remote sensing

MDPI

Technical Note

PolarFormer: A Registration-Free Fusion Transformer with Polar Coordinate Position Encoding for Multi-View SAR Target Recognition

Xiang Yu [1,*], Ying Qian [1], Guodong Jin [2], Zhe Geng [2] and Daiyin Zhu [2]

[1] School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China; qiany@njit.edu.cn
[2] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; jinguodong@nuaa.edu.cn (G.J.); zhegeng@nuaa.edu.cn (Z.G.); zhudy@nuaa.edu.cn (D.Z.)
* Correspondence: yx@njit.edu.cn; Tel.: +86-13601582084

Highlights

What are the main findings?

- A novel multi-view polar coordinate position encoding is proposed to accurately model the complex geometric relationships among unaligned SAR images.
- A spatially aware self-attention mechanism is designed to inject this geometric information as an inductive bias into the Transformer, enhancing its ability to perceive spatial structures.

What is the implication of the main finding?

- The proposed registration-free paradigm demonstrates that precise geometric modeling can completely supplant physical image registration, simplifying the processing pipeline and avoiding feature distortion.
- This work provides a more effective and robust pathway for early fusion of multi-view SAR data, significantly improving target recognition accuracy in complex scenarios.

Abstract

Multi-view Synthetic Aperture Radar (SAR) provides rich information for target recognition. However, fusing features from unaligned multi-view images presents challenges for existing methods. Conventional early fusion methods often rely on image registration, a process that is computationally intensive and can introduce feature distortions. More recent registration-free approaches based on the Transformer architecture are constrained by standard position encodings, which were not designed to represent the rotational relationships among multi-view SAR data and thus can cause spatial ambiguity. To address this specific limitation of position encodings, we propose a registration-free fusion framework based on a spatially aware Transformer. The framework includes two key components: (1) a multi-view polar coordinate position encoding that models the geometric relationships of patches both within and across views in a unified coordinate system; and (2) a spatially aware self-attention mechanism that injects this geometric information as a learnable inductive bias. Experiments were conducted on our self-developed FAST-Vehicle dataset, which provides full 360° azimuthal coverage. The results show that our method outperforms both registration-based strategies and Transformer baselines that use conventional position encodings. This work indicates that for multi-view SAR fusion, explicitly modeling the underlying geometric relationships with a suitable position encoding is an effective alternative to physical image registration or the use of generic, single-image position encodings.

check for updates

Academic Editor: Stefano Tebaldini

Received: 28 August 2025
Revised: 14 October 2025
Accepted: 23 October 2025
Published: 28 October 2025

Citation: Yu, X.; Qian, Y.; Jin, G.; Geng, Z.; Zhu, D. PolarFormer: A Registration-Free Fusion Transformer with Polar Coordinate Position Encoding for Multi-View SAR Target Recognition. *Remote Sens.* **2025**, *17*, 3559. https://doi.org/10.3390/rs17213559

Keywords: multi-view SAR target recognition; vision transformer; position encoding; registration-free fusion; spatially aware self-attention

1. Introduction

Multi-view Synthetic Aperture Radar (SAR) provides the capability to capture comprehensive scattering characteristics of a target by observing it from various azimuths, which is essential for robust target recognition under complex conditions [1–7]. Among the numerous fusion strategies, early fusion is often preferred as it enables interaction at the most fundamental feature level, thereby maximizing the exploitation of complementary information across views while maintaining a concise network architecture [3,8–12]. However, the realization of effective early fusion is confronted by two fundamental challenges, which have constrained the performance of existing methods.

The first challenge is the dependency on registration and the consequent feature distortion. Traditional early fusion methods, such as those proposed in [13–15], typically stack multi-view features along the channel dimension. This operation implicitly assumes that the input views are perfectly aligned spatially. Consequently, pre-processing steps like image registration or view reprojection are mandatory. However, the unique imaging mechanism of SAR makes precise, pixel-level registration exceedingly difficult [16,17]. Inevitable registration errors are amplified at the feature level, leading to the forced alignment of scattering centers that are not spatially coincident. This results in severe feature blurring and information contamination, ultimately degrading the recognition performance [18,19].

The second, and more profound, challenge lies in the inadequacy of conventional position encodings. In recent years, some works have attempted to employ the Transformer architecture to process spatially concatenated features, aiming to relax the stringent requirement for pixel-perfect registration [20–23]. Such methods rely on position encodings to reconstruct the spatial relationships between patches. However, existing position encodings—be they absolute, relative, or learnable—are fundamentally designed for single, rigid, two-dimensional images. They are inherently incapable of expressing the complex geometric correspondence between, for instance, "patch-p in image A" and "patch-q in the rotated image B." Therefore, when presented with unaligned multi-view images, these position encodings fail. They are unable to establish effective intra-view and inter-view spatial associations, causing spatial ambiguity and leaving the full potential of early fusion untapped [24].

In light of these limitations, the field urgently requires a novel, registration-free early fusion paradigm. Such a paradigm must be capable of fundamentally addressing the problem of modeling the intricate geometric relationships among multi-view images.

To this end, we propose a novel fusion paradigm that integrates feature-driven self-attention with spatial perception. Instead of relying on image registration, our approach resolves the aforementioned challenges by directly and precisely modeling the spatial geometry of multi-view data within the attention mechanism itself. Our core contributions are summarized as follows:

1. We introduce a novel multi-view polar coordinate position encoding. This encoding scheme decouples and precisely describes the local position of any patch within its view, as well as the position and rotational offset of its host view relative to a global reference frame. This allows us, for the first time, to establish a unified and consistent spatial coordinate system for unaligned multi-view images at the feature level.
2. We design a spatially aware self-attention mechanism. This mechanism converts the polar encoding into a position-aware score matrix, which quantifies the true spatial

distance between any two patches, regardless of whether they originate from the same view. By integrating this score matrix as a learnable bias term into the Transformer's self-attention computation, we inject a powerful and explicit spatial inductive bias into the model.

3. We build an end-to-end, registration-free early fusion framework. Leveraging these innovations, we construct a complete recognition network that requires no image preprocessing. The network learns features directly from spatially concatenated multi-view images, thereby retaining the advantages of early fusion while fundamentally circumventing the problems associated with registration.

In the experimental section, we construct the FAST-Vehicle dataset, comprising nine classes of typical targets. Through comprehensive comparative and ablation studies, we demonstrate the significant superiority of our proposed method over existing state-of-the-art (SOTA) approaches and traditional registration-based strategies, achieving an accuracy improvement of up to 4.8%.

The remainder of this paper is organized as follows. Section 2 reviews the related work on multi-view SAR target recognition and Transformer-based methods. Section 3 elaborates on our proposed methodology, detailing the polar coordinate position encoding and the spatially aware self-attention mechanism. Section 4 presents the experimental setup, dataset, and comprehensive results, including comparisons with state-of-the-art methods and in-depth ablation studies. Finally, Section 5 concludes the paper with a summary of our findings and an outlook on future work.

