# Multiscale Complex-Valued Feature Attention Convolutional Neural Network for SAR Automatic Target Recognition

Xiaoqian Zhou, Cai Luo, *Senior Member, IEEE*, Peng Ren, *Senior Member, IEEE*, and Bin Zhang

***Abstract*—Synthetic aperture radar (SAR) images often lack sufficient attention to target features, inadequately express target feature information, and neglect phase information in conventional Convolutional Neural Network (CNN) recognition methods. These limitations lead to reduced recognition accuracy and slower processing speeds, critical drawbacks for SAR Automatic Target Recognition (ATR) systems. To overcome these challenges, this paper proposes the multi-scale complex-valued feature attention CNN (MsCvFA-CNN) for SAR ATR. MsCvFA-CNN is a model specifically designed for amplitude and phase information of SAR images. A novel complex-valued attention module (CAM) is proposed in this work to focus on the amplitude and phase characteristics of the target separately. By decoupling the amplitude and phase features, the CAM reduces the training time of the network, while preserving the relevant information. Furthermore, the MsCvFA-CNN employs multiple branches for feature extraction with different kernel sizes, which are then combined with CAM in the fusion stage to improve the network's representation of target features. The proposed MsCvFA-CNN is evaluated on both the complex-valued moving and stationary target acquisition and recognition (MSTAR) dataset, as well as the more challenging dataset for urban interpretation (OpenSARUrban). The results demonstrate that it outperforms traditional networks in terms of recognition accuracy and computational efficiency. Specifically, the use of complex-valued networks results in a 2.23% improvement in recognition accuracy compared to traditional real-valued networks. When CAM is added, the network's accuracy is further improved by 3.21%, and the number of epochs required to achieve the highest accuracy is reduced by nearly half.**

***Index Terms*—Attention mechanism, automatic target recognition (ATR), complex-valued convolutional neural network (Cv-CNN), multiscale structure, synthetic aperture radar (SAR).**

## I. Introduction

SYNTHETIC aperture radar (SAR) is a highly advanced and sophisticated active microwave imaging sensor that employs radar technology to generate high-resolution ground images from a great distance. SAR coherently processes the echo signals, extracts surface information by phase differences, and integrates the information to form a high-resolution image with exceptional accuracy, resolution, and less susceptibility to weather conditions [1]. The SAR is extensively applied in diverse domains, such as military, civilian, and environmental monitoring owing to its distinct advantages. In the field of remote sensing big data, artificial intelligence interpretation of SAR images presents a significant scientific application challenge.

### A. Review of SAR Target Recognition

Automatic target recognition (ATR) is a crucial research field in SAR image processing. The standard SAR ATR typically comprises three stages: detection, recognition, and classification [2]. Its development dates back to the 1990s [3], with early research methods focused mainly on SAR feature extraction methods, including amplitude features [4], polarization features [5], texture features [6], and shape features [7]. The primary methods used were support vector machine [8] and K-nearest neighbors [9]. With the rise of machine learning, Emanuele et al. [10] combined their models with machine learning methods to estimate water content in SAR images. Despite the smaller number of SAR image problems due to the difficulty of SAR data acquisition and limited use, the development of SAR ATR has progressed steadily over the years. Lang et al. [11] introduced a transfer learning method to classify SAR images, which reduces the use of sample size by using pretrained convolutional neural network (CNN) models and midlevel feature extraction. Subsequently, Liu et al. [12] improved the ability of SAR target recognition by utilizing knowledge from other electromagnetic signal datasets through the use of electromagnetic feature transfer learning. These methods paved the way for deep learning (DL)-based SAR ATR.

With the improvement of computer processing power, SAR ATR method based on DL has become a hot topic. The DL algorithms have the unique ability to integrate target detection and recognition into a single step, enabling automatic feature extraction from SAR images through a multilayer convolutional network structure, thereby significantly reducing the workload involved in feature design. The proposed approach eliminates the need for prior knowledge-based feature design, thereby enhancing the detection and recognition capabilities of the model [12].

Manuscript received 10 July 2023; revised 23 November 2023; accepted 12 December 2023. Date of publication 18 December 2023; date of current version 29 December 2023. This work was supported in part by the National Key R&D Program of China under Grant 2021YFE0111600, and in part by the Fundamental Research Funds for the Central Universities under Grant 22CX01004A. *(Corresponding author: Cai Luo.)*

Xiaoqian Zhou, Cai Luo, and Peng Ren are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: zhouxiaoqian_upc@163.com; luo_cai@upc.edu.cn; pengren@upc.edu.cn).

Bin Zhang is with the Beijing Research Institute of Telemetry, Beijing 100081, China (e-mail: 13581568198@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3342986

The earliest DL applications in SAR ATR were primarily based on stacked autoencoder [13], data augmentation [14], deep belief network [15], deep convolutional neural network (DCNN) [16], and generative adversarial networks [17]. Since then, DL has made notable advancements in SAR image processing, with the potential to revolutionize the field.

### B. Problems and Solutions of CNN Methods

Recent studies have demonstrated that CNN-based SAR image processing has achieved remarkable performance in tasks, such as land classification, change detection, and vehicle identification [18]. Nevertheless, there remain several challenges when applying CNN in the SAR field. The CNN-based SAR ATR methods mentioned previously only use image data that is real-valued and primarily concentrate on the representation of target features in the amplitude domain. However, SAR images are complex-valued data that encompass both amplitude and phase information [19]. This is because SAR images are formed by considering the electromagnetic vector sum of the interaction between the electromagnetic wave and the target or scene [20]. The SAR images contain rich electromagnetic scattering information of terrain and targets, such as texture, shape, and reflectivity [21]. Therefore, the phase component of SAR images plays a significant role in enhancing detection performance, particularly in terms of geometric shape [22]. In addition, Fu et al. [23] validated that multiscale CNNs outperform traditional single-scale CNNs in target recognition. Furthermore, Woo et al. [24] recommended that integrating attention mechanisms into CNNs could enhance the model's interpretability, improve its recognition accuracy and computational efficiency. However, ordinary attention mechanisms are not suitable for SAR images due to their complex-valued nature, which incorporates phase information.

To overcome these challenges, several solutions have been proposed. Since real-valued CNN cannot extract rich phase information from SAR images, El-Darymli et al. [25] modeled SAR images based on phase information characterization. Tygert et al. [26] developed a mathematical motivation for the complex-valued CNN (Cv-CNN) structure, while Wilmanski et al. [27] used a complex-valued convolutional layer in the first layer and real-valued calculations for the subsequent layers, achieving a recognition performance improvement from 87.3% to 99.21%. Zhang et al. [28] extended the use of complex-valued neural networks (Cv-NNs) to polarimetric SAR ATR, which has been shown to reduce recognition errors when compared to traditional real-valued CNN. Moreover, multiscale CNNs have been proposed to enrich feature representations [29]. Zeng et al. [30] pioneered the use of full Cv-CNN in SAR ATR, achieving a remarkably high recognition rate in target identification. Nonetheless, the full complex-valued computation poses challenges, such as high computational cost, redundant feature extraction, and prolonged computation time. To improve computational efficiency, Gao et al. [31] employed an attention model to improve the attention of CNNs on crucial target features. However, the phase information of SAR targets remains underutilized.

