A Novel Sequence Modeling Network for Multiview SAR Target Recognition

Wei Zhu, Weiguo Xiao, Xin Li, Shunping Xiao, Xiaolin Hu, and Linbin Zhang

***Abstract*—Synthetic aperture radar (SAR) is an active remote sensing system that utilizes radar to produce images of the Earth's surface. Due to its ability to operate under diverse weather conditions and throughout the day, SAR has gained significant attention in both civilian and military domains. The utilization of multiview SAR sequences enables the acquisition of a more comprehensive range of information than a single image, and facilitates adaptation to diverse scenarios, thereby enhancing the ability to accommodate variations in samples. Drawing inspiration from the Transformer architecture, this article proposes a multiview SAR target recognition method, called Res-Xformer, that not only deconstructs the deep learning procedure into single image feature extraction and sequence feature fusion but also divides the task of sequence feature extraction into sequence information fusion and feature channel fusion. Different from the Transformers focusing on the attention mechanism to fuse sequence information, alternative fusion methods such as multilayer perceptron (MLP) and pooling are also proposed in this study. Experimental results using the Moving and Stationary Target Acquisition and Recognition dataset demonstrate that the proposed method performs well across various operational conditions, with MLP and pooling as sequence token mixers yielding comparable performance to the attention mechanism.**

***Index Terms*—Multiview, Res-Xformer, sequence feature fusion, sequence modeling architecture, synthetic aperture radar (SAR), token mixer.**

I. Introduction

SYNTHETIC aperture radar (SAR) [1] is an active remote sensing system that utilizes radar to generate high-resolution images of the Earth's surface. SAR operates by emitting microwave pulses toward the ground and subsequently capturing the reflected signals. By measuring the time delay between signal transmission and reception, SAR can calculate the distance between the microwave sensors and the Earth's surface, ultimately producing a range image. SAR images represent an estimate of the radar backscatter on the ground.

As SAR has attracted increasing attention for its ability to perform in all weather conditions and all day long, not to mention its increasingly high resolution, it has a variety of civilian and military applications [2].

Simultaneously, due to the distinct imaging mechanism of SAR images compared to optical images familiar to the human visual system, interpreting SAR imagery requires specialized knowledge in areas such as SAR imaging mechanisms and microwave scattering characteristics. Consequently, SAR interpretation poses challenges that demand expertise. As a fundamental issue within the SAR image interpretation system, automatic target recognition (ATR) for SAR holds direct implications for practical applications such as ground object identification and battlefield situational awareness, playing a crucial role in national defense [3].

In recent years, with the development of high-performance computing and the geometric growth of data scale, deep learning methods for automatic feature extraction have played an important role in lots of fields [5]. In particular, computer vision has witnessed remarkable progress, attaining state-of-the-art outcomes in image denoising, semantic segmentation, classification, and other related fields.

The deep learning methods based on convolutional neural networks (CNNs) have also exhibited significant advantages in SAR target detection and recognition tasks. One aspect of the research focuses on addressing the limited number of SAR image annotation samples [26], [27], [28], [29]. On another front, researchers focus on designing network structures suitable for SAR ATR problems. CNNs [6] have revolutionized computer vision by enabling accurate recognition and classification of images by machines. CNN-based networks are increasingly popular in SAR recognition due to their ability to effectively extract features from images. For example, Chen et al. [11] proposed a fully convolutional network for accomplishing target recognition in SAR images, and Geng et al. [12] designed a deep supervised contractive CNN for segmenting SAR images against speckle interference. Under standard operating conditions (SOCs), these methods demonstrate excellent results. However, their performances decline sharply when confronted with disparities between training and testing samples, thereby limiting their practical applicability. To address this limitation, it is necessary to incorporate additional information, such as multiview SAR sequences into deep networks.

Manuscript received 15 March 2024; revised 22 May 2024; accepted 28 May 2024. Date of publication 4 June 2024; date of current version 14 June 2024. *(Corresponding author: Wei Zhu.)*

Wei Zhu is with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha 410073, China, and also with the National Key Laboratory of Intense Pulsed Radiation Simulation and Effect, Northwest Institute of Nuclear Technology, Xi'an 710024, China (e-mail: zhuwei@nint.ac.cn).

Weiguo Xiao, Xin Li, and Xiaolin Hu are with the National Key Laboratory of Intense Pulsed Radiation Simulation and Effect, Northwest Institute of Nuclear Technology, Xi'an 710024, China (e-mail: xiaoweiguo@nint.ac.cn; lixin@nint.ac.cn; huxiaolin@nint.ac.cn).

Shunping Xiao and Linbin Zhang are with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha 410073, China (e-mail: xiaoshunping_nudt@163.com; zhanglinbin14@gfkd.mtn).

Digital Object Identifier 10.1109/JSTARS.2024.3409170

The SAR target is highly sensitive to the change of the observation azimuth, and the multiview SAR image sequence of the same observation target contains the scattered echoes from multiple angles, providing more information than the single-view image [1]. It can capture detailed information about targets from different perspectives while minimizing errors due to occlusion or shadowing effects. It exhibits superior resistance to the sensitivity of SAR images and enables more precise characterization of target scattering properties, thereby enhancing algorithmic accuracy and robustness.

The neural networks utilized for target recognition using multiaspect SAR image sequences primarily consist of recurrent neural networks (RNNs) [9], such as long short-term memory (LSTM) [10], and CNNs. For multiview SAR processing, the authors in [19], [20], [21], and [22] used a CNN-based network to extract features, and then combined sequence information. As to the use of RNNs, the authors in [23], [24], and [25] fused sequence information by implementing recurrent units.

Designed to process sequence problems, the Transformer [31] is a powerful deep learning model architecture that has made significant contributions to the field of natural language processing. By decomposing single image into smaller patches and processing them individually, a Vision Transformer (ViT) [34] revolutionarily applies the Transformer in the field of computer vision.

The application of Transformers in SAR domains remains limited. Li et al. [42] employed a basic transformer for multiaspect SAR recognition and utilized attention to integrate information within the sequence. Wang et al. [41] introduced convolution into a patch embedding procedure inspired by ViT to address few-shot learning challenges in SAR ATR tasks. The Transformer model exhibits its potential for multiview SAR recognition tasks. It is highly likely that the Transformers will assume an even more crucial role in SAR image interpretation.

These initial applications of the basic Transformer primarily focus on using an attention mechanism to integrate information while neglecting the impact of other components such as its structure, which is crucial for optimizing its performance in SAR recognition tasks from our perspective. Motivated by previous studies on the role of the attention mechanism in Transformers for natural image classification [35], [36], [37], [38], [39], [40], we aim to explore the application of the Transformer architecture for image sequence in the field of multiview SAR recognition.

In this article, we propose a novel target recognition method for multiview SAR image sequence, which is the first time to systematically analyze the validity of the Transformer structure in a multiview SAR area. The main contributions of this article are as follows.

1) We propose a Res-Xformer method that combines convolution and the sequence modeling structure to solve multiview SAR recognition. The method uses ResNet [43] to embed single images into feature tokens and Xformer, inspired by the Transformer structure, to fuse the sequence of feature tokens, resulting in an output type for the input sample. "X" represents the changeable token mixer.
2) Instead of treating the network as a black box, the proposed approach deconstructs the deep learning processes of multiview SAR into single image feature extraction and sequence feature fusion. The sequence feature fusion module is further divided into sequence token fusion (STF) and feature channel fusion (FCF), presenting a framework for multiview SAR recognition.
3) In contrast to existing applications of the Transformer in SAR ATR focusing on attention mechanisms, this article focuses on an aggregation procedure of sequence features by proposing a framework for sequence signal feature extraction.
4) The concept of class token for sequence information representation is innovatively proposed to duel with multiview SAR ATR, corresponding to the structure of sequence feature fusion network.

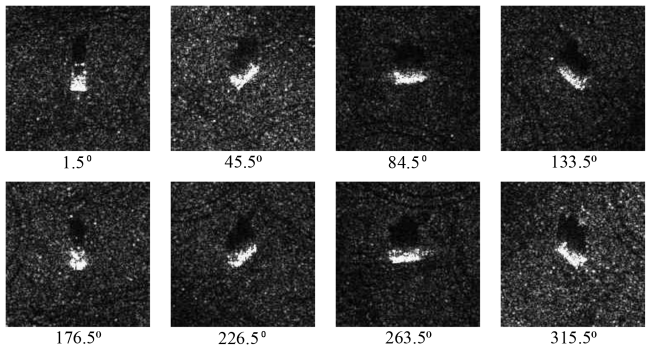

Fig. 1. SAR images of target T62 acquired at varying azimuth angles.

This article is organized as follows. The proposed Res-Xformer network for multiview SAR ATR is described in detail in Section II. Section III presents the formation of the dataset for the proposed method. The experiments and result discussions are given in Section IV, while taking into account various factors such as network parameters. Finally, Section V concludes this article.

II. Proposed Method

The proposed Res-Xformer method for multiview SAR image classification is introduced in this section. The overall framework of the method is presented, along with the architecture and details of each module.

A. Multiview SAR Target Recognition Framework

Multiview SAR is essentially a combination of multiple SAR images taken from different azimuth angles. Multiview SAR can provide a richer set of information for target feature analysis and classification than single-view SAR.

However, working with multiaspect images also presents challenges due to their sensitivity to aspect angles. As shown in Fig. 1, the backscatter imaging mechanism employed in SAR makes extracted features highly dependent on the acquisition geometry [4]. To effectively process multiview SAR image sequences, two dimensions need to be considered.

First, feature information needs to be extracted from each individual SAR image within the sequence. Second, these extracted

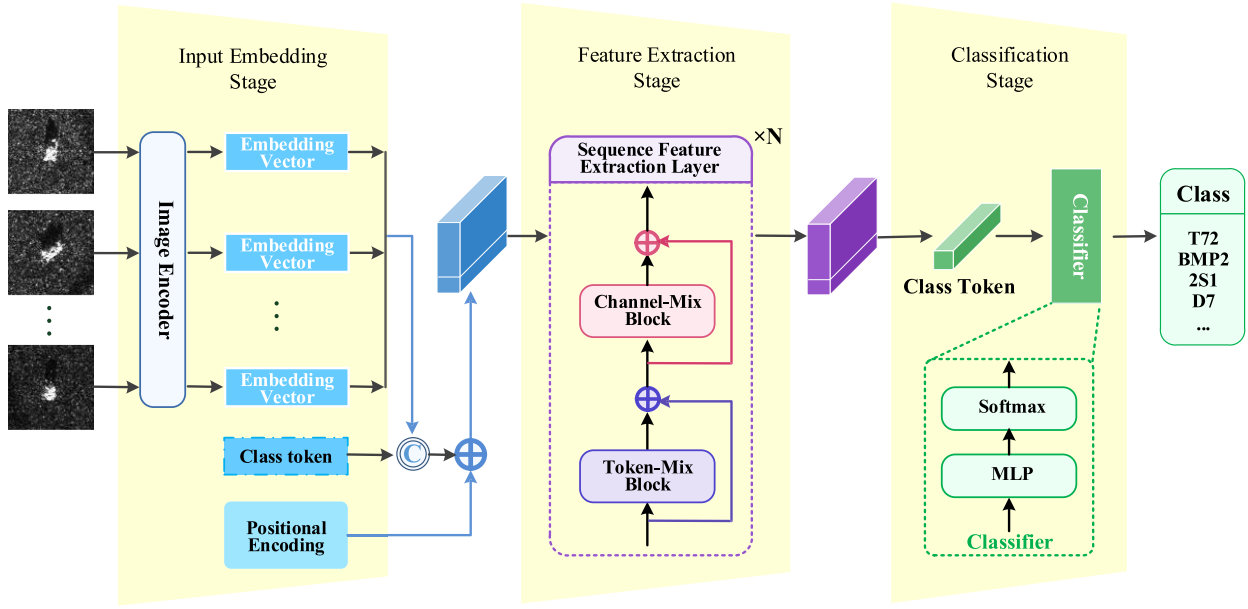

Fig. 2. Overall framework of the proposed Res-Xformer for multiview SAR ATR.

features must be fused together coherently across the entire sequence to create a comprehensive representation. This fusion process aims at integrating complementary information from different views while minimizing redundancy or conflicting details.