2. Related Works

2.1. Multi-View SAR Target Recognition

Multi-view SAR target recognition aims to enhance performance by fusing information from diverse viewing angles [2,7,12]. Based on the level at which fusion is performed, existing methods can be broadly classified into decision-level, intermediate-level, and feature-level (early) fusion.

Decision-level fusion. After classifying each view independently, this category of methods integrates the outputs from individual classifiers via mechanisms such as voting or weighted averaging [6,25–27]. Although these methods are straightforward and robust, their performance is limited because the complete independence of feature extraction across views prevents the full exploitation of underlying inter-view correlations.

Intermediate-level fusion. This approach typically occurs after feature extraction but before classification, for instance, by fusing the feature maps from various views or by constructing a joint feature representation [28,29]. This strategy strikes a balance between feature interaction and model complexity. However, the design of the fusion module tends to be intricate, and it can still be susceptible to challenges in feature alignment.

Feature-level (Early) Fusion. Forming the focus of this work, early fusion's core advantage lies in its ability to capture inter-view complementarity at the most fundamental feature level. Traditional methods, such as [13–15], typically stack post-registration multi-view images along the channel dimension, which are then fed into a CNN for feature extraction [30,31]. While theoretically capable of achieving deep fusion, the performance of such methods is critically dependent on registration accuracy. However, the coherent speckle noise and geometric distortions inherent in SAR imagery make precise, pixel-level registration exceedingly challenging [32–35]. Inevitable registration errors lead to severe feature distortion, often nullifying the gains afforded by the multi-view perspective. Recently, some efforts have explored using the Transformer architecture to process spatially stacked views, thereby relaxing the strict alignment requirement [36–38]. This, however,

gives rise to a more fundamental question: how to devise an effective position encoding for unaligned multi-view data.

2.2. Position Encoding in Vision Transformers

The Vision Transformer (ViT) [39] partitions an image into a sequence of patches and models the dependencies among them using a self-attention mechanism. As the self-attention mechanism is permutation-invariant, ViT must incorporate Position Encoding to provide the model with information about the spatial arrangement of these patches. The predominant position encoding schemes include:

Absolute Position Encoding (APE). This includes learnable encodings, as used in the original ViT, or sinusoidal encodings, as proposed in [40,41]. APE assigns a unique, fixed encoding to each patch based on its absolute coordinates within the image. However, when dealing with multiple, unaligned, and rotated images, the absolute coordinates of the same target feature will differ across views. Consequently, APE not only fails to provide inter-view correspondence but also introduces spatial confusion.

Relative Position Encoding (RPE). RPE and its variants [41–43] model relative spatial relationships by introducing a bias term into the attention calculation, which represents the offset between two patches. While RPE performs admirably within a single image, it was originally designed to handle two-dimensional translational shifts. For the complex geometric transformations involving both rotation and translation in multi-view SAR images, conventional RPE is likewise inadequate for accurate representation.

This analysis reveals a critical disconnect: the very architectures (Vision Transformers) that hold the potential to bypass registration are fundamentally limited by position encodings ill-suited for the complex geometric relationships across multiple views. Addressing this specific inadequacy is the central motivation for our work. By designing a novel multi-view polar coordinate position encoding, we aim to provide the Transformer model with an accurate spatial prior for multi-view geometry under a registration-free condition, thereby enabling truly effective early feature fusion.

3. Proposed Method

3.1. Spatial-Dimension Concatenation and the Global Feature Extraction Framework

To maximize the utilization of multi-view SAR information at the early fusion stage while fundamentally circumventing the feature distortion induced by traditional image registration, we introduce a feature extraction framework based on Spatial-Dimension Concatenation. The core principle of this strategy involves spatially stitching multiple independent views to form a single, wide composite image. This approach preserves the spatial integrity of each original view. It also reframes the problem of inter-view feature interaction as one of unified, global feature modeling.

As illustrated in Figure 1a, our processing pipeline adheres to the standard Vision Transformer (ViT) paradigm and comprises the following three steps:

3.1.1. Image Patching and Linear Embedding

First, we partition the concatenated N-view composite image, $\mathrm{I} \in R^{H \times (N \times W) \times C}$, into a series of non-overlapping two-dimensional patches. Subsequently, these 2D patches are flattened into one-dimensional vectors and embedded into a higher-dimensional feature space via a learnable linear projection layer, yielding a sequence of patch tokens $T = \{t_1, t_2, \ldots, t_M\}$, where M is the total number of patches. Furthermore, we prepend a learnable class token to the beginning of the sequence, which serves to aggregate global information for the final classification task.