### C. Our Novel Contributions

There is still considerable work to be done to improve the application of Cv-CNNs in SAR ATR. Since the computation of Cv-CNNs is about twice as large as that of ordinary CNNs, efficient and accurate Cv-CNNs become the focus of this article. Inspired by previous work [30], [32], the multiscale complex-valued feature attention CNN (MsCvFA-CNN), a multiscale complex-valued feature attention network, is proposed to address the issues of incomplete feature representation, missing phase information, and slow processing speed in SAR target recognition. The network demonstrates excellent performance in experiments, as detailed in the following.

First, this article proposes a novel complex-valued channel attention module, called CAM, which simultaneously focuses on both amplitude and phase features of SAR targets and solves the problem of increased channel numbers due to cascade features. In addition, the CAM enhances the network's robustness by enabling it to prioritize important features even in small sample scenarios.

Next, a multiscale complex-valued feature extraction and fusion module (MEFM) is proposed, which consists of branches using convolution kernels of varying sizes to extract complex-valued features. The MEFM is combined with CAM to enhance the network's ability to express complex-valued SAR target features comprehensively.

Finally, considering the distinctive imaging mechanism of SAR that encompasses amplitude as well as phase components, the proposed network modules employ complex-valued calculations.

To provide a clear structure, this article is organized as follows. Section II introduces the computational background of Cv-NNs. Section III elaborates on the framework of the MsCvFA-CNN and explains each module in detail. The experimental results on the complex-valued moving and stationary target acquisition and recognition (MSTAR) and OpenSARUrban datasets, as well as the comparison and analysis of the results, are described in Section IV. Finally, Section V concludes this article.

## II. Background Knowledge of Cv-CNNs

The Cv-NN is a type of neural network that uses complex-valued parameters and variables for identifying targets from complex-valued input data. Input data is represented as $Z = A \cdot e^{i\theta} = a + ib$, where $A$ denotes the amplitude of the signal and $\theta$ denotes the phase. Cv-NN research can be traced back to the 1970s when Aizenberg et al. [33] proposed the possibility of using complex-valued data to process information and constructed Cv-NNs by extending the saturated output value of neurons to the unit circle in the complex-valued domain.

The SAR images not only contain amplitude information of the echo signal but also record phase information, which can be written in the form of $Z = A \cdot e^{i\theta} = a + ib$. This study aims to improve SAR target recognition accuracy by fully utilizing the phase information in images through the use of a complex-valued network.

The input data are assumed to be a complex-valued vector $X = (x_1, x_2, \ldots, x_n)^T$, where the first term $x_0 = -1 - j$

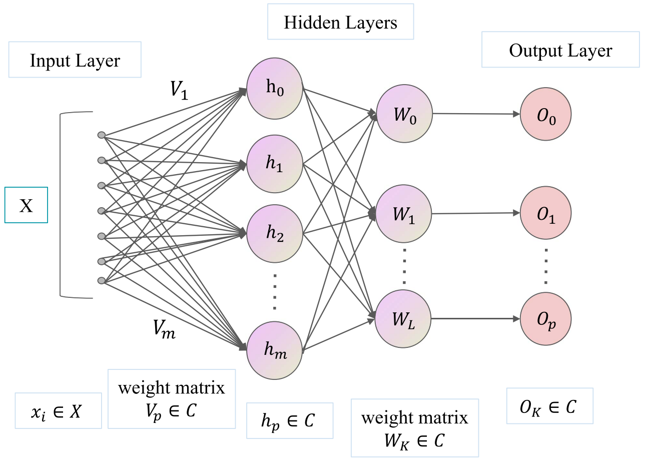


Fig. 1. Basic architecture of Cv-NN.

imports the bias term $\theta$ into the hidden layer. The output complex-valued vector of the hidden layer is denoted as $H = (h_1, h_2, \ldots, h_m)^T$, with the first element $h_0 = -1 - j$ providing the bias term λ for the output layer. The output layer produces the result vector $O = (o_1, o_2, \ldots, o_p)^T$, while the label vector is denoted by $L = (l_1, l_2, \ldots, l_p)^T$, where $\{x_i, y_i, o_i, l_i\} \in C$, with $C$ representing the complex-valued field. The weight matrix is expressed by $V = (V_1, V_2, \ldots, V_p, \ldots, V_m)$ describes the weight connections between the input and hidden layers. Each neuron in the hidden layer is associated with a weight vector $V_p$, with each $V_p$ having a length of $n$. Thus, the neurons in the hidden layer are directly linked to each neuron in the input layer via the weight matrix $V$. The basic complex-valued network calculation module employed in this work is based on [28], and the next subsections provide detailed explanations of each part's computational aspects.

### A. Complex-Valued Convolution Layer

The architecture of Cv-NN, which uses complex-valued forms for the input data, bias data, and hyperparameters, can be seen in Fig. 1. During forward computation, the nonlinear function operates on the amplitude and phase components of the input and weight elementwise, producing complex-valued output. In back propagation, the stochastic gradient descent (SGD) algorithm operates on complex-valued forms by multiplying elementwise.

Inspired by Cv-NN, the MsCvFA-CNN architecture is based on complex-valued convolution operation. Convolution computation can automatically extract deep features of the target. The neurons in the convolution layer use a set of filters (weight matrix $V$) to perform convolution calculations with some neurons in the previous layer. This convolution structure exploits the shift-invariance and spatial correlation of target features [34], where shift-invariance refers to the property of the convolution operation that it produces the same result regardless of the position of the input features in the receptive field, and spatial correlation refers to the statistical dependence between nearby features in the image. To reduce redundancy, a set of weight matrices is shared among all receptive fields on the same layer during convolution operations, while different feature maps employ distinct weight matrices. The weight matrix $V$ is designed to identify specific features of the input data, enabling each feature map in the preceding layer to represent a unique feature at different positions.

The $l$th layer's feature map is represented by $H_f^{(l)} \in C^{w_2 \times H_2 \times I}$. It is convolved with the previous layer feature map $H_f^{(l-1)} \in C^{w_1 \times H_1 \times K}$ and the convolution kernel $w_{ki}^{(l)} \in C^{F \times F \times K \times I}$. The result is then added with the bias $b_{li} \in C^l$ and finally subjected to a nonlinear function $f(\cdot)$. The calculation equation is given as follows:

$$\begin{aligned}
V_p^{(l)} &= \sum_{k=1}^{k} H_f^{(l-1)} * w_{ki}^{(l)} + b_i^{(l)} \\
&= \sum_{k=1}^{k} \left\{ \left[ \Re\left(Y_k^{(l-1)}\right) \cdot \Re\left(w_{ki}^{(l)}\right) - \Im\left(H_f^{(l-1)}\right) \right.\right. \\
&\quad \left. \Im\left(w_k^{(l)}\right) \right] + j \left[ \Re\left(H_f^{(l-1)}\right) \cdot \Im\left(w_{ki}^{(l)}\right) \right. \\
&\quad \left.\left. + \Im\left(H_f^{(l-1)}\right) \cdot \Re\left(w_{ki}^{(l)}\right) \right] \right\} + b_i^{(l)} \qquad (1)
\end{aligned}$$

$$\begin{aligned}
H_f^{(l)} &= f\left\{ \Re\left[V_p^{(l)}\right] \right\} + jf\left\{ \Im J\left[V_p^{(l)}\right] \right\} \\
&= \frac{1}{1 + e^{-\Re\left[v_p^{(l)}\right]}} + j\frac{1}{1 + e^{-\Im\left[v_p^{(l)}\right]}} \qquad (2)
\end{aligned}$$

in which, the symbol * denotes the convolution calculation. $V_p^{(l)}$ denotes the weight matrix of the $i$th feature map in the $l$th layer. $H_f^{(l-1)}$ denotes the $k$th feature map of the $l-1$th layer, and $f(\cdot)$ represents the nonlinear function.

