sequence T. Thus, the correspondence between tokens and input image patches is represented by

$$T[i] \leftrightarrow Img[\left\lfloor \frac{i}{n} \right\rfloor, i\%d], \tag{16}$$

where Img represents the image block set, $\lfloor\rfloor$ represents rounding down, and % is the remainder symbol. According to the correspondence between tokens and the input image as described in Equation (16), the distribution of contribution values for each token can be mapped to the corresponding region of the input image. This 2D contribution distribution map, obtained through further inference from the attribution inference matrix, is referred to as the attribution mask AM. Specifically, given the token sequence contribution value vector

$C_l \in \mathbb{R}^{n^2}$ for the l-th layer, normalizing it, and then applying the mapping relationship described in Equation (16), we can obtain the attribution mask AM_i for the l-th layer:

$$AM_l[\left\lfloor \frac{i}{n} \right\rfloor, i\%d] = Norm(C_l)[i], \tag{17}$$

where $AM_l \in \mathbb{R}^{n \times n}$ reflects the attention of the network's l-th layer to the input features. However, model decisions are the result of the combined effects of all layers. Therefore, it is necessary to aggregate the contribution distribution of all layers to comprehensively provide the results of salient feature position. Specifically, by multiplying the attribution masks of all layers together, we can obtain a heatmap of the regions where the model focuses on features:

$$HM = \prod_{l=1}^{L} AM_l. \tag{18}$$

Finally, the heatmap is resized to match the dimensions of the input image using a linear interpolation algorithm, and then overlaid on the input image to generate the visualization. Intuitively, it can be understood as the contribution distribution of input features to the model's decision-making process. Regions with higher heatmap values indicate that features in corresponding positions of the original image contribute more to the network's decision-making process. The proposed salient feature position method effectively visualizes the feature regions on which the model relies to make decisions. This enhances the interpretability of Transformer networks' operational mechanisms.

3. Results

3.1. Experimental Dataset

(1) Flevoland Dataset: The Flevoland dataset is an L-band fully PolSAR image acquired by the AIR-SAR airborne platform in 1989 over an agricultural region in the Netherlands. The image has a spatial resolution of 6.6 m × 12.1 m and a size of 750 × 1024 pixels, containing 15 land-cover categories. The corresponding Pauli RGB image is shown in Figure 5a.

(2) San Francisco Dataset: The San Francisco dataset is an L-band fully PolSAR image acquired by the AIR-SAR airborne platform in 1989 over the San Francisco Bay area, USA. The image has a spatial resolution of approximately 10 m × 10 m and a size of 900 × 1024 pixels, containing 5 land-cover categories. The corresponding Pauli RGB image is shown in Figure 5b.

(3) Oberpfaffenhofen Dataset: The Oberpfaffenhofen dataset is an L-band fully PolSAR image acquired by the ESAR airborne platform in 2002 over Oberpfaffenhofen, Germany. The image has a spatial resolution of approximately 3.0 m × 2.2 m and a size of 1300 × 1200 pixels, containing 3 land-cover categories. The corresponding Pauli RGB image is shown in Figure 5c.

3.2. Implementation Details

To verify the effectiveness of the proposed AIIA method and improve the interpretability of the Transformer feature classification model, experiments are designed from three aspects: attention head pruning, key token measurement, and feature visualization. The test model uses the PolSAR image classification model based on the hierarchical scattering–spatial interaction proposed in [43], and the interpretability analysis of the Transformer encoder network is performed. The model takes a 20 × 20 PolSAR image patch centered on the target pixel as input for classification. The total number of Transformer layers L is 6, and the number of attention heads H of each Transformer layer is 16. The output dimension of each attention head is 16, and the number of input tokens at each layer n is 324 (excluding

class-token). All experiments were repeated five times, with the mean results reported to ensure reliable evaluation.



(a) (b) (c)

Figure 5. Pauli RGB images of the three benchmark PolSAR datasets. (**a**) Flevoland, an agricultural scene containing 15 land-cover categories. (**b**) San Francisco, a coastal urban scene containing 5 land-cover categories. (**c**) Oberpfaffenhofen, a mixed urban and natural scene containing 3 land-cover categories. The Pauli RGB representation visualizes different polarimetric scattering characteristics through pseudo-color mapping, providing an intuitive illustration of the spatial structures and land-cover distributions across the three datasets.

We compared the current mainstream interpretable methods of the Transformer model, including the evaluation method based on the attention matrix (AttMatrix) [44] and the evaluation method based on Taylor expansion (Taylor) [45]. The Overall Accuracy (OA) is used to evaluate the classification performance of the model after pruning, while the class average accuracy is calculated to assess the model's classification performance across individual classes after pruning. In feature visualization experiments, heat maps are employed to illustrate the attention of visualization methods on input sample images.

3.3. Attention Head Pruning

The effectiveness of attention head contribution measurement methods is quantitatively assessed through attention head pruning. In the experiments, attention heads with lower importance scores are pruned based on a specified percentage. The overall contribution score of the head is then obtained by summing all elements of this matrix. Under the same pruning ratio, a more effective evaluation method results in the smaller drop in classification accuracy. Specifically, AttMatrix [44] calculates the maximum value of the attention matrix of each head in the Transformer model as a measure of contribution. Taylor [45] measures the contribution of each attention head by first computing the dot product between the activation value and the gradient of each token, which forms an importance matrix for that head.

Table 1 shows the attention head pruning results on three PolSAR image classification datasets, reporting the OA (%) of each method under different pruning ratios. The row with a gray background indicates the baseline results without pruning, while the highest classification accuracy under each pruning rate setting is highlighted in bold. The detailed analysis based on the experimental results is as follows:

(1) On three PolSAR image datasets, the proposed AIIA generally outperforms other methods across various pruning rates, which achieves superior attention head pruning performance. Notably, at a pruning rate of 0.4, AIIA results in only 0.74%, 0.88%, and 0.58% reductions in OA compared to the baseline on the three datasets, respectively. In contrast, the second-best method (Taylor) leads to reductions of 1.17%, 1.25%, and 1.17%, respectively. These results highlight the effectiveness of the attention causality measurement model. Specifically, AIIA is capable of accurately identifying

high-contribution attention heads through attribution inference, which provides an effective structural interpretation of Transformer models.

(2) For different pruning ratios, the proposed AIIA and the Taylor expansion-based methods consistently outperform the original attention matrix-based methods on all three datasets. It demonstrates that the interpretability method that directly analyzes the attention weights has certain limitations, which are insufficient for accurately evaluating the attribution relationship between individual attention heads and the model's classification performance.

(3) From the overall accuracy variations in each method across the three datasets, it can be observed that when the pruning rate is below 0.5, the network's classification performance does not degrade significantly. However, once the pruning rate exceeds 0.5, the three evaluation methods exhibit different declining trends, among which the proposed AIIA is relatively gentle and better than other methods. It indicates that approximately 50% of the attention heads in the baseline network have limited impact on the network model decision, and the complexity of the network can be more effectively reduced based on the evaluation of AIIA.