As depicted in Fig. 2, we propose the Res-Xformer for multiview target recognition, which consists of three stages: input embedding, sequence feature extraction, and classification.

Initially, the images in the input multiview sequence are independently embedded into feature vectors using ResNet. In addition, a learnable embedding called class token is concatenated with these vectors. Drawing inspiration from ViT, the proposed sequence feature extraction network (SFEN) for multiview SAR employs N stacked feature extraction layers. Each layer incorporates two primary blocks, referred to as Xformer, for the fusion of sequence information. The term "X" denotes its replicable token mixer design. During the classification stage, a straightforward classifier consisting of MLP and softmax takes the state of the class token as a representation for decision-making of the multiview SAR sequence.

The subsequent sections provide a comprehensive depiction of the specific modules employed in the proposed method.

B. Input Embedding

The initial step of the proposed methodology involves converting the image sequence into feature tokens, thereby enabling subsequent stages to effectively extract and integrate both sequence tokens and feature channels for accurate classification.

1) Main Flow: Sequence models, such as LSTM and the Transformer, typically require one-dimensional (1-D) features as input for each time step. Similarly, in our proposed method, the SFEN also requires a 1-D vector for each image as input during the sequence encoder stage to conveniently process sequence information. This component aims to reduce the dimensionality of the input by transforming the 2-D images in the sequence into 1-D feature tokens, as illustrated in Fig. 3.

The input sample consists of a sequence of images, where each image is considered as a token with the same shape (H, W). Since our focus is on the amplitude images, the gray image can be represented as $\mathbf{x}_{in}^i \in \mathbb{R}^{H\times W}, i = 1, \ldots, L$, where L denotes the length of the sequence sample. After passing through the image encoders, each image token's embedding output $\mathbf{x}_t^i \in \mathbb{R}^{d_{\text{emb}}}$ has a dimensionality of d_{emb}.

To incorporate classification information, we prepend a learnable embedding $\mathbf{x}_t^0$ as the class token and add positional encodings PE to retain positional information. This procedure can be described as follows:

$$\begin{aligned}\mathbf{X}_{\text{emb}} &= \text{Concat}(\mathbf{x}_t^0, \text{InputEncoder}(\mathbf{X}_{in})) + \text{PE} \\ &= \left[\mathbf{x}_t^0, \mathbf{x}_t^1, \ldots, \mathbf{x}_t^L\right] + \text{PE}.\end{aligned} \tag{1}$$

where $\mathbf{X}_{in} \in \mathbb{R}^{L\times H\times W}$ with image shape (H, W) and sequence length L, and $\mathbf{x}_t^i \in \mathbb{R}^{d_{\text{emb}}}$ with embedding dimension d_{emb}.

Consequently, the output embedding sequence $\mathbf{X}_{\text{emb}}$ has a dimensionality of $(L + 1) \times d_{\text{emb}}$. The class token is a learnable embedding to serve as the representation of the sequence. The positional encodings are to retain order information.

2) Single Image Encoding: The encoder in the input embedding stage plays a crucial role in reducing dimensionality and extracting features from individual images within the multiview SAR sequence. This enables the sequence to be seamlessly transmitted to the subsequent stage for sequence feature extraction. This procedure can be described as follows:

$$\mathbf{x}_t^i = \text{InputEncoder}(\mathbf{x}_{in}^i). \tag{2}$$

where $\mathbf{x}_{in}^i \in \mathbb{R}^{H\times W}, i = 1, \ldots, L$ with image shape (H, W) and sequence length L, and $\mathbf{x}_t^i \in \mathbb{R}^{d_{\text{emb}}}$ with embedding dimension d_{emb}.

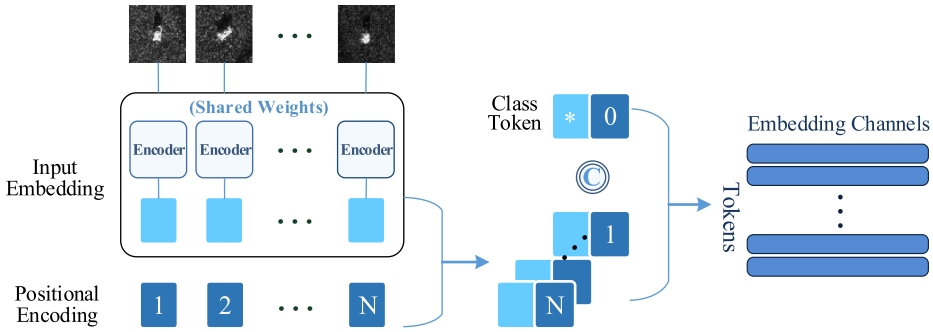

Fig. 3. Illustration of the input embedding procedure for multiview SAR ATR.

TABLE I
ARCHITECTURES OF INPUT IMAGE ENCODER BASED ON RESNET18 [43]

Layer Name	Layer Detail
Conv 1	$3\times3, 64, /2$ 3×3 max pool, /2
Conv 2	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$
Conv 3	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$
Conv 4	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$
Conv 5	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$
Out	Average Pool

In this proposed method, ResNet18 [43] is chosen as the backbone for input embedding instead of deeper versions since SAR image datasets typically have much smaller scales compared to natural images. Given that the input SAR images are grayscale and relatively small in size, the first convolution layer of ResNet18 is modified to employ a 3×3 convolution kernel with one input channel. Furthermore, we remove the classification head from ResNet18 as the encoder solely provides feature vectors rather than classification results. For more details on the encoder module, refer to Table I, where 3×364 means the size of the convolution filters in the layer is 3×3 and the number of filters is 64, and /2 means stride is 2.

The class token, denoted as $\mathbf{x}_t^0 \in \mathbb{R}^{d_{\text{emb}}}$, is a learnable variable with identical dimensionality to that of the image embedding and is transmitted into positional encoding along with the image embedding tokens for further processing.

3) Positional Encoding: In the subsequent stage, the sequence fusion module should be capable of processing the relative positions of tokens within the input sequence.

Hence, positional encodings are incorporated into the image token embeddings within a sequence at the input layer of the SFEN to establish a sense of sequential order among tokens. The positional encodings possess equivalent dimensionality d_{emb} as token embeddings and can either be learnable or assigned [26]. In [27], both approaches for positional encoding were examined and yielded nearly identical outcomes. We choose the fixed version of sinusoidal encoding functions in our proposed method. These functions encompass sine and cosine components due to their inherent resilience toward variations in sequence length, which would facilitate our future research. The encoding functions are defined as follows:

$$\begin{cases} \text{PE}_{(\text{pos},2i)} = \sin\left(\text{pos}/10000^{2i/d_{\text{emb}}}\right) \\ \text{PE}_{(\text{pos},2i+1)} = \cos\left(\text{pos}/10000^{2i/d_{\text{emb}}}\right) \end{cases} \tag{3}$$

where pos is the position in the sequence and *i* denotes the index in the embedding vector's dimension.

In the absence of information regarding the sequential relationships of the input tokens, the network treats them as an unordered collection, disregarding their inherent order. By incorporating 1-D positional encoding for each token, the inputs are perceived as a sequence, enabling the model to leverage the order information.

C. Sequence Feature Extraction Network

The features of tokens derived from the input sequence are propagated to this stage of Res-Xformer for subsequent feature extraction and fusion at the sequence level. The output dimensions remain consistent with the input.

1) Main Structure: The SFEN is presented in Fig. 4, providing an overview of its functionality. After incorporating the 1-D embedding of each token within the sequence as input, which is generated by the preceding stage of encoding, the output of SFEN maintains identical dimensions to that of the input. Moreover, each token encompasses integrated information from other tokens throughout the process. The procedure can be written as

$$\begin{aligned} \mathbf{X}_F &= \left[\mathbf{x}_F^0, \mathbf{x}_F^1, \ldots, \mathbf{x}_F^L\right] \\ &= \text{SFEN}(\mathbf{X}_{\text{emb}}) = \left[FCF\left(STF\left(\mathbf{X}_{\text{emb}}\right)\right)\right]^N \end{aligned} \tag{4}$$

where $\mathbf{X}_{\text{emb}} \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$ and $\mathbf{X}_F \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$ with sequence length L and embedding dimension d_{emb}.

The proposed SFEN comprises N repeated layers for encoding sequence features, each of which consists of two primary blocks: the STF block and the FCF block. These blocks are dedicated

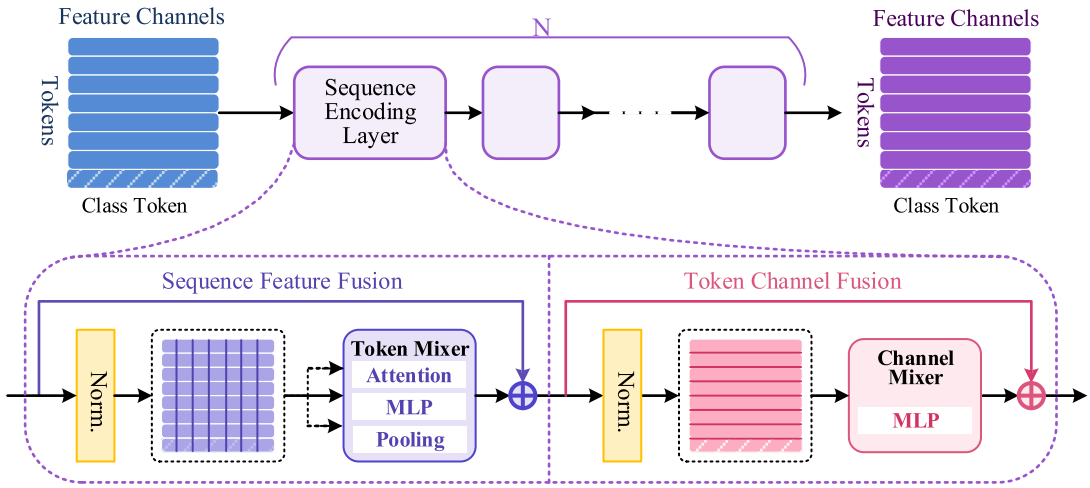

Fig. 4. Architecture of the proposed SFEN block.

to integrating sequence information and fusing feature channels for individual tokens, respectively. Consequently, following this stage, the class token in the input embeddings can assimilate information from all tokens within the sequence.

2) Sequence Token Fusion: The STF block, as depicted in Fig. 4, takes the sequence image embeddings as input and comprises three components: layer normalization (Norm), token information mixer, and a skip connection to facilitate training and prevent the occurrence of vanishing gradients. The procedure can be described as follows:

$$\begin{aligned}\mathbf{X}_{\text{STF}} &= STF\left(\mathbf{X}_{\text{emb}}\right) \\ &= \text{TokenMixer}\left(\text{Norm}\left(\mathbf{X}_{\text{emb}}\right)\right) + \mathbf{X}_{\text{emb}}\end{aligned} \tag{5}$$

where $\mathbf{X}_{\text{emb}} \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$, $\mathbf{X}_{SFF} \in \mathbb{R}^{(L+1)\times d_{emb}}$.