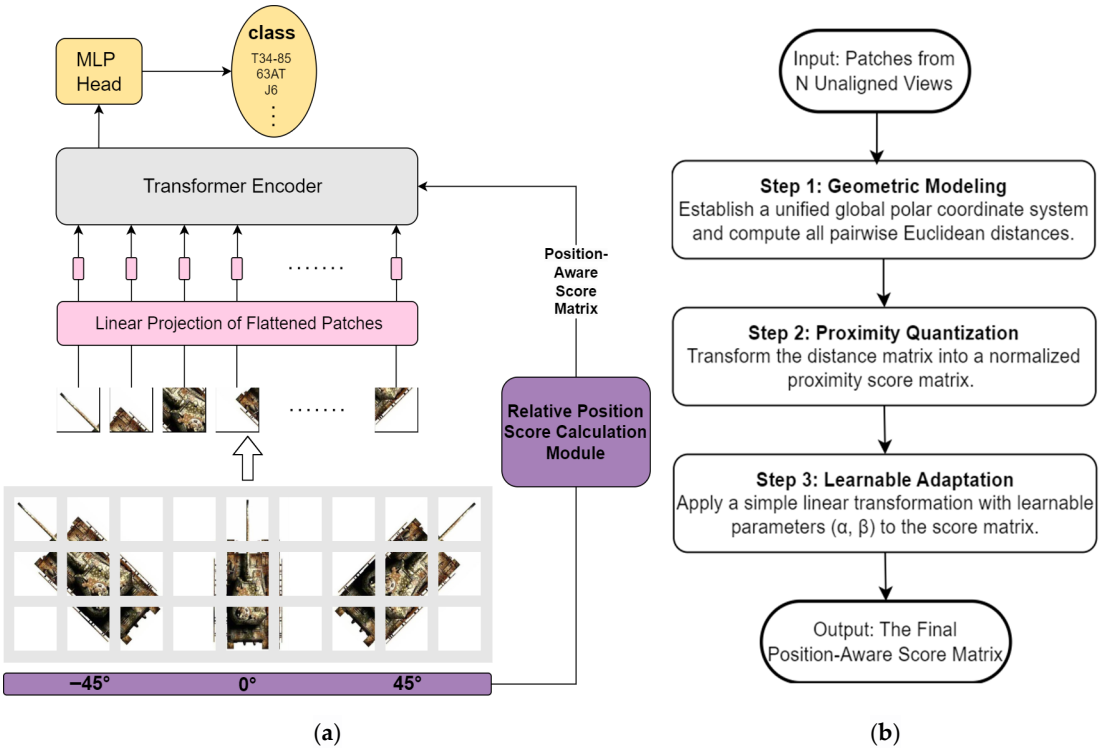

Figure 1. The architecture of our proposed registration-free fusion framework: (**a**) An overview of the framework. Multiple unaligned SAR images are concatenated spatially, partitioned into patches, and then linearly embedded into a sequence of tokens; (**b**) Core logic of the Relative Position Score Calculation Module. The proposed module transforms patches from unaligned views into a position-aware score matrix through three key steps. First, the geometric relationships between all patches are modeled within a unified polar coordinate system. Next, these physical relationships are quantified into a normalized proximity score matrix. Finally, a learnable linear transformation adapts this matrix to produce the final, dynamic position-aware score.

3.1.2. Feature and Positional Information Fusion

In a departure from the standard ViT, which directly adds position encodings to the patch tokens, we decouple the feature and positional information. This design choice is motivated by the unique challenges of multi-view fusion, where entangling spatial and semantic information can be suboptimal. By injecting the positional information as an independent bias term into the attention score matrix, we achieve three key theoretical advantages.

First, this approach decouples the representation of "what" (feature content) from "where" (spatial geometry). The self-attention scores computed from Query (Q) and Key (K) matrices primarily reflect semantic similarity, while our position-aware score matrix then modulates these scores based on geometric proximity. This allows the model to learn spatial relationships and feature similarities in separate pathways, preventing mutual interference. Second, it enhances flexibility, as the model can learn to weigh the importance of semantic content versus spatial priors. Third, it mitigates feature contamination. Directly adding a view-variant geometric encoding to a feature embedding, which should ideally be view-invariant, can contaminate the feature representation, making it harder for the model to learn a consistent semantic understanding of the target across different views. As detailed in the subsequent sections, we design a novel multi-view polar coordinate position encoding and transform it into a Relative Position Score Matrix, which is fused with the feature-based attention scores within the self-attention layer.

3.1.3. Transformer Encoder

The sequence of embedded patch tokens is then fed into a standard Transformer Encoder, as depicted in Figure 1b. The encoder is composed of multiple stacked attention blocks, where each block contains a Multi-Head Self-Attention (MHSA) layer and a Feed-Forward Network (FFN) layer. The MHSA layer is the core of this framework, as it computes the relevance weights between any two patch tokens in the sequence. This enables the model to capture long-range dependencies, including both intra-view relationships (e.g., the association between a tank's turret and its chassis) and inter-view relationships (e.g., the correspondence between the front of a vehicle in view A and its rear in view B). Through such global feature correlation and aggregation, our model not only achieves feature extraction within each view but also promotes the exchange of complementary information between views, ultimately leading to a highly consistent semantic representation in the fused features.

Finally, the feature vector corresponding to the class token output by the Transformer Encoder is passed to an MLP head for the final classification prediction.

3.2. Polar Coordinate Relative Position Encoding

To accurately describe the complex spatial geometry among unaligned multi-view images, we introduce a novel relative position encoding method based on polar coordinates. The choice of polar coordinates is motivated by their natural advantage in representing rotational transformations. multi-view SAR imagery inherently involves observing a target from different azimuth angles, which corresponds to a rotation. In a polar coordinate system (r, θ), a rotation around the origin is expressed as a simple addition to the angle component θ, while the radius r remains invariant. This property allows for a more direct and computationally simpler modeling of the geometric relationships between views compared to Cartesian coordinates, where rotation requires a more complex transformation of both x and y components. This method leverages the natural advantages of the polar coordinate system in representing rotational and translational relationships, enabling the construction of a unified and consistent global coordinate system for all patches after spatial-dimension concatenation. Our encoding process is structured in two steps: first, we define a local polar coordinate system within each individual view, and second, we extend it to a global coordinate system for the multi-view context.

3.2.1. Local Polar Coordinates Within a Single View

As shown in Figure 2, we first establish a local polar coordinate system for each individual view i. Let the image center be the origin, which we define as the local pole O_i. The horizontal axis extending to the right serves as the polar axis X_i. The position of the p-th patch's center, i_p, is then uniquely represented by the polar coordinates (r_{ip}, θ_{ip}). Here, r_{ip} is the polar radius, corresponding to the Euclidean distance between i_p and O_i. θ_{ip} is the polar angle, defined as the counter-clockwise angle from the polar axis X_i to the vector $\overrightarrow{O_i i_p}$, spanning the range $[0, 2\pi)$. This tuple, (r_{ip}, θ_{ip}), provides a precise description of the patch's absolute position within its host view.

3.2.2. Global Polar Coordinates for Multiple Views

To unify the spatial relationships across all views, a global reference frame is required. As depicted in Figure 3, we designate the center of the first view (e.g., view 1), O_1, as the global pole O. A fixed reference direction, such as the radar's line-of-sight, is defined as the global polar axis X. Building upon this frame, we define two key relative transformation parameters for every other view i, a target center offset $\widetilde{r}_i$ and a view rotation offset $\widetilde{\theta}_i$. $\widetilde{r}_i$ represents the displacement vector from the global pole O to the center of view i, O_i. $\widetilde{\theta}_i$

denotes the counter-clockwise angle of rotation from the global polar axis X to the local polar axis of view i, X_i.

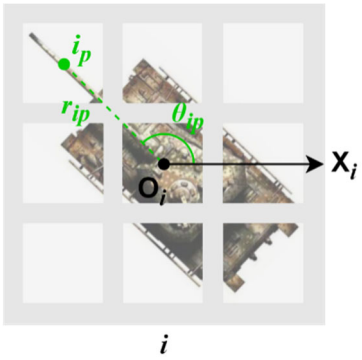

Figure 2. Definition of the local polar coordinate system within a single view. The local coordinate system for each view i is established with the image center as the pole O_i and the horizontal rightward direction as the polar axis X_i. The position of any patch p is uniquely determined by its polar radius r_{ip} (Euclidean distance to the pole) and its polar angle θ_{ip} (counter-clockwise angle from the axis).