(4) On the Flevoland and San Francisco datasets, when the pruning rate is set to 0.1, the accuracy of the proposed method slightly exceeds that of the baseline (pruning rate of 0). It demonstrates that removing some redundant attention heads can enhance network performance, which further validates the effectiveness of AIIA in accurately measuring the contribution of attention heads.

Figure 6 intuitively presents the distribution of attention head contribution scores across different layers on the Flevoland dataset. Specifically, subfigures (a)–(f) correspond to the visualization results of Transformer layers (1)–(6), respectively. The values in the figure represent the normalized contribution scores. Darker colors indicate higher values, suggesting that the corresponding attention heads contribute more significantly to the model decision. In contrast, lighter colors indicate lower contributions, implying that removing those heads would have minimal impact on the network's prediction performance. It can be observed that in lower layers, the distribution of attention head contributions exhibits greater variance, with significant differences in contribution scores and a relatively larger number of low-contributing heads. In contrast, higher layers show a more uniform distribution, with smaller differences in contribution scores among attention heads. This phenomenon may be attributed to the feature learning characteristics of deep networks, where features extracted in deeper layers are typically more abstract and semantically rich. Consequently, attention heads are required to operate across multiple subspaces to capture such high-level representations, leading to a greater need for diverse and active attention heads in higher layers.

3.4. Key Token Measurement

The effectiveness of the proposed information flow inference of key tokens is quantitatively validated through token pruning experiments. Additionally, contribution analysis across different layers is conducted to assess their impact on network decisions, and the information flow paths of key tokens are visualized using heatmaps. These analyses enhance the interpretability of Transformer-based models. Specifically, for the evaluation experiment of information flow inference, 1000 samples are randomly selected from each class in the test set to construct a sub-test set. Based on the proposed method, non-key tokens in each sample are pruned. The prediction accuracy is then computed on the sub-test set to analyze the effectiveness of the information flow inference of key tokens. AttMatrix [44] uses the sum of each row in the attention matrix as the contribution score of the corresponding

token. Taylor [45] computes the contribution score of each token by taking the dot product between its activation value and corresponding gradient.

Table 1. Experimental results of attention head pruning using different methods on three datasets OA (%). The gray background indicates the classification results without pruning, while bold font highlights the best classification results among all compared methods at the current pruning ratio.

Ratio	Flevoland			San Francisco			Oberpfaffenhofen		
	AttMatrix [44]	Taylor [45]	AIIA	AttMatrix [44]	Taylor [45]	AIIA	AttMatrix [44]	Taylor [45]	AIIA
0	98.48 ± 0.03	98.48 ± 0.03	98.48 ± 0.03	97.23 ± 0.05	97.23 ± 0.05	97.23 ± 0.05	98.53 ± 0.01	98.53 ± 0.01	98.53 ± 0.01
0.1	97.45 ± 0.14	98.45 ± 0.09	**99.07** ± 0.05	96.89 ± 0.18	97.21 ± 0.15	**97.29** ± 0.08	98.18 ± 0.16	**98.53** ± 0.04	98.52 ± 0.03
0.2	96.12 ± 0.16	98.12 ± 0.14	**98.16** ± 0.01	96.12 ± 0.07	97.09 ± 0.13	**97.26** ± 0.01	97.97 ± 0.06	98.41 ± 0.02	**98.48** ± 0.02
0.3	92.92 ± 0.23	97.92 ± 0.17	**98.26** ± 0.08	94.21 ± 0.13	97.01 ± 0.18	**97.13** ± 0.05	97.18 ± 0.07	98.01 ± 0.16	**98.23** ± 0.06
0.4	91.31 ± 0.25	97.31 ± 0.16	**97.74** ± 0.07	92.48 ± 0.09	95.98 ± 0.13	**96.35** ± 0.08	96.89 ± 0.26	97.36 ± 0.18	**97.95** ± 0.11
0.5	88.68 ± 0.22	96.68 ± 0.15	**97.21** ± 0.09	91.29 ± 0.12	95.2 ± 0.14	**95.87** ± 0.04	96.08 ± 0.17	97.04 ± 0.13	**97.21** ± 0.09
0.6	83.88 ± 0.23	92.88 ± 0.15	**93.69** ± 0.07	90.36 ± 0.23	94.13 ± 0.15	**95.01** ± 0.01	88.54 ± 0.27	90.18 ± 0.23	**92.31** ± 0.12
0.7	60.19 ± 0.21	73.19 ± 0.12	**78.19** ± 0.09	69.32 ± 0.27	72.69 ± 0.16	**80.26** ± 0.05	82.21 ± 0.17	86.89 ± 0.16	**89.65** ± 0.07
0.8	40.79 ± 0.16	60.79 ± 0.15	**69.41** ± 0.11	45.69 ± 0.15	49.64 ± 0.11	**58.21** ± 0.05	65.24 ± 0.16	71.51 ± 0.18	**76.21** ± 0.03
0.9	31.8 ± 0.13	44.8 ± 0.16	**54.85** ± 0.04	28.12 ± 0.11	28.26 ± 0.14	**35.78** ± 0.03	47.96 ± 0.16	49.56 ± 0.17	**56.21** ± 0.08
1.0	25.11 ± 0.16	25.11 ± 0.16	25.11 ± 0.16	16.29 ± 0.27	16.29 ± 0.27	16.29 ± 0.27	21.86 ± 0.14	21.86 ± 0.14	21.86 ± 0.14