The token mixer should possess the capability of infiltration and assimilation, implying its ability to incorporate feature information from all other tokens and integrate it into the current token. While the basic Transformer employs an attention mechanism as its token mixer, we propose that in the SAR ATR field, the Transformer can be effective due to its structure, thereby operations with the capability of aggregating information have the potential to function as a token mixer. In our SFEN, multilayer perceptron (MLP) and pooling are considered as viable alternatives for token mixing.

a) Attention mechanism: The attention mechanism [32], [33] is a fundamental concept that enables the model to selectively focus on specific aspects of input data, exerting a more pronounced influence on the output. It computes the correlations between the sequence tokens and generates attention weights, subsequently utilized for integrating information within the sequence.

For an input sequence $\mathbf{X} \in \mathbb{R}^{L\times D}$, the scaled dot-product attention used in the Transformer can be written as

$$\begin{cases}\mathbf{A}\left(\mathbf{X}\right) = \text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \\ [\mathbf{Q},\mathbf{K},\mathbf{V}] = \mathbf{X}\mathbf{W}_{QKV}\end{cases} \tag{6}$$

where $\mathbf{A} \in \mathbb{R}^{L\times d_v}$ and $\mathbf{W}_{QKV} \in \mathbb{R}^{D\times(2d_k+d_v)}$; queries $\mathbf{Q}$ and keys $\mathbf{K}$ have the feature dimension d_k, and values $\mathbf{V}$ have the feature dimension d_v

The particular attention used in the Transformer is the multihead self-attention (MSA), which projects inputs into multiple queries, keys, and values h times. This approach facilitates the model's ability to capture diverse relationships among tokens within the sequence and acquire more comprehensive representations. Mathematically, the multihead attention can be formulated as follows:

$$\begin{cases}\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = Concat(head_1,\ldots,head_h)\mathbf{W}^O \\ \text{head}_i = \text{A}_i\left(\mathbf{X}\right) = \text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})\end{cases} \tag{7}$$

where $\mathbf{W}^O \in \mathbb{R}^{hd_v\times D}$ are the projection matrices.

The procedure of MSA can be written as

$$\begin{cases}\text{TokenMixer}_{\text{MSA}}\left(\mathbf{X}_n\right) = \text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) \\ \mathbf{X}_n = \text{Norm}\left(\mathbf{X}_{\text{emb}}\right)\end{cases} \tag{8}$$

where $\mathbf{X}_n \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$ and $d_k = d_v = d_{\text{emb}}/h$.

b) MLP: The MLP is a type of artificial neural network that consists of multiple layers of interconnected nodes. Each node receives inputs from the previous layer, performs a weighted sum of these inputs, and generates an output through an activation function. This allows the network to learn complex patterns and relationships in input samples.

The MLP utilized as the token mixer comprises of two fully connected (FC) layers and activation functions. The formulation for the token mixer based on MLP can be expressed as follows:

$$\begin{cases}\text{TokenMixer}_{\text{MLP}}\left(\mathbf{X}_n\right) = \sigma\left(\sigma\left(\mathbf{X}_n^T\mathbf{W}_1^{\text{MLP}_t}\right)\mathbf{W}_2^{\text{MLP}_t}\right)^T \\ \mathbf{X}_n = \text{Norm}\left(\mathbf{X}_{\text{emb}}\right)\end{cases} \tag{9}$$

where $\mathbf{X}_n \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$, $\mathbf{W}_1^{\text{MLP}_t} \in \mathbb{R}^{(L+1)\times d_{\text{MLP}_t}}$, and $\mathbf{W}_2^{\text{MLP}_t} \in \mathbb{R}^{d_{\text{MLP}_t}\times(L+1)}$; $\sigma(\cdot)$ is the activation function; d_{MLP_t} is the dimension of the inner layer, which is usually larger than the input dimension $(L+1)$.

The two-layer design facilitates interdimensional expansion. Due to the FC layer, each node in the MLP has a receptive field encompassing all tokens, enabling the MLP to effectively amalgamate information from all tokens and function as a token mixer.

c) Pooling: Pooling can be seen as a concatenation method within its reception field, where it combines neighboring tokens to form higher level representations. This helps capture local dependencies and patterns within the sequence. However, due to the nature of pooling operations, it cannot consider all the individual tokens' relationships simultaneously.

To address this limitation, we introduce an additional FC layer without an activation function before the pooling module. This layer serves as a bridge between the pooled representations and the original input tokens by connecting each token with every other token through weighted connections.

By adding this FC layer, we enable the pooling module to have access to information from all input tokens directly. The weights assigned to these connections allow for capturing global dependencies and long-range relationships among different parts of the sequence.

The pooling method utilized in this article is the average pool. The token mixer based on pooling can be expressed as

$$\begin{cases} \text{TokenMixer}_{\text{Pooling}}\left(\mathbf{X}_n\right) = Pool_t\left(\mathbf{X}_n^T\right)^T \\ Pool_t\left(\mathbf{X}\right) = \text{AvgPool}\left(\mathbf{X}\mathbf{W}^{\text{Pool}_t}\right) + \mathbf{X} \\ \mathbf{X}_n = \text{Norm}\left(\mathbf{X}_{\text{emb}}\right) \end{cases} \tag{10}$$

where $\mathbf{X}_n \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$ and $\mathbf{W}^{\text{Pool}_t} \in \mathbb{R}^{(L+1)\times(L+1)}$.

3) Feature Channel Fusion: Similar to the STF block, the FCF block also utilizes layer normalization and a residual shortcut. However, instead of a token information mixer, it employs a feature channel mixer.

The procedure for the FCF block can be expressed as follows:

$$\begin{aligned} \mathbf{X}_F &= \text{FCF}\left(\mathbf{X}_{\text{STF}}\right) \\ &= \text{ChannelMixer}\left(\text{Norm}\left(\mathbf{X}_{\text{STF}}\right)\right) + \mathbf{X}_{\text{STF}} \end{aligned} \tag{11}$$

where $\mathbf{X}_{\text{STF}} \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$ and $\mathbf{X}_F \in \mathbb{R}^{(L+1)\times d_{\text{emb}}}$.

The channel mixer is defined by a channel MLP in most transformer-like models, which consists of two FC layers and an activation function. The MLP-based block can be written as

$$\begin{aligned} &ChannelMixer_{\text{MLP}}\left(\mathbf{X}\right) = \sigma\left(\mathbf{X}\mathbf{W}_1^{\text{MLP}_c}\right)\mathbf{W}_2^{\text{MLP}_c} \\ &\mathbf{X} = \text{Norm}\left(\mathbf{X}_{\text{STF}}\right) \end{aligned} \tag{12}$$

where $\mathbf{W}_1^{\text{MLP}_c} \in \mathbb{R}^{d_{\text{emb}}\times d_{\text{MLP}_c}}$ and $\mathbf{W}_2^{\text{MLP}_c} \in \mathbb{R}^{d_{\text{MLP}_c}\times d_{\text{emb}}}$; $\sigma(\cdot)$ is the activation function and GELU [44] is applied here. d_{MLP_c} is the dimension of the inner layer, usually larger than the input.

D. Classification Mechanism

By incorporating both sequence information and channel features, the SFEN introduces a specialized token that is concatenated to the sequence embeddings, effectively representing the comprehensive information of the entire sequence as a singular entity, which serves as the class token within the ViT architecture.

The primary function of the class token is to facilitate classification tasks by aggregating information from all tokens, enabling the classifier to generate the prediction for the entire sequence. Concatenated at the beginning of image embedding tokens with equal dimensions as other tokens, it serves as part of the input to SFEN. As these input tokens traverse through SFEN's sequence encoding layers, information from image tokens is incorporated into the class token. Ultimately, this class token encapsulates comprehensive sequence information and proceeds through a classifier for final decision-making. In addition, it should be noted that positional encoding also encompasses the class token.

The classifier comprises an MLP layer followed by a softmax function, where MLP converts class tokens into predicted classes and softmax further transforms predictions into probability estimates. The procedure can be written as

$$\begin{cases} \text{out} = CLASS(\mathbf{x}_F^0) = \text{softmax}\left(MLP_{\text{class}}\left(\mathbf{x}_F^0\right)\right) \\ MLP_{\text{class}} = \sigma\left(\text{Norm}\left(\mathbf{x}_F^0\right)\mathbf{W}_1^{\text{MLP}_{\text{class}}}\right)\mathbf{W}_2^{\text{MLP}_{\text{class}}} \end{cases} \tag{13}$$

where $\mathbf{W}_1^{\text{MLP}_{\text{class}}} \in \mathbb{R}^{d_{\text{emb}}\times d_{\text{MLP}_{\text{class}}}}$, $\mathbf{W}_2^{\text{MLP}_{\text{class}}} \in \mathbb{R}^{d_{\text{MLP}_{\text{class}}}\times NC}$, $\mathbf{x}_F^0 \in \mathbb{R}^{1\times d_{\text{emb}}}$, and $\sigma(\cdot)$ is the activation function. $d_{\text{MLP}_{\text{class}}}$ is the dimension of the MLP layer, usually larger than the number of target classes *NC*.

The class token plays an important role in comprehending multiview SAR sequences through Transformer-based networks, serving as a condensed representation that summarizes the entire sequence and enables the model to grasp context and relationships between different images in the sequence rather than treating them individually, thus facilitating holistic reasoning about the sequence.

III. Recognition Problem Formation

This section first outlines the methodology for forming multiview SAR sequences. Subsequently, the recognition tasks for experiments are constructed under different operating conditions.

A. Multiview SAR Sequence Formation

These multiple-view SAR images provide more information for recognition than single-view, which can be extremely useful in various applications, such as object detection, tracking, and classification. By analyzing these images, important attributes can be extracted to represent the target.

In the practical context of multiview SAR signal acquisition, the SAR platform captures multiple images of a ground target from various depression and aspect angles, facilitating a comprehensive understanding of its features and characteristics. The geometric model of multiview SAR ATR is illustrated in Fig. 5(a), where sequential view images are acquired at a given view interval θ with $k > 1$ views to obtain diverse SAR images from different aspect angles.

The multiview SAR image sequences, generated following the procedure outlined in [17], are utilized as input for the proposed method in this article, as depicted in Fig. 5(b). For $\mathbf{X}_S = [\mathbf{x}_s^1, \mathbf{x}_s^2, \ldots, \mathbf{x}_s^L]$ as a multiview SAR sequence with class label *S* and sequence length *L*, the corresponding aspect angles

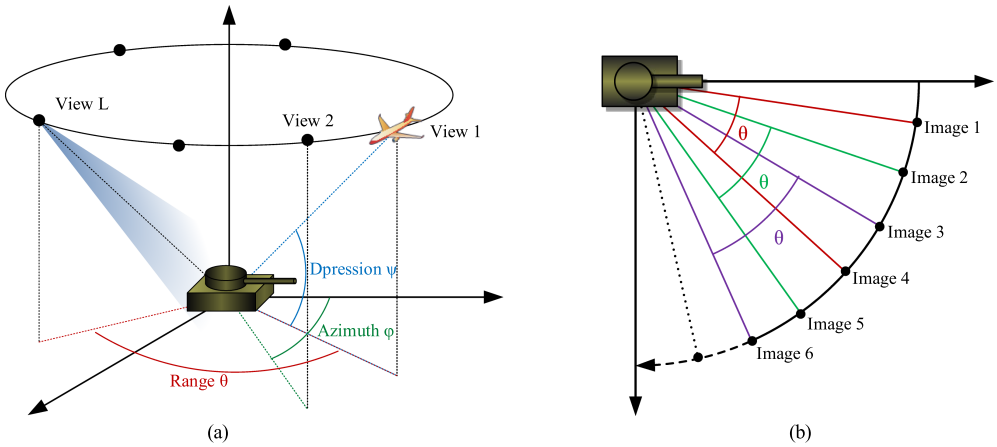

Fig. 5. Multiview SAR image sequence acquisition. (a) Schematic diagram of the multiview SAR imaging of a ground target. (b) Image sequence construction.

