




Article

Multi-Aspect Convolutional-Transformer Network for SAR Automatic Target Recognition

Siyuan Li ^{1,2,3}, Zongxu Pan ^{1,2,3,*}  and Yuxin Hu ^{1,2,3}¹ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China² Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China³ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zxpan@mail.ie.ac.cn

Abstract: In recent years, synthetic aperture radar (SAR) automatic target recognition (ATR) has been widely used in both military and civilian fields. Due to the sensitivity of SAR images to the observation azimuth, the multi-aspect SAR image sequence contains more information for recognition than a single-aspect one. Nowadays, multi-aspect SAR target recognition methods mainly use recurrent neural networks (RNN), which rely on the order between images and thus suffer from information loss. At the same time, the training of the deep learning model also requires a lot of training data, but multi-aspect SAR images are expensive to obtain. Therefore, this paper proposes a multi-aspect SAR recognition method based on self-attention, which is used to find the correlation between the semantic information of images. Simultaneously, in order to improve the anti-noise ability of the proposed method and reduce the dependence on a large amount of data, the convolutional autoencoder (CAE) used to pretrain the feature extraction part of the method is designed. The experimental results using the MSTAR dataset show that the proposed multi-aspect SAR target recognition method is superior in various working conditions, performs well with few samples and also has a strong ability of anti-noise.

Keywords: synthetic aperture radar (SAR); automatic target recognition (ATR); multiview; self-attention; convolutional autoencoder (CAE)

**Citation:** Li, S.; Pan, Z.; Hu, Y.

Multi-Aspect Convolutional-Transformer Network for SAR Automatic Target Recognition.

Remote Sens. **2022**, *14*, 3924.<https://doi.org/10.3390/rs14163924>

rs14163924

Academic Editors: Deliang Xiang, Ying Luo, Xueru Bai, Gangyao Kuang and Xiaolan Qiu

Received: 19 July 2022

Accepted: 10 August 2022

Published: 12 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Synthetic aperture radar (SAR) is a high-resolution coherent imaging radar. As an active microwave remote sensing system, which is not affected by light and climatic conditions, SAR can achieve all-weather day-and-night earth detection [1]. At the same time, SAR adopts synthetic aperture technology and matched filtering technology, which can realize long-distance high-resolution imaging. Therefore, SAR is of great significance in both military and civilian fields [2].

In recent years, with the development of SAR technology, the ability to obtain SAR data has been greatly improved. The early methods of manually interpreting SAR images cannot support the rapid processing of large amounts of SAR data due to their low time efficiency and high cost. How to quickly mine useful information from massive high-resolution SAR image data and apply it to military reconnaissance, agricultural and forestry monitoring, geological survey and many other fields has become an important problem in SAR applications that needs to be solved urgently. Therefore, the automatic target recognition (ATR) [3] of SAR images to solve this problem has become a research hotspot.

Since a SAR image shows the scattering characteristics of the target to electromagnetic waves, the SAR image is very different from the optical image, which also has a great impact on the SAR ATR. Classical SAR ATR methods include a template-based method and model-based method. The template-based method is one of the earliest proposed

methods, including direct template matching methods that calculate the similarity between the template formed by processing the training sample itself and the test sample for classification [4], and feature template matching methods that use classifiers such as SVM [5], KNN [6] and Bayes classifier [7] after extracting various features [8,9] for classification. The template-based method is simple in principle and easy to implement, but requires large and diverse training data to build a complete template library. To make up for the shortcomings of the template-based method, the model-based method [10] is proposed, which includes two parts: model construction and online prediction.

With the development of machine learning and deep learning technology, the automatic feature extraction ability of a neural network has attracted the attention of researchers, and then deep learning has been applied in SAR ATR. Initially, the neural network model in traditional computer vision (CV) was directly applied to SAR target recognition. For instance, Kang et al. transferred existing pretrained networks after fine-tuning [11]. Unsupervised learning methods such as autoencoder [12] and deep belief network (DBN) [13] were also used to automatically learn SAR image features. Afterward, the network structure and loss function were designed for the specific task of target recognition using the amplitude information of SAR images, which were more in line with the requirements of the SAR target recognition task and undoubtedly achieved better recognition performance. Chen et al. designed a fully convolutional network (A-ConvNets) for recognition on the MSTAR target dataset [14]. Lin et al. proposed the deep convolutional Highway Unit for ATR of a small number of SAR targets [15]. Du et al. recommended the application of multi-task learning to SAR ATR to learn and share useful information from two auxiliary tasks designed to improve the performance of recognition tasks [16]. Gao et al. proposed to extract polarization features and spatial features, respectively, based on a dual-branch deep convolution neural network (Dual-CNN) [17]. Shang et al. designed deep memory convolution neural networks (M-Net), including an information recorder to remember and store samples' spatial features [18]. Recently, the characteristics brought by the special imaging mechanism of SAR are being focused on, and some methods combining deep learning with physical models have appeared. Zhang et al. proposed a domain knowledge-powered two-stream deep network (DKTS-N), which incorporates a deep learning network with SAR domain knowledge [19]. Feng et al. combined electromagnetic scattering characteristics with a depth neural network to introduce a novel method for SAR target classification based on an integration parts model and deep learning algorithm [20]. Wang et al. recommended an attribute-guided multi-scale prototypical network (AG-MsPN) that obtains more complete descriptions of targets by subband decomposition of complex-valued SAR images [21]. Zhao et al. proposed a contrastive-regulated CNN in the complex domain to obtain a physically interpretable deep learning model [22]. Compared with traditional methods, deep learning methods have achieved better recognition results in SAR ATR. However, most of these current SAR ATR methods based on deep learning are aimed at single-aspect SAR images.

In practical applications, due to the special imaging principle of SAR, the same target will show different visual characteristics under different observation conditions, which also makes the performance of SAR ATR methods affected by various factors, such as environment, target characteristics and imaging parameters. The observation azimuth is also one of the influencing factors. The sensitivity of the scattering characteristics of artificial targets to the observation azimuth leads to a large difference in the visual characteristics of the same target at different aspects. Therefore, the single-aspect SAR image loses the scattering information related to the observation azimuth [23]. The target recognition performance of single-aspect SAR is also affected by the aspect.

With the development of SAR systems, multi-aspect SAR technologies such as Circular SAR (CSAR) [24] can realize continuous observation of the same target from different observation azimuth angles. The images of the same target under different observation azimuth angles obtained by multi-angle SAR contain a lot of identification information. The multi-aspect SAR target recognition technology uses multiple images of the target

obtained from different aspects and combines the scattering characteristics of different aspects to identify the target category. Compared with single-aspect SAR images, multi-aspect SAR image sequences contain spatially varying scattering features [25] and provide more identification information for the same target under different aspects. On the other hand, multi-aspect SAR target recognition can improve the target recognition performance by fully mining the intrinsic correlation between multi-aspect SAR images.

The neural networks that use multi-aspect SAR image sequences for target recognition mainly include recurrent neural networks (RNN) and convolutional neural networks (CNN). Zhang et al. proposed multi-aspect-aware bidirectional LSTM (MA-BLSTM) [26], which extracts features from each multi-aspect SAR image through a Gabor filter, and further uses LSTM to store the sequence features in the memory unit and transmit through learnable gates. Similarly, Bai et al. proposed a bidirectional convolutional-recurrent network (BCRN) [27], which uses Deep CNN to replace the manual feature extraction process of MA-BLSTM. Pei et al. proposed a multiview deep convolutional neural network (MVD-CNN) [28], which uses a parallel CNN to extract the features of each multi-aspect SAR image, and then merges them one by one through pooling. Based on MVDCNN, Pei et al. improved the original network with the convolutional gated recurrent unit (ConvGRU) and proposed a multiview deep feature learning network (MVDFLN) [29].

Although these methods have obtained good recognition results, there are still the following problems:

1. When using RNN or CNN to learn the association between multi-aspect SAR images, the farther the two images are in a multi-aspect SAR image sequence, the more difficult it is to learn the association between them. That is, the association will depend on the order of the image in the sequence.
2. All current studies require a lot of data for training the deep networks, and the accuracy will drop sharply in the case of few samples.
3. The existing approaches do not consider the influence of noise, which leads to a poor anti-noise ability of the model.

To address these problems, in this paper, we propose a multi-aspect SAR target recognition method based on convolutional autoencoder (CAE) and self-attention. After pre-training, the encoder of CAE will be used to extract the features of single-aspect SAR images in the multi-aspect SAR image sequence, and then the intrinsic correlation between images in the sequence will be mined through a transformer based on self-attention.

In this paper, it is innovatively proposed to mine the correlation between multi-aspect SAR images through a transformer [30] based on self-attention. Vision transformer (ViT) [31] for optical image classification and the networks based on attention for single-aspect SAR ATR, such as the mixed loss graph attention network (MGA-Net) [32] and the convolutional transformer (ConvT) [33], extract representative features by determining the correlation between various parts of an image itself. Unlike them, the ideas of natural language processing (NLP) tasks are leveraged to mine the association between the semantic information of each image in the multi-aspect SAR image sequence. Because each image is correlated with other images in the same way in the calculation process of self-attention, the order dependence problem faced by existing methods will be avoided. Considering that self-attention loses local details, the CNN pre-trained by CAE with shared parameters is designed to extract local features for each image in the sequence. On the one hand, effective feature extraction provided by CNN can diminish the requirement for sample size. On the other hand, by minimizing the gap between the reconstructed image and the original input, the autoencoder ensures that the features extracted by the encoder can effectively represent the principal information of the original image. Thus CAE plays a vital role in anti-noise.