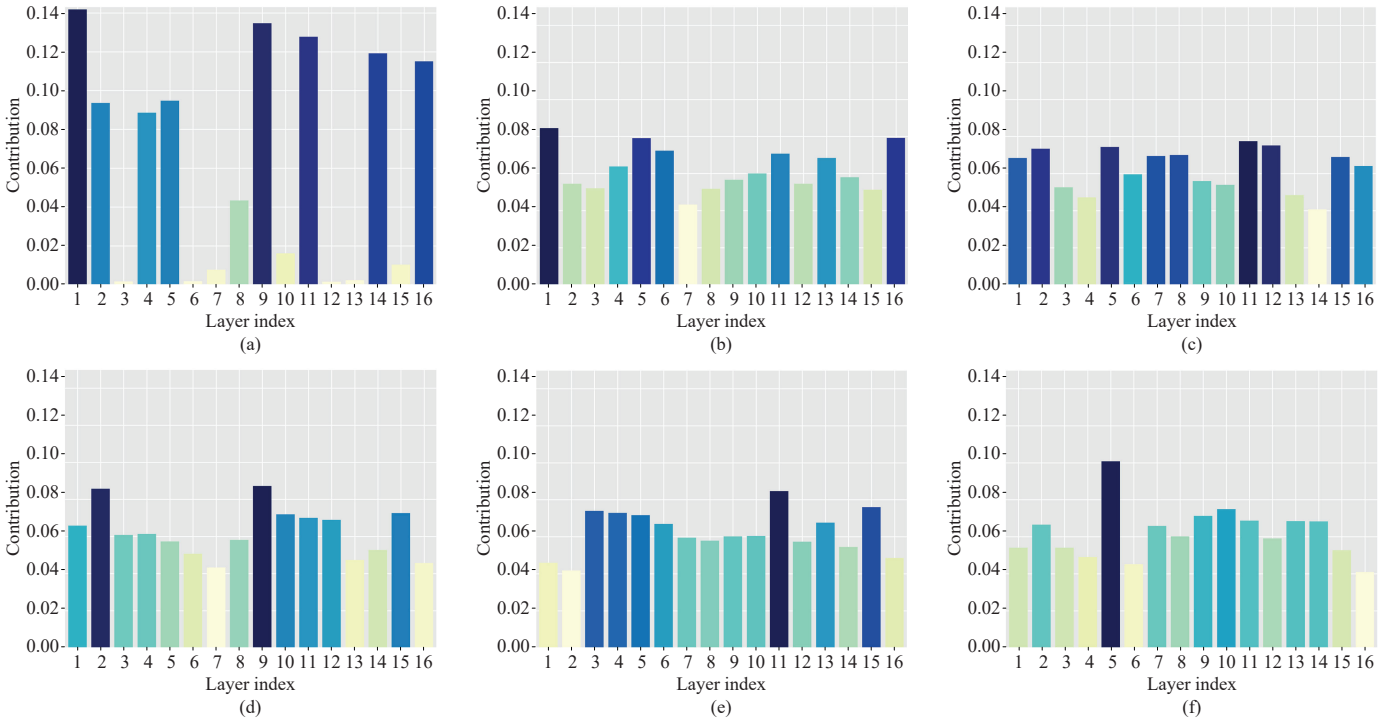

Figure 6. Visualization of the contribution value of attention heads at each layer. (**a**–**f**) correspond to Transformer layers 1–6, respectively. Darker colors indicate higher values, indicating that the corresponding attention heads contribute more significantly to the model decision. In contrast, lighter colors indicate lower contributions.

Table 2 presents the token pruning results of the proposed AIIA and the comparison methods on three PolSAR image classification datasets. The table reports the OA (%) of each method under different pruning rates across the three datasets. The baseline results without pruning are shown in the row with a gray background, and the highest classification accuracy at each pruning rate is marked in bold for emphasis. The following analysis can be obtained from the experimental results:

(1) Across different token pruning rates, AIIA consistently achieves higher overall accuracy than other comparison methods on all three PolSAR image classification datasets.

Analyzing the pruning results while maintaining over 90% classification accuracy, AIIA demonstrates superior performance across all datasets. On the Flevoland dataset, with a pruning rate 0.7, AIIA outperforms AttMatrix and Taylor by 15.2% and 6.8%, respectively. On the San Francisco dataset at the same pruning rate, it exceeds AttMatrix and Taylor by 2.53% and 1.56%. Similarly, for the Oberpfaffenhofen dataset, AIIA achieves gains of 3.92% over AttMatrix and 1.97% over Taylor. It demonstrates that the proposed method can effectively evaluate the contribution of tokens to the model's decision-making and provides a more accurate measurement.

(2) When the pruning rate is below 0.7, all three methods maintain relatively high overall accuracy across the datasets, achieving 97.46%, 97.21%, and 94.06% on the Flevoland, San Francisco, and Oberpfaffenhofen datasets, respectively. The performance degradation is minimal, indicating that approximately 70% of the tokens in the forward have little effect on the prediction results. It provides a theoretical foundation for the lightweight design of Transformer-based networks.

(3) When the pruning rate is relatively low, the overall accuracy of AIIA and Taylor remains close across the three datasets, whereas AttMatrix shows a more significant drop in performance. It indicates that relying solely on the attention matrix is insufficient for accurately evaluating the importance of tokens. However, as the pruning rate increases (above 70%), the proposed AIIA significantly outperforms the other methods, demonstrating its superior accuracy in attributing and quantifying token contributions across different layers of the Transformer network.

To provide a more intuitive understanding of the feature transmission across layers within the Transformer network, a heat map is used to display the contribution values of tokens and the retained key information flow structure. The contribution distribution of tokens at each level before and after sample pruning and the visualization results of the network information flow are shown in Figure 7. The vertical axis represents the layer index of the Transformer network, while the horizontal axis denotes the token index. The contribution values at each layer are normalized, with darker colors indicating higher contribution scores and lighter colors representing lower scores. A contribution value of zero indicates that the corresponding token has been pruned. It can be observed that tokens with higher contribution values are retained at each layer, forming the critical paths for model inference. Notably, the Class Token has a unique role in model prediction, which is always preserved and not shown in the heatmap.

Table 2. Experimental results of key token measurement using different methods on three datasets OA (%). The gray background indicates the classification results without pruning, while bold font highlights the best classification results among all compared methods at the current pruning ratio.

Ratio	Flevoland			San Francisco			Oberpfaffenhofen		
	AttMatrix [44]	Taylor [45]	AIIA	AttMatrix [44]	Taylor [45]	AIIA	AttMatrix [44]	Taylor [45]	AIIA
0	99.53 ± 0.02	99.53 ± 0.02	99.53 ± 0.02	98.84 ± 0.08	98.84 ± 0.08	98.84 ± 0.08	99.56 ± 0.03	99.56 ± 0.03	99.56 ± 0.03
0.1	99.32 ± 0.04	99.41 ± 0.07	**99.53** ± 0.02	98.1 ± 0.11	98.68 ± 0.08	**98.87** ± 0.09	99.52 ± 0.13	**99.55** ± 0.02	99.54 ± 0.11
0.2	99.03 ± 0.08	99.21 ± 0.09	**99.4** ± 0.03	97.26 ± 0.02	98.34 ± 0.07	**98.61** ± 0.07	99.06 ± 0.09	99.48 ± 0.17	**99.52** ± 0.06
0.3	98.13 ± 0.17	99.01 ± 0.15	**99.42** ± 0.07	97.1 ± 0.06	98.04 ± 0.05	**98.43** ± 0.04	98.18 ± 0.19	99.01 ± 0.15	**99.24** ± 0.02
0.4	95.81 ± 0.13	98.86 ± 0.15	**99.21** ± 0.06	97.01 ± 0.16	97.94 ± 0.08	**98.22** ± 0.09	97.74 ± 0.12	98.48 ± 0.13	**99.03** ± 0.03
0.5	96.86 ± 0.12	97.93 ± 0.15	**98.85** ± 0.04	96.86 ± 0.15	96.87 ± 0.11	**97.69** ± 0.03	96.21 ± 0.17	97.63 ± 0.08	**98.69** ± 0.07
0.6	85.93 ± 0.14	95.13 ± 0.15	**98.33** ± 0.07	96.35 ± 0.13	96.24 ± 0.21	**97.1** ± 0.07	94.25 ± 0.14	96.59 ± 0.13	**97.28** ± 0.06
0.7	82.26 ± 0.23	90.66 ± 0.16	**97.46** ± 0.08	94.68 ± 0.12	95.65 ± 0.16	**97.21** ± 0.11	90.14 ± 0.18	92.09 ± 0.15	**94.06** ± 0.01
0.8	67.46 ± 0.16	80.53 ± 0.19	**85.93** ± 0.04	68.27 ± 0.07	74.29 ± 0.21	**81.09** ± 0.08	60.21 ± 0.15	71.24 ± 0.2	**79.89** ± 0.17
0.9	49.53 ± 0.22	55.13 ± 0.18	**72.13** ± 0.14	54.29 ± 0.15	56.24 ± 0.16	**62.17** ± 0.11	49.68 ± 0.12	56.38 ± 0.19	**65.89** ± 0.08
1.0	13.06 ± 0.22	13.06 ± 0.22	13.06 ± 0.22	7.24 ± 0.17	7.24 ± 0.17	7.24 ± 0.17	15.24 ± 0.16	15.24 ± 0.16	15.24 ± 0.16