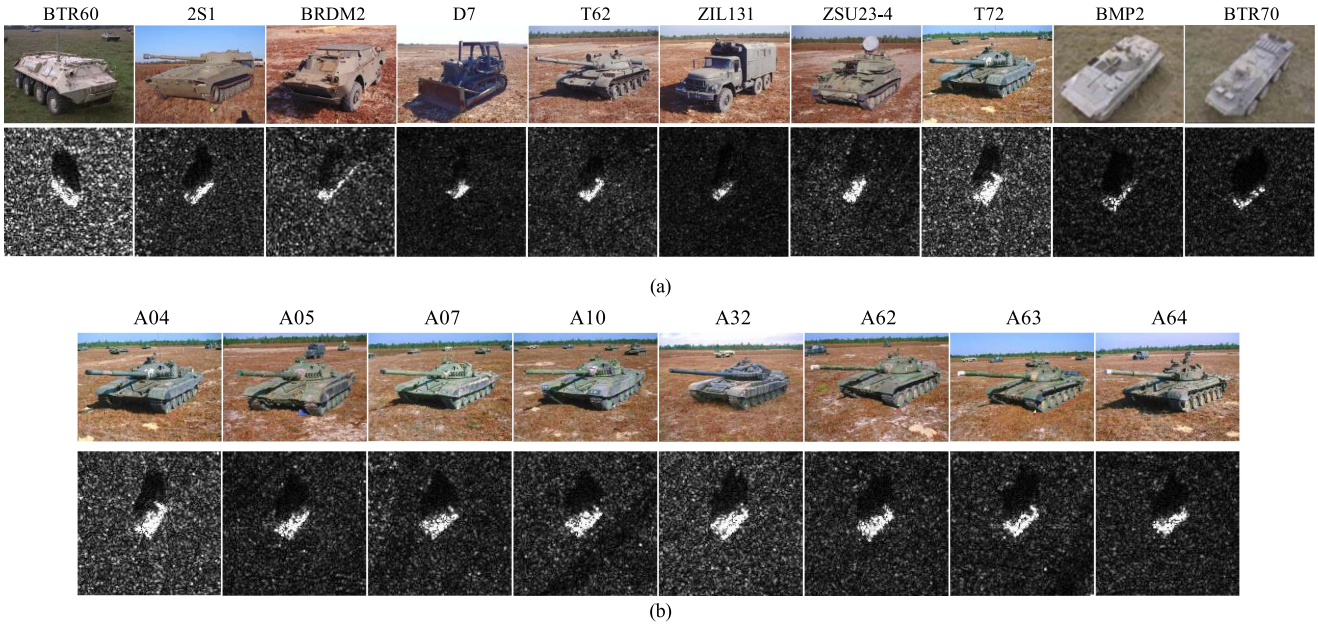

Fig. 6. Optical image of the targets in the MSTAR dataset and corresponding SAR image. (a) Ten targets. (b) T72 variances.

are $\varphi(\mathbf{x}_s^i)$, which should follow the rules in the following:

$$\begin{cases} \varphi\left(\mathbf{x}_s^1\right) < \left(\mathbf{x}_s^2\right) < \cdots < \left(\mathbf{x}_s^L\right) \quad \text{or} \quad \varphi\left(\mathbf{x}_s^1\right) > \left(\mathbf{x}_s^2\right) \\ \quad > \cdots > \left(\mathbf{x}_s^L\right) \\ \left|\varphi\left(\mathbf{x}_s^1\right) - \varphi\left(\mathbf{x}_s^L\right)\right| \leq \theta. \end{cases} \tag{14}$$

The raw SAR images are organized into sequence samples for the purpose of training and evaluating the proposed Res-Xformer in this study. These multiview SAR images contain richer information for recognition compared to single-view ones, so important attributes can be extracted to represent the ground target, which enables better discrimination between similar objects.

B. Dataset

Provided by Sandia National Laboratory, the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [7] is utilized for the purpose of conducting recognition experiments in this study. Many studies have been conducted based on this particular dataset. As such, it has become a benchmark for evaluating the performance of different ATR algorithms and techniques.

The images in this dataset have a resolution of 0.3 m $\times$ 0.3 m, captured within the X-band frequency range and employing HH polarization mode. As depicted in Fig. 6, this dataset comprises image data related to ten types of ground vehicles at different depression angles and azimuth angles. It includes vehicle target images captured at each azimuth angle ranging from 0° to 360°, making it suitable for establishing a multiview image target recognition dataset.

This dataset also provides diverse vehicle types and varying operating conditions. The composition of raw data in the MSTAR dataset is presented in Tables II and III, providing

TABLE II
MAIN CONTENT OF THE MSTAR DATASET

Target Type	Serial No.	Depression			
		15°	17°	30°	45°
BTR60		195	256		
2S1		274	299	288	303
BRDM2		274	298	287	303
D7		274	299		
T62		273	299		
ZIL131		274	299		
ZSU23/4		274	299	288	303
BTR70		196	233		
	9563	195	233		
BMP2	9566	196	232		
	C21	196	233		
	132	196	232		
T72	812	195	231		
	S7	191	228		

TABLE III
VARIANTS OF T72 IN THE MSTAR DATASET

Serial No.	Depression	
	15°	17°
A04	274	299
A05	274	299
A07	274	299
A10	271	296
A32	274	298
A62	274	299
A63	274	299
A64	274	299

detailed information on the quantity of raw SAR images for each target categorized by different depression angles and type serials.

By establishing diverse experimental datasets under varying operating conditions and sample sizes, we can investigate the influence of factors such as depression angles, target serials, and training set size on the proposed method.

C. Classification Mission Construct

The MSTAR dataset consists of two categories for acquisition conditions: SOCs and extended operating conditions (EOC). SOC includes images from the training set and testing set that have the same target type and similar imaging configuration. On the other hand, EOC data in the training set and testing set display greater dissimilarity and pose increased challenges for identification. EOC can further be divided into configuration-variant (EOC-C), depression-variant (EOC-D), as well as version-variant EOC (EOC-V), which can offer different challenges for testing image recognition algorithms.

The types of targets and their quantities in the dataset under SOC condition are presented in Table IV. Samples for training and testing within the same category share a common model with slight variations in depression angles, which represents the most comparable working conditions between training and test data.

In comparison to SOC, EOC presents a more realistic testing environment due to the larger data difference between the training set and test set, not only in depression angle but also in image configuration and target type. As such, EOC more closely approximates real-world testing conditions. In EOC-C, the primary difference between the training and test sets lies in the configuration type of images. As given in Table V, the main training set comprises BTR70, BRDM2, BMP2, and T72 targets, while the test set includes two different configurations of BMP2 and T72 targets.

Meanwhile, EOC-V aims to assess the performance of the method for different versions of a singular target category, with its primary training dataset being identical to that of EOC-C, while its test dataset comprises distinct versions of the T72. Table V presents the distributions of the training and test sets under EOC-V conditions.

Unlike other operating conditions, EOC-D is associated with larger variations in depression angle. There are four types of targets in EOC-D, as given in Table VI. The training set has a depression angle of 17°, while the test set has a depression angle of 30°.

IV. EXPERIMENTS AND RESULTS

To validate the efficacy of the proposed approach, we first specify the network architecture setup. Subsequently, the performance of the proposed method is evaluated through experiments conducted under diverse conditions. Finally, we compare the performance of the proposed method with other methods.

A. Experimental Setup

1) Environment: All the experiments are conducted on a laptop with an 11th Gen Intel(R) Core (TM) i9-11950H at 2.60 GHz, and a NVIDIA RTX A4000 Laptop GPU.

The proposed method in this study is implemented using the deep learning framework PyTorch (version 1.12.1), along with the Torchvision library (version 0.13.1), under Python 3.10. The attention mechanism in SFEN is implemented with the help of the vision_transformer.Encoder module in the torchvision.models. The Xformer with alternative token mixer MLP and Pooling is implemented with the help of layers in timm 0.4 library. ResNet18 is implemented with the help of torchvision.models.ResNet. To accommodate SAR grayscale images, modifications are made to ResNet18 by rewriting its first layer Conv. Since ResNet serves solely for feature extraction purposes, the subsequent dimension transformation and recognition components are excluded.

2) Dataset Setup: We generate experimental sequences in accordance with the methodology outlined in Section III-A,

TABLE IV
DATASET FOR EXPERIMENT UNDER SOC

Target Type	Serial No.	No. of Samples							
		Train(17°)				Test(15°)			
		Image	2-View	3-View	4-View	Image	2-View	3-View	4-View
BTR60		256	509	760	1009	195	387	577	765
2S1		299	595	889	1181	274	545	814	1081
BRDM2		298	593	886	1177	274	545	814	1081
D7		299	595	889	1181	274	545	814	1081
T62		299	595	889	1181	273	543	811	1077
ZIL131		299	595	889	1181	274	545	814	1081
ZSU23/4		299	595	889	1181	274	545	814	1081
BTR70		233	463	691	917	196	389	580	769
BMP2	9563	233	463	691	917	195	387	577	765
T72	132	232	461	688	913	196	389	580	769

TABLE V
DATASET FOR EXPERIMENT UNDER EOC-C AND EOC-V

Operating Condition	Depression	Target Type	Serial No.	No. of Samples			
				Image	2-View	3-View	4-View
EOC_C & EOC_V	Train (17°)	BTR70		256	463	691	917
		BRDM2		299	593	886	1177
		BMP2	9563	233	463	691	917
		T72	132	232	461	1241	913
EOC_C	Test (15°,17°)	BMP2	9566	428	850	1268	1682
			C21	429	852	1271	1686
			812	426	846	1262	1674
		T72	A04	573	1140	1703	2262
			A05	573	1140	1703	2262
			A07	573	1140	1703	2262
			A10	567	1128	1685	2238
EOC_V	Test (15°,17°)	T72	S7	419	832	1241	1646
			A32	572	1138	1700	2258
			A62	573	1140	1703	2262
			A63	573	1140	1703	2262
			A64	573	1140	1703	2262

TABLE VI
DATASET FOR EXPERIMENT UNDER EOC-D

Depression	Target Type	Serial No.	No. of Samples			
			Image	2-View	3-View	4-View
Train (17°)	2S1		299	595	889	1181
	BRDM2		298	593	889	1177
	ZSU23/4		299	593	886	1181
	T72	132	232	461	688	913
Test (30°)	2S1		288	573	856	1137
	BRDM2		287	571	856	1137
	ZSU23/4		288	573	856	1137
	T72	A64	288	573	853	1133

with a sequence length of $L = 4$. Within each sequence, the maximum interval between images' view angles is restricted to 45°. Moreover, the images within the same sequence are arranged in ascending or descending order based on their view angles to demonstrate the evolving characteristics with varying azimuth angles in the multiview sequence. The number of sequence samples under different working conditions and sequence lengths is given in Tables IV–VI.

3) Network Setup: The input embedding dimension $d_{\text{embedding}}$ is set as either 512 or 1024, while the middle layer dimensions of feature channel mixer MLP d_{MLP_c} is defined as 1024 and $d_{\text{MLP}_t} = d_{\text{Pool}} = 10$ in token mixer. GELU [44] is applied here as an activation function.