### B. Complex-Valued Pooling Layer

The pooling layer is a crucial element in CNNs that consolidates similar features identified by the convolutional layer. This operation reduces the dimensionality of the convolutional results, which is commonly referred to as downsampling [35]. The pooling operation divides the feature map derived from convolution into various local regions, and the pooling function computes the statistical characteristics of each region. In the MsCvFA-CNN, two pooling methods are used: CvMax-Pool and CvAvg-Pool, which perform maximum and average pooling operations in the complex-valued domain.

When executing the CvMax-Pool operation, it is worth noting that, unlike real-valued CNNs, the maximum value of the real or imaginary part features cannot be simply selected in the complex-valued network, otherwise, it will cause the problem of mismatch between amplitude and phase. This mismatch can significantly hinder the model's ability to recognize SAR targets effectively by accurately matching their amplitude and phase information. To tackle this issue, we calculate the position coordinates of the maximum amplitude value in a local area, which represents the maximum amplitude information. Then, we employ the position coordinates to obtain the corresponding phase information. Finally, the CvMax-Pool result is obtained by selecting the maximum value in the region and its corresponding phase information. The calculation process can be described as follows.

Assuming that feature map for $l$th layer is $H_{(x)}^{(l)} = \Re(x) + j\Im(x)$, the output result of the maximum pooling operation is

$$\mathcal{M}_{(x,y)}^{(l)} = \text{MaxPooling}\left\{\left|H_{(x)}^{(l)}\right|\right\} \tag{3}$$

where $|\cdot|$ represents the modular calculation of complex-valued characteristics. Next, the position coordinates of the output result

$$P(x,y) \longleftarrow \text{coor}\left\{\mathcal{M}_{(x,y)}^{(l)}\right\}. \tag{4}$$

The coordinate of the target location on the original feature map resulting from the CvMax-Pool calculation is denoted by coor. The pooling outcomes of the target's real and imaginary parts, determined from the aforementioned position coordinates, are represented by

$$\begin{aligned} \Re'(x) &= \Re(P(x,y)) \\ \Im'(x) &= \Im(P(x,y)). \end{aligned} \tag{5}$$

From the abovementioned calculation formula, the result of CvMax-Pool is

$$\text{CvMax\_}H_{(x)}^{(l)} = \Re'(x) + j\Im'(x). \tag{6}$$

Similarly, as the average pooling operation calculates the average of both the amplitude and phase of the local region, there is no risk of encountering a mismatch in amplitude and phase characteristics. Regarding CvAvg-Pool, we compute the mean values of the SAR target's real and imaginary parts separately, as demonstrated in the following:

$$\text{CvAvg\_}H_{(x)}^{(l)} = \text{AvgPooling}\{\Re(x)\} + j\,\text{AvgPooling}\{\Im(x)\}. \tag{7}$$

### C. Complex-Valued Fully Connected (Cv- FC) Layer

The final layer of the network is a fully connected layer, in which neurons are arranged in a linear structure and connected to the previous layer's neurons through a weight matrix. The fully connected layer is commonly utilized as a classifier in neural networks, owing to its efficient feature extraction capabilities. In the MsCvFA-CNN, the Cv-FC layer considers both the real and imaginary characteristics of targets, which distinguishes it from real-valued CNNs. This is achieved through the use of a specific formula, as shown in the following:

$$O_k^{(l)} = \omega_k^{(l)} * H_k^{(l-1)} + b_i^{(l)}. \tag{8}$$

### D. Complex-Valued Activation Function

Activation functions increase the learning capacity of models by introducing nonlinearity. As such, they are critical components in the architecture of these networks. In the network, we use the complex-valued rectified linear unit (Cv-ReLU) as the activation function. This function is particularly effective in addressing the problem of gradient instability and is widely used in DL. To design the Cv-ReLU, it is important to consider the amplitude and phase relationship of the SAR targets. Therefore, we apply the Cv-ReLU separately to the real and imaginary parts, using the following calculation:

$$\text{ReLU}(\Re(x)) = \begin{cases} \text{Re}(x) & \text{Re}(x) \geq 0 \\ 0 & \text{Re}(x) < 0 \end{cases} \tag{9}$$

$$\text{Cv\_ReLU}(x) = \text{ReLU}(\Re(x)) + j\,\text{ReLU}(\Im(x)). \tag{10}$$

### E. Complex-Valued Domain Back Propagation Algorithm and Weight Update

The supervised training method employed by CNNs minimizes the error between predicted and true values to obtain optimal weight matrices and bias parameters. The SGD method is commonly used for adjusting network parameters to minimize the error or loss function [36], allowing the error to gradually approach 0 [37]. The cross-entropy (CE) function is a widely used loss function $\mathcal{L}_{\text{CE}}$ in neural networks [38]. Assuming that the training dataset can be represented as $\{I_{[s]}, L_{[s]}\}_{s=1}^{S}$, the predicted label data output from the network can be represented as $\{P_{[s]}\}_{s=1}^{S}$, where $S$ is the total number of training samples. Here, $I_{[s]}$, $L_{[s]}$, and $P_{[s]}$ represent the input data, label data, and predicted label data of the $s$th training sample, respectively. As the input data are in the complex-valued domain, the computation of the real part and imaginary part for the predicted data can be expressed as follows:

$$|\text{P}| = \sqrt{\Re(P)^2 + \Im(\text{P})^2}. \tag{11}$$

The loss function can be calculated using the following formula:

$$\mathcal{L}_{\text{CE}} = \sum_{s=1}^{S} \left|P_{[s]}\right| \cdot \log \frac{\left|P_{[s]}\right|}{L_{[s]}}. \tag{12}$$

The iterative process of adjusting weights and biases is employed to minimize the aforementioned loss function

$$\begin{aligned} w_{ki}^{(l)}(T) &= w_{ki}^{(l)}(T-1) + \Delta w_{ki}^{(l)}(T-1) \\ &= w_{ki}^{(l)}(T-1) - \eta \frac{\partial \mathcal{L}}{\partial w_{ki}^{(l)}(T-1)} \end{aligned} \tag{13}$$

$$\begin{aligned} b_i^{(l)}(T) &= b_i^{(l)}(T-1) + \Delta b_i^{(l)}(T-1) \\ &= b_i^{(l)}(T-1) - \eta \frac{\partial \mathcal{L}}{\partial b_i^{(l)}(T-1)} \end{aligned} \tag{14}$$

where $w_{ki}^{(l)}(T)$ and $b_i^{(l)}(T)$ are the weight and bias, respectively, used for the $T$th iteration update, while $\eta$ represents the learning rate for the SGD optimizer. In the complex-valued field, the back propagation algorithm also follows the chain rule approach to derivation, which gives the following formulas for the partial derivatives of the weight and bias:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{CE}}}{\partial w_{ki}^{(l)}} &= \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \Re\left[w_{ki}^{(l)}\right]} + \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \Im\left[w_{ki}^{(l)}\right]} \\ &= \left\{ \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \Im\left[V_i^{(l)}\right]} \frac{\partial \Re\left[V_i^{(l)}\right]}{\partial \Re\left[w_{ki}^{(l)}\right]} + \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \Im\left[V_i^{(l)}\right]} \frac{\partial \Im\left[V_i^{(l)}\right]}{\partial \Re\left[w_{ki}^{(l)}\right]} \right\} \end{aligned}$$