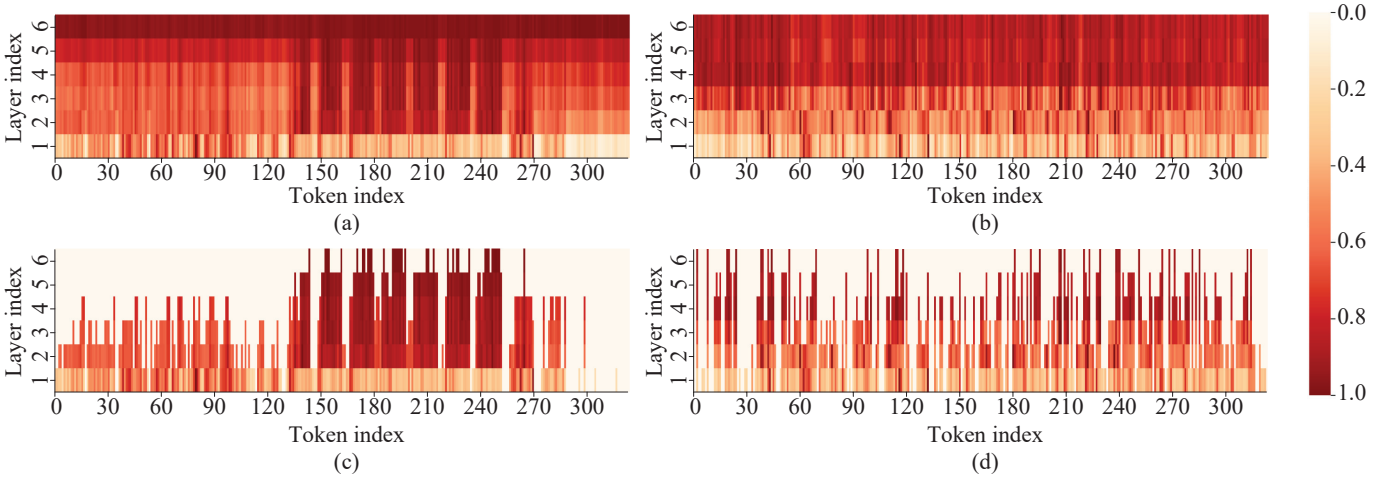

Figure 7. Visualization of Token contribution distributions and key information flows for two PolSAR samples. (**a**) Token contribution distribution of Sample #1. (**b**) Token contribution distribution of Sample #2. (**c**) Key Token information flow of Sample #1 at a pruning rate of 0.9. (**d**) Key Token information flow of Sample #2 at a pruning rate of 0.9. The horizontal and vertical axes represent the Token index and Transformer layer index, respectively. Darker colors indicate higher contribution scores. In contrast, lighter colors indicate lower scores.

The visualization results demonstrate that the proposed information flow inference of key tokens conforms to the feature propagation patterns of Transformer networks. The pruned tokens are excluded from subsequent layer inputs, and the contribution value changes step by step across layers. As shown in Figure 7a,b, the contribution values in lower layers exhibit greater variation and are more sparsely distributed, whereas in higher layers, the normalized contribution values tend to increase, become more uniform, and are more densely distributed. This phenomenon may be attributed to the feature learning behavior of deep networks, where high-level token representations are formed by weighted combinations of low-level features, leading to more abstract and semantically rich information that is generally more relevant to the final classification decision. Figure 7c,d demonstrate noticeable differences in the positions of retained key Tokens and information flow structures across different samples. These results further validate the ability of the proposed method to identify sample-specific key Tokens and reveal the dynamic variations in critical information flow across different inputs, thereby providing deeper insights into the input-dependent decision-making mechanisms of Transformer networks.

To further evaluate the effectiveness of the proposed key token measurement method, we conduct token perturbation experiments on two randomly selected test samples from different classes in the Flevoland dataset. Specifically, negative perturbation progressively removes tokens in ascending order of their attribution scores, while positive perturbation removes tokens in descending order. The Taylor method is employed as a baseline, and the prediction confidence for the target class is recorded under different token pruning ratios. As shown in Figure 8, AIIA consistently maintains higher prediction confidence than Taylor under negative perturbation, indicating that the key tokens identified by AIIA can better preserve the model's predictive capability. For sample #1, at a pruning ratio of 80%, AIIA maintains a prediction confidence of approximately 0.65, compared with 0.56 for Taylor. When the pruning ratio increases to 90%, AIIA still achieves confidence of approximately 0.42, whereas that of Taylor drops below 0.20. Conversely, under positive perturbation, AIIA exhibits a more pronounced decrease in prediction confidence, demonstrating that removing its identified key tokens has a greater impact on model predictions. Furthermore, sample #2 generally exhibits a sharper confidence decline than sample #1, suggesting that the importance of key tokens varies across different input samples. These results demonstrate

that the attribution scores derived from AIIA can reflect the influence of individual tokens on classification decisions, further validating the effectiveness and reliability of the proposed key token measurement and information flow inference method.