The batch size is set to 32, and the probability of dropout is 0.2. Adam optimizer is adopted for training the network along with cross-entropy loss function. The learning rate is set to 0.00005, and the amount of training epochs is 30.

In addition to evaluating the performance of different token mixers in SFEN, our experiments also investigate the impact of varying the number of stacked layers on SFEN's performance. Specifically, we examine how setting the number of experimental layers to 1, 3, 6, 8, and 16 affects performance. Furthermore, we explore the influence of different numbers of heads on multihead attention's effectiveness. For a single layer, we configure it with four heads; for the three-layer network, it is equipped with either four or eight heads; and for the eight-layer network, it is configured with either 8 or 16 heads.

B. Overall Performance

The performances of the proposed Res-Xformer with different token mixers under different operating conditions are given in Table VII. The results of the normal-scale model generally demonstrate a slight advantage over those of the larger-scale model.

In terms of the number of layers in SFEN, an increase in the number of layers initially leads to an improvement in the overall recognition rate, followed by a decline. The optimal performance is observed at the third and fourth layers, indicating that insufficient layers impede feature extraction while excessive layers result in overfitting. Conversely, for large-scale models, superior results are achieved with one or two layers, highlighting the occurrence of accelerated overfitting when model size is increased.

Like other methods, Res-Xformer performs well under SOC and the accuracy rates are over 99% while obtaining near-perfect accuracy for each mixer type under EOC-V, demonstrating that the proposed method can effectively adapt version-variant of targets. The accuracies are slightly lower under EOC-C but still achieve an accuracy rate exceeding 98%. On the other hand, it performs the poorest under EOC-D, indicating that variants of depression angles pose a greater challenge for the proposed method.

While the performances of the proposed Res-Xformer with various types of parameters under EOC-D are comparatively worse than those observed in other operation conditions, it is important to note that achieving 97% accuracy still demonstrates a reasonably high level of performance.

C. Results and Discussions Under SOC

As given in Table VII, there is not much fluctuation caused by changes in model parameters, with a recognition rate maintained at 99%. Among different token mixers, pooling shows the best performance, while attention performs the worst. In terms of model size, larger models have worse performance, and stacking more layers leads to poorer results. This suggests that recognition tasks under SOC exhibit a relatively low level of complexity, thereby causing complex models to suffer from overfitting, resulting in a decline in recognition accuracy.

When considering the layer count of SFEN, the proposed method with MLP as a token mixer achieves optimal results with one–three layers, and the most favorable outcome is observed with a single layer. Due to its utilization of FC layers for information aggregation, MLP exhibits relatively high computational complexity. Consequently, the highest recognition rate is attained using a one-layer encoder. Simultaneously, the performance of large-scale models starts to significantly deteriorate beyond three layers. The performance of pooling is slightly degraded by a single layer, but as the number of stacked layers increases, the performance initially improves and then deteriorates. The optimal performance is achieved when four layers are stacked. This is because pooling is a relatively simple calculation, so when the number of layers in the encoder is small, the network fails to effectively capture the characteristics of the data. Therefore, it is necessary to stack more layers to increase the complexity of the feature extraction network. Attention exhibits a similar tendency, yielding optimal outcomes when employing a three-layer encoder and four heads of MSA. Overall, the impact of the number of heads in MSA is relatively modest; however, an excessive number of heads can lead to a decline in recognition accuracy. Utilizing either four or eight heads yields superior performance.

When examining the model scale and overall size effect, it is often found that using one or two layers can yield optimal results due to the higher complexity of the entire model, allowing for a better representation of target characteristics and mitigating overfitting issues.

In order to compare the effects of different token mixers, the confusion matrix is calculated using a three- or four-layer SFEN that is generally considered optimal. Table VIII displays the confusion matrices of the Res-Xformer model with MSA, MLP, and pooling as token mixers, respectively. The rows represent the true category of the target, while the columns indicate the predicted label generated by the Res-Xformer.

When four-head MSA is used as a token mixer, the Res-Xformer demonstrates good recognition performance for BTR60, BRDM2, ZSU23/4, BTR70, and T72 without any false identification. However, a small number of 2S1 instances are mistakenly identified as T62 and BTR70, with the maximum misidentification being 2S1 as T72. D7 and T62 exhibited occasional misidentifications as ZIL131; moreover, they are more frequently mistaken for ZSU23/4. In addition, there are sporadic

TABLE VII
PERFORMANCES FOR EXPERIMENTS UNDER DIFFERENT WORKING CONDITIONS

Model Scale	Token Mixer	Layers	Heads	Recognition Rate (%)			
				SOC	EOC-V	EOC-C	EOC-D
Normal	MSA	1	4	99.5	100.0	97.1	96.2
			8	99.7	100.0	97.3	96.4
		3	4	99.8	100.0	97.5	95.5
			8	99.6	100.0	98.5	96.6
			16	99.1	99.9	97.6	96.6
		8	8	98.7	99.8	98.4	97.2
			16	99.0	99.8	98.2	97.1
	MLP	1	-	99.8	100.0	99.1	97.7
		2	-	99.7	100.0	98.3	96.3
		3	-	99.8	100.0	98.1	97.2
		4	-	99.2	100.0	97.8	95.1
		6	-	99.6	99.8	98.1	97.8
		8	-	99.6	100.0	98.2	95.9
		16	-	99.3	99.9	98.1	95.4
	Pooling	1	-	99.3	100.0	98.5	94.8
		2	-	99.9	100.0	98.4	94.9
		3	-	99.9	100.0	98.9	95.2
		4	-	99.9	100.0	98.8	91.8
		6	-	99.4	100.0	98.1	96.1
		8	-	99.6	99.9	98.6	96.9
		16	-	98.9	100.0	97.0	96.0
Big	MSA	1	4	98.9	100.0	96.7	96.3
			8	99.7	100.0	96.7	95.7
		3	4	98.1	100.0	97.0	96.6
			8	98.9	99.8	98.3	95.6
			16	99.0	100.0	98.5	95.5
		8	8	97.7	99.9	97.7	95.3
			16	97.0	99.8	97.9	95.7
	MLP	1	-	99.6	100.0	98.0	97.0
		2	-	99.8	99.8	97.8	95.5
		3	-	98.7	100.0	97.7	97.8
		4	-	99.2	100.0	98.0	96.4
		6	-	99.1	100.0	97.9	96.6
		8	-	99.4	99.9	98.0	96.8
		16	-	99.1	99.8	98.4	96.1
	Pooling	1	-	99.6	99.9	97.1	94.0
		2	-	99.9	100.0	98.6	95.1
		3	-	99.0	100.0	99.0	96.4
		4	-	99.1	100.0	98.0	95.7
		6	-	99.1	100.0	97.6	95.2
		8	-	99.3	99.9	98.3	97.2
		16	-	98.7	99.9	97.7	95.9

cases where ZIL131 is incorrectly recognized as ZSU23/4. The recognition accuracy is 99.76%.

When MLP is a token mixer, the proposed method achieves an overall recognition rate of 99.82%. However, the misidentified categories are relatively scattered. BTR60, BRDM2, D7, BMP2, and T72 are mistakenly identified in small quantities as BRDM2, ZIL131, ZSU23/4, BTR60, and T62 respectively. 2S1 is occasionally misidentified as BTR60 and T62. ZIL131 is occasionally misidentified as T62 and ZSU23/4. BTR70 is occasionally misidentified as 2S1 and BMP2.When pooling is used as a token mixer, Res-Xformer achieves an overall recognition rate of 99.92%, surpassing all other token mixers in performance. The experimental results demonstrate that the Res-Xformer incorporating a superposition of three layers of SFEN closely approximates the recognition rate achieved by four-layer SFEN. The types of mistakenly identified samples are more concentrated in the four-layer condition, with no false identification observed for ZIL131, ZSU23/4, BMP2, BTR70, and T72. In the condition of three-layer SFEN, S1, BRDM2, D7, ZIL131, and T7 are, respectively, misidentified as T62, ZIL131, ZSU23/4, ZSU23/4, T62, BTR70, and a few BTR70 samples are misidentified as BTR60 and BRDM2. The general recognition

TABLE VIII
CONFUSION MATRICES OF RES-XFORMER WITH DIFFERENT TOKEN MIXERS UNDER SOC (RECOGNITION RATE: ATTENTION 99.76%, MLP 99.82%, POOLING 99.92%)

Token Mixer (No. of Layers)	Target Type	Recognition Rate (%)										
		BTR60	2S1	BRDM2	D7	T62	ZIL131	ZSU23/4	BMP2	BTR70	T72	ALL
MSA (3-4)	BTR60	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.76
	2S1	0.00	99.54	0.00	0.00	0.09	0.00	0.00	0.00	0.09	0.28	
	BRDM2	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	D7	0.00	0.00	0.00	99.63	0.00	0.09	0.28	0.00	0.00	0.00	
	T62	0.00	0.00	0.00	0.00	99.26	0.19	0.56	0.00	0.00	0.00	
	ZIL131	0.00	0.00	0.00	0.00	0.00	99.81	0.19	0.00	0.00	0.00	
	ZSU23/4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	
	BMP2	0.26	0.26	0.00	0.00	0.00	0.00	0.00	99.48	0.00	0.00	
	BTR70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	
	T72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	
MLP (3)	BTR60	99.87	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.82
	2S1	0.09	99.72	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	
	BRDM2	0.00	0.00	99.81	0.00	0.00	0.19	0.00	0.00	0.00	0.00	
	D7	0.00	0.00	0.00	99.81	0.00	0.00	0.19	0.00	0.00	0.00	
	T62	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	
	ZIL131	0.00	0.00	0.00	0.00	0.09	99.72	0.19	0.00	0.00	0.00	
	ZSU23/4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	
	BMP2	0.13	0.00	0.00	0.00	0.00	0.00	0.00	99.87	0.00	0.00	
	BTR70	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.39	99.48	0.00	
	T72	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	99.87	
Pool (3)	BTR60	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.88
	2S1	0.00	99.91	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	
	BRDM2	0.00	0.00	99.91	0.00	0.00	0.09	0.00	0.00	0.00	0.00	
	D7	0.00	0.00	0.00	99.81	0.00	0.00	0.19	0.00	0.00	0.00	
	T62	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	
	ZIL131	0.00	0.00	0.00	0.00	0.00	99.63	0.37	0.00	0.00	0.00	
	ZSU23/4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	
	BMP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	
	BTR70	0.13	0.00	0.13	0.00	0.00	0.00	0.00	0.00	99.74	0.00	
	T72	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	99.87	
Pool (4)	BTR60	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.92
	2S1	0.00	99.72	0.09	0.00	0.19	0.00	0.00	0.00	0.00	0.00	
	BRDM2	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	D7	0.00	0.00	0.00	99.63	0.00	0.09	0.28	0.00	0.00	0.00	
	T62	0.00	0.09	0.00	0.00	99.91	0.00	0.00	0.00	0.00	0.00	
	ZIL131	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	
	ZSU23/4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	
	BMP2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	
	BTR70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	
	T72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	

rate reaches 99.88%. Regarding four-layer SFEN, only three types of samples are misidentified: 2S1 as BRDM2 and T62; D7 as ZIL131 and ZSU23/4; and T62 as 2S1. This results in an overall recognition rate of 99.92%.