Compared with available multi-aspect SAR target recognition methods, the novelty as well as the contribution of the proposed method can be summarized as follows.

1. A multi-aspect SAR target recognition method based on self-attention is proposed. Compared with existing methods, the calculation process of self-attention makes it not affected by the order of images in the sequence. To the best of our knowledge,

this is the first attempt to apply a transformer based on self-attention to complete the recognition task of multi-aspect SAR image sequences.

2. CAE is introduced for feature extraction in our method, which is due to the additional consideration of the cases with few samples and noise compared with other methods and is created to improve the ability of the network to effectively extract the major features through pre-training and fine-tuning.
3. Compared with the existing methods designed for multi-aspect SAR target recognition, our network obtains higher recognition accuracy on the MSTAR dataset and exhibits more robust recognition performance in version and configuration variants. Furthermore, our method demonstrates better in the recognition task with a small number of samples. Our method achieves stronger performance in anti-noise assessment as well.

The remainder of this paper is organized as follows: Section 2 describes the proposed network structure in detail. Section 3 presents the experimental details and results. Section 4 discusses the advantages and future work of the proposed method. Section 5 summarizes the full paper.

2. Multi-Aspect SAR Target Recognition Framework

2.1. Overall Structure

As shown in Figure 1, the proposed multi-aspect SAR target recognition method consists of five parts, i.e., multi-aspect SAR image sequence construction, single-aspect feature extraction, multi-aspect feature learning, feature dimensionality reduction and target classification. Among them, feature extraction is implemented using CNN pretrained by CAE, and multi-aspect feature learning uses the transformer encoder [31] structure based on self-attention.

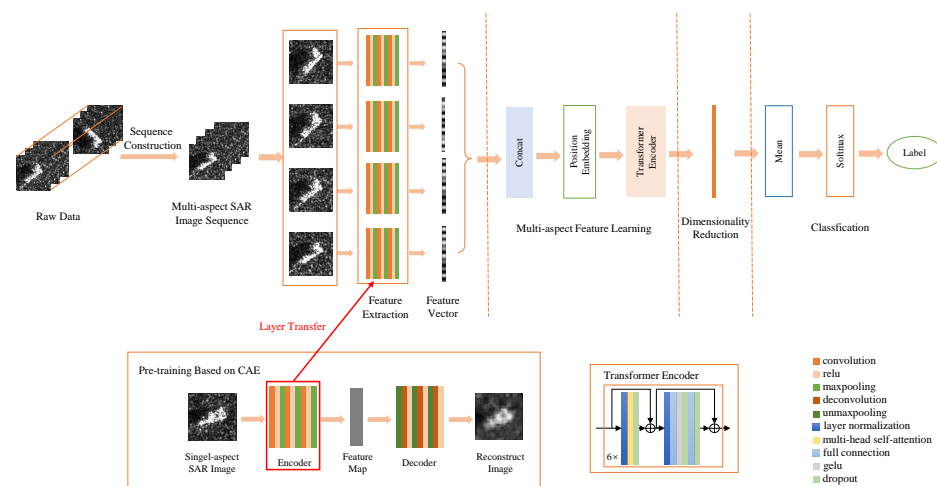


Figure 1. Basic architecture of the proposed multi-aspect SAR target recognition method.

Before feature extraction, single-aspect SAR images are used to construct multi-aspect SAR image sequences and also serve as the input to pre-train CAE, which includes the encoder that utilizes the multi-layer convolution-pooling structure to extract features and the decoder that utilizes the multi-layer deconvolution-unpooling structure to reconstruct images. The encoder of CAE after pre-training will be transferred to CNN for feature extraction of each image in the input multi-aspect SAR image sequence. In particular, the output of feature extraction for an image is a 1-D feature vector. Then, in the multi-aspect feature learning structure, the vectorized features extracted from each image are spliced and added to position embedding to be the input of the transformer encoder based on multi-head self-attention. All the output features of the transformer encoder are reduced

in dimension by 1×1 convolutions and then averaged. Finally, the softmax classifier gives the recognition result.

In the training process, the whole network obtains errors from the output and propagates back along the network to update parameters. It should be noted that the parameters of multi-layer CNN used for feature extraction can be frozen or updated with the entire network for fine-tuning.

In the following discussions, the details of each part of the proposed method and the training process will be introduced in turn, such as the loss function and so forth.

2.2. Multi-Aspect SAR Image Sequence Construction

Multi-aspect SAR images can be obtained by imaging the same target from different azimuth angles and different depression angles by radars on one or more platforms. Figure 2 shows a simple geometric model for multi-aspect SAR imaging.

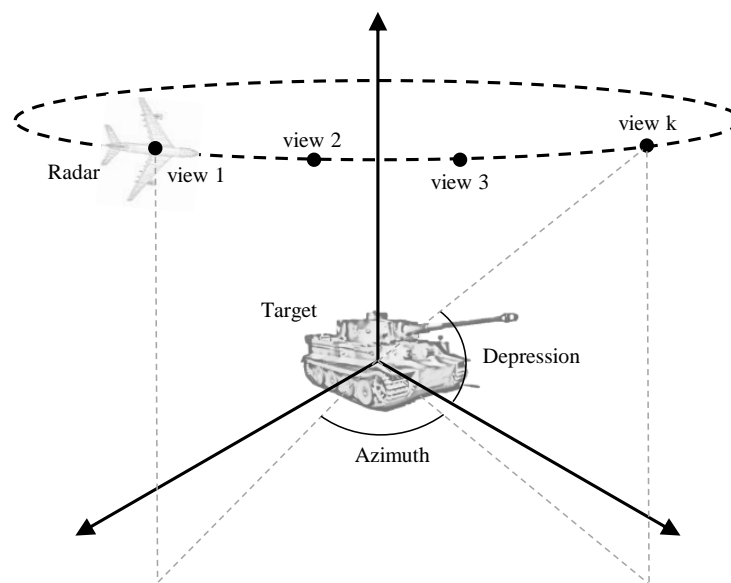


Figure 2. A simple geometric model for multi-aspect SAR imaging.

On this basis, multi-aspect SAR image sequences are built based on the following rules. Suppose $X^r = \{X^1, X^2, \dots, X^C\}$ is the raw SAR image set, where $X^i = \{x_1^i, x_2^i, \dots, x_{n_i}^i\}$ is the image set sorted by azimuth angle for a specific class c_i . C is the number of classes and n_i is the number of images contained in one class. The azimuth of the image is $\varphi(x_j^i)$. For a given angle range θ and sequence length k , a window with fixed length $k + 1$ is placed along the original image set with a stride of 1 and the images in the window form k sequences of length k by permutation, then the sequence whose azimuth difference between any two images is smaller than θ is reserved as the training sample of the network. In addition, the final retained sequence samples are required to not contain duplicate samples. The process of multi-aspect SAR image sequence construction is summarized in Algorithm 1. In Algorithm 1, $X^S = \{X_S^1, X_S^2, \dots, X_S^C\}$ is the multi-aspect SAR image sequence, where $X_S^i = \{X_{s_1}^i, X_{s_2}^i, \dots, X_{s_{N_i}}^i\}$ is the sequence set for a specific class c_i . N_i is the number of sequences contained in one class.

Figure 3 shows an example of multi-aspect SAR image sequence construction. When the sequence length is set to 4, ideally, 12 sequence samples can be received from only 7 images. In this way, enough training samples can be obtained from limited raw SAR images.

Algorithm 1 Construct multi-aspect SAR image sequence

Input: angle range θ and sequence length k , raw SAR images $X^r = \{X^1, X^2, \dots, X^C\}$, and class labels $c_i \in \{1, 2, \dots, C\}$

Output: multi-aspect SAR image sequence $X^S = \{X_S^1, X_S^2, \dots, X_S^C\}$

```

for  $i = 1$  to  $C$  do
  for  $j = 1$  to  $n_i - k$  do
    if  $|\varphi(x_j^i) - \varphi(x_{j+k}^i)| \leq \theta$ 
      Construct all possible sequence except  $\{x_{j+1}^i, x_{j+2}^i, \dots, x_{j+k}^i\}$ 
    else if  $|\varphi(x_j^i) - \varphi(x_{j+k-1}^i)| \leq \theta$ 
      Construct the sequence  $\{x_j^i, x_{j+1}^i, \dots, x_{j+k-1}^i\}$ 
  end for
  if  $|\varphi(x_{j+1}^i) - \varphi(x_{n_i}^i)| \leq \theta$ 
    Construct the sequence  $\{x_{j+1}^i, x_{j+2}^i, \dots, x_{n_i}^i\}$ 
  Get  $X_S^i = \{X_{s_1}^i, X_{s_2}^i, \dots, X_{s_{N_i}}^i\}$ 
end for

```

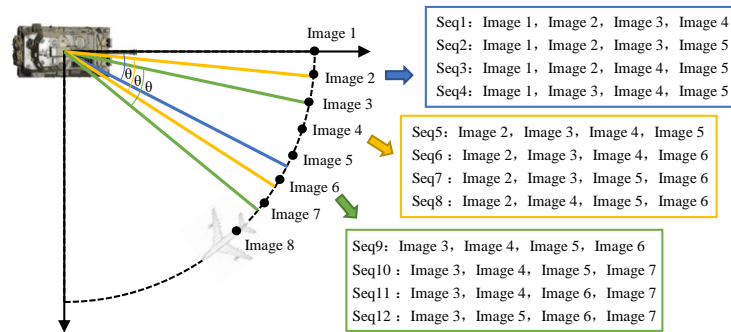


Figure 3. Example of multi-aspect SAR image sequence construction.