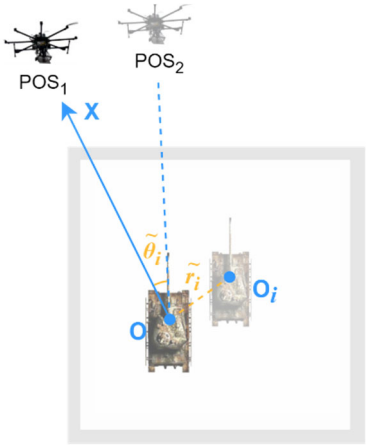

Figure 3. Establishment of the global coordinate system from local view coordinates. The global reference frame is defined by the pole (O) and axis (X) of the first view. Any other view i is oriented relative to this frame by a rotational offset $\tilde{\theta}_i$. The global coordinates of a patch are then derived by combining its local coordinates with this global offset, unifying all views into a common spatial framework.

By combining the local coordinates with these global transformation parameters, we can derive the final position encoding, $PE(i, p)$, for the p-th patch in the i-th view under the global coordinate system. According to the principle of vector addition, its global polar coordinates are:

$$PE(i, p) = (r_{ip} + \tilde{r}_i, \theta_{ip} + \tilde{\theta}_i), \tag{1}$$

In our spatial-dimension concatenation strategy, the centers of all views are aligned within the stitched composite image. We can therefore assume the target center offset $\tilde{r}_i = 0$. Under this condition, the formula simplifies to:

$$PE(i, p) = (r_{ip}, \theta_{ip} + \tilde{\theta}_i), \tag{2}$$

This final encoding rule accomplishes two objectives: it uses r_{ip} and θ_{ip} to describe the fine-grained spatial distribution of patches within a single view, while simultaneously using the global rotation offset $\tilde{\theta}_i$ to precisely characterize the rotational relationship between different views. This rule unifies patches from different views and rotational states into

a common coordinate framework. It extends the concept of adjacency from within a single view to the entire multi-view image set, thereby laying a mathematical foundation for the subsequent attention mechanism to enhance the correlation of patches across different views.

3.3. The Position-Aware Score Matrix and Its Injection Method

With the precise multi-view global position encoding established, the pivotal challenge becomes how to effectively inject this geometric information into the Transformer's self-attention mechanism. Standard self-attention is spatially agnostic; its scores are determined solely by the similarity between Query and Key feature content. To endow the model with an awareness of the spatial relationships between patches during attention computation, we design a relative position score matrix. This matrix acts as a spatial prior bias that is directly incorporated into the attention calculation.

3.3.1. Distance Matrix Construction from Position Encodings

The score matrix is designed to quantify the spatial proximity between any two patches, irrespective of whether they originate from the same view. We begin by calculating their Euclidean distance within the global polar coordinate system.

Let us consider any two patches: the p-th patch from view k and the q-th patch from view j. According to Equation (1), their global position encodings are given by:

$$PE(k,p) = (r_{kp} + \widetilde{r}_k, \theta_{kp} + \widetilde{\theta}_k), \tag{3}$$

$$PE(j,q) = (r_{jq} + \widetilde{r}_j, \theta_{jq} + \widetilde{\theta}_j), \tag{4}$$

Based on the law of cosines in a polar coordinate system, the straight-line distance d_{pq} between them can be expressed as:

$$d_{pq} = \sqrt{(r_{kp} + \widetilde{r}_k)^2 + (r_{jq} + \widetilde{r}_j)^2 - 2(r_{kp} + \widetilde{r}_k)(r_{jq} + \widetilde{r}_j)\cos[(\theta_{kp} + \widetilde{\theta}_k) - (\theta_{jq} + \widetilde{\theta}_j)]}, \tag{5}$$

When $\widetilde{r}_k = \widetilde{r}_j = 0$, the above expression degenerates to Equation (6):

$$d_{pq} = \sqrt{r_{kp}^2 + r_{jq}^2 - 2r_{kp}r_{jq}\cos[(\theta_{kp} + \widetilde{\theta}_k) - (\theta_{jq} + \widetilde{\theta}_j)]}, \tag{6}$$

From this, we can compute the pairwise distances between all M patches, forming a symmetric distance matrix $D \in R^{M \times M}$:

$$D(p,q) = \begin{bmatrix} d_{1,1} & \mathrm{d}_{1,2} & \cdots & d_{1,M} \\ d_{2,1} & \ddots & & \\ \vdots & & \ddots & \\ d_{M,1} & & & d_{M,M} \end{bmatrix}, \tag{7}$$

3.3.2. Distance Matrix Normalization

To transform the distance information into a score suitable for integration with attention scores and to enhance the association between nearby patches, we perform a linear normalization on the distance matrix D. This process scales its element values to the range $[-1, 1]$, where a higher score indicates closer proximity. The normalized score $\widetilde{d}_{pq}$ in the resulting matrix $\widetilde{D}$ is calculated as follows:

$$\widetilde{d}_{pq} = \frac{-2d_{pq}}{\max(D)} + 1, \tag{8}$$

where max(D) is the maximum value among all distances in the matrix. As visualized in Figure 4, the normalized score matrix $\widetilde{D}$ exhibits a clear spatial structure: elements near the main diagonal have the highest scores (close to 1), which smoothly decay as the distance increases. This intuitively reflects the spatial proximity prior that we have constructed.

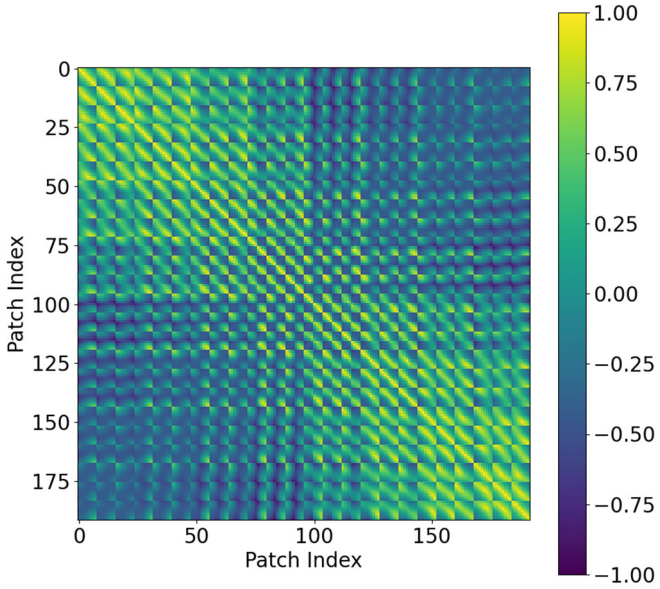

Figure 4. Visualization of Normalized Relative Position Score Matrix. A heatmap of the score matrix $\widetilde{D}$, where axes are patch indices and color intensity represents spatial proximity (brighter is closer). The strong diagonal pattern, with scores decaying with distance, visually confirms the effective encoding of our spatial proximity prior.

3.3.3. Spatially Aware Self-Attention Mechanism

Finally, we inject the positional information into the core of the multi-head self-attention. Instead of adding the position information directly to the feature embeddings, we introduce it as an independent bias term to be summed with the feature-based attention scores, forming the final hybrid attention score. This decoupled design helps to mitigate interference between the feature and position modalities, thereby strengthening the modeling capability of each.