These results demonstrate that the three types of token mixers yield comparable outcomes. When attention is employed as the token mixer, increasing the number of layers leads to improved performance. Conversely, pooling as a token mixer produces comparable outcomes across different layer numbers. In general, MLP and MSA exhibit similar recognition results, while pooling performs slightly better and has the lowest computational complexity. Although MLP's computation is influenced by the dimension of its middle layer, it remains relatively higher compared to MSA.

D. Results and Discussions Under EOC

1) EOC-C: As given in Table VII, the recognition rate under EOC-C is generally a bit lower than SOC. The gap between large-scale models and regular ones in terms of recognition rate is minor, but large-scale models with different numbers of layers slightly outperform regular ones when pooling is used as a token mixer.

Compared to under SOC, where the Res-Xformer with pooling as the token mixer achieves the best performance, models with pooling as the token mixer slightly lag behind those with MLP under EOC-C, and larger-scale models perform even better than regular ones. This fully demonstrates that EOC-C is more complex than SOC and requires more sophisticated components or larger structures of Res-Xformer to achieve better results.

TABLE IX
CONFUSION MATRIX OF RES-XFORMER WITH DIFFERENT TOKEN MIXERS UNDER EOC-C (RECOGNITION RATE: ATTENTION 98.51%, MLP 99.15%, POOLING 98.89%)

Token Mixer (No. of Layers)	**Target Type**	Serial No.	**Recognition Rate (%)**				
			BTR70	**BRDM2**	**BMP2** 9563	**T72** 132	**ALL**
MSA (3-4)	BMP2	C21	0.00	0.00	87.90	12.10	97.52
		9566	0.00	0.00	94.89	5.11	
	T72	812	0.00	0.00	0.18	99.82	
		A04	0.00	0.00	1.81	98.19	
		A05	0.00	0.00	0.62	99.38	
		A07	0.00	0.00	0.04	99.96	
		A10	0.00	0.00	0.00	100.00	
MSA (3-8)	BMP2	C21	0.71	0.77	97.57	0.95	98.51
		9566	2.32	0.42	94.23	3.03	
	T72	812	0.30	0.00	0.12	99.58	
		A04	0.00	0.71	1.41	97.88	
		A05	0.00	0.00	0.71	99.29	
		A07	0.00	0.00	0.00	100.00	
		A10	0.00	0.00	0.00	100.00	
MLP (1)	BMP2	C21	0.12	1.90	97.63	0.36	99.15
		9566	1.43	0.00	96.55	2.02	
	T72	812	0.18	0.00	0.66	99.16	
		A04	0.00	0.00	0.35	99.65	
		A05	0.00	0.00	0.00	100.00	
		A07	0.00	0.00	0.00	100.00	
		A10	0.00	0.00	0.00	100.00	
MLP (3)	BMP2	C21	0.12	0.24	97.63	2.02	98.14
		9566	0.00	0.65	97.44	1.90	
	T72	812	0.00	0.00	0.90	99.10	
		A04	0.00	0.00	4.64	95.36	
		A05	0.00	0.00	0.44	99.56	
		A07	0.00	0.00	2.08	97.92	
		A10	0.00	0.00	0.04	99.96	
Pooling (3)	BMP2	C21	0.00	1.96	96.50	1.54	98.89
		9566	0.06	0.65	96.37	2.91	
	T72	812	0.00	0.00	0.00	100.00	
		A04	0.00	0.18	1.41	98.41	
		A05	0.00	0.00	0.00	100.00	
		A07	0.00	0.00	0.00	100.00	
		A10	0.00	0.00	0.00	100.00	

The confusion matrices of the proposed method with three-layer SFEN under EOC-C are given in Table IX. The rows represent the true category of the target, while the columns indicate the predicted label generated by the trained Res-Xformer using the four-class dataset. Since the one-layer Res-Xformer with MLP as a token mixer achieves the best performance, we have also included the confusion matrix of this model in Table IX.

When MSA is used as a token mixer, the MSA of eight heads outperforms that of four heads. For comparison with the SOC mode, we also give the confusion matrix of MSA of four heads.

When taking four-head MSA as a token mixer, the Res-Xformer has the lowest recognition rate, but its misidentifications are relatively concentrated and only occur between BMP2 and T72, indicating that the model struggles to distinguish BMP2 from T72 very well with these parameters. Comparatively, SN_9566 in BMP2 has a higher recognition rate than SN_C21. The overall accuracy of T72 recognition is relatively high, with A04 being mistakenly identified as BMP2 most frequently. When using eight-head MSA, the recognition rate reaches 98.51% with the improved classification of SN_C21 in BMP2, thus classified better. However, BMP2 and T72 can be misidentified as BTR70 and BRDM2, and SN_9566 in BMP2 is most mistaken for T72 and BTR70. The A04 still remains the worst identified among T72.

When taking MLP as a token mixer with one-layer SFEN, the recognition situation of Res-Xformer is similar to that of

eight-head MSA but with a higher accuracy of 99.15%. The A05, A07, and A10 are all correctly recognized, and the SN_9566 is identified better than when using MSA. When the number of layers in SFEN increases to three, the recognition rate, on the contrary, decreases to 98.15%. The different versions of BMP2 are more accurately recognized, whereas the different versions of T72 are less effectively recognized with misrecognition toward BMP2. This indicates that the recognition complexity of different versions of BMP2 is higher, and more complex models exhibit superior recognition capabilities. However, for T72, an excessively complex model leads to overfitting and results in a decline in recognition accuracy.

When pooling is used as a token mixer in Res-Xformer, the performance is a little better than MSA, and slightly worse than MLP, reaching 98.89%. The model with these parameters recognizes T72 well, only A04 exists misrecognition. However, in the case of BMP2, the misrecognition occurs for all four target types. Given that the complexity of pooling operations is lower than that of MLP, this also suggests that recognizing the different versions of T72 is easier than BMP2. It is also worth noticing that the recognition rate of large-scale models is slightly higher than that of ordinary models under this condition, reaching 99%, which indicates that increasing the complexity of Res-Xformer slightly helps with the recognition under EOC-C.

2) EOC-V: When it comes to EOC-V, as given in Table VII, Res-Xformers with all three token mixers achieve excellent results, with a recognition rate reaching 100%. Therefore, we do not provide confusion matrices here. Only a few models exhibit minimal misidentifications, and their recognition performance even surpassed that under the SOC condition, thereby demonstrating that the proposed method is suitable for recognition tasks under EOC-V.

3) EOC-D: The results under EOC-D of different types of the proposed Res-Xformer with different numbers of layers are given in Table VII. Compared to other operation conditions, the performances of the proposed Res-Xformer with various types of parameters under EOC-D are the worst, achieving only about 97% accuracy. This indicates a significant discrepancy between the training set and test set, making it more challenging for the proposed method.

When considering the layer count of SFEN, the proposed method with MLP as a token mixer achieves optimal results with one–three layers, but best with six layers. As the number of layers increases, there is generally a decline in performance. But when MSA and pooling are taken as token mixers, they exhibit opposite patterns and perform best when the layer count is eight. The complexity of MLP is higher than MSA and pooling, which leads to overfitting and affects the performance of Res-Xformer when the number of layers is too large. As task under EOC-D is more complex than SOC and EOC-C, the Res-Xformer needs to increase its complexity when using MSA and pooling as a token mixer.

To compare the effects of different token mixers and maintain consistency with other conditions, we calculated the confusion matrix of the proposed method using three-layer SFEN under EOC-D. The results are given in Table X. The rows represent the true category of the target, while the columns indicate the predicted label generated by the Res-Xformer. Since the one-layer Res-Xformer with MLP as token a mixer and eight-layer Res-Xformer with MSA and pooling achieve the best performances, we have also included their confusion matrices in Table X.

When using MSA as a token mixer, a three-layer Res-Xformer has similar discriminative performance on the targets with an eight-layer one but the accuracy is lower. They can accurately identify BRDM2 and ZSU23/4, but their identification accuracy for 2S1 is poor. They often mistake 2S1 for T72 and T72 for ZSU23/4. However, the recognition performance of the eight-layer model for T72 is slightly better, resulting in an overall recognition rate of 97.21%, which is a little better than Pooling, and slightly worse than MLP.

When MLP is used as a token mixer, the recognition performance of the three-layer Res-Xformer is similar to that of the one-layer one, but with a slightly higher rate of misidentification. Compared to using MSA, the types of misidentifications are a little more diverse. The one-layer model with MLP has a higher accuracy in recognizing 2S1 than when with MSA, resulting in the highest recognition rate achieved at 97.69%.

When pooling is used as a token mixer in Res-Xformer, the eight-layer model achieves a recognition rate of 98.89%, which is the lowest among the three kinds of token mixers. The three-layer model performs well in identifying 2S1 and ZSU23/4 but poorly for the other two targets. As the eight-layer model improves accuracy in recognizing BRDM2 and T72, the recognition rate for the other two targets decreases slightly, resulting in a final recognition rate of 96.92%. This indicates that due to the lower complexity of Pool, it cannot effectively differentiate all categories in EOC-D problems.

E. Comparison With Other Methods

The performance of the proposed Res-Xformer is compared with eight other methods, including iterative graph thickening (IGT) [16], conditional Gaussian model (CGM) [15], sparse representation-based classification (SRC) [17], multiview deep convolutional neural network (MVDCNN) [20], bidirectional convolutional-recurrent network (BCRN) [24], multiview deep feature learning network (MVDFLN) [25], deep feature extraction and fusion network (FEF-Net) [22], and convolutional-Transformer network (CTN) [42], which are classical or recently published in SAR recognition field. IGT and CGM are methods for single SAR images, so the performances cannot surpass sequence ones. The BCRNs with sequence lengths of 5 and 15 are denoted as BCRN-S5 and BCRN-S15, respectively. CTN-L6 denotes CTN using a six-layer-Transformer encoder. Our Res-Xformers with MSA, MLP, and pooling as token mixers are denoted as Res-MSAformer, Res-MLPformer, and Res-Poolformer, respectively. The Res-Xformers with one-layer and three-layer SFEN are denoted as Res-Xformer-1 and Res-Xformer-3, respectively.