2.3. Feature Extraction Pre-Trained by CAE

Drawing on the idea of NLP, our method considers each image in a multi-aspect SAR image sequence to be equivalent to each word in a sentence. To effectively extract major features from each image in a multi-aspect SAR image sequence in parallel, CNN pre-trained by CAE with shared parameters is designed, which can reduce the number of learning parameters as well. Figure 4 shows the network structure of CAE, which consists of an encoder and decoder.

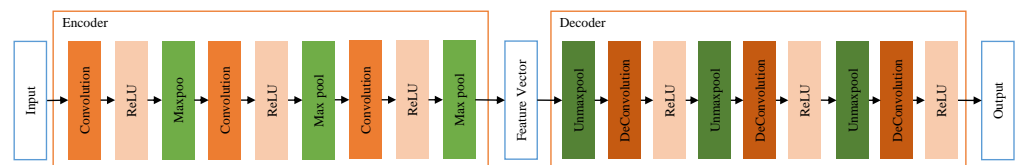


Figure 4. The structure of CAE.

The encoder is comprised of convolutional layers, pooling layers and the nonlinear activation function. The convolutional layer is the core structure, which extracts image features through convolution operations. The convolution layer in the neural network initializes a learnable convolution kernel, which is convolved with the input vector to obtain a feature map. Each convolution kernel has a bias, which is also trainable. The activation function is a mapping from the input to the output of the neural network, which increases the nonlinear properties of the neural network. ReLU, which is simple to calculate and can speed up the convergence of the network, is used as the activation function in the encoder of CAE. The convolutional layer is usually followed by the pooling layer, which plays

the role of downsampling. Common pooling operations mainly include max pooling and average pooling. In this method, maximum pooling is selected; that is, it takes the largest value in the pooling window as the value after pooling.

For the l th convolution-maxpooling layer of the encoder, suppose x^{l-1} is the input and x^l is the output feature map, and the input of the first layer x^0 is the input image x . Suppose W^l is the convolution kernel in the l th layer, and b^l is its bias. The feedforward propagation process of a convolution-maxpooling layer in the encoder can be expressed as:

$$a^l = x^{l-1} * W^l + b^l \quad (1)$$

$$x^l = f_{\text{DOWN}}(\sigma(a^l)) \quad (2)$$

where $*$ and f_{DOWN} denote the convolution and the pooling operation, respectively. σ represents the ReLU activation function, which is defined as:

$$\sigma(z) = \max(0, z) \quad (3)$$

The decoder reconstructs the image according to the feature map. The decoder is also a multi-layer structure, which contains unmaxpooling layers, deconvolution layers and the ReLU activation function. Unpooling is the reverse operation of pooling, which restores the original information to the greatest extent by complimenting. In this work, unmaxpooling is chosen; that is, it is assigned the value to the position of the maximum value in the pooling window recorded during pooling, and we supplement 0 for the other positions in the pooling window. The deconvolution layer performs convolution between the feature map and the transposed convolution kernel so as to reconstruct the image based on the feature map.

For the l th unpooling-deconvolution layer of the decoder, suppose \hat{x}^{l-1} is the input and \hat{x}^l is the output. The input of the decoder's first layer \hat{x}^0 is the output of the encoder x^L . Suppose \hat{W}^l is the convolution kernel in the l th layer, \hat{W}^T represents the transpose of the convolution kernel, and \hat{b}^l is the bias in the l th layer. The feedforward propagation process of an unpooling-deconvolution layer in the decoder can be expressed as:

$$\hat{a}^l = f_{\text{UP}}(\hat{x}^{l-1}) \quad (4)$$

$$\hat{x}^l = \sigma(\hat{a}^l * (\hat{W}^l)^T + \hat{b}^l) \quad (5)$$

where f_{UP} represents the unpooling operation.

The output of the whole CAE \hat{x} is the output of the decoder's last layer \hat{x}^L . CAE takes a single-aspect SAR image as input, and the output is a reconstructed image of the same size as the input image. The training of the CAE module will be detailed in Section 2.7.1. After the training, the encoder is transferred to extract features of each image in the sequence, the output feature vector is the output of the last layer of the encoder $x_{(i)}^L$, $i = 1, \dots, k$.

2.4. Multi-Aspect Feature Learning Based on Self-Attention

The multi-aspect feature learning part of the proposed method is modified based on the transformer encoder to make it suitable for multi-aspect SAR target recognition. The feature vectors extracted from each image in the sequence are combined as $X_F = [x_{(1)}^L, \dots, x_{(k)}^L]$, which is the input of position embedding. Positional embedding is proposed because self-attention does not consider the order between input vectors, but in translation tasks in NLP, the position of a word has an impact on the meaning of a sentence. As described in Section 2.2, multi-aspect SAR images are constructed into sequences according to azimuth angles, and the angle information of images in the sequence also needs to be recorded by position embedding. In this work, sine and cosine functions are used to calculate the positional embedding [30]. The output of positional embedding X_{PE} is input to the transformer encoder next.

The transformer encoder, which is the kernel structure of this part, is shown in Figure 5. The transformer encoder is composed of multiple layers, and each layer contains two residual blocks, which mainly include the multi-head self-attention (MSA) unit and multi-layer perceptron (MLP) unit. Supposed there are N layers in the transformer encoder, the details of these two residual blocks in each layer will be introduced below.

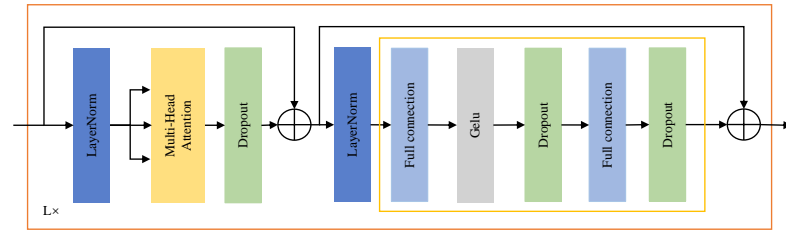


Figure 5. The structure of the transformer encoder.

The first residual block is formed by adding the result of the input vector going through layer normalization (LN) [34] and MSA unit to itself. LN is used to implement normalization. Different from the commonly used batch normalization (BN), all the hidden units in a layer share the same normalization terms under LN. Thus, LN does not impose any constraint on the size of a mini-batch. MSA, the core of the transformer encoder, is to calculate the correlation as a weight by multiplying the query and the key and then using this weight to weighted sum the value to increase the weight of related elements in a sequence and reduce the weight of irrelevant elements.

Here the calculation process of the first residual block is described. Suppose $Z^{n-1} \in \mathbb{R}^{k \times m}$ is the input of the first residual block on the n th layer of the transformer encoder, and $Z^n \in \mathbb{R}^{k \times m}$ is the output. Among them, k is the number of images in a multi-aspect SAR image sequence, m represents the channel number of each image's feature vectors, and the input of the first layer is the output of the position-embedding X_{PE} . The input vector $Z^{n-1} \in \mathbb{R}^{k \times m}$ first passes through LN to get $Y^{n-1} \in \mathbb{R}^{k \times m}$; that is:

$$Y^{n-1} = \Phi_{LN}(Z^{n-1}) \quad (6)$$

where Φ_{LN} indicates the process of LN. Then, Y^{n-1} is divided into H parts along the channel dimension. Suppose $d_h = m/H$, each part is recorded as $Y_h^{n-1} \in \mathbb{R}^{k \times d_h}$, $h = 1, \dots, H$, and corresponds to a head of self-attention. In other words, $Y^{n-1} = [Y_1^{n-1}, \dots, Y_H^{n-1}]$, where H is the number of heads in MSA. For each head, the input Y_h^{n-1} is multiplied by three learnable matrices, the query matrix $W_{qh}^n \in \mathbb{R}^{k \times k}$, the key matrix $W_{kh}^n \in \mathbb{R}^{k \times k}$ and the value matrix $W_{vh}^n \in \mathbb{R}^{k \times k}$ to obtain the query vector $Q_h^n \in \mathbb{R}^{k \times d_h}$, the key vector $K_h^n \in \mathbb{R}^{k \times d_h}$ and the value vector $V_h^n \in \mathbb{R}^{k \times d_h}$, which can be formulated as:

$$Q_h^n = W_{qh}^n Y_h^{n-1} \quad (7)$$

$$K_h^n = W_{kh}^n Y_h^{n-1} \quad (8)$$

$$V_h^n = W_{vh}^n Y_h^{n-1} \quad (9)$$

Then, the transposed matrix of K_h^n and Q_h^n are multiplied to obtain the initial correlation matrix $\hat{A}_h^n \in \mathbb{R}^{d_h \times d_h}$, and then a softmax operation is performed on \hat{A}_h^n column by column to obtain the correlation matrix $A_h^n \in \mathbb{R}^{d_h \times d_h}$. The calculation process is written as:

$$\hat{A}_h^n = (K_h^n)^T Q_h^n \quad (10)$$

$$A_h^n(i, j) = \text{Softmax}(\hat{A}_h^n(i, j)) = \frac{e^{\hat{A}_h^n(i, j)}}{\sum_{s=1}^{d_h} e^{\hat{A}_h^n(s, j)}} \quad (11)$$