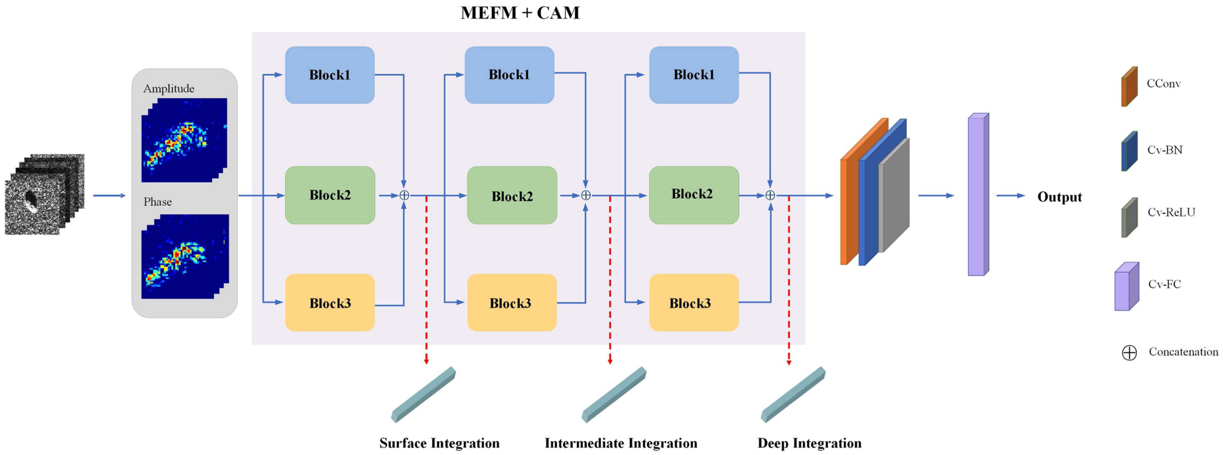


Fig. 2. Fundamental structure of MsCvFA-CNN.

$$+j\left\{\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Re\left[V_i^{(l)}\right]}\frac{\partial\Re\left[V_i^{(l)}\right]}{\partial\Im\left[w_{ki}^{(l)}\right]}+\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Im\left[V_i^{(l)}\right]}\frac{\partial\Im\left[V_i^{(l)}\right]}{\partial\Im\left[w_{ki}^{(l)}\right]}\right\} \quad (15)$$

$$\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial b_i^{(l)}}=\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Re\left[b_i^{(l)}\right]}+\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Im\left[b_i^{(l)}\right]}$$

$$=\left\{\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Re\left[V_i^{(l)}\right]}\frac{\partial\Re\left[V_i^{(l)}\right]}{\partial\Re\left[b_i^{(l)}\right]}+\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Im\left[V_i^{(l)}\right]}\frac{\partial\Im\left[V_i^{(l)}\right]}{\partial\Re\left[b_i^{(l)}\right]}\right\}$$

$$+j\left\{\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Re\left[V_i^{(l)}\right]}\frac{\partial\Re\left[V_i^{(l)}\right]}{\partial\Im\left[b_i^{(l)}\right]}+\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial\Im\left[V_i^{(l)}\right]}\frac{\partial\Im\left[V_i^{(l)}\right]}{\partial\Im\left[b_i^{(l)}\right]}\right\} \quad (16)$$

where $V_i^{(l)}$ denotes the weight matrix of the $i$th feature map of $l$th layer.

## III. MsCvFA-CNN Classification Methodology

The overall structure of MsCvFA-CNN is shown in Fig. 2. This network takes complex-valued SAR images containing both amplitude and phase information as input, and subsequently conducts feature extraction and feature attention through the MEFM and CAM modules, with specific details elaborated as follows.

### A. Multiscale Feature Extraction and Fusion Module

As a deep neural network grows deeper, it becomes more proficient in feature representation and nonlinear fitting. However, the limited size of existing SAR datasets presents a challenge as utilizing a deep network with such datasets may result in overfitting. Consequently, it is essential to explore other approaches to improve feature representation capabilities. The convolution kernel size is a crucial hyperparameter in the convolution operation as it determines the scale of the resulting feature map. Smaller kernel sizes are suitable for extracting local rich texture feature information, while larger kernel sizes are more appropriate for capturing global contour features. However, traditional CNNs employing fixed-size convolution kernels may discard important local or global feature information, leading to suboptimal classification outcomes.

Inspired by the GoogLeNet [39] and MDL-RS fusion framework [40] models, this article proposes MsCvFA-CNN, employing a parallel network structure called MEFM. MEFM consists of three cascaded branches, each extracting features from multiple channels using different kernel sizes. This design provides different receptive fields at the same layer, enabling the extraction of multiscale features of SAR targets and enhancing the network's feature representation capability. The network structure of MEFM is illustrated in Fig. 3, with specific details provided as follows.

Initially, SAR images are processed via complex-value computations to obtain the real and imaginary components. These real and imaginary parts are then simultaneously passed through three branches with convolutional kernels of sizes 3 × 3, 7 × 7, and 11 × 11 to extract target features. The resulting real and imaginary feature maps are respectively denoted as Feature_r and Feature_i. Each branch focuses on learning features from different local regions of varying scales. This enables the network to pay attention to features of different sizes, thereby enhancing the network's multiscale feature representation capability. The Feature_r and Feature_i obtained from the three branches are resized to the same dimensions and concatenated together to form new complex-valued features. This feature extraction strategy is repeated three times. The complex-valued feature maps obtained from these three rounds of feature extraction and fusion are termed shallow features maps, middle features maps, and deep features maps, respectively. Through this multiscale, multichannel feature extraction and fusion approach, the network can better capture the complex features of targets in SAR images, improving the accuracy and robustness of target recognition.

### B. Complex-Valued Attention Module

The convolution of SAR targets at multiple scales generates numerous amplitude and phase features. However, these

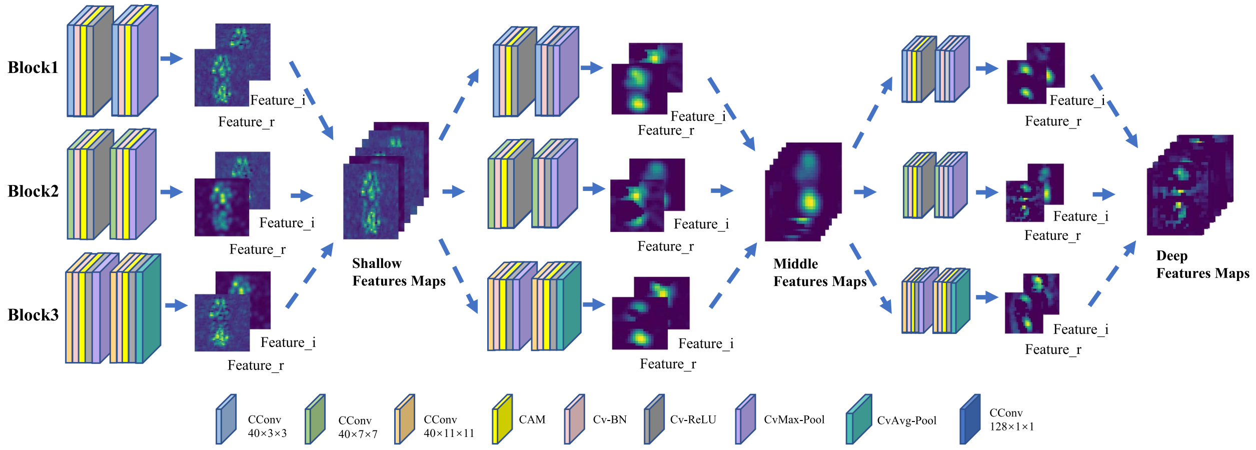


Fig. 3. Architecture of the proposed MEFM.