As given in Table XI, the proposed Res-Xformer achieves better or comparable recognition rates under different working conditions compared to other methods, only lagging behind BCRN-S15. However, the performance of BCRN-S15 should

TABLE X
CONFUSION MATRIX OF RES-XFORMER WITH DIFFERENT TOKEN MIXERS UNDER EOC-D (RECOGNITION RATE: ATTENTION 97.21%, MLP 97.69%, POOLING 96.92%)

Token Mixer (No. of Layers)	Target Type	Recognition Rate (%)				
		2S1	BRDM2	ZSU23/4	T72-132	ALL
MSA (3-8)	2S1	92.26	1.85	0.18	5.72	96.57
	BRDM2	0.09	99.82	0.00	0.09	
	ZSU23/4	0.00	0.00	99.82	0.18	
	T72-A64	1.14	0.00	4.49	94.37	
MSA (8-16)	2S1	92.08	0.35	0.00	7.56	97.21
	BRDM2	0.09	98.50	0.00	1.15	
	ZSU23/4	0.00	0.00	99.65	0.35	
	T72-A64	0.18	0.09	1.14	98.59	
MLP (3)	2S1	94.99	0.00	0.09	4.93	97.16
	BRDM2	3.97	95.85	0.00	0.18	
	ZSU23/4	0.09	0.09	99.21	0.62	
	T72-A64	0.18	0.00	1.23	98.59	
MLP (1)	2S1	97.63	0.97	0.00	1.41	97.69
	BRDM2	0.62	98.94	0.09	0.35	
	ZSU23/4	0.97	0.00	98.77	0.26	
	T72-A64	2.55	0.00	2.02	95.43	
Pool (3)	2S1	99.56	0.00	0.00	0.44	95.62
	BRDM2	6.35	92.85	0.44	0.35	
	ZSU23/4	0.70	0.00	98.86	0.44	
	T72-A64	5.28	0.00	3.52	91.20	
Pool (8)	2S1	95.69	1.14	0.00	3.17	96.92
	BRDM2	0.26	99.29	0.26	0.18	
	ZSU23/4	0.00	0.00	93.58	6.42	
	T72-A64	0.26	0.00	0.62	99.12	

TABLE XI
PERFORMANCE COMPARISON WITH OTHER METHODS

Method	Recognition Rate (%)			
	SOC	EOC-C	EOC-V	EOC-D
IGT	95	-	85	80
CGM	97.18	81.22	79.30	-
SRC	98.94	96.78	-	-
MVDCNN	98.52	95.45	95.46	94.61
BCRN-S5	97.82	90.84	95.77	93.3
BCRN-S15	99.97	99.81	98.93	99.54
MVDFLN	99.62	97.84	99.10	97.12
FEF-Net	99.34	-	-	-
CTN-L6	99.9	98.5	99.78	-
Res- MSAformer-3	99.76	98.51	100.00	96.57
Res-MLPformer-1	99.82	99.15	100.00	97.16
Res-MLPformer-3	99.80	98.14	100.00	97.69
Res-Poolformer-3	99.91	98.89	100.00	95.62
Res-Poolformer-8	99.65	98.63	99.91	96.92

not be directly compared due to its significantly longer sequence length than other methods. Under SOC, Res-Poolformer-3 outperforms other methods, while Res-MSAformer and Res-MLPformer only slightly trail behind CTN-L6. Under EOC-D, although our Res-Xformers do not perform well and do not exhibit exceptional performance compared to other methods, we still achieve comparable results. Notably, Res-MLPformers slightly outperform other methods, while Res-Poolformer-8 and Res-MSAformer-3 have slightly lower performance than MVDFLN. Under EOC-C and EOC-V, all three types of Res-Xformers demonstrate superiority over other methods, thereby demonstrating the superiority of our proposed method and validating the effectiveness of our proposed structure for multiview SAR sequence recognition.

TABLE XII
PERFORMANCE COMPARISON WITH DIFFERENT NUMBER OF VIEWS OF SAR AS INPUT

Token Mixer	No. of Views	No. of Layers	Recognition Rate (%)			
			SOC	EOC-V	EOC-C	EOC-D
Attention	2	3-4	98.5	99.8	97.8	92.5
		3-8	99.0	99.7	97.2	96.1
	3	3-4	98.9	99.8	97.0	94.5
		3-8	98.2	99.9	95.9	94.0
MLP	2	1	99.3	99.9	97.7	91.5
		3	98.7	99.6	96.4	94.0
	3	1	99.7	100.0	98.2	96.9
		3	99.2	99.8	98.0	96.7
Pooling	2	1	99.0	99.6	95.9	90.6
		3	98.7	99.9	97.1	93.1
	3	1	98.6	99.7	98.4	95.5
		3	99.0	99.9	98.8	92.1

F. Comparison With Different Numbers of Views

The larger the number of views in the SAR image sequence, the more information the Res-Xformer can obtain, which is expected to result in improved recognition outcomes. In order to evaluate the effectiveness of our approach, we conducted experiments under all four operating conditions, specifically investigating how varying the number of views impacts the performance. The results are given in Table XII.

The results demonstrate that the overall recognition performance slightly decreases when using 2-view or 3-view SAR sequences as inputs compared to the 4-view case. This suggests that having access to more views enhances the model's ability to accurately classify objects. In addition, more complex tasks, such as those under EOC-D, are influenced to a greater extent by the number of views. This can be easily understood because when there is less input information and taking into account the inherent difficulty of recognition problems, missing information has a larger impact leading to decreased recognition performance. Moreover, this also proves that our proposed method can indeed effectively integrate sequence information and reflect the changes in the input sequence information.

G. Discussion

Previous studies have primarily focused on feature extraction based on CNN variants while lacking in-depth research on the fusion of features from multiview SAR sequences. From the various types of experiment results mentioned above, our designed Res-Xformer effectively represents the information of multiview SAR and achieves a state-of-the-art recognition rate. This is because, on the one hand, our method is based on ResNet for extracting features of a single image, which lays a solid foundation for sequence feature extraction in the next step. On the other hand, The SFEN module, consisting of a token mixer and a channel mixer, is designed to integrate features of tokens within the sequence, allowing the class token to represent the entire sequence. As a result, the proposed methods have achieved favorable outcomes.

Unlike previous studies that focus on using attention mechanisms to fuse features, our emphasis lies in the combination of token mix and channel mix as the key to sequence fusion. The introduction of our replaceable token mixer further verifies that although MSA has an effect, it is not the main focus of sequence feature extraction and fusion.

Our proposed model fits well under SOC and EOC-V conditions, achieving a recognition rate close to 100%. Despite experiencing a slight decrease in recognition accuracy under EOC-C, it still surpasses 98.5%, establishing itself as the state-of-the-art method for SAR target recognition performance. However, our method falls slightly short when dealing with tasks under EOC-D.

The performance of the proposed Res-Xformer under EOC-D, although being the worst among all the operating conditions, still manages to achieve a recognition rate that is comparable to other approaches. While it may be considered relatively low compared to other operation conditions where higher accuracies have been achieved by the proposed Res-Xformer, this does not necessarily imply poor overall performance.

For the three types of token mixers, pooling, MSA, and MLP, their complexity increases in turn. At the same time, increasing the number of layers in the SFEN also enhances its complexity. However, for simple tasks, overly complex networks can lead to overfitting and a decrease in recognition performance. For complex tasks, on the other hand, overly simple networks cannot effectively model the characteristics of each target category and fail to accurately represent them.

Therefore, for tasks under SOC, EOC-C, and EOC-V conditions, one-layer Res-MLPformer, three-layer Res-MSAformer, and three-layer Res-Poolformer can achieve optimal results, while too many layers may cause overfitting and reduce generalization performance leading to lower recognition rates. As for EOC-D, which is relatively more complicated than others, a model that is too simple may not have enough learning capacity, so a less layered Res-Poolformer performs slightly worse while Res-MLPformer achieves better recognition results.

Further investigation into optimizing parameters, specifically for EOC-D and the augmentation method of the training set,

TABLE XIII
PERFORMANCES OF RESNET18 UNDER DIFFERENT WORKING CONDITIONS

No. of Views	Recognition Rate (%)			
	SOC	EOC-V	EOC-C	EOC-D
2	99.3	98.8	96.6	90
3	99.1	98.9	97.5	95.1
4	99.7	99.1	97.8	94.9

could potentially improve the generalization performance of the model.

V. ABLATION EXPERIMENTS AND DISCUSSION

The ablation experiments are conducted under SOC and EOCs to compare the effect of the SFEN module, class token, positional embedding (PE), and the design of the FC layer in the pooling mixer.

A. ResNet Without Xformer

The main parts related to the representation of the multiview SAR sequence in our proposed method are the single image encoding and SFEN-based Xformer.

ResNet has been widely applied in the field of computer vision and has achieved excellent results. Therefore, we adopted ResNet8 in the single-image encoding stage to enhance the overall performance of the network. To verify the role of our proposed SFEN, we conduct ablation experiments by removing the Xformer structure and retaining only ResNet18. As a result of removing the sequence fusion part, we concatenate the images within multiview SAR sequence as the input of ResNet18. The results are depicted in Table XIII.

As given in Table XIII, ResNet18 achieves a good recognition rate but is inferior to Res-Xformer. It shows similar results compared to other algorithms in Table XI. For tasks under SOC, achieving a recognition accuracy of over 99% with ResNet18 is relatively easy, similar to other CNN-based methods. The performance of ResNet18 is similar under EOC-V as it is under SOC. However, when it comes to the EOC-C condition, the recognition rate is weaker compared to models with Xformer, and it decreases even more noticeably under EOC-D, which presents greater challenges.

While ResNet can extract image features well, it performs decently under SOC and EOC-V conditions. However, its performance declines for more complex tasks. The lack of design for fusing sequence features prevents the effective utilization of sequential information, limiting its ability to capture spatial correlations among images within a sequence and thus affecting its effectiveness.

This demonstrates that our Res-Xformer design, on the one hand, builds a solid foundation for extracting sequence features by leveraging ResNet to extract features of images within the sequence. On the other hand, through the design of the SFEN module, it integrates sequence information and achieves good results. In contrast to previous studies focusing on using attention mechanisms to fuse sequence information, our designed replaceable token mixer further verifies that MSA is not the key factor in sequence feature fusion. Instead, the design of a combination of token mixer and channel mixer should be emphasized in the fusing sequence feature.

TABLE XIV
ABLATION EXPERIMENTS OF RES-POOLFORMER ON FC IN POOLING MIXER UNDER DIFFERENT WORKING CONDITIONS

No. of Layers	FC	Recognition Rate (%)			
		SOC	EOC-V	EOC-C	EOC-D
1	×	99.4	100	98.5	93.6
	✓	99.3	100	98.5	94.8
3	×	99.4	99.9	98.3	93.6
	✓	99.9	100	98.9	95.2
8	×	99.6	100	97.9	95.8
	✓	99.6	99.9	98.6	96.9

B. Fully Connected Layer in Pool-Mixer

The purpose of adding an FC layer in the pooling token mixer is to address the limitation of the reception field of pooling operations. This means that each time we perform pooling, the information of all tokens within the sequence is related directly to the current token. This is because when the sequence length is greater than the pooling size, not all token information can be associated. Here, ablation experiments are conducted to test the role of the FC, and the results are given in Table XIV.

In Table XIV, the "√" denotes that is involved in the token mixer of pooling, while the "✗" denotes that the FC is not involved. For more complex recognition tasks, such as under the EOC-D condition, it can be observed that FC notably improves recognition rates. Under other conditions, the presence of FC slightly enhances classification performance but its effect is not significant.

Although the effect of FC is not obvious, we still believe that this design is necessary. Because in principle, we added FC layers to consider the impact of receptive fields when pooling operations are limited by pooling size. For example, in 4-view SAR recognition tasks, for a pooling operation with a pooling size of 3, there is no direct interaction and fusion of information between the first and last tokens. Of course, as the number of SFEN layers increases, the information between these two tokens will also be perceived implicitly through pooling with other tokens. However, this also means that these two tokens are not treated equally compared to other tokens. Perhaps when the sequence length is long enough, FC layer can play a better role and demonstrate its effectiveness.

C. Class Token and Positional Embedding

PE and class token are designed for multiview SAR recognition problems. PE provides sequential information to the feature extraction network, while the class token is used to synthesize sequence features. Unlike methods that concatenate or average sequence features after feature extraction of tokens within the

TABLE XV
ABLATION EXPERIMENTS ON CLASS TOKEN AND PE UNDER DIFFERENT WORKING CONDITIONS

Token Mixer	No. of Layers	PE	Class Token	Recognition Rate (%)			
				SOC	EOC-V	EOC-C	EOC-D
MSA	3-4	×	×	99.5	99.3	96.2	87.5
		×	✓	99.1	99.5	94.3	90.8
		✓	×	99.5	99.3	95.7	82.9
		✓	✓	99.8	100.0	97.5	95.5
	3-8	×	×	99.0	99.9	95.9	89.8
		×	✓	99.3	99.9	95	95.4
		✓	×	98.9	99.9	95.8	93.1
		✓	✓	99.6	100.0	98.5	96.6
MLP	1	×	×	99.8	99.5	97.6	95.7
		×	✓	99.8	99.7	98	90.0
		✓	×	99.8	99.6	97.6	90.0
		✓	✓	99.8	100.0	99.1	97.7
	3	×	×	99.3	99.5	97.1	86.1
		×	✓	99.5	99.6	98.4	92.8
		✓	×	99.6	99.6	98.1	95.6
		✓	✓	99.8	100.0	98.1	97.2
Pooling	1	×	×	99.1	99.8	96.7	90.6
		×	✓	99.2	99.9	98.3	92.5
		✓	×	99.3	99.8	97.4	94.4
		✓	✓	99.3	100.0	98.5	94.8
	3	×	×	99.8	99.9	98.4	95.1
		×	✓	99.9	99.9	96.9	94.3
		✓	×	99.9	99.9	97.4	94.9
		✓	✓	99.9	100.0	98.9	95.2

sequence, the class token is involved in the entire sequence feature extraction process and gradually learns the characteristics of the sequence during the fusion of sequence features. This design has a more explicit role and a delicate feature fusion process.