V_h^n is multiplied by the weight matrix A_h^n to get the output Y_h^n ; that is:

$$Y_h^n = V_h^n A_h^n \quad (12)$$

Then, the output vectors of each head $Y_h^n, h = 1, \dots, H$ are concatenated along the channel dimension to obtain $\tilde{Y}^n = [Y_1^n, \dots, Y_H^n] \in \mathbb{R}^{k \times m}$. The output of MSA $Y^n \in \mathbb{R}^{k \times m}$ is obtained by multiplying \tilde{Y}^n with a trainable matrix $W_o^n \in \mathbb{R}^{m \times m}$; that is $Y^n = \tilde{Y}^n W_o^n$. Finally, to get the output of the first residual block $\tilde{Z}^n \in \mathbb{R}^{k \times m}$, the residual operation is applied to compute the summation of this block Z^{n-1} and the output of MSA Y^n , which can be formulated as:

$$\tilde{Z}^n = Z^{n-1} + Y^n \quad (13)$$

The second residual block of each layer in the transformer encoder is composed of adding the results of the input vector through MLP to itself. The input vector of the second residual block \tilde{Z}^n goes through LN, a fully connected sublayer with the nonlinear activation function and another fully connected sublayer in turn. The two fully connected sublayers expand and restore the vector dimension, respectively, thereby enhancing the expressive ability of the model. The number of neurons of the two fully connected sublayers is N_{fc1} and m , respectively. GELU is employed as the activation function, which performs well in transformer-based networks, and its definition is given as follows:

$$\text{GELU}(x) = x\Phi(x) \quad (14)$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Finally, through the residual operation, the output of the second residual block Z^n , which is also the output of the n th layer of the transformer encoder, is obtained. The calculation process of the second residual block is described as:

$$Z^n = \tilde{Z}^n + \Phi_{FC2}(\Phi_{FC1}(\Phi_{LN}(\tilde{Z}^n))) \quad (15)$$

where Φ_{FC1}, Φ_{FC2} represent the operation of the two fully connected sublayers separately.

To alleviate the overfitting issue that is prone to occur when the training sample is insufficient, dropout is added at the end of each fully connected sublayer. Dropout is implemented by ignoring part of the hidden layer nodes in each training batch, and to put it simply, p percent neurons in the hidden layer stop working during each training iteration.

2.5. Feature Dimensionality Reduction

After the multi-aspect feature learning process based on self-attention, features that contain intrinsic correlation information of multi-aspect SAR image sequence have been extracted with the size of $k \times m$. Before using these features for classification, their dimension needs to be reduced.

MLP is the most commonly used feature dimensionality reduction method but one that lacks cross-channel information integration. Our proposed method uses a 1×1 convolution [35] for dimensionality reduction, which uses the 1×1 convolution kernel and adjusts the feature dimension by the number of convolution kernels. The 1×1 convolution realizes the combination of information between channels, thus reducing the loss of information during dimensionality reduction. At the same time, compared with MLP, the 1×1 convolution reduces the number of parameters.

The input of the 1×1 convolutional layer is the output of the transformer encoder Z^N , and the output size is $k \times C$, where C is the number of classes. Therefore, the number of convolution kernels of the 1×1 convolution layer is C . Before the convolution operation, dimension transformation is performed from Z^N to $Z_r^N \in \mathbb{R}^{k \times 1 \times m}$. Suppose the output of dimensionality reduction is Z_c , $W^{1 \times 1}$ is the convolution kernel of the 1×1 convolutional layer and $b^{1 \times 1}$ is the bias. The dimensionality reduction process is described as:

$$Z_c = Z_r^N * W^{1 \times 1} + b^{1 \times 1} \quad (16)$$

After dimensionality reduction, dimensional transformation is performed on $Z_c \in \mathbb{R}^{k \times 1 \times C}$ to obtain $Z \in \mathbb{R}^{k \times C}$ for subsequent classification.

2.6. Classification

The classification process uses the softmax classifier. After dimensionality reduction, Z is averaged along the sequence dimension to get $\bar{Z}_{mean} = [\bar{z}_1, \dots, \bar{z}_C]^T$. Finally, the softmax operation is applied upon \hat{Z}_{mean} to get the probability output $Z_{mean} = [z_1, \dots, z_C]^T$.

2.7. Training Process

2.7.1. Pre-Train of CAE and Layer Transfer

CAE is an unsupervised learning method. As described in Section 2.3, taking the single-aspect SAR images as input, after the forward propagation, reconstructed images will be obtained. Network optimization of CAE is achieved by minimizing the mean square error (MSE) between the reconstructed image and the original image; that is, using the MSE loss function. Suppose the total number of samples is N_s . x is the input image, and \hat{x} is the reconstructed image. The MSE loss function is defined as

$$L_{MSE} = \frac{1}{N_s} \sum_{i=1}^{N_s} \|\hat{x}^{(i)} - x^{(i)}\|_2^2 \quad (17)$$

In this work, backpropagation (BP) is selected to minimize the loss function and optimize the CAE parameters. The aim of BP is to calculate the gradient and update the network parameters to achieve the global minimum of the loss function. The process of BP is to calculate the error of each parameter from the output layer to the input layer through the chain rule, which is related to the partial derivative of the loss function relative to the trainable parameters, and then the parameters are updated according to the gradient. When the network converges, it means that the encoder of CAE can extract enough information to reconstruct the single-aspect SAR image, which can prove the effectiveness of feature extraction.

After the CAE network converges, the parameters of the encoder are saved for subsequent layer transfer. The specific layer transfer operation is first when initializing network parameters; the CNN for feature extraction loads the parameters of the trained encoder of CAE. Next, in the process of network training, the parameters of each layer of the CNN can be frozen, which means that they will remain unchanged during the training process or continue to be optimized; that is, fine-tuning. It should be noted that only the first two layers of the CNN are frozen, and the last layer of the CNN is fine-tuned along with the overall network in our method. This is because the pre-training of CAE is carried out for single-aspect SAR images. Therefore, in order to effectively extract sufficient internal correlation information to support multi-aspect feature learning, it is necessary to fine-tune during the whole network training. At the same time, the parameters of the first two layers are frozen to maintain the effective extraction of the main features of each single-aspect image to ensure the noise resistance of the network.

2.7.2. Training of Overall Network

Taking the multi-aspect SAR image sequence as input, after the forward propagation, the predicted results given by the proposed network will be obtained. Network optimization is achieved by minimizing the cross entropy between data labels and the predicted result given by the network; that is, using the cross entropy loss function. Suppose the total number of the sequence samples is N_m . For the i th sample, let $[z_1^i, \dots, z_C^i]^T$ denote the label, and $[\hat{z}_1^i, \dots, \hat{z}_C^i]^T$ represent the probability output predicted by the network. When the sample belongs to the j th class, $z_j^i = 1$ and $z_k^i = 0$ ($k \neq j$). Then the cross entropy loss function can be written as:

$$L = -\frac{1}{N_m} \sum_{i=1}^{N_m} \sum_{j=1}^C z_j^{(i)} \log(\hat{z}_j^{(i)}) \quad (18)$$

The proposed method also minimizes the loss function and optimizes the network parameters by BP, which is the same as most SAR ATR methods. The parameter updating in the BP process is affected by the learning rate, which determines how much the model parameters are adjusted according to the gradient in each parameter update. In the early stage of network training, there is a large difference between the network parameters and the optimal solution. Thus the gradient descent can be carried out faster with a larger learning rate. However, in the later stage of network training, gradually reducing the value of the learning rate will help the convergence of the network and make the network's approach to the optimal solution easier. Therefore, in this work, the learning rate is decreased by piecewise constant decay.

3. Experiments and Results

To verify the effectiveness of our proposed method, first, the network architecture setup is specified, and then the multi-aspect SAR image sequences are constructed using the MSTAR dataset under the standard operating condition (SOC) and extended operating condition (EOC), respectively. Finally, the performance of the proposed method has been extensively assessed by conducting experiments under different conditions.