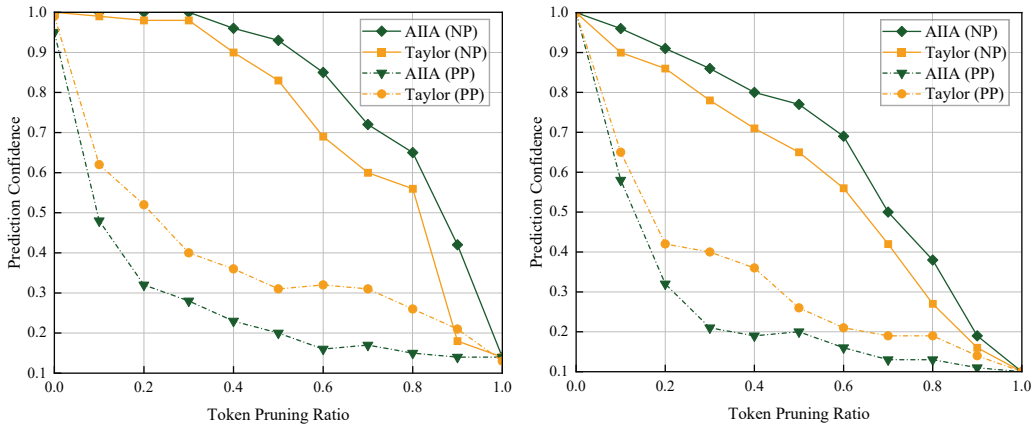

Figure 8. Comparison of AIIA and Taylor in key Token perturbation experiments on the Flevoland dataset. Prediction confidence under different Token pruning ratios is evaluated using negative perturbation (NP) and positive perturbation (PP) for (**a**) Sample #1 and (**b**) Sample #2.

3.5. Feature Visualization

To further validate the effectiveness of the proposed salient feature position method, we conduct feature visualization analyses on Transformer models using three PolSAR image datasets. The key regions in the input images that the model focuses on for prediction are highlighted using heatmaps. We compare our method with widely adopted feature visualization techniques, including Grad-CAM [26], which is extensively used for interpreting remote sensing image features, and Attention Rollout [31], a method based on attention weight propagation. The experimental results of each visualization method are shown in Figure 9, where regions with higher heatmap values indicate areas of greater model attention as inferred by the corresponding method. The rightmost column displays the class label of the central pixel for each sample, serving as a reference for the model's predicted target category.

For the target category of the test sample image, the model is expected to assign higher attention, which corresponds to higher heatmap values. As shown in Figure 9, the three methods exhibit different spatial distributions of salient features across the selected PolSAR samples. Grad-CAM generally highlights localized feature responses, while Attention Rollout produces broader attention distributions that may include surrounding regions unrelated to the target class. In comparison, AIIA exhibits a more consistent correspondence between the highlighted regions and the target land-cover areas. For instance, in the grass sample, AIIA concentrates its responses within the grass region, whereas the comparison methods exhibit less localized activations. Similar observations can be found in the forest and open areas samples, where AIIA highlights regions that are more consistent with the corresponding target classes. These results indicate that AIIA can effectively identify image regions relevant to classification decisions, providing more intuitive and reliable interpretations of Transformer-based PolSAR image classification.

To complement the qualitative visualization results in Figure 9, we conduct quantitative experiments on three PolSAR datasets to evaluate the reliability of salient feature attribution. Specifically, the saliency rankings generated by AIIA, Grad-CAM [26], and Attention Rollout [31] are used to progressively remove input regions with the lowest contribution scores. The classification performance is evaluated using OA at different removal ratios. If a

method successfully identifies less relevant regions, removing these regions should have a relatively limited impact on classification performance, allowing the model to retain the discriminative information necessary for accurate predictions. The results of the quantitative experiments are presented in Table 3.

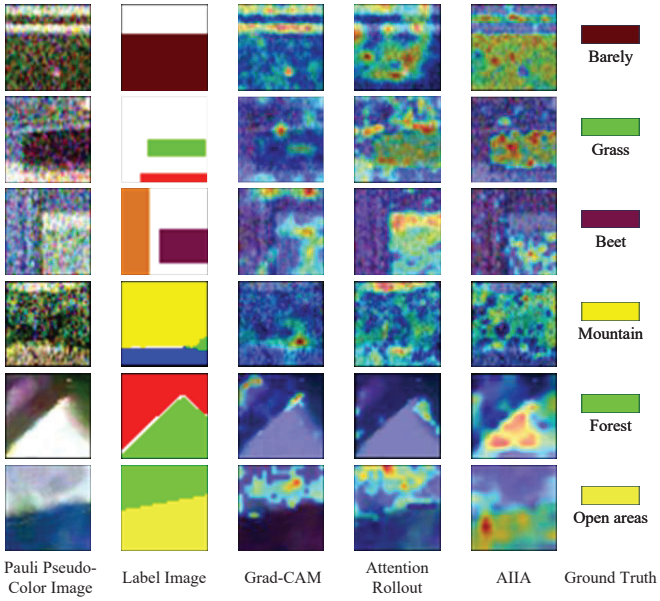

Figure 9. Feature visualization experimental results. The rightmost column represents the central pixel of the current sample image. In the heatmaps, darker colors indicate higher model attention.

The experimental results show that AIIA generally achieves higher classification accuracy than Grad-CAM and Attention Rollout across the three datasets at different removal ratios. Notably, at the high removal ratio of 0.9, AIIA achieves OA improvements of 2.79%, 2.76%, and 1.69% over Attention Rollout on the Flevoland, San Francisco, and Oberpfaffenhofen datasets, respectively. Compared with Grad-CAM, the improvements reach 6.47%, 9.13%, and 3.95%, respectively. Even after 90% of the low-contribution regions are removed, AIIA maintains relatively high classification accuracy, indicating that its attribution rankings can effectively distinguish low-contribution regions from those containing important discriminative information.

Table 3. Experimental results of salient feature position using different methods on three datasets OA (%). The gray background indicates the classification results without pruning, while bold font highlights the best classification results among all compared methods at the current pruning ratio.