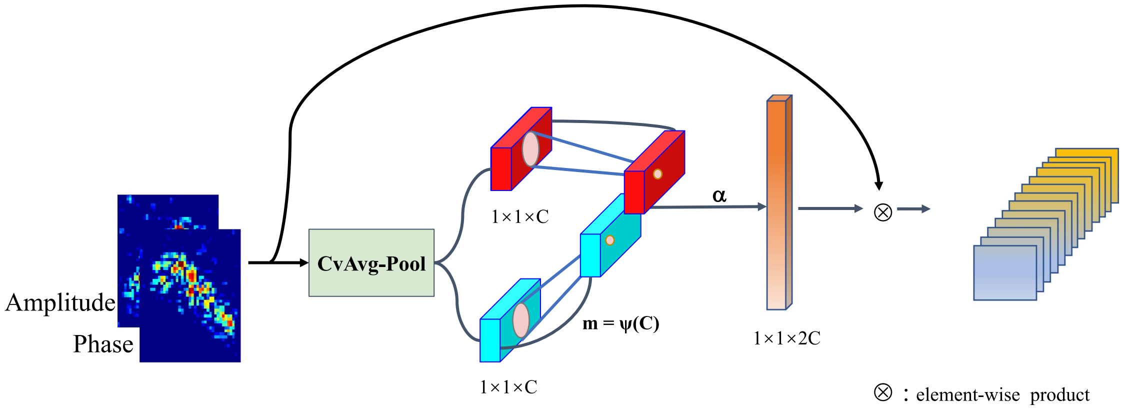


Fig. 4. Network structure of the CAM module.

unprocessed multiscale features often contain many redundant and irrelevant features, which can lead to insufficient focus on salient information. To address this issue, we introduce a complex-valued channel attention module for SAR ATR, which aims to enhance the extraction of salient features and suppress irrelevant ones. Unlike the ordinary real-valued attention module, CAM is designed for complex-valued SAR images. This method attends to both the amplitude and phase information simultaneously, thereby enhancing the network's utilization of phase information. Due to the cascaded feature fusion process, the number of channels in the feature maps is doubled. Therefore, CAM adopts the channel attention mechanism to overcome the performance-complexity contradiction, involving only a few parameters. By simply adding the computationally efficient CAM, the network can achieve significant performance improvement.

Fig. 4 illustrates the network architecture of CAM. We perform global complex-valued average pooling without reducing the channel dimension, allowing local cross-channel interaction to be considered for each channel and its corresponding $m$ neighborhoods. The kernel size $m$ represents the range of local coverage for cross-channel interaction, which determines the size of the neighborhood considered when predicting the attention of a channel. As SAR images are complex-valued data, we make the interactive coverage of the phase channel equal to that of the magnitude. Thus, we only need to calculate the size of $m$ for the amplitude information of the images. To avoid manually adjusting the value of $m$, we employ an adaptive method to determine the size of $m$. Specifically, we use the banded matrix $W_m$ to learn channel attention, which can be expressed as follows:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,m} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,m+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-m+1} & \cdots & w^{C,C} \end{bmatrix}. \tag{17}$$

For the convenience of calculation, all channels share the same learning parameters, namely

$$w^{p,q} = \sigma\left(\sum_{j=1}^{\mathrm{m}} w_j y_j^{(l)}\right), y_j^{(l)} \in C_m^{(l)} \tag{18}$$

where $m$ represents the coverage of cross-channel interaction, $w_j$ denotes the number of channels, $y_j^{(l)}$ denotes the $j$th neuron

**Algorithm 1:** Computational process of CAM.

```
Input: $C$ = channel dimension of the feature map;
  default: λ = 2, b = 1.
Output: channel complex-valued attention weight matrix
  $\boldsymbol{W}_k^r + j\boldsymbol{W}_k^i$.
Initialize the band matrix $\boldsymbol{W}_k^r$, $\boldsymbol{W}_k^i$;
Compute the coverage of cross-channel interaction $m$
  by (21);
Enter the feature diagram extracted from the previous
  layer to the complex-valued average pooling layer;
Perform one-dimensional complex-valued
  convolution;
Then use the Sigmoid function and obtain $\boldsymbol{y}_i^{r,j}$, $\boldsymbol{y}_i^{i,j}$;
Compute (18) and obtain the band matrix $\boldsymbol{W}_k^r$, $\boldsymbol{W}_k^i$;
Return $\boldsymbol{W}_k^r + j\boldsymbol{W}_k^i$.
```

in the $l$th layer, and $C_m^{(l)}$ represents the set of $m$ adjacent channels of $y^{(l)}$.

Next, we need to determine the coverage range of cross-channel interaction (that is, 1-D convolution kernel size $m$). The most common approach is to manually adjust the coverage of interactions for the number of channels in different CNN architectures, but this method consumes a significant amount of computational resources. Inspired by the idea that group convolution can improve the CNN architecture [41], we hypothesize that there may be a mapping $\phi$ between the coverage range $m$ of cross-channel interaction and the channel dimension $C$

$$C = \phi(m). \tag{19}$$

The simplest mapping relationship can be represented by a linear function, that is, $\phi(m) = \lambda \cdot m - b$, but the relationship that a linear function can describe is limited. According to the channel dimension $C$, it is generally set to the power of 2. Therefore, extending the function $\phi(m)$ to a nonlinear function can provide a more effective solution.

Then the following relationship:

$$C = \phi(m) = 2^{(\lambda \cdot m - b)} \tag{20}$$

when the size of the channel dimension $C$ is given, the interaction coverage $m$ can be determined by the following formula:

$$m = \psi(C) = \left| \frac{\log_2(C)}{\lambda} + \frac{b}{\lambda} \right|_{\text{odd}} \tag{21}$$

where $|\cdot|_{\text{odd}}$ denotes the nearest odd number. Based on experimental results, it has been observed that when $\lambda = 2$ and $b = 1$, the high-dimensional channels have longer interaction coverage, while the low-dimensional channels have shorter interaction distance.

The detailed process of the CAM module is presented in Algorithm 1. The module quickly generates channel attention through 1-D convolution, and nonlinear mapping adaptively terminates the range of cross-channel interaction. With the introduction of CAM, the channelwise filtering of SAR target features is enhanced, allowing the model to focus more on the critical complex-valued features of the target and suppress background interference, leading to more accurate target localization and reduced confusion.