3.1. Network Architecture Setup

In the experiment, the network instances were deployed whose input can be single-aspect images or multi-aspect sequences to comprehensively evaluate the recognition performance of the network. In this instance, the size of the input SAR image is 64×64 . The encoder of CAE includes three convolution-maxpooling layers, which obtain 64, 64 and 256 feature maps, respectively. In each layer, the convolution operation with kernel size 7×7 and stride size 2×2 is followed by the max-pooling operation with kernel size 3×3 and stride size 2×2 . The decoder of CAE includes three unpooling-deconvolution layers with the number of channels 64, 64 and 1, respectively. In each layer, the unpooling operation with kernel size 3×3 and stride size 2×2 is followed by the deconvolution operation with kernel size 7×7 and stride size 2×2 . In the transformer encoder with 6 layers, MSA has 4 heads and the number of neurons of the 2 fully connected sublayers is 512 and 256 in turn.

Our proposed network is implemented by the deep learning toolbox Pytorch 1.9.1. All the experiments are conducted on a PC with an Intel Core i7-9750H CPU at 2.60 GHz, 16.0 G RAM, and a NVIDIA GeForce RTX 2060 GPU. The learning rate is 0.00001 when training CAE and starts from 0.001 with decay rate 0.9 every 30 epochs when training the whole network. The mini-batch size is set to 32, and the probability of dropout is 0.1.

3.2. Dataset

The MSTAR dataset [36], which was jointly released by the U.S. Defense Advanced Research Projects Agency (DARPA) and the U.S. Air Force Research Laboratory (AFRL), consists of high-quality SAR image data collected from ten stationary military vehicles (i.e., rocket launcher: 2S1; tank: T72 and T62; bulldozer: D7; armored personnel carrier: BMP2, BRDM2, BTR60 and BTR70; air defense unit: ZSU234; truck: ZIL131) through the X-band high-resolution Spotlight SAR by Sandia National Laboratory between 1996 and 1997. All images in the MSTAR dataset have a resolution of 0.3×0.3 m with HH polarization. The azimuth aspect range of imaging each target covers $0^\circ \sim 360^\circ$ with an interval of $5^\circ \sim 6^\circ$. The optical images of ten targets and their corresponding SAR images are illustrated in Figure 6.

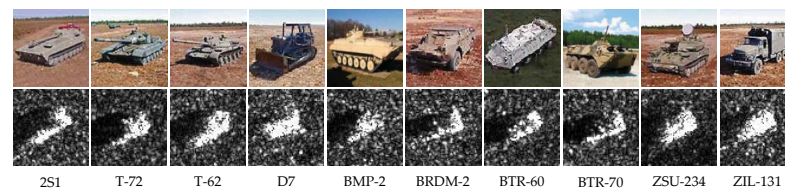


Figure 6. The optical images of ten targets and their corresponding SAR images.

The acquisition conditions of the MSTAR dataset include two categories: Standard Operating Condition (SOC) and Extended Operating Condition (EOC). Specifically, SOC refers to the images of the training set and testing set that have the same target type and similar imaging configuration. Compared with SOC, the data of the training set and testing set of EOC are more different and have greater difficulty in identification. Generally, the EOC includes configuration-variant (EOC-C) and version-variant EOC (EOC-V).

First, the images are cropped to 64×64 and normalized. Next, the experiments are conducted using the sequences constructed according to the steps described in Section 2.2 under SOC and EOC and the results are shown in Sections 3.3 and 3.4. Then, the performance of the proposed method is compared with other existing methods in Section 3.5. Finally, some further discussions are shown in Section 3.6 about the influence of convolution kernel size in feature extraction and the performance of the network with few samples and noise.

3.3. Results under SOC

SOC means that the training and test datasets have the same target type and similar imaging configuration. The experiment under SOC is a classical ten-class classification problem of vehicle targets. Among the raw SAR images, the images collected at 17° depression angle are set as the training set, and the SAR images collected at 15° depression angle as the testing set. By applying the method of constructing sequences described in Section 2.2, the training sequences and testing sequences are obtained when the angle range is set as 45° as in other multi-aspect SAR ATR methods, considering the actual radar imaging situation and the tradeoff between the cost of data acquisition and network training [28]. Table 1 shows the class types and the number of training samples and test samples used in the experiment when the sequence length is set to 2, 3 and 4, respectively.

Table 1. Dataset of experiment under SOC.

Class Type	Image Samples	Training			Image Samples	Testing		
		2-Aspect Sequences	3-Aspect Sequences	4-Aspect Sequences		2-Aspect Sequences	3-Aspect Sequences	4-Aspect Sequences
2S1	299	578	840	1084	274	525	755	956
BRDM2	298	576	837	1080	274	525	755	956
BTR60	256	489	709	906	195	362	508	589
D7	299	578	840	1084	274	525	755	959
T72	232	443	640	814	196	364	499	569
BMP2	233	443	642	812	195	362	494	558
BTR70	233	442	639	817	196	363	496	568
T62	299	578	840	1084	273	522	749	954
ZIL131	299	578	840	1084	274	525	755	956
ZSU234	299	578	846	1087	274	525	755	959
Total	2747	5283	7673	9852	2425	4598	6521	8024

Table 2 shows the classification confusion matrix when the input is a single-aspect image as control, and Tables 3–5 show the confusion matrix when each input sequence sample contains 2, 3 and 4 images, respectively. Confusion matrix is widely used in SAR target recognition to evaluate the recognition performance of the method. Each element of the confusion matrix represents the number of samples of each class recognized as a certain

class. The rows of the confusion matrix correspond to the actual class of the target, and the columns show the class predicted by the network.

Table 2. Confusion matrix of a single-aspect experiment under SOC.

Class	2S1	BRDM2	BTR60	D7	T72	BMP2	BTR70	T62	ZIL131	ZSU234	Acc (%)
2S1	258	1	3	0	3	0	0	8	1	0	94.16
BRDM2	0	269	1	0	1	0	0	0	0	3	98.18
BTR60	0	3	187	0	1	0	2	0	0	2	95.90
D7	0	0	0	272	0	0	0	0	2	0	99.27
T72	0	0	0	0	196	0	0	0	0	0	100.00
BMP2	0	0	2	0	5	188	0	0	0	0	96.41
BTR70	0	0	1	0	0	0	195	0	0	0	99.49
T62	1	0	0	0	0	0	0	267	3	2	97.80
ZIL131	0	0	0	2	0	0	0	3	269	0	98.18
ZSU234	0	0	0	2	0	0	0	0	0	272	99.27
Total											97.86

Table 3. Confusion matrix of a two-aspect experiment under SOC.

Class	2S1	BRDM2	BTR60	D7	T72	BMP2	BTR70	T62	ZIL131	ZSU234	Acc (%)
2S1	512	0	0	0	2	0	1	10	0	0	97.52
BRDM2	0	523	1	0	0	0	0	0	0	1	99.62
BTR60	0	2	360	0	0	0	0	0	0	0	99.45
D7	0	0	0	524	0	0	0	0	1	0	99.81
T72	0	0	0	0	364	0	0	0	0	0	100.00
BMP2	0	0	0	0	3	359	0	0	0	0	99.17
BTR70	0	0	0	0	0	0	363	0	0	0	100.00
T62	5	0	0	0	0	0	0	517	0	0	99.04
ZIL131	0	0	0	0	0	0	0	3	522	0	99.43
ZSU234	0	0	0	1	0	0	0	0	0	524	99.81
Total											99.35

Table 4. Confusion matrix of a three-aspect experiment under SOC.

Class	2S1	BRDM2	BTR60	D7	T72	BMP2	BTR70	T62	ZIL131	ZSU234	Acc (%)
2S1	739	0	0	0	0	0	0	15	0	1	97.88
BRDM2	0	755	0	0	0	0	0	0	0	0	100.00
BTR60	0	1	507	0	0	0	0	0	0	0	99.80
D7	0	0	0	751	0	0	0	0	4	0	99.47
T72	0	0	0	0	499	0	0	0	0	0	100.00
BMP2	0	0	0	0	5	489	0	0	0	0	98.99
BTR70	0	0	0	0	0	0	496	0	0	0	100.00
T62	3	0	0	0	0	0	0	746	0	0	99.60
ZIL131	0	0	0	0	0	0	0	5	750	0	99.34
ZSU234	0	0	0	1	0	0	0	0	0	754	99.87
Total											99.46

From Tables 3–5, it can be observed that the recognition rate of our proposed method with 2, 3 and 4-aspect SAR image input sequences are all higher than 99.00% under SOC in the ten-class problem. Compared with the recognition rate shown in Table 2, it is proven that the multi-aspect SAR image sequence contains more recognition information than the single-aspect SAR image. In addition, from the improvement of the recognition rate in Tables 3–5 from 99.35%, 99.46% to 99.90%, it can be concluded that the self-attention process of our proposed method can effectively extract more internal correlation information of multi-aspect SAR images, so as to improve the recognition rates with the increase in the sequence length of multi-aspect SAR image sequence samples.

Table 5. Confusion matrix of a four-aspect experiment under SOC.