Ratio	Flevoland			San Franciso			Oberpfaffenhofen		
	Grad-CAM [26]	Attention Rollout [31]	AIIA	Grad-CAM [26]	Attention Rollout [31]	AIIA	Grad-CAM [26]	Attention Rollout [31]	AIIA
0	99.53 ± 0.02	99.53 ± 0.02	99.53 ± 0.02	98.84 ± 0.08	98.84 ± 0.08	98.84 ± 0.08	99.56 ± 0.03	99.56 ± 0.03	99.56 ± 0.03
0.1	97.12 ± 0.03	98.97 ± 0.05	**99.25** ± 0.06	97.39 ± 0.08	98.27 ± 0.11	**98.27** ± 0.14	98.47 ± 0.07	98.88 ± 0.11	**99.11** ± 0.04
0.2	96.74 ± 0.07	97.89 ± 0.12	**98.87** ± 0.02	95.25 ± 0.17	96.93 ± 0.11	**97.87** ± 0.09	97.51 ± 0.16	97.55 ± 0.08	**98.78** ± 0.09
0.3	95.73 ± 0.17	97.57 ± 0.13	**98.09** ± 0.01	94.29 ± 0.12	96.13 ± 0.16	**97.83** ± 0.03	96.35 ± 0.11	97.27 ± 0.17	**98.65** ± 0.07
0.4	95.7 ± 0.13	96.64 ± 0.17	**98.02** ± 0.09	93.62 ± 0.18	96.02 ± 0.16	**97.13** ± 0.11	95.18 ± 0.13	97.19 ± 0.19	**98.39** ± 0.12
0.5	94.53 ± 0.18	96.13 ± 0.07	**97.9** ± 0.05	92.58 ± 0.16	95.38 ± 0.13	**96.63** ± 0.03	94.39 ± 0.18	96.46 ± 0.08	**97.45** ± 0.03
0.6	92.13 ± 0.16	94.98 ± 0.17	**96.98** ± 0.11	92.48 ± 0.24	94.6 ± 0.27	**96.38** ± 0.17	92.45 ± 0.14	95.14 ± 0.15	**96.63** ± 0.04
0.7	87.79 ± 0.21	89.84 ± 0.19	**91.59** ± 0.05	85.11 ± 0.13	88.69 ± 0.27	**91.02** ± 0.1	80.57 ± 0.16	88.41 ± 0.09	**89.55** ± 0.08
0.8	80.39 ± 0.14	85.39 ± 0.26	**87.29** ± 0.07	80.28 ± 0.25	85.62 ± 0.19	**87.6** ± 0.13	72.81 ± 0.29	76.66 ± 0.24	**79.94** ± 0.14
0.9	68.4 ± 0.23	72.08 ± 0.14	**74.87** ± 0.08	62.77 ± 0.17	69.14 ± 0.15	**71.9** ± 0.09	58.66 ± 0.15	60.92 ± 0.24	**62.61** ± 0.11
1.0	23.78 ± 0.15	23.78 ± 0.15	23.78 ± 0.15	18.71 ± 0.21	18.71 ± 0.21	18.71 ± 0.21	20.97 ± 0.12	20.97 ± 0.12	20.97 ± 0.12

4. Discussion

4.1. Ablation Experiment

To investigate the contributions of the key components of the proposed attribution inference module, we conduct an ablation study using attention head pruning as the evaluation protocol. Four variants are considered, including attention weights alone, gradient information alone, the combination of attention weights and gradients without path integration, and the complete AIIA method. By comparing their classification performance under different pruning ratios, we analyze the effects of individual components on attention head importance estimation. We report the mean OA across pruning ratios of 0.1–0.9 (Mean OA) and the OA at a pruning ratio of 0.9 (OA@0.9) to evaluate the effectiveness of the proposed attribution inference module and its key components. This study aims to clarify the contributions of attention weights, gradient information, and path integration to the effectiveness of the proposed attribution inference module. The detailed experimental results are presented in Table 4.

Table 4. Ablation study of the attribution inference module based on attention head pruning (%). The ✓ indicates the corresponding component is included, while × indicates exclusion.

Variant	Attention	Gradient	Path Integration	Flevoland		San Francisco		Oberpfaffenhofen	
				Mean OA	OA@0.9	Mean OA	OA@0.9	Mean OA	OA@0.9
Attention-only	✓	×	×	75.9	31.8	78.28	28.12	85.58	47.96
Gradient-only	×	✓	×	84.46	44.8	80.8	28.26	87.5	49.56
Attention-Gradient w/o Integration	✓	✓	×	85.97	51.42	82.38	32.4	88.02	53.54
AIIA	✓	✓	✓	**87.39**	**54.85**	**83.68**	**35.78**	**89.42**	**56.21**

The experimental results show that the complete AIIA method achieves the highest Mean OA and OA@0.9 on all three datasets, demonstrating the effectiveness of the proposed attribution inference module. Specifically, compared with the attention-weight-based method (Attention-only), AIIA improves Mean OA by 11.49%, 5.4%, and 3.84% on Flevoland, San Francisco, and Oberpfaffenhofen, respectively. The corresponding improvements in OA@0.9 are 23.05%, 7.66%, and 8.25%. Compared with the gradient-based method (Gradient-only), AIIA improves Mean OA by 2.93%, 2.88%, and 1.92%, respectively, while the improvements in OA@0.9 are 10.05%, 7.52%, and 6.65%. These results indicate that relying solely on attention weights or gradient information has certain limitations, whereas combining the two sources of information enables more effective estimation of attention head importance.

Furthermore, a comparison between Attention-Gradient w/o Integration and the complete AIIA method demonstrates the contribution of the path integration mechanism. By incorporating path integration into the combination of attention weights and gradient information, AIIA improves Mean OA by 1.42%, 1.3%, and 1.4% on Flevoland, San Francisco, and Oberpfaffenhofen, respectively. The corresponding improvements in OA@0.9 are 3.43%, 3.38%, and 2.67%. Notably, even at a pruning ratio of 0.9, the complete AIIA method maintains OA values of 54.85%, 35.78%, and 56.21% on the three datasets, respectively, outperforming all other variants. These findings suggest that path integration further improves the estimation of attention head contributions, enabling the model to maintain higher classification accuracy when only a small proportion of attention heads is retained.

4.2. Generalization Experiment

The proposed AIIA framework primarily focuses on interpreting the internal decision-making mechanisms of Transformer models by integrating self-attention weights with gradient information for attribution analysis. Since this mechanism primarily analyzes

the internal attention representations of Transformer models, it exhibits good applicability and can be further extended to the interpretability analysis of PolSAR image classification models under different sensor platforms, frequency bands, and imaging conditions.

To further validate the applicability of the proposed method under different imaging conditions and on recently released PolSAR datasets, we selected C-band fully polarimetric SAR imagery from the AIR-MDSAR-Map dataset [46] published in 2024 for generalization experiments. Compared with the original L-band benchmark datasets, this dataset presents different frequency bands, spatial resolutions, and land-cover characteristics, providing additional experimental evidence for evaluating the interpretability of AIIA under different data acquisition conditions.

The results of the generalization experiments, including attention head pruning, key Token pruning, and salient feature analysis, are presented in Table 5. Notably, we introduced an additional quantitative pruning experiment for salient feature analysis, in which input regions with the lowest contribution scores were progressively removed according to their saliency rankings. The classification performance of different methods was then evaluated under varying removal ratios. The experimental results demonstrate that AIIA consistently outperforms the corresponding comparison methods across all three experimental settings at pruning ratios ranging from 0.1 to 0.9. Specifically, at an extreme pruning ratio of 0.9, AIIA achieves classification accuracy improvements of 3.58% and 8.53% over Taylor and AttMatrix, respectively, in the attention head pruning experiment. In the key Token pruning experiment, AIIA outperforms Taylor and AttMatrix by 2.89% and 5.05%, respectively. In the salient feature analysis experiment, AIIA achieves improvements of 2.28% and 7.63% over Attention Rollout and Grad-CAM, respectively. These results demonstrate that AIIA can identify important attention heads, key Tokens, and salient features that contribute significantly to model decisions, maintaining better classification performance even when only a limited amount of critical information is retained under high pruning ratios.