The high recognition accuracy of MsCvFA-CNN for SAR targets is mainly due to three factors as follows.

1) *Cv-CNN:* The network is specifically designed for complex-valued SAR images. All the functions and modules used in the network utilize both the amplitude and phase information of the image. Different from real-valued CNNs, the network increases the utilization of phase information, thereby improving the recognition performance.
2) *MEFM:* The MEFM uses convolution kernels of various sizes to extract complex-valued information from SAR targets. The amplitude and phase characteristics of the target are fused in a cascade manner. It can enhance the model's ability to represent features and fit nonlinear functions.
3) *CAM:* CAM is an attention mechanism designed specifically for complex-valued images. Adding CAM to the network helps to improve attention to the crucial amplitude and phase features of the target. CAM overcomes the contradiction between performance and computational complexity, improving the network's attention to target features.

From the perspective of SAR imaging mechanism, MsCvFA-CNN is a target recognition framework specially designed for SAR images.

## IV. Experiments and Analysis

This section mainly verifies the effectiveness of the proposed MsCvFA-CNN in SAR ATR. We use the measured data to conduct experiments, and propose experimental comparison, analysis, discussion. All experiments are conducted on an Intel Xeon Silver 4208 CPU and Nvidia GeForce GTX 3090 GPU. As the proposed network is specifically designed for SAR ATR, our primary objective is to achieve accurate and fast target identification. Therefore, we evaluate the network using the probability of correct recognition (PCR) and the epoch at the highest recognition rate. The PCR is calculated by dividing the number of correctly identified targets ($N_c$) by the total number of targets ($N_t$), as shown in the equation

$$\text{PCR} = \frac{N_c}{N_t} \times 100\%. \tag{22}$$

To demonstrate the superior performance of the proposed network, we compare it with three widely recognized SAR ATR algorithms: A-ConvNets [42], which is a traditional real-part DCNN; Re_MsCvFA-CNN, which is a real-valued network with the same architecture as the proposed MsCvFA-CNN in this article; and MS-CVNets [30], which is a Cv-NN.

### A. Experimental Data Description

The SAR images utilized in this study were obtained from the MSTAR dataset, which comprises SAR images of ten distinct types of military ground vehicles. These vehicles vary in target type, azimuth angles, depression angles, and shape configurations. The types of vehicles included in the dataset are: armored vehicles, tanks, rocket launcher, antiaircraft unit,

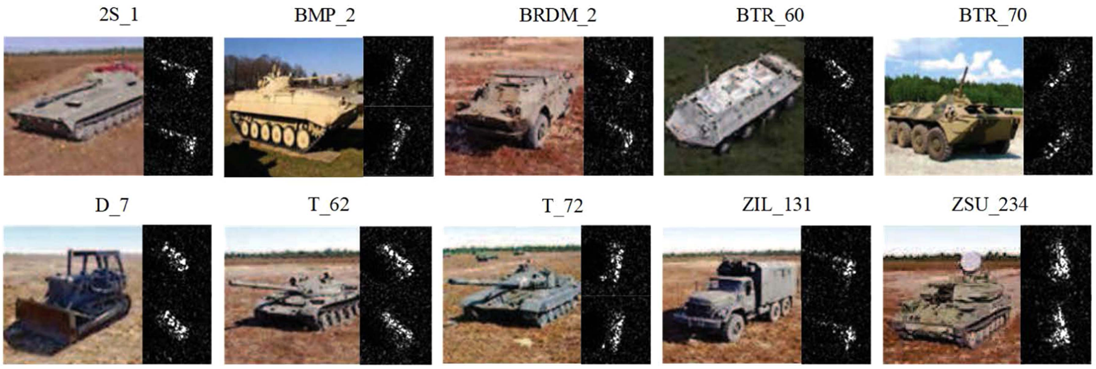


Fig. 5. MSTAR dataset contains ten types of vehicle target instances. Left: Optical images. Right: SAR image S (top right: Amplitude; bottom right: Phase).

TABLE I
Detailed Configuration of the MSTAR Dataset Used in the SOC-10 Experiment

| Class | Serial no. | Train | | Test | |
|---|---|---|---|---|---|
| | | Depression | Quantity | Depression | Quantity |
| 2S_1 | B_01 | 17° | 299 | 15° | 274 |
| BMP_2 | 9563 | 17° | 233 | 15° | 196 |
| BRDM_2 | E_71 | 17° | 298 | 15° | 274 |
| BTR_60 | K_10yt7532 | 17° | 256 | 15° | 195 |
| BTR_70 | C_71 | 17° | 233 | 15° | 196 |
| D_7 | 92v13015 | 17° | 299 | 15° | 274 |
| T_62 | A_51 | 17° | 299 | 15° | 273 |
| T_72 | 132 | 17° | 232 | 15° | 196 |
| ZIL_131 | E_12 | 17° | 299 | 15° | 274 |
| ZSU_234 | D_08 | 17° | 299 | 15° | 274 |

military truck, bulldozer. The MSTAR dataset has been commonly used as a standard for assessing and comparing the effectiveness of various SAR ATR algorithms. However, the current state-of-the-art ATR method using CNNs [32], [43] typically only use the amplitude component of the MSTAR dataset, and do not fully exploit the phase component. To fully exploit both magnitude and phase parts of data, we employ the complex-valued MSTAR dataset compiled by Zeng et al. [30]. They partitioned the MSTAR dataset into amplitude and phase parts. Fig. 5 on the left shows optical images of ten different types of ground military vehicles, while on the right are the corresponding SAR amplitude and phase components. We comprehensively evaluate the performance of the MsCvFA-CNN for SAR ATR using complex-valued datasets under both standard 10-class operating conditions (SOC-10) and two types of extended operating conditions (EOC), providing a comprehensive performance evaluation.

### *B. Experimental Results Under Standard Operating Conditions (SOC)*

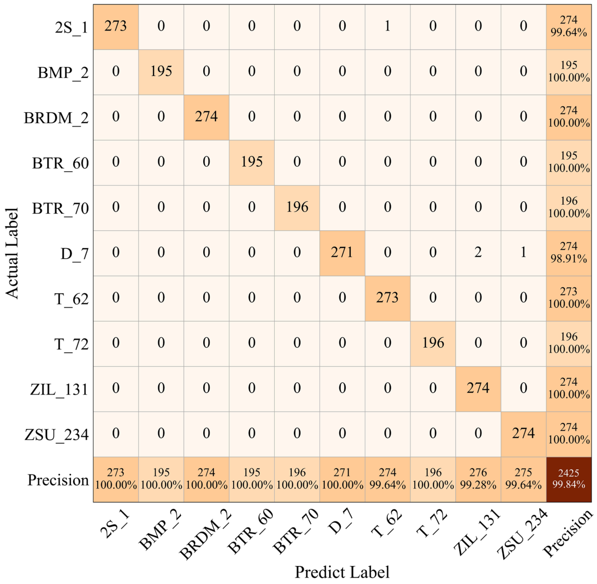
In the MSTAR dataset, SOC refers to the scenario where the target shapes and configurations in the training set are similar to those in the test set, but the imaging is performed at different azimuth and depression angles. Under SOC, we test the results of the proposed network for ten target classifications. Table I summarizes the target model, imaging depression angles, and sample sizes for SOC-10. The training dataset comprises 2743 SAR images, while the test set comprises 2431. In this experiment, training images were taken at 17°depression angle, while test images were taken at 15°depression angle. We use SAR images with a size of 64 × 64.