Class	2S1	BRDM2	BTR60	D7	T72	BMP2	BTR70	T62	ZIL131	ZSU234	Acc (%)
2S1	951	0	0	0	0	0	0	5	0	0	99.48
BRDM2	0	956	0	0	0	0	0	0	0	0	100.00
BTR60	0	0	589	0	0	0	0	0	0	0	100.00
D7	0	0	0	956	0	0	0	0	3	0	99.69
T72	0	0	0	0	569	0	0	0	0	0	100.00
BMP2	0	0	0	0	0	558	0	0	0	0	100.00
BTR70	0	0	0	0	0	0	568	0	0	0	100.00
T62	0	0	0	0	0	0	0	954	0	0	100.00
ZIL131	0	0	0	0	0	0	0	0	956	0	100.00
ZSU234	0	0	0	0	0	0	0	0	0	959	100.00
Total											99.90

3.4. Results under EOC

Compared with SOC, the experiment under EOC is more difficult for target recognition due to the structure difference between the training set and testing set, which is often used to verify the robustness of the target recognition network. The experiments under EOC mainly include two experimental schemes, configuration variation (EOC-C) and version variation (EOC-V).

According to the original definition, EOC-V refers to targets of the same class that were built to different blueprints, while EOC-C refers to targets that were built to the same blueprints but had different post-production equipment added. The training sets under EOC-C and EOC-V are the same, which consist of four classes of targets (BMP2, BRDM2, BTR70 and T72), and the depression angle is 17°. The testing set under EOC-C consists of images of two classes of targets (BMP2 and T72) with seven different configuration variations acquired at both 17° and 15° depression angles, and the testing set under EOC-V consists of images of T72 with five different version variations acquired at both 17° and 15° depression angles. The training and testing samples for the experiment under EOC-C and EOC-V are listed in Tables 6–8.

Table 6. The training set of experiment under EOC-C and EOC-V.

Class Type	Depression Angle	Image Samples	2-Aspect Sequences	3-Aspect Sequences	4-Aspect Sequences
BMP2	17°	233	443	642	812
BRDM2	17°	298	576	837	1080
BTR70	17°	233	442	639	817
T72	17°	232	443	640	814
Total	17°	996	1904	2758	3523

Table 7. The testing set of experiment under EOC-C.

Class Type	Depression Angle	Image Samples	2-Aspect Sequences	3-Aspect Sequences	4-Aspect Sequences
T72/A04	17°&15°	573	1105	1598	2044
T72/A05	17°&15°	573	1105	1598	2050
T72/A07	17°&15°	573	1105	1604	2050
T72/A10	17°&15°	567	1092	1577	2001
T72/812	17°&15°	426	803	1133	1369
BMP2/9566	17°&15°	428	807	1145	1401
BMP2/C21	17°&15°	429	811	1143	1381
Total	17°&15°	3569	6828	9798	12,296

Table 8. The testing set of experiment under EOC-V.

Class Type	Depression Angle	Image Samples	2-Aspect Sequences	3-Aspect Sequences	4-Aspect Sequences
T72/A32	17°&15°	572	1103	1595	2046
T72/A62	17°&15°	573	1105	1604	2050
T72/A63	17°&15°	573	1105	1598	2044
T72/A64	17°&15°	573	1105	1598	2050
T72/S7	17°&15°	419	789	1116	1349
Total	17°&15°	2710	5207	7511	9539

The confusion matrices of experiments under EOC-C and EOC-V with single-aspect input images, 2, 3 and 4-aspect input sequences are summarized in Tables 9 and 10, respectively.

Table 9 shows the superior recognition performance of the proposed network in identifying BMP2 and T72 targets with configuration differences. The recognition rates of the proposed method reach 96.91%, 97.66% and 98.50% with 2, 3 and 4-aspect input sequences, respectively, which are all higher than 94.65% for the single-aspect input image. It can prove that, under EOC-C, the network can still learn more recognition information from multi-aspect images through self-attention so as to obtain better recognition performance.

Table 9. Confusion matrix of experiments under EOC-C.

Instances	Class	BMP2	BRDM2	BTR70	T72	Acc (%)	Total (%)
single-aspect	T72/A04	41	2	3	527	91.97	94.65
	T72/A05	0	1	0	572	99.83	
	T72/A07	5	0	1	567	98.95	
	T72/A10	2	0	2	563	99.29	
	T72/812	4	1	11	410	96.24	
	BMP2/9566	371	1	26	29	86.68	
	BMP2/C21	368	7	19	35	85.78	
2-aspect	T72/A04	63	5	10	1027	92.94	96.91
	T72/A05	1	0	0	1104	99.91	
	T72/A07	7	1	1	1096	99.19	
	T72/A10	0	0	0	1092	100.00	
	T72/812	1	0	0	802	99.88	
	BMP2/9566	761	3	13	30	94.30	
	BMP2/C21	735	10	7	59	90.63	
3-aspect	T72/A04	75	0	21	1502	93.99	97.66
	T72/A05	0	0	0	1598	100.00	
	T72/A07	4	0	2	1598	99.63	
	T72/A10	0	0	0	1577	100.00	
	T72/812	0	0	0	1133	100.00	
	BMP2/9566	1114	0	9	22	97.29	
	BMP2/C21	1047	1	6	89	91.60	
4-aspect	T72/A04	24	0	23	1997	97.70	98.50
	T72/A05	0	0	0	2050	100.00	
	T72/A07	1	0	2	2047	99.85	
	T72/A10	0	0	0	2001	100.00	
	T72/812	0	0	0	1369	100.00	
	BMP2/9566	1382	0	0	0	98.64	
	BMP2/C21	1266	12	0	103	91.67	

Table 10 shows the excellent performance of the proposed network in identifying T72 targets with version differences. With the increase in the input sequence length, the recognition rate of the network rises as well, from 98.12% for single-aspect input to 99.78% for four-aspect input.

Table 10. Confusion matrix of experiments under EOC-V.

Instances	Class	BMP2	BRDM2	BTR70	T72	Acc (%)	Total (%)
single-aspect	T72/A32	10	0	1	561	98.08	98.12
	T72/A62	0	0	5	569	98.95	
	T72/A63	8	2	4	558	97.38	
	T72/A64	3	2	7	561	97.91	
	T72/S7	4	0	2	410	98.32	
2-aspect	T72/A32	13	0	0	1090	98.82	99.35
	T72/A62	0	0	0	1105	100.00	
	T72/A63	11	0	2	1092	98.82	
	T72/A64	0	2	5	1098	99.37	
	T72/S7	1	0	0	788	99.87	
3-aspect	T72/A32	7	0	0	1588	99.56	99.61
	T72/A62	0	0	0	1604	100.00	
	T72/A63	19	0	0	1579	98.81	
	T72/A64	0	3	0	1595	99.81	
	T72/S7	0	0	0	1116	100.00	
4-aspect	T72/A32	8	0	0	2038	99.61	99.78
	T72/A62	0	0	0	2050	100.00	
	T72/A63	3	0	2	2039	99.76	
	T72/A64	0	8	0	2042	99.61	
	T72/S7	0	0	0	1349	100.00	

The above experiments indicate that the proposed network can achieve a high recognition rate when tested under different operating conditions, which confirms the application value of the proposed network in actual SAR ATR tasks.

3.5. Recognition Performance Comparison

In this section, our proposed network is compared with six other methods, i.e., joint sparse representation (JSR) [37], sparse representation-based classification (SRC) [38], data fusion [39], multiview deep convolutional neural network (MVDCNN) [28], bidirectional convolutional-recurrent network (BCRN) [27] and multiview deep feature learning network (MVDFLN) [29], which have been widely cited or recently published in SAR ATR. Among them, the first three are classical multi-aspect SAR ATR methods. JSR and SRC are two classical methods based on sparse representation, and data fusion refers to the fusion of multi-aspect features based on principal component analysis (PCA) and discrete wavelet transform (DWT). The last three are all deep learning multi-aspect SAR ATR methods.

Here, first, the recognition performance under SOC and EOC is compared between these methods. The recognition rates for each method under SOC and EOC are listed in Table 11. It should be noted that in order to objectively evaluate the performance of the method, it should be ensured that the datasets are as much the same as possible. Among the six methods, the original BCRN uses the image sequences with a sequence length of 15 as the input, which contains more identification information and requires a larger computational burden. That is, the original BCRN is difficult to compare with the other methods with an input sequence length of 3 or 4. Therefore, BCRN is implemented, and the results in Table 11 are obtained using the same four-aspect training and testing sequences as our method.

From Table 11, it is obvious that our method has a higher recognition rate than the other six methods in multi-aspect SAR ATR tasks, which proves that the combination of CNN and self-attention can learn the recognition information more effectively in multi-aspect SAR target recognition.

Table 11. Performance comparison between our method and other methods.