The standard feature-based attention score matrix $\mathrm{A}(Q, K)$, is computed as follows:

$$\mathrm{A}(Q, K) = \frac{QK^T}{\sqrt{d_k}}, \tag{9}$$

We then transform the normalized distance matrix $\widetilde{D}$ into the final position-aware score matrix $S(\widetilde{D})$, via a learnable linear transformation:

$$S(\widetilde{D}) = \alpha\widetilde{D} + \beta, \tag{10}$$

where α and β are learnable scalar parameters. They allow the model to adaptively adjust the importance (scaling and shifting) of the spatial information during training.

The final Spatially Aware Attention is then computed as:

$$\mathrm{Attention}(Q, K, V, \widetilde{D}) = \mathrm{softmax}[A(Q, K) + S(\widetilde{D})]V, \tag{11}$$

In this manner, the position-aware score matrix $S(\widetilde{D})$ acts as a spatial relation map, guiding the model at every step of the attention computation. It imposes a spatial constraint on the attention distribution, encouraging the model to prioritize interactions between

spatially closer patches and to maintain geometric consistency. Simultaneously, because α and β are learnable, the model retains the flexibility to model long-range dependencies based on feature content. This design endows the model with an explicit capability for spatial generalization, enabling it to better understand and utilize the spatial structure of multi-view data.

4. Experiments and Analysis

4.1. Dataset and Experimental Setup

The experiments are conducted on the FAST-Vehicle dataset, which was collected by a Mini-SAR system developed in-house at the Nanjing University of Aeronautics and Astronautics. Unlike public datasets such as MSTAR, which only contain a limited range of aspect angles, our FAST-Vehicle dataset provides full 360° azimuthal coverage. It also includes a richer set of depression angle variations and comprises nine classes of typical military targets, offering a more ideal platform for studying the continuous variations and rotational invariance of multi-view features (see Figure 5). The dataset is partitioned into a training set and a test set according to the data collection campaigns (see Table 1 for details).

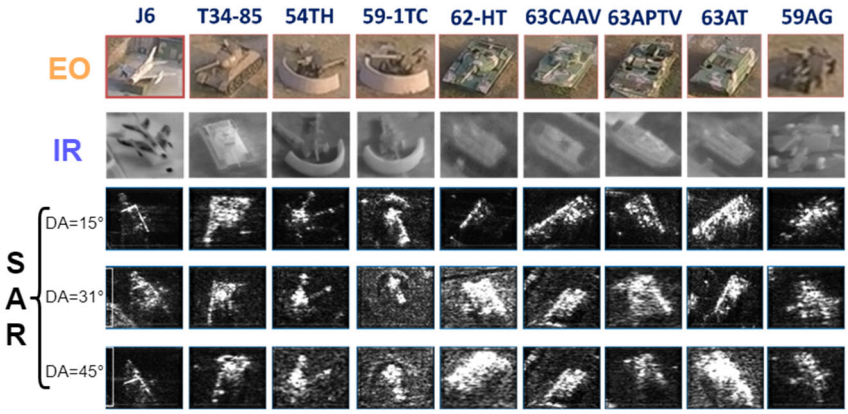

Figure 5. Electro-optical, IR, and SAR image chips of the targets in the FAST-vehicles dataset. Examples of the nine target classes in the FAST-Vehicle dataset. The rows display images from Electro-Optical (EO), Infrared (IR), and SAR sensors at different depression angles (DA), illustrating the multi-modal and multi-condition nature of the data.

Table 1. Number of training and test images for the FAST-Vehicles experimental setup.

Types	Training Set (July)		Test Set (March)	
	Depression Angle	Number	Depression Angle	Number
62LT	26°, 31°, 37°, 45°	1103	15°, 31°, 45°	385
63APTV	26°, 31°, 37°, 45°	1101	15°, 31°, 45°	414
63CAAV	26°, 31°, 37°, 45°	1104	15°, 31°, 45°	357
63AT	26°, 31°, 37°, 45°	1102	15°, 31°, 45°	488
T34-85	26°, 31°, 37°, 45°	1084	15°, 31°, 45°	1810
59-1TC	26°, 31°, 37°, 45°	1105	15°, 31°, 45°	854
54TH	26°, 31°, 37°, 45°	1094	15°, 31°, 45°	446
59AG	26°, 31°, 37°, 45°	1093	15°, 31°, 45°	2027
J6	26°, 31°, 37°, 45°	1101	15°, 31°, 45°	416

The original image resolution is 64 × 64. To be compatible with the ViT model's input size, all images are uniformly resized to 224 × 224. We use three-view images with equal

azimuthal spacing, which are then concatenated along the spatial dimension to generate a composite sample of size 224 × 672. The experiments are implemented using the PyTorch (2.3.1) framework on an NVIDIA GeForce RTX 2080Ti (11GB) GPU. The model is optimized using the SGD optimizer with a momentum of 0.9. The initial learning rate is set to 0.001 and is coupled with a cosine annealing schedule. We use a weight decay of 5×10^{-5}, a DropPath rate of 0.2, and a batch size of 32.

4.2. Comparative Experiments and Analysis

4.2.1. Comparison of Different Position Encoding Strategies

To evaluate the performance of our proposed method, we conducted a series of comparative experiments, with the results presented in Table 2. All models in this comparison are based on the Vision Transformer (ViT) [39] architecture and were designed to systematically analyze the effects of view fusion and position encoding.

Table 2. Performance comparison of different position encoding strategies.

View(s)	Position Encoding	Training Samples	Test Samples	Parameters (M)	Accuracy (%)
Single-view	Learnable	9592	7197	9.546	54.6
3-view	Polar Coordinates (ours)	8284	1654	43.102	57.6
3-view	Absolute (2D sin/cos)	8284	1654	43.102	53.4
3-view	Relative (2D)	8284	1654	43.104	53.7
3-view	Learnable	8284	1654	43.255	52.8
3-view	MV-DCN	8284	1654	35.536	56.8

To assess the benefits of both multi-view fusion and our specific encoding, we established two key comparisons. First, a standard single-view ViT served as a performance baseline. Second, within the three-view fusion framework, we isolated the effect of the encoding strategy by benchmarking our novel polar coordinate encoding against three conventional schemes: absolute (sin/cos) [44], relative (2D) [45], and standard learnable absolute encoding [39].

The experimental results reveal several key phenomena. First, it is noted that all multi-view models based on spatial-dimension concatenation have a significantly larger number of learnable parameters (~43 M) compared to the single-view baseline (9.5 M). This is primarily due to the threefold increase in the input sequence length, which leads to a corresponding growth in the parameter count of the self-attention mechanism in the Transformer encoder. However, despite having nearly identical parameter counts (all ~43 M), the multi-view models employing traditional absolute, relative, and learnable position encodings exhibit no significant performance advantage (53.4%, 53.7%, 52.8%, respectively), and are even outperformed by the single-view baseline (54.6%), which has nearly five times fewer parameters.