Fig. 6 and Table II demonstrate the recognition performance of different networks on the SOC-10 of the MSTAR dataset. Fig. 6 shows the confusion matrix of MsCvFA-CNN, which achieved an overall target recognition accuracy of 99.84%. The recognition accuracy for eight target types was 100%, while the target 2S_1 has a recognition accuracy of 99.64%, the target D_7 has a recognition accuracy of 98.91%. Fig. 7 compares the recognition accuracy of four different networks at different epochs. A-ConvNets, which performs data augmentation during training, achieves the highest accuracy of 93.24% at epoch 39. Re_MsCvFA-CNN achieves the highest accuracy of 96.12% at epoch 125, while MS-CVNets achieves the highest accuracy of 99.79% at epoch 280. MsCvFA-CNN achieves the highest accuracy of 99.84% at epoch 158, indicating that it achieves the highest accuracy with the least amount of training time. The results in Table II indicate that MsCvFA-CNN has the highest

TABLE II
RECOGNITION ACCURACY OF FOUR DIFFERENT METHODS FOR TARGETS IN SOC-10 (%)

| Models | 2S_1 | BMP_2 | BRDM_2 | BTR_60 | BTR_70 | D_7 | T_62 | T_72 | ZIL_131 | ZSU_234 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A-ConvNets | 83.94 | 83.08 | 91.61 | 91.79 | 98.47 | 97.08 | 94.87 | 98.98 | 97.08 | 95.26 | 93.24 |
| Re_MsCvFA-CNN | 94.16 | 98.60 | 98.77 | 97.95 | 90.33 | 98.56 | 94.18 | 94.90 | 95.20 | 98.55 | 96.12 |
| MS-CVNets | 99.30 | 100 | 99.60 | 99.50 | **100** | 100 | 99.60 | 100 | 100 | 100 | 99.79 |
| MsCvFA-CNN | **99.64** | **100** | **100** | **100** | 98.91 | **100** | **100** | **100** | **100** | **100** | **99.84** |

The bold values represent the maximum values among the numbers compared.



| Actual Label \ Predict Label | 2S_1 | BMP_2 | BRDM_2 | BTR_60 | BTR_70 | D_7 | T_62 | T_72 | ZIL_131 | ZSU_234 | Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2S_1 | 273 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 274<br>99.64% |
| BMP_2 | 0 | 195 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 195<br>100.00% |
| BRDM_2 | 0 | 0 | 274 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 274<br>100.00% |
| BTR_60 | 0 | 0 | 0 | 195 | 0 | 0 | 0 | 0 | 0 | 0 | 195<br>100.00% |
| BTR_70 | 0 | 0 | 0 | 0 | 196 | 0 | 0 | 0 | 0 | 0 | 196<br>100.00% |
| D_7 | 0 | 0 | 0 | 0 | 0 | 271 | 0 | 0 | 2 | 1 | 274<br>98.91% |
| T_62 | 0 | 0 | 0 | 0 | 0 | 0 | 273 | 0 | 0 | 0 | 273<br>100.00% |
| T_72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 196 | 0 | 0 | 196<br>100.00% |
| ZIL_131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 274 | 0 | 274<br>100.00% |
| ZSU_234 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 274 | 274<br>100.00% |
| Precision | 273<br>100.00% | 195<br>100.00% | 274<br>100.00% | 195<br>100.00% | 196<br>100.00% | 271<br>100.00% | 274<br>99.64% | 196<br>100.00% | 276<br>99.28% | 275<br>99.64% | 2425<br>99.84% |

Fig. 6. Confusion matrix obtained by MsCvFA-CNN in the SOC-10 experiment of the MSTAR dataset.

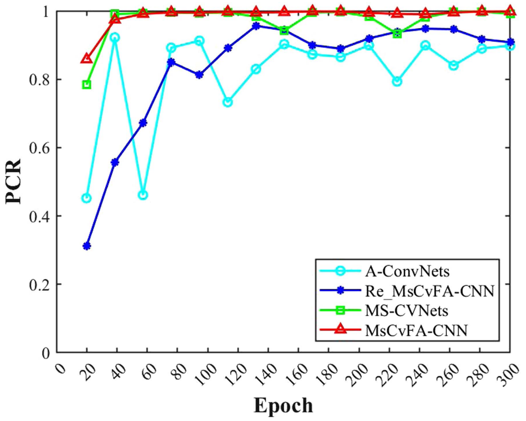


Fig. 7. Performance of the four networks in terms of recognition accuracy on the MSTAR dataset.

recognition accuracy among the four different networks for SAR images.

### C. Experimental Results Under EOC

In this experiment, the MSTAR dataset is divided into EOC-VV and EOC-NV subsets to evaluate the network's performance under changes in target imaging angle and noise, where EOC denotes the significant variations in target shape, imaging angle, shape configuration and model between the training and test sets of SAR images.

*1) EOC-VV:* With the increasing deployment of spaceborne, airborne, and missileborne SAR systems, the imaging angle of SAR systems is subject to continuous variation. Consequently, it is critical to investigate ATR networks that can maintain robust performance in the face of these imaging angle variations. Table III lists the imaging depression angles for four types of targets and the corresponding number of images for each depression angle. As SAR imaging is angle-sensitive, the accuracy of networks decreases with increasing imaging angles. The test results in Table IV demonstrate a sharp decrease in recognition accuracy for the four networks, especially when the imaging angle changes from 30°to 45°. The MsCvFA-CNN achieved the highest recognition accuracy among the four networks when the SAR imaging angles were 15°, 30°, and 40°, which were 99.87%, 99.65%, and 58.16%, respectively. Moreover, the recognition performance of MsCvFA-CNN remained stable as the depression angle varied between 15°and 30°.

*2) EOC-NV:* To test the network's robustness to noise variation in SAR images, we introduced Gaussian noise to the SOC-10 dataset to form the EOC-NV dataset, with signal-to-noise ratios (SNR) between 0 and 10 dB. The experimental results are presented in Table V, which show that the A-ConvNets model is highly sensitive to changes in noise levels. Specifically, for every 2.5 dB change in noise, the accuracy rate decreased by approximately 10%. After integrating the attention mechanism, the network's sensitivity to variations in noise significantly diminishes. When the noise varied within 7.5 dB, the accuracy of MS-CVNets fluctuated considerably. In contrast, the recognition accuracy of MsCvFA-CNN was 88.82% when the SNR was 0 dB, and remained relatively stable when the SNR ranged from 2.5 to 10 dB. In the EOC-NV experiment, MsCvFA-CNN achieved the highest accuracy.

The abovementioned two experiments prove the robustness of the proposed MsCvFA-CNN to changes in SAR imaging angles and noise levels. In addition, our proposed network demonstrated the highest recognition accuracy among the comparison networks, suggesting that the use of complex-valued data, multiscale structures, and complex-valued attention mechanisms can improve the recognition performance of ATR networks.