Method	SOC	Accuracy (%)	
		EOC-C	EOC-V
JSR	94.69	-	-
SRC	98.94	96.78	-
Data Fusion	98.32	-	-
MVDCNN	98.52	95.45	95.46
BCRN	99.50	97.21	98.59
MVDFLN	99.62	97.84	99.10
Our Method	99.90	98.50	99.78

Then, as shown in Table 12, compared with BCRN, our method greatly reduces the model size; that is, it greatly reduces the number of parameters and is in the same order of magnitude as MVDCNN. Considering the network structure of MVDCNN, when the sequence length increases, the network depth and the number of parallel branches will increase correspondingly. On the contrary, our method does not change the network structure when the sequence length increases, so it is more flexible, and the number of parameters increases slowly with the sequence length. As for the FLOPs, which represent the speed of network reasoning, it can be seen that our method still needs to be optimized. It is speculated that the amount of floating point operations mainly comes from the large convolution kernels for feature extraction and matrix operations for self-attention.

Table 12. Model size comparison with four-aspect input sequences.

Method	BCRN	MVDCNN	Our Method
Model Size (M)	135.25	11.49	15.94
FLOPs (G)	1.894	2.654	2.873

3.6. Discussion

For further discussion, the experiments on the network structure of feature extraction are carried out first. In order to obtain 1-D feature vectors, when a smaller convolution kernel is selected, the number of layers of CNN will increase accordingly. The recognition rates compared between the 6-layer CNN for feature extraction with the convolution kernel size of 3×3 and 3-layer CNN with the kernel size of 7×7 are shown in Table 13. From the results under EOC, it can be seen that the recognition performance of the larger convolution kernel network is better. Such a result is obtained because the larger convolution kernel can better extract the global information in raw images, which is beneficial to self-attention to learn common features in image sequences as a basis for classification. On the contrary, small convolution kernels are more concentrated on local information, which varies greatly from different angles.

Table 13. Performance comparison between different kernel sizes.

Convolution Kernel Size	SOC	Accuracy (%)	
		EOC-C	EOC-V
3×3	99.90	97.28	98.70
7×7	99.90	98.50	99.78

When the number of transformer encoder layers is different, the recognition performance of the whole network will also be different. The recognition accuracy under different operating conditions with different transformer encoder layers is shown in Table 14. When the transformer encoder has more layers, it means that the self-attention calculation has been carried out more times, so it is possible to mine more correlation information, which is also confirmed by the higher recognition accuracy achieved in the experiment.

Table 14. Performance comparison between different layers of the transformer encoder.

Number of Layers	SOC	Accuracy (%)	
		EOC-C	EOC-V
2	99.66	97.99	99.57
4	99.80	98.21	99.69
6	99.90	98.50	99.78

Next, in order to verify the recognition ability of the method with few samples, the training sequence is downsampled and the recognition rates of BCRN and MVDCNN are compared with our method when the number of training sequences is 50%, 25%, 10%, 5% and 2% of the original under SOC. As shown in Table 15, the recognition accuracy only decreases by 5% when the number of training sequence samples decreases to 2%, which is much less than 21.8% for BCRN and 11.03% for MVDCNN. As is known to all, the transformer needs a lot of data for training, which is mainly due to the lack of prior experience contained in the convolution structure, such as translation invariance and locality [30]. In our proposed method, we use CNN to extract features first so as to make up for the lost local information. Therefore, the network also has excellent performance in the case of few samples.

Table 15. Recognition performance comparison with few samples.

Method	Accuracy of Downsampling Training Samples in Different Proportions (%)					
	100%	50%	25%	10%	5%	2%
MVDCNN	99.09	98.75	98.45	97.33	95.00	87.99
BCRN	99.50	99.43	98.99	94.34	91.43	77.70
Our Method	99.90	99.71	99.70	99.21	98.11	94.90

Finally, considering that in actual SAR ATR tasks, SAR images often contain noise, which has a great impact on the performance of SAR ATR because of the sensitivity of SAR images to noise, experiments are carried out to test the anti-noise performance of the network under SOC. As shown in Figure 7, the output of input image reconstruction by convergent CAE can reflect the main characteristics of the target in the input image but blur some other details. This indicates that when the image contains noise, the feature extraction network can filter the noise and extract the main features of the target. This is proven by the test image with noise with variance from 0.01 to 0.05 and the results of its convergent CAE reconstruction shown in Figure 8.

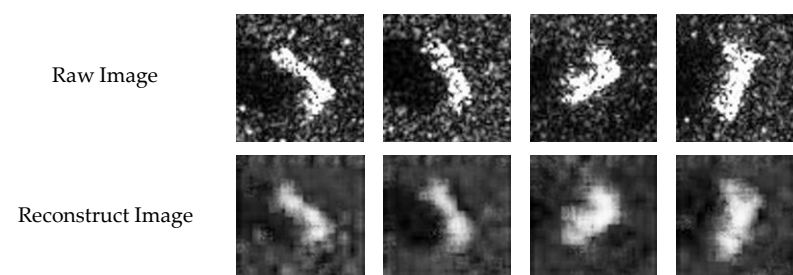
**Figure 7.** Comparison between raw image and image reconstructed by CAE.

Table 16 shows the recognition rates of the methods to be compared when the variance of noise increases from 0.01 to 0.05. Obviously, after pre-training, our method has excellent anti-noise ability. When the input image is seriously polluted by noise, the recognition rate of BCRN and the network without CAE is low, but the proposed method with CAE still maintains a high recognition rate, which shows that CAE plays an important role in anti-noise.

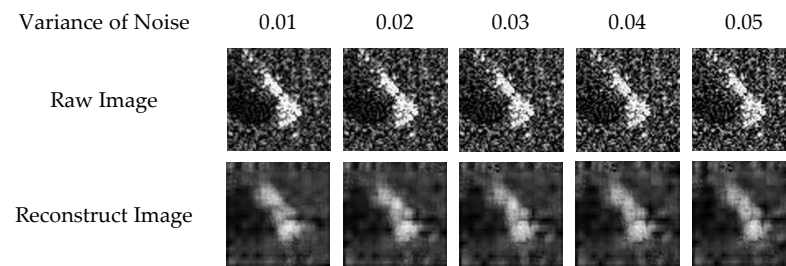


Figure 8. Comparison between raw image with noise with different variances and image reconstructed by CAE.

Table 16. Comparison of anti-noise performance.

Method	Accuracy of Different Variances of Noise Added to Testing Samples				
	0.01	0.02	0.03	0.04	0.05
BCRN	98.46	88.73	70.31	54.78	44.63
Our Method without CAE	98.19	88.97	74.00	46.73	33.97
Our Method with CAE	98.98	98.97	98.37	96.83	94.08

In addition, to further verify the effectiveness of the pre-trained CAE, as shown in Table 17, CAE is replaced by other structures for experimental comparison. DS-Net [40] in the table is a feature extraction structure composed of dense connection and separable convolution. The experimental results show that compared with other structures, the proposed method with CAE does show better recognition performance under a variety of complex conditions, which proves the advantage of CAE in extracting major features.

Table 17. Comparison between different structures for feature extraction.

Feature Extraction Structure	Accuracy under Different Cases (%)		
	SOC	2% Downsample	Noise with 0.05 Variances
Resnet18	99.19	85.41	42.33
DS-Net	97.78	88.29	55.33
CAE	99.90	97.33	94.08

4. Discussion

4.1. Advantages

From the experiments in Section 3, it can be seen that compared with the existing methods, the proposed method has achieved higher recognition accuracy under both SOC and EOC. This proves the feasibility of self-attention under various complex conditions in multi-aspect SAR target recognition.

The experiment carried out with few samples in Section 3 shows that the proposed method can still achieve a higher recognition rate than other methods. It proves the advantages of the whole method in the case of small datasets, which make the proposed method more practical, considering the high cost of radar image acquisition.

The anti-noise experimental results in Section 3 verify the advantages of the proposed method. After pre-training, the anti-noise ability of the whole method is greatly enhanced, and a high recognition rate is obtained on the testing samples containing noise. Due to the characteristics of the coherent imaging system, the radar images almost certainly contain noise, so the excellent anti-noise performance makes the method more valuable for practical application.

4.2. Future Work

The future subsequent research will mainly focus on two directions. One is to reduce floating-point operands and improve reasoning speed, which is the main disadvantage of our proposed method. In order to achieve the goal, the two main structures of the network will be optimized, namely, the CNN structure whose FLOPs can be reduced by applying separable convolution [41] and the transformer encoder that can be accelerated by pruning [42] or improving the structure [43].

The other is to further improve the recognition performance of multi-aspect SAR ATR, for which some attempts will be made to explore the combination of deep learning and the physical characteristics brought by the special imaging mechanism of SAR. The method proposed in this paper only uses the amplitude information of SAR images for network training and testing. However, the complex SAR images contain more identification information, which can be used to train deeper networks or make up for the lack of information in small datasets. It is perhaps to extract and fuse the amplitude and phase information of complex SAR images with reference to Deep SAR-Net [44] or expand the convolutional neural network to the complex domain with reference to CV-CNN [45] so as to make full use of the information contained in complex SAR images.

5. Conclusions

In this paper, a multi-aspect SAR target recognition network based on CNN and self-attention has been presented. The overall network consists of the feature extraction layers, the multi-aspect feature learning layers, feature dimensionality reduction and classification. Specifically, after pre-training by single-aspect SAR images, the encoder of CAE is transferred for feature extraction of images in the multi-aspect SAR image sequence separately, and then the internal correlation between images in the sequence is learned by self-attention. Finally, after dimensionality reduction by 1×1 convolution, the feature vectors of images in the sequence are averaged and fed into the softmax layer for classification. Experiments on the MSTAR dataset show that the proposed method can obtain satisfactory recognition performance under a variety of complex conditions. At the same time, the recognition rate in the case of few training samples can still be close to that of the complete training set. Besides that, the anti-noise ability of the whole network is greatly enhanced after pre-training.