This outcome substantiates our core argument: in a registration-free, early fusion scenario, standard position encodings designed for single images introduce severe spatial confusion. They are unable to effectively guide the model to leverage multi-view information, consequently leading to performance degradation. In contrast, our PolarFormer, which explicitly models the multi-view geometry, achieves a top accuracy of 57.6%. This performance gain is not merely due to an increase in model scale, but directly stems from the precise geometric guidance provided by our novel polar coordinate encoding. The comparison with the state-of-the-art method, MV-DCN [4], further substantiates this conclusion. Although our PolarFormer (43.1 M parameters) is larger than the more lightweight

MV-DCN (35.5 M), it still achieves a superior accuracy (57.6% vs. 56.8%). This result is significant, as it suggests that for this task, the architectural paradigm is more critical than model size alone. It indicates that investing model capacity into a standard Transformer backbone, when guided by an explicit and precise geometric prior as in our method, is a more effective strategy than relying on a smaller, specialized CNN architecture like MV-DCN that only implicitly adapts to local deformations.

To further analyze the classification performance of our method, we present the confusion matrix on the test set in Figure 6. Overall, the values on the main diagonal of the confusion matrix are significantly higher than the off-diagonal values, indicating that our algorithm achieves a high level of accuracy across most categories. Furthermore, the generally low off-diagonal values suggest that our method exhibits minimal confusion between different classes, enabling effective discrimination. When considered in conjunction with the lower accuracies of other encoding strategies shown in Table 2 and the potential for feature space confusion, Figure 6 visually corroborates the effectiveness of our algorithm in enhancing the classification accuracy of multi-view images.

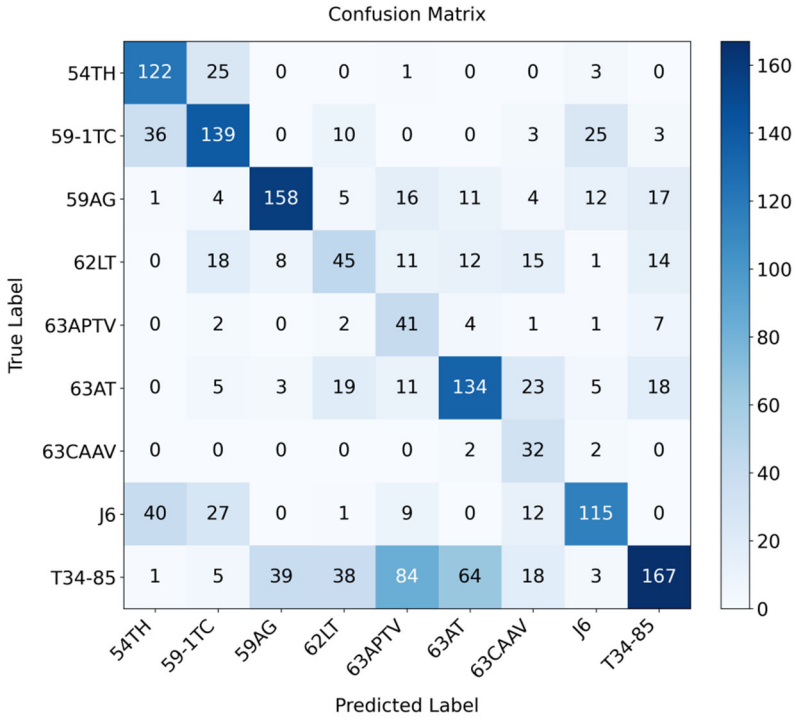

Figure 6. Confusion matrix of our proposed method on the FAST-Vehicle test set. The matrix shows the classification performance across all nine target classes. The high values along the main diagonal and low off-diagonal values demonstrate the method's high accuracy and effective inter-class discrimination.

4.2.2. Comparison with Image Registration Strategies

To further substantiate the superiority of our registration-free paradigm, we compare it against a stronger, more traditional pipeline that involves registration followed by recognition. This baseline strategy first aligns the three-view images using an image registration algorithm and then feeds the aligned stack into a ViT model that employs standard absolute position encoding.

The experimental results, as shown in Table 3, indicate that the performance of our proposed method (57.6%) surpasses that of the traditional image registration strategy (54.2%). Although image registration can achieve a preliminary alignment of the views, its processing pipeline inevitably introduces information loss and feature distortion due

to interpolation and resampling. This, in turn, adversely affects the model's recognition performance. In contrast, our method requires no image preprocessing and learns features directly from the raw multi-view data. This results in more robust and discriminative feature representations, demonstrating its superiority in utilizing multi-view information and affirming that modeling precise geometric relationships is a feasible and more effective alternative to physical image registration.

Table 3. Performance comparison between image registration and polar coordinate encoding strategies.

Position Encoding	Training Samples	Test Samples	Parameters (M)	Accuracy (%)
Polar Coordinates (ours)	8284	1654	43.102	57.6
Image Registration (Absolute)	8284	1654	43.102	54.2

4.3. Ablation Study

To isolate and evaluate the independent contribution of each innovative component within our method, we designed the following ablation studies.

4.3.1. Efficacy of the Polar Coordinate Position Encoding

To validate the intrinsic advantages of our polar coordinate encoding, we designed a controlled experiment. The experimental group employs our proposed encoding to generate the position score matrix. For the control group, we replace this with a standard, randomly initialized Learnable Positional Embedding, as used in ViT. This embedding is then passed through a small MLP to be transformed into a bias term of the same dimensionality as our score matrix and subsequently injected into the attention mechanism. Crucially, apart from the source of the position encoding, all other settings were kept identical.

As presented in Table 4, the model using our polar coordinate encoding achieves an accuracy of 57.6%, significantly outperforming the 53.6% achieved by the model using the standard learnable position encoding. This result indicates that, within an identical spatial feature injection framework, our proposed polar coordinate encoding provides more effective spatial guidance for the model in learning the spatial information of multi-view images. This indicates that the ability of our encoding to more precisely capture and utilize the view-dependent geometric information is a key determinant of the model's performance enhancement.

Table 4. Impact of position encoding on recognition rate.

Position Encoding Source	Training Samples	Test Samples	Parameters (M)	Accuracy (%)
Polar Coordinates (ours) + Score Matrix	8284	1654	43.102	57.6
Learnable + Score Matrix	8284	1654	43.104	53.6

4.3.2. Analysis of the Impact of View-Angle Separation

To investigate the effect of azimuthal angle separation on the recognition performance of our method, we conducted a series of comparative experiments, testing the model's accuracy at four different intervals: 6°, 12°, 18°, and 24°.