The generalization experiments on the AIR-MDSAR-Map dataset further validate AIIA's effectiveness on C-band PolSAR imagery and provide experimental support for its use in image classification models across different frequency bands, sensor platforms, and imaging conditions. These findings indicate that the proposed method has promising potential for application across diverse data acquisition conditions, while its attribution analysis mechanism provides a unified basis for interpreting the decision-making processes of Transformer models under varying imaging conditions.

Table 5. Quantitative comparison of OA (%) on the AIR-MDSAR-Map C-band fully PolSAR dataset under different pruning ratios for attention head pruning, key Token pruning, and salient feature analysis. The gray background indicates the classification results without pruning, while bold font highlights the best classification results among all compared methods at the current pruning ratio.

Ratio	Attention Head			Key Token			Salient Feature		
	AttMatrix [44]	Taylor [45]	AIIA	AttMatrix [44]	Taylor [45]	AIIA	Grad-CAM [26]	Attention Rollout [31]	AIIA
0	75.63	75.63	75.63	75.63	75.63	75.63	75.63	75.63	75.63
0.1	73.61	74.51	**74.79**	74.47	75.32	**75.61**	74.5	74.6	**74.89**
0.2	73.07	73.52	**74.16**	74.04	75.01	**75.42**	73.78	74.37	**74.7**
0.3	71.83	72.82	**73.43**	73.97	74.45	**75.15**	72.83	74.04	**74.54**
0.4	71.26	71.84	**73.12**	73.34	74.38	**75.02**	72.45	73.59	**74.06**
0.5	70.97	71.07	**72.91**	73.03	73.36	**74.33**	71.21	73.34	**73.94**
0.6	64.74	66.77	**69.19**	67.23	72.79	**73.56**	69.92	71.41	**72.46**
0.7	43.51	48.41	**51.01**	64.21	71.72	**72.36**	65.52	68.26	**70.84**
0.8	38.28	46.43	**47.13**	59.75	62.2	**67.24**	54.42	60.94	**62.11**
0.9	24.35	29.3	**32.88**	55.98	58.14	**61.03**	41.43	46.78	**49.06**
1.0	19.64	19.64	19.64	13.85	13.85	13.85	18.05	18.05	18.05

5. Conclusions

This paper proposes a systematic interpretability method for Transformer-based PolSAR image classification models, reducing the complexity of designing separate interpretation schemes for different model components. Specifically, an Attribution Inference Interpretability Analysis (AIIA) framework is proposed. The main innovation lies in constructing a unified attribution inference matrix by integrating attention weights with gradient information, which quantitatively characterizes the contributions of Token interactions to model predictions rather than relying solely on attention weights. Based on this matrix, a comprehensive interpretability analysis method is established from three complementary perspectives, including attention head contribution measurement, key Token information flow inference, and salient feature position. By providing a unified attribution basis for network structure analysis, hierarchical feature propagation, and spatial feature interpretation, the proposed framework enables a more systematic understanding of the internal decision-making mechanisms of Transformer models.

Extensive experiments on three PolSAR image classification datasets demonstrate the effectiveness of the proposed AIIA framework from both quantitative and qualitative perspectives. The interpretability of AIIA is quantitatively evaluated through attention head and Token pruning experiments, together with quantitative assessments of salient feature position, demonstrating its ability to identify critical network components and key Tokens while accurately localizing discriminative image regions relevant to classification decisions. Qualitative visualizations further illustrate the distribution of attention head contributions, key Token information flows, and salient image regions, providing intuitive explanations of the model's classification decisions.

In future work, we will extend the proposed AIIA framework to pre-trained models, focusing on the interpretability of their learned feature representations and internal decision-making mechanisms. By analyzing the contributions of different network components and feature interactions during pre-training and downstream task adaptation, we aim to reveal how pre-trained knowledge influences model predictions, thereby enhancing the interpretability and reliability of pre-trained models in PolSAR image classification.

Author Contributions: Conceptualization, L.D., J.G., Y.Z. (Yuhan Zhou) and W.J.; methodology, L.D. and Y.Z. (Yuhang Zhang); software, J.G.; validation, L.D., Y.Z. (Yuhang Zhang) and Y.Z. (Yuhan Zhou); formal analysis, L.D. and J.G.; investigation, J.G.; resources, L.D. and W.J.; data curation, L.D. and J.G.; writing—original draft preparation, L.D.; writing—review and editing, L.D., Y.Z. (Yuhang Zhang) and Y.Z. (Yuhan Zhou); visualization, L.D. and Y.Z. (Yuhan Zhou); supervision, W.J.; project administration, W.J.; funding acquisition, L.D. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Basic Research Program of Shaanxi (2026JC-YXQN-214), the Shaanxi Province Sangin Outstanding Young Talents Program, and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX2026049).

Data Availability Statement: The data used in this study are publicly available benchmark datasets. The Flevoland and San Francisco datasets were acquired by the AIRSAR platform, and the Oberpfaffenhofen dataset was acquired by the E-SAR airborne platform. The datasets are available from their corresponding public data repositories.