Author Contributions: S.L. and Z.P. conceived and designed the experiments; Y.H. and Z.P. contributed materials and computing resources; S.L. performed the experiments and analyzed the data; S.L. wrote the original draft preparation; Z.P. checked the experimental data, examined the experimental results and revised the original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Youth Innovation Promotion Association, CAS under number 2022119.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://www.sdms.afrl.af.mil/>, accessed on 25 June 2022.

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

References

1. Wang, P.; Liu, W.; Chen, J.; Niu, M.; Yang, W. A high-order imaging algorithm for high-resolution spaceborne SAR based on a modified equivalent squint range model. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1225–1235. [CrossRef]
2. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
3. Bhanu, B. Automatic target recognition: State of the art survey. *IEEE Trans. Aerosp. Electron. Syst.* **1986**, *AES-22*, 364–379. [CrossRef]
4. Novak, L.M.; Owirka, G.J.; Brower, W.S.; Weaver, A.L. The automatic target-recognition system in SAIP. *Linc. Lab. J.* **1997**, *10*, 187–201.

5. Zhao, Q.; Principe, J.C. Support vector machines for SAR automatic target recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2001**, *37*, 643–654. [\[CrossRef\]](#)
6. Hou, B.; Kou, H.; Jiao, L. Classification of polarimetric SAR images using multilayer autoencoders and superpixels. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 3072–3081. [\[CrossRef\]](#)
7. Ma, W.; Wu, Y.; Gong, M.; Xiong, Y.; Yang, H.; Hu, T. Change detection in SAR images based on matrix factorisation and a Bayes classifier. *Int. J. Remote Sens.* **2019**, *40*, 1066–1091. [\[CrossRef\]](#)
8. Papson, S.; Narayanan, R.M. Classification via the Shadow Region in SAR Imagery. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 969–980. [\[CrossRef\]](#)
9. Lu, X.; Han, P.; Wu, R. Research on mixed PCA/ICA for SAR image feature extraction. In Proceedings of the 2008 9th International Conference on Signal Processing, Beijing, China, 26–29 October 2008; pp. 2465–2468.
10. Ikeuchi, K.; Wheeler, M.D.; Yamazaki, T.; Shkunaga, T. Model-based SAR ATR system. Algorithms for Synthetic Aperture Radar Imagery III. *Int. Soc. Opt. Photonics* **1996**, 2757, 376–387.
11. Kang, C.; He, C. SAR image classification based on the multi-layer network and transfer learning of mid-level representations. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1146–1149.
12. Xie, H.; Wang, S.; Liu, K.; Lin, S.; Hou, B. Multilayer feature learning for polarimetric synthetic radar data classification. In Proceedings of the 2014 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Quebec City, QC, Canada, 13–18 July 2014; pp. 2818–2821.
13. Lv, Q.; Dou, Y.; Niu, X.; Xu, J.; Xu, J.; Xia, F. Urban land use and land cover classification using remotely sensed SAR data through deep belief networks. *J. Sens.* **2015**, *2015*, 538063. [\[CrossRef\]](#)
14. Chen, S.; Wang, H.; Xu, F.; Jin, Y.Q. Target classification using the deep convolutional networks for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4806–4817. [\[CrossRef\]](#)
15. Lin, Z.; Ji, K.; Kang, M.; Leng, X.; Zou, H. Deep convolutional highway unit network for SAR target classification with limited labeled training data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1091–1095. [\[CrossRef\]](#)
16. Du, W.; Zhang, F.; Ma, F.; Yin, Q.; Zhou, Y. Improving SAR target recognition with multi-task learning. In Proceedings of the 2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Waikoloa, HI, USA, 26 September–2 October 2020; pp. 284–287.
17. Gao, F.; Huang, T.; Wang, J.; Sun, J.; Hussain, A.; Yang, E. Dual-branch deep convolution neural network for polarimetric SAR image classification. *Appl. Sci.* **2017**, *7*, 447. [\[CrossRef\]](#)
18. Shang, R.; Wang, J.; Jiao, L.; Stolkin, R.; Hou, B.; Li, Y. SAR targets classification based on deep memory convolution neural networks and transfer parameters. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2834–2846. [\[CrossRef\]](#)
19. Zhang, L.; Leng, X.; Feng, S.; Ma, X.; Ji, K.; Kuang, G.; Liu, L. Domain knowledge powered two-stream deep network for few-shot SAR vehicle recognition. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [\[CrossRef\]](#)
20. Feng, S.; Ji, K.; Zhang, L.; Ma, X.; Kuang, G. SAR target classification based on integration of ASC parts model and deep learning algorithm. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10213–10225. [\[CrossRef\]](#)
21. Wang, S.; Wang, Y.; Liu, H.; Sun, Y. Attribute-guided multi-scale prototypical network for few-shot SAR target classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12224–12245. [\[CrossRef\]](#)
22. Zhao, J.; Datcu, M.; Zhang, Z.; Xiong, H.; Yu, W. Contrastive-regulated CNN in the complex domain: A method to learn physical scattering signatures from flexible PolSAR images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10116–10135. [\[CrossRef\]](#)
23. Zhang, X.Z.; Huang, P.K. Multi-aspect SAR target recognition based on combined time-frequency feature and HMM. *Syst. Eng. Electron.* **2010**, *32*, 712–717.
24. Soumekh, M. Reconnaissance with slant plane circular SAR imaging. *IEEE Trans. Image Process.* **1996**, *5*, 1252–1265. [\[CrossRef\]](#)
25. Zhao, P.; Liu, K.; Zou, H.; Zhen, X. Multi-stream convolutional neural network for SAR automatic target recognition. *Remote Sens.* **2018**, *10*, 1473. [\[CrossRef\]](#)
26. Zhang, F.; Hu, C.; Yin, Q.; Li, W.; Li, H.C.; Hong, W. Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition. *IEEE Access* **2017**, *5*, 26880–26891. [\[CrossRef\]](#)
27. Bai, X.; Xue, R.; Wang, L.; Zhou, F. Sequence SAR image classification based on bidirectional convolution-recurrent network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9223–9235. [\[CrossRef\]](#)
28. Pei, J.; Huang, Y.; Huo, W.; Zhang, Y.; Yang, J.; Yeo, T.S. SAR automatic target recognition based on multiview deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2196–2210. [\[CrossRef\]](#)
29. Pei, J.; Huo, W.; Wang, C.; Huang, Y.; Zhang, Y.; Wu, J.; Yang, J. Multiview deep feature learning network for SAR automatic target recognition. *Remote Sens.* **2021**, *13*, 1455. [\[CrossRef\]](#)
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the NIPS 2017, Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Yang, M.; Bai, X.; Wang, L.; Zhou, F. Mixed loss graph attention network for few-shot SAR target classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [\[CrossRef\]](#)

33. Wang, C.; Huang, Y.; Liu, X.; Pei, J.; Zhang, Y.; Yang, J. Global in local: A convolutional transformer for SAR ATR FSL. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4509605. [[CrossRef](#)]
34. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
35. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
36. Ross, T.D.; Worrell, S.W.; Velten, V.J.; Mossing, J.C.; Bryant, M.L. Standard SAR ATR evaluation experiments using the MSTAR public release data set. *Algorithms Synth. Aperture Radar Imag.* **1998**, *3370*, 566–573.
37. Zhang, H.; Nasrabadi, N.M.; Zhang, Y.; Huang, T.S. Multi-view automatic target recognition using joint sparse representation. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 2481–2497. [[CrossRef](#)]
38. Ding, B.; Wen, G. Exploiting multi-view SAR images for robust target recognition. *Remote Sens.* **2017**, *9*, 1150. [[CrossRef](#)]
39. Ruohong, H.; Keji, M.; Yanjing, L.; Jiming, Y.; Ming, X. SAR target recognition with data fusion. In Proceedings of the 2010 WASE International Conference on Information Engineering, Beidai, China, 14–15 August 2010; Volume 2, pp. 19–23.
40. Shang, R.; He, J.; Wang, J.; Xu, K.; Jiao, L.; Stolkin, R. Dense connection and depthwise separable convolution based CNN for polarimetric SAR image classification. *Knowl.-Based Syst.* **2020**, *194*, 105542. [[CrossRef](#)]
41. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
42. Tang, Y.; Han, K.; Wang, Y.; Xu, C.; Guo, J.; Xu, C.; Tao, D. Patch slimming for efficient vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12165–12174.
43. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
44. Huang, Z.; Datcu, M.; Pan, Z.; Lei, B. Deep SAR-Net: Learning objects from signals. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 179–193. [[CrossRef](#)]
45. Zhang, Z.; Wang, H.; Xu, F.; Jin, Y.Q. Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7177–7188. [[CrossRef](#)]