The ablation experiments on PE and class token are conducted in a 4-view scenario. Based on the results analysis in the previous section, for target recognition problems under different conditions, both three-layer Res-MSAformer and Res-Poolformer and one-layer Res-MLPformer show the best recognition performances. Therefore, this part of the ablation experiment is also based on models with these parameters. When PE is not involved in the model, the features of all the sequence tokens are averaged as the sequence feature for classification. The results are summarized in Table XV, where "$\surd$" denotes that the module is involved, while "✗" denotes that the module is not involved.

We design PE to enable the model to establish a sense of sequential order among tokens from a theoretical perspective. However, its effectiveness may not be as apparent when dealing with sequences not long enough.

In most cases, the models generally perform poorly when neither PE nor class token is involved, whereas the recognition rate is highest when both are included. The ablation experiments show that PE and class token can bring different degrees of performance improvement, with class token having a greater impact. It is worth noting here that the effect of involving individual class token or PE alone is not as obvious. However, when they are used together in SFEN, significant performance improvement can be achieved. This indicates that their combination works better than when they are involved separately.

We believe that this is because the addition of individual PE allows the network to perceive the location information of tokens during the process of sequence feature extraction and fusion. However, due to the averaging operation of the tokens' feature for final token feature aggregation, this diminishes the impact of such positional information. On the other hand, without involving PE, there is no consideration of position information for class token during feature fusion. Therefore, combining both modules highlights their respective roles and leads to improved accuracy in recognition.

VI. CONCLUSION

In this study, we focus on deconstructing the deep learning processes of multiview SAR image recognition tasks and providing a more flexible way to model the multiview SAR sequence.

We divide this task into feature extraction of the single image within the multiview SAR sequence and feature extraction of the entire sequence. We decompose the SFEN into an STF module that fuses information across different images within each sequence, and an FCF module that combines information from different feature channels.

Unlike previous studies that focus on using attention mechanisms to fuse features, the emphasis of our study lies in the combination of token mixer and channel mixer as the key to sequence fusion. We focus on analyzing the structure of the

Transformer architecture itself rather than MSA and propose a replaceable token mixer by analyzing the functions of each module. By analyzing and decomposing multi-SAR ATR problems, we reason the design of each module in the proposed network, and the experiments conducted in this article have proven the effectiveness of our proposed method.

Our future work will involve exploring other modules as potential alternatives for sequence token mixers in Xformer and investigating their impact on the SAR image sequence recognition tasks.

REFERENCES

[1] W. M. Brown, "Synthetic aperture radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-3, no. 2, pp. 217–229, Mar. 1967.

[2] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013.

[3] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 3, pp. 56–81, Mar. 2021.

[4] G. F. Brendel and L. L. Horowitz, "Benefits of aspect diversity for SAR ATR: Fundamental and experimental results," *Proc. SPIE*, vol. 4053, pp. 567–578, Aug. 2000.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[7] *The Air Force Moving and Stationary Target Recognition Database*, Accessed: 2014. [Online]. Available: https://www.sdms.afrl.af.mil

[8] M. Z. Brown, "Analysis of multiple-view Bayesian classification for SAR ATR," in *Proc. SPIE 5095, Algorithms Synthetic Aperture Radar Imagery X*, 2003, pp. 265–274.

[9] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990, doi: 10.1207/s15516709cog1402_1.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016, doi: 10.1109/TGRS.2016.2551720.

[12] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017, doi: 10.1109/TGRS.2016.2645226.

[13] D. Ho Tong Minh et al., "Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 464–468, Mar. 2018, doi: 10.1109/LGRS.2018.2794581.

[14] E. Ndikumana, E. Ndikumana, D. Ho Tong Minh, N. Baghdadi, D. Courault, and L. Hossard, "Deep recurrent neural network for agricultural classification using multitemporal SAR Sentinel-l for Camargue, France," *Remote Sens.*, vol. 10, no. 8, pp. 464–468, Aug. 2018, doi: 10.3390/rs10081217.

[15] J. A. O'Sullivan, M. D. DeVore, V. Kedia, and M. I. Miller, "SAR ATR performance using a conditionally Gaussian model," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 1, pp. 91–108, Jan. 2001, doi: 10.1109/7.913670.

[16] U. Srinivas, V. Monga, and R. G. Raj, "SAR automatic target recognition using discriminative graphical models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 1, pp. 591–606, Jan. 2014, doi: 10.1109/TAES.2013.120340.

[17] B. Ding and G. Wen, "Exploiting multi-view SAR images for robust target recognition," *Remote Sens.*, vol. 9, no. 11, Nov. 2017, Art. no. 1150, doi: 10.3390/rs9111150.

[18] F. Zhang, Z. Z. Fu, Y. S. Zhou, W. Hu, and W. Hong, "Multi-aspect SAR target recognition based on space-fixed and space-varying scattering feature joint learning," *Remote Sens. Lett.*, vo. vol. 10, no. 10, pp. 998–1007, Jul. 2019, doi: 10.1080/2150704X.2019.1635287.

[19] H. Zou et al., "Research on multi-aspect SAR images target recognition using deep learning," *J. Signal Process.*, vol. 34, no. 5, pp. 513–522, 2018, doi: 10.16798/j.issn.1003-0530.2018.05.002.

[20] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T.-S. Yeo, "SAR automatic target recognition based on multiview deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2196–2210, Apr. 2018, doi: 10.1109/TGRS.2017.2776357.

[21] P. Zhao, K. Liu, H. Zou, and X. Zhen, "Multi-stream convolutional neural network for SAR automatic target recognition," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 1473, doi: 10.3390/rs10091473.

[22] J. Pei et al., "FEF-Net: A deep learning approach to multiview SAR image target recognition," *Remote Sens.*, vol. 13, no. 17, Sep. 2021, Art. no. 3493, doi: 10.3390/rs13173493.

[23] F. Zhang, C. Hu, Q. Yin, W. Li, H.-C. Li, and W. Hong, "Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition," *IEEE Access*, vol. 5, pp. 26880–26891, 2017, doi: 10.1109/ACCESS.2017.2773363.

[24] X. Bai, R. Xue, L. Wang, and F. Zhou, "Sequence SAR image classification based on bidirectional convolution-recurrent network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9223–9235, Nov. 2019, doi: 10.1109/TGRS.2019.2925636.

[25] J. Pei et al., "Multiview deep feature learning network for SAR automatic target recognition," *Remote Sens.*, vol. 13, no. 8, Apr. 2021, Art. no. 1455, doi: 10.3390/rs13081455.

[26] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016, doi: 10.1109/LGRS.2015.2513754.

[27] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, "Improving SAR automatic target recognition models with transfer learning from simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1484–1488, Sep. 2017, doi: 10.1109/LGRS.2017.2717486.

[28] Z. Cui, M. Zhang, Z. Cao, and C. Cao, "Image data augmentation for SAR sensor via generative adversarial nets," *IEEE Access*, vol. 7, pp. 42255–42268, 2019, doi: 10.1109/ACCESS.2019.2907728.

[29] X. Bao, Z. Pan, L. Liu, and B. Lei, "SAR image simulation by generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 9995–9998, doi: 10.1109/IGARSS.2019.8899286.

[30] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," 2017, *arXiv:1705.03122v3.*

[31] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473.*

[33] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," 2017, *arXiv:1702.00887.*

[34] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929.*

[35] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," 2021, *arXiv:2105.03824.*

[36] P. H. Martins, Z. Marinho, and A. F. T. Martins, "∞-former: Infinite memory transformer," 2021, *arXiv:2109.00301.*

[37] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," 2021, *arXiv:2105.01601.*

[38] H. Touvron et al., "ResMLP: Feedforward networks for image classification with data-efficient training," 2021, *arXiv:2105.03404.*

[39] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10819, doi: 10.1109/CVPR52688.2022.01055.

[40] A. Trockman and J. Zico Kolter, "Patches are all you need?" 2022, *arXiv:2201.09792.*

[41] C. Wang, Y. Huang, X. Liu, J. Pei, Y. Zhang, and J. Yang, "Global in local: A convolutional transformer for SAR ATR FSL," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4509605, doi: 10.1109/LGRS.2022.3183467.

[42] S. Li, Z. Pan, and Y. Hu, "Multi-aspect convolutional-transformer network for SAR automatic target recognition," *Remote Sens.*, vol. 14, no. 16, Aug. 2022, Art. no. 3924, doi: 10.3390/rs14163924.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[44] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415.*

Wei Zhu was born in Luoyang, Henan, China, in 1985. She received the B.Eng. degree in electronic and information engineering from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2008, and the M.Eng. degree in signal and information processing from the Northwest Institute of Nuclear Technology (NINT), in 2011. She is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha, China.

She is currently an Associate Professor with NINT. Her research interests include signal processing and machine learning.

Weiguo Xiao received the B.S. degree in remote sensing from Zhejiang University, Hangzhou, China, in 1994, the M.S. degree in engineering mechanics from the Northwest Institute of Nuclear Technology (NINT), Xi'an, China, in 2000, and the Ph.D. degree in engineering mechanics from the University of Science and Technology, Hefei, China, in 2010.

He is currently a Professor with NINT. His research interests include seismic monitoring, seismic signal processing, explosion sources theory, and discrimination between explosions and earthquakes.

Xin Li was born in Tongren, Qinghai, China, in 1982. He received the B.S. degree in mathematics and applied mathematics from the School of Faculty of Science, Xi'an Jiaotong University, Xi'an, China, in 2005, and the M.Eng. degree in signal and information processing from the Northwest Institute of Nuclear Technology (NINT), Xi'an, in 2008.

He is currently an Associate Professor with NINT. His research interests include signal processing and development of data processing software.

Shunping Xiao received the B.S. degree in electronic engineering in 1986 and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 1995.

Since 2000, he has been a Professor with NUDT. He is currently the Chief Engineer of Hunan Advanced Technology Research Institute and a Professor with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, NUDT. His research interests include radar target recognition and space information acquisition and processing.

Xiaolin Hu was born Huaihua, Hunan, China, in 1991. He received the B.S. and Ph.D. degrees in geophysics from the University of Science and Technology of China, Hefei, Anhui, China, in 2013 and 2018, respectively.

He is currently an Associate Researcher with the North West Institute of Nuclear Technology. His research interests include signal processing and geophysical inversion.

Linbin Zhang received the B.S. degree in information engineering in 2018 from the National University of Defense Technology, Changsha, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the State Key Laboratory of Complex Electromagnetic Environment Effects.

His research interests include object detection, image classification, few-shot learning, and machine learning and its applications to remote sensing images.