Acknowledgments: The authors are grateful for the constructive comments from anonymous reviewers and the editors. During the preparation of this manuscript, the authors used ChatGPT (GPT-6, OpenAI) for language editing and grammar checking. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Tsokas, A.; Rysz, M.; Pardalos, P.M.; Dipple, K. SAR data applications in earth observation: An overview. *Expert Syst. Appl.* **2022**, *205*, 117342. [CrossRef]
2. Taricco, G. Physics-Aware Deep Learning for SAR and InSAR Remote Sensing: Models, Methods, and Open Challenges. *Remote Sens.* **2026**, *18*, 2567. [CrossRef]
3. Li, G.; Wang, J.Y.; Chen, Z.S.; Wang, Z.P.; Xiong, S.H.; Pedrycz, W. Prediction of airport runway subsidence using SBAS-InSAR and LSTM networks optimized by EnKF. *Inf. Sci.* **2026**, *729*, 122872. [CrossRef]
4. Dong, H.; Si, L.; Qiang, W.; Zhang, L.; Yu, J.; Wu, Y.; Zheng, C.; Sun, F. A Novel Causal Inference-Guided Feature Enhancement Framework for PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5200816. [CrossRef]
5. Das, N.; Bortiew, A.; Patra, S.; Bruzzone, L. Dual-Branch CNN Incorporating Multiscale SVD Profile for PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5223612. [CrossRef]
6. Shi, J.; Wang, W.; Jin, H.; Nie, M.; Ji, S. A Lightweight Riemannian Covariance Matrix Convolutional Network for PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4411017. [CrossRef]
7. 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]
8. Dong, H.; Zhang, L.; Zou, B. Exploring Vision Transformers for Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5219715. [CrossRef]
9. Ni, K.; Yuan, C.; Zheng, Z.; Zhang, B.; Wang, P. MPT-SFANet: Multiorder Pooling Transformer-Based Semantic Feature Aggregation Network for SAR Image Classification. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, *60*, 4923–4938. [CrossRef]
10. Geng, J.; Zhang, Y.; Jiang, W. Polarimetric SAR Image Classification Based on Hierarchical Scattering-Spatial Interaction Transformer. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5205014. [CrossRef]
11. Shen, D.; Ma, X.; Kong, W.; Liu, J.; Wang, J.; Wang, H. Hyperspectral Target Detection Based on Interpretable Representation Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5519416. [CrossRef]
12. Zhao, C.; Chen, S.; Zhang, J. MSCIAIG-Net: Multi-scale spatial-channel interactive attention and interpretable guidance for fine-grained ship classification and visual inference network in remote sensing images. *Eng. Appl. Artif. Intell.* **2025**, *157*, 111520. [CrossRef]
13. Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
14. Verhoeven, V.B.; Raumonen, P.; Akerblom, M. FuzzyQSM: A Novel Tree Structure Reconstruction Approach Using Uncertainty. *Remote Sens.* **2026**, *18*, 3244. [CrossRef]
15. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
16. Strumbelj, E.; Kononenko, I. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.* **2010**, *11*, 1–18.
17. Yellinek, N.; Karlinsky, L.; Giryes, R. 3VL: Using Trees to Improve Vision-Language Models' Interpretability. *IEEE Trans. Image Process.* **2025**, *34*, 495–509. [CrossRef] [PubMed]
18. Huang, Z.; Li, Y. Interpretable and Accurate Fine-grained Recognition via Region Grouping. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8659–8669. [CrossRef]
19. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: Deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8930–8941.
20. Zhao, G.; Wu, H.; Li, B.; Zhang, X.; Vasilakos, A.V. SensMark: Robust and interpretable model watermarking via contextual sensitivity estimation and adaptive trigger insertion. *Inf. Sci.* **2026**, *756* , 123869. [CrossRef]
21. Wang, S.; Yuan, G.; Li, J. RTDU: Interpretable Region-Aware Transformer-Based Deep Unfolding Network for Pan-Sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 12897–12914. [CrossRef]
22. Zhang, Q.; Yang, Y.; Ma, H.; Wu, Y.N. Interpreting CNNs via Decision Trees. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6254–6263. [CrossRef]
23. Xu, K.; Park, D.H.; Yi, C.; Sutton, C. Interpreting deep classifier by visual distillation of dark knowledge. *arXiv* **2018**, arXiv:1803.04042.
24. He, Y.; Liu, P.; Wang, Z.; Hu, Z.; Yang, Y. Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4335–4344. [CrossRef]

25. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
26. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
27. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. [CrossRef]
28. Khalid, M.I.; Mi, J.X.; Ali, G.; Ali, T.; Hijji, M.; Ayaz, M.; ur Rehman, Z. Interpretable object detection via integrated heatmap, concept attribution, and sobol sensitivity analysis. *Inf. Sci.* **2026**, *738*, 123133. [CrossRef]
29. Hao, Y.; Dong, L.; Wei, F.; Xu, K. Self-attention attribution: Interpreting information interactions inside transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Voliume 35, pp. 12963–12971.
30. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 782–791. [CrossRef]
31. Abnar, S.; Zuidema, W. Quantifying attention flow in transformers. *arXiv* **2020**, arXiv:2005.00928.
32. Chen, T.; Zhang, L.; Guo, W.; Zhang, Z.; Datcu, M. Analyzing the Adversarial Robustness and Interpretability of Deep SAR Classification Models: A Comprehensive Examination of Their Reliability. *Remote Sens.* **2025**, *17*, 1943. [CrossRef]
33. Huang, Z.; Yao, X.; Liu, Y.; Dumitru, C.O.; Datcu, M.; Han, J. Physically explainable CNN for SAR image classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 25–37. [CrossRef]
34. Fan, J.; Sui, Z.; Wang, X. Multiphysical Interpretable Deep Learning Network for Oil Spill Identification Based on SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4202015. [CrossRef]
35. Feng, S.; Ji, K.; Wang, F.; Zhang, L.; Ma, X.; Kuang, G. PAN: Part Attention Network Integrating Electromagnetic Characteristics for Interpretable SAR Vehicle Target Recognition. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5204617. [CrossRef]
36. Ahmad, O.; Béreux, N.; Baret, L.; Hashemi, V.; Lecue, F. Causal Analysis for Robust Interpretability of Neural Networks. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 4673–4682. [CrossRef]
37. Petsiuk, V.; Jain, R.; Manjunatha, V.; Morariu, V.I.; Mehra, A.; Ordonez, V.; Saenko, K. Black-box Explanation of Object Detectors via Saliency Maps. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11438–11447. [CrossRef]
38. Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; Chan, A.B. Gradient-based Visual Explanation for Transformer-based CLIP. In *Proceedings of the 41st International Conference on Machine Learning*; Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F., Eds.; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2024; Volume 235, pp. 61072–61091.
39. Zhao, C.; Hsiao, J.H.; Chan, A.B. Gradient-Based Instance-Specific Visual Explanations for Object Specification and Object Discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5967–5985. [CrossRef] [PubMed]
40. Kovaleva, O.; Romanov, A.; Rogers, A.; Rumshisky, A. Revealing the dark secrets of BERT. *arXiv* **2019**, arXiv:1908.08593.
41. Brunner, G.; Liu, Y.; Pascual, D.; Richter, O.; Ciaramita, M.; Wattenhofer, R. On identifiability in transformers. *arXiv* **2019**, arXiv:1908.04211.
42. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2017; pp. 3319–3328.
43. Geng, J.; Dong, L.; Zhang, Y.; Jiang, W. Masked auto-encoding and scatter-decoupling transformer for polarimetric SAR image classification. *Pattern Recognit.* **2025**, *166*, 111660. [CrossRef]
44. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What does bert look at? An analysis of bert's attention. *arXiv* **2019**, arXiv:1906.04341.
45. Michel, P.; Levy, O.; Neubig, G. Are sixteen heads really better than one? *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14014–14024.
46. Zheng, N.; Yang, Z.; Shi, X.; Yang, H.; Sun, Y.; Wang, F. Airborne Multi-Dimensional SAR Land Cover Dataset and Fusion Classification Method. *Natl. Remote Sens. Bull.* **2024**, *28*, 2209–2222. [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.