The results in Table 5 show that the recognition accuracy exhibits a non-linear relationship with the view-angle separation, peaking at 18° (57.6%). This suggests that a moderate angular separation (increasing from 6° to 18°) is beneficial for enlarging the view-dependent disparity, thereby enhancing the complementarity of the target's features across different views and enabling the model to capture a more comprehensive target representation.

However, when the separation becomes too large (e.g., 24°), the performance begins to decline. This is likely because excessive view disparity weakens the semantic consistency of the features, making it difficult for the model to focus on the core semantic characteristics of the target while simultaneously leveraging the complementary information from the views.

Table 5. Impact of azimuthal angle separation on recognition rate.

Azimuthal Angle Separation (°)	Accuracy (%)
6	53.8
12	55.3
18	57.6
24	54.4

5. Conclusions

This paper addresses the prevalent challenges of spatial ambiguity and registration dependency in early fusion methods for multi-view SAR target recognition. We propose a novel fusion paradigm based on spatially aware self-attention. The core innovation of our work lies in the design of a multi-view polar coordinate position encoding, which accurately represents the complex geometric relationships among multi-view images, thereby ensuring spatial consistency. Based on this encoding, we further construct a position-aware score matrix and integrate it into the Transformer's self-attention mechanism as a spatial inductive bias, which significantly strengthens the model's ability to perceive multi-view spatial structures.

We conducted comprehensive experiments on our self-developed FAST-Vehicle dataset. The results demonstrate that our method achieves a significant accuracy improvement of up to 4.8% and, more importantly, its performance markedly surpasses that of traditional "register-then-recognize" strategies. In-depth comparative and ablation studies further confirm that our proposed polar coordinate encoding, when compared to standard position encodings, can more effectively guide the model in utilizing multi-view information and is the key driver of the performance enhancement. Additionally, we investigated the impact of view-angle separation on performance, finding that the model achieves optimal performance at an 18° interval, a finding that holds significant reference value for practical applications.

In summary, this work successfully realizes a registration-free, end-to-end early feature fusion paradigm. It demonstrates that through precise modeling of underlying geometric relationships, physical image registration can be completely supplanted. This provides a promising new avenue for the efficient utilization of multi-view remote sensing image information. We acknowledge the computational complexity of our method as a primary limitation. The self-attention mechanism's complexity scales quadratically with the sequence length, which is directly proportional to the number of concatenated views. While this cost is manageable for the typical range of views (e.g., 3 to 10) employed in multi-view SAR target recognition, we recognize that it would become a significant bottleneck in scenarios requiring a much larger number of views. Therefore, a key direction for future work will be to mitigate this scalability issue by exploring more efficient Transformer architectures, such as those based on sparse attention or linear attention mechanisms. Furthermore, extending and validating our method on a broader range of public datasets, including MSTAR, is essential to further assess its generalizability. Future work will also continue to explore adaptive strategies for view-angle separation and more refined spatial feature embedding methods.

Author Contributions: X.Y.: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing—Original Draft, Visualization. (Corresponding Author). Y.Q.: Conceptualization, Methodology, Software, Validation, Visualization, Supervision, Project Administration, Writing—Review and Editing. G.J.: Software, Validation, Investigation. Z.G.: Software, Validation, Investigation. D.Z.: Conceptualization, Resources, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The FAST-Vehicle dataset supporting the conclusions of this article is not publicly available. This dataset was specifically developed in-house to provide full 360° azimuthal coverage under more challenging and realistic clutter conditions than are typically found in existing public benchmarks, which was necessary for the research objectives of this study. Due to the proprietary nature of the data and the sensitive information related to military targets, the dataset cannot be made public at this time. However, data may be made available from the corresponding author upon reasonable request for academic research purposes, subject to institutional approval and data usage agreements.

Acknowledgments: The authors would like to express their sincere gratitude to the research team at the Radar Imaging Technology Laboratory, College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, for their invaluable work in developing the Mini-SAR system and collecting the FAST-Vehicle dataset, which was fundamental to this study.

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

References

1. Tang, Y.; Chen, J. A multi-view SAR target recognition method using feature fusion and joint classification. *Remote Sens. Lett.* **2022**, *13*, 631–642. [CrossRef]
2. Ding, B.; Wen, G. Exploiting multi-view SAR images for robust target recognition. *Remote Sens.* **2017**, *9*, 1150. [CrossRef]
3. Wang, Z.; Zhang, G.; Zhu, D.; Dai, Q. Multi-view rotation double-layer fusion CNN-LSTM for SAR target recognition. In Proceedings of the 2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV), Zhenjiang, China, 12–14 July 2024; IEEE: New York, NY, USA, 2024; pp. 349–353.
4. Wang, Z.; Wang, C.; Pei, J.; Huang, Y.; Zhang, Y.; Yang, H.; Xing, Z. Multi-view SAR automatic target recognition based on deformable convolutional network. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; IEEE: New York, NY, USA, 2021; pp. 3585–3588.
5. Lv, J.; Zhu, D.; Geng, Z.; Han, S.; Wang, Y.; Ye, Z.; Zhou, T.; Chen, H.; Huang, J. Recognition for SAR deformation military target from a new MiniSAR dataset using multi-view joint transformer approach. *ISPRS J. Photogramm. Remote Sens.* **2024**, *210*, 180–197. [CrossRef]
6. Huan, R.; Pan, Y. Decision fusion strategies for SAR image target recognition. *IET Radar Sonar Navig.* **2011**, *5*, 747–755. [CrossRef]
7. Hu, Z.; Zhang, G.; Zhu, D. Multi-view SAR target recognition using bidirectional Conv-LSTM network. In Proceedings of the 2022 14th International Conference on Signal Processing Systems (ICSPS), Zhenjiang, China, 18–20 November 2022; IEEE: New York, NY, USA, 2022; pp. 410–413.
8. Zijun, W.; Gong, Z.; Daiyin, Z.; Qijun, D. MV-ResFPN: A residual network deeply combining multi-view and multiscale fusion for SAR target recognition. In Proceedings of the SPIE 13539, Sixteenth International Conference on Graphics and Image Processing (ICGIP 2024), Nanjing, China, 8–10 November 2024; p. 135390V.
9. Xiao, Z.; Zhang, G.; Dai, Q.; Wang, Z. A multi-view stable feature extraction network for SAR target recognition. In Proceedings of the 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Chengdu, China, 3–5 November 2023; IEEE: New York, NY, USA, 2023; pp. 114–117.
10. Yifei, S.; Sihang, D.; Boda, Q.; Shuliang, G.; Xiaoyue, J.; Xiaoyi, F. Synthetic aperture radar target recognition using multi-view feature enhancement-based contrastive clustering. *J. Appl. Remote Sens.* **2024**, *19*, 16503. [CrossRef]
11. Dai, Q.; Zhang, G.; Xue, B.; Fang, Z. Capsule-guided multi-view attention network for SAR target recognition with small training set. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
12. Tang, Y.; Wang, L.; Li, Y.; Zhu, D. Relation aware network for multi-view SAR target recognition. In Proceedings of the 2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Zhuhai, China, 22–24 November 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.

13. He, L.; Ohbuchi, R.; Jiang, M.; Furuya, T.; Zhang, M. Cascaded multi-channel feature fusion for object detection. In Proceedings of the 3rd International Conference on Control and Computer Vision, Macau, China, 23–25 August 2020; pp. 11–16.
14. Teepe, T.; Wolters, P.; Gilg, J.; Herzog, F.; Rigoll, G. EarlyBird: Early-fusion for multi-view tracking in the bird's eye view. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 1–6 January 2024; IEEE: New York, NY, USA, 2024; pp. 102–111.
15. Mao, S.; Yang, J.; Gou, S.; Jiao, L.; Xiong, T.; Xiong, L. Multi-scale fused SAR image registration based on deep forest. *Remote Sens.* **2021**, *13*, 2227. [CrossRef]
16. Huang, X.; Ding, J.; Guo, Q. Unsupervised image registration for video SAR. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1075–1083. [CrossRef]
17. Paul, S.; Pati Umesh, C. Automatic optical-to-SAR image registration using a structural descriptor. *IET Image Process.* **2020**, *14*, 62–73. [CrossRef]
18. Li, B.; Guan, D.; Xie, Y.; Zheng, X.; Chen, Z.; Pan, L.; Zhao, W.; Xiang, D. Global optical and SAR image registration method based on local distortion division. *Remote Sens.* **2025**, *17*, 1642. [CrossRef]
19. Pan, B.; Jiao, R.; Wang, J.; Han, Y.; Hang, H. SAR image registration based on KECA-SAR-SIFT operator. In Proceedings of the 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Nanjing, China, 23–25 September 2022; IEEE: New York, NY, USA, 2022; pp. 114–119.
20. Geng, J.; Zhang, Y.; Jiang, W. Polarimetric SAR image classification based on hierarchical scattering-spatial interaction transformer. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
21. Deng, J.; Zhu, Y.; Zhang, S.; Chen, S. SAR image recognition Using ViT network and contrastive learning framework with unlabeled samples. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
22. Qin, Y.; Xu, W.; Yao, Y.; Huang, X. SAR-3DTR: A novel feature hybrid transformer network for end-to-end 3-D target reconstruction from SAR images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
23. Zhang, B.; Wu, Q.; Wu, F.; Huang, J.; Wang, C. A Lightweight pyramid transformer for high-resolution SAR image-based building classification in port regions. *Remote Sens.* **2024**, *16*, 3218. [CrossRef]
24. Yataka, R.; Wang, P.P.; Boufounos, P.; Takahashi, R. Multi-view radar detection transformer with differentiable positional encoding. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
25. Wang, L.; Tang, M.; Rong, Y.; Ni, M.; Li, F. Multi-view SAR image classification through decision fusion of adaptive dictionary learning and CNN. In Proceedings of the 2024 Photonics & Electromagnetics Research Symposium (PIERS), Chengdu, China, 21–25 April 2024; IEEE: New York, NY, USA, 2024; pp. 1–7.
26. Zhang, T. A Multi-view SAR target recognition method based on adaptive weighted decision fusion. *Remote Sens. Lett.* **2023**, *14*, 1196–1205. [CrossRef]
27. Juan, L. Synthetic aperture radar target recognition based on adaptive decision fusion of multiple views. *J. Electron. Imaging* **2024**, *33*, 23015. [CrossRef]
28. Chen, Y.; Bruzzone, L. Self-supervised SAR-optical data fusion of sentinel-1/-2 Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
29. Zhang, Y.; Guo, X.; Ren, H.; Wan, Q.; Shen, X. Multi-View fusion based on expectation maximization for SAR target recognition. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; IEEE: New York, NY, USA, 2020; pp. 778–781.
30. Ettinger, G.J.; Snyder, W.C. Model-based fusion of multi-look SAR for ATR. In Proceedings of the Algorithms for Synthetic Aperture Radar Imagery IX, Orlando, FL, USA, 1–5 April 2002; SPIE: Bellingham, WA, USA, 2002; pp. 277–289.
31. Chang, W.; Chen, N. Pixel level fusion approach based on optimizing visual perception for multi-band SAR image. *Syst. Eng. Electron.* **2004**, *9*, 1299–1301.
32. Sun, Y.L.; Wang, J. Performance analysis of SIFT feature extraction algorithm in application to registration of SAR image. In Proceedings of the MATEC Web of Conferences, Lucerne, Switzerland, 6–10 July 2016; EDP Sciences: London, UK, 2016; p. 01063.
33. Sreeja, G.; Saraniya, O. A comparative study on image registration techniques for SAR images. In Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; IEEE: New York, NY, USA, 2019; pp. 947–953.
34. Yu, Q.; Pang, B.; Wu, P.; Zhang, Y. Automatic coarse-to-precise subpixel multi-band SAR images co-registration based affine SIFT and radial base function(RBF). In Proceedings of the 2015 IEEE 5th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Singapore, 1–4 September 2015; IEEE: New York, NY, USA, 2015; pp. 684–687.
35. Li, N.; Hu, X. UltraWideband Mutual RFI Mitigation Between SAR Satellites: From the Perspective of European Sentinel-1A. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–20. [CrossRef]
36. Fan, Y.; Wang, F.; Wang, H. A transformer-based coarse-to-fine wide-swath SAR image registration method under weak texture conditions. *Remote Sens.* **2022**, *14*, 1175. [CrossRef]

37. Zhou, R.; Wang, G.; Xu, H.; Zhang, Z. A sub-second method for SAR image registration based on hierarchical episodic control. *Remote Sens.* **2023**, *15*, 4941. [CrossRef]
38. Deng, X.; Mao, S.; Yang, J.; Lu, S.; Gou, S.; Zhou, Y.; Jiao, L. Multi-class double-transformation network for SAR image registration. *Remote Sens.* **2023**, *15*, 2927. [CrossRef]
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.
41. Foumani, N.M.; Tan, C.W.; Webb, G.I.; Salehi, M. Improving position encoding of transformers for multivariate time series classification. *Data Min. Knowl. Discov.* **2024**, *38*, 22–48. [CrossRef]
42. Wu, K.; Peng, H.; Chen, M.; Fu, J.; Chao, H. Rethinking and improving relative position encoding for vision transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 10013–10021.
43. Ma, Y.; Wang, R. Relative-position embedding based spatially and temporally decoupled transformer for action recognition. *Pattern Recognit.* **2024**, *145*, 109905. [CrossRef]
44. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 261–272.
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 9992–10002.