### D. Recognition Accuracy With Small-Size Training Datasets

To showcase the superior recognition capability of the MsCvFA-CNN model when handling limited training samples, we conducted a series of comparative tests on the SOC-10 experiment dataset using varying numbers of targets in the

TABLE III
EXPERIMENTAL CONFIGURATION OF EOC-VV

| Datasets | Class | Depression | No. of images |
|---|---|---|---|
| Train | 2S_1, BRDM_2, T_72, ZSU_234 | 17° | 986 |
| Test | 2S_1, BRDM_2, T_72, ZSU_234 | 15°, 30°, 45° | 822, 863, 863 |

TABLE IV
RECOGNITION ACCURACY OF FOUR DIFFERENT METHODS FOR TARGETS IN EOC-VV (%)

| Models | Depression15°(%) | Depression30°(%) | Depression45°(%) |
|---|---|---|---|
| A-ConvNets | 99.39 | 87.21 | 35.39 |
| Re_MsCvFA-CNN | 96.85 | 93.17 | 49.56 |
| MS-CVNets | 99.76 | 96.26 | 57.72 |
| MsCvFA-CNN | **99.87** | **99.65** | **58.16** |

The bold values represent the maximum values among the numbers compared.

TABLE V
RECOGNITION ACCURACY OF FOUR DIFFERENT METHODS FOR TARGETS IN EOC-NN (%)

| Models | 0 dB (%) | 2.5 dB (%) | 5 dB (%) | 7.5 dB (%) | 10 dB (%) |
|---|---|---|---|---|---|
| A-ConvNets | 50.69 | 58.32 | 65.21 | 72.67 | 84.23 |
| Re_MsCvFA-CNN | 65.73 | 75.19 | 88.94 | 90.25 | 94.32 |
| MS-CVNets | 72.87 | 80.96 | 93.98 | 97.12 | 99.46 |
| MsCvFA-CNN | **88.82** | **96.46** | **98.56** | **99.51** | **99.75** |

The bold values represent the maximum values among the numbers compared.

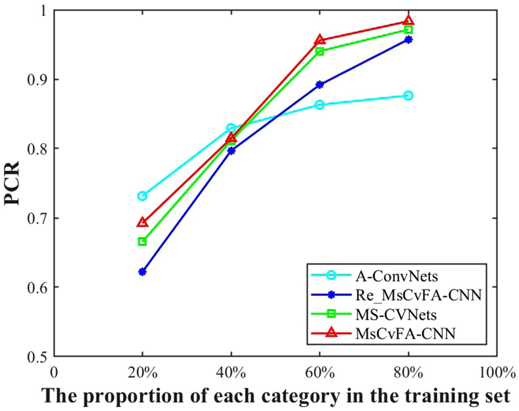


Fig. 8. Accuracy curves of the four networks on the small-size sample dataset.

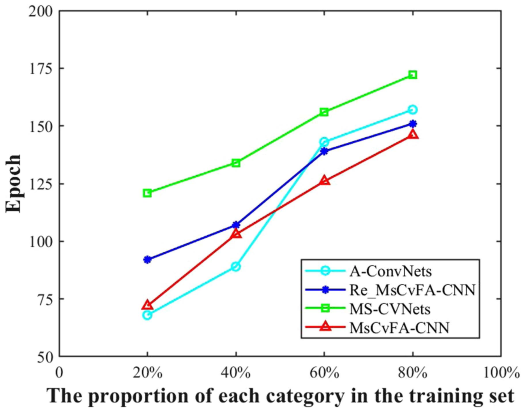


Fig. 9. Epoch when the four networks achieve the highest accuracy on the small-size sample dataset.

training set. We randomly selected different proportions of samples from each category in the SOC-10 dataset as the training set. Since the A-ConvNets is first data augmentation before training, it can achieve a fairly high recognition accuracy even with a small sample proportion, and the epoch to reach the highest accuracy is small, all because of the contribution of data augmentation to this. However, it can be seen from Figs. 8 and 9 that the proposed MsCvFA-CNN has a huge advantage over other networks when the sample proportion reaches 50%. Even when only 60% of the training samples are used, MsCvFA-CNN achieves a recognition accuracy of 95.58%, with the smallest epoch. These results demonstrate that the MsCvFA-CNN model offers high stability and fast recognition speed in scenarios with limited training samples, thanks to the integration of CAM, which enables the network to focus on critical target characteristics, increases the feature difference between each category, and avoids the underfitting problem, ultimately leading to higher recognition accuracy.

### *E. Detection of Target Recognition Performance in Large-Scale Scenes*

In addition, to evaluate the applicability of the proposed network to large-scale scenes with complex terrain in SAR imagery,

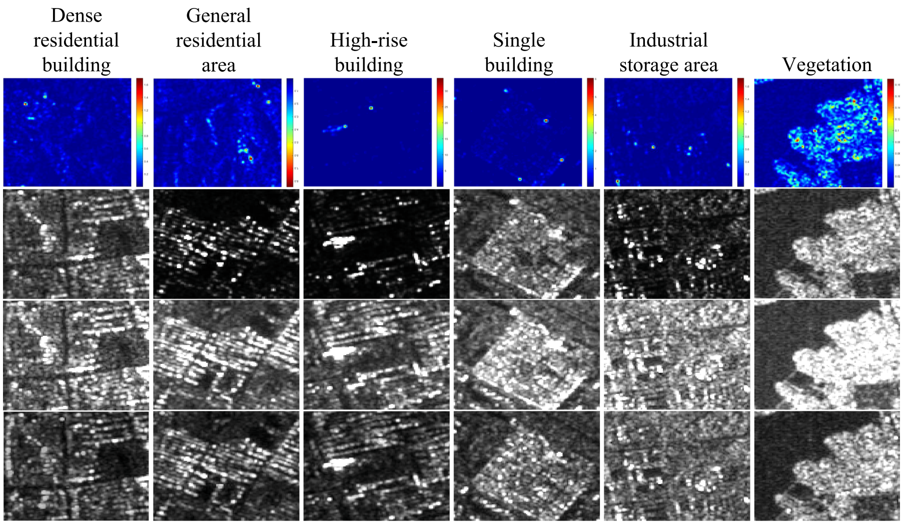
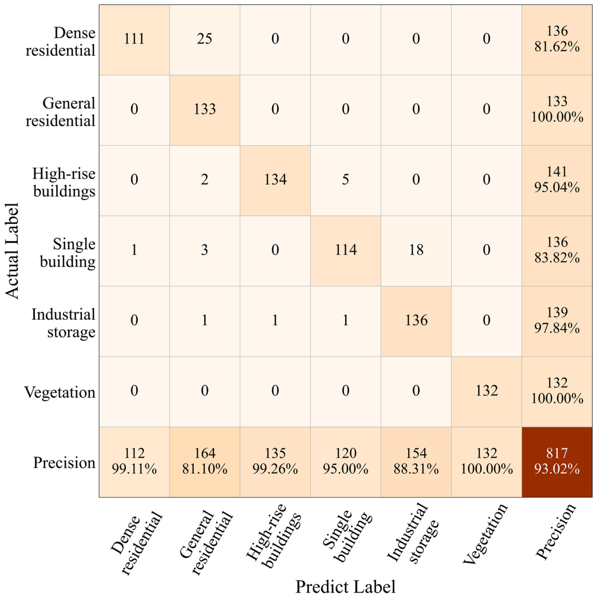


Fig. 10. Example images of OpenSARUrban include six classifications. From top to bottom, they are: RGB, SAR, amplitude map, and phase map.

we conducted validation on the OpenSARUrban dataset [44]. This dataset is specifically designed for urban interpretation using SAR images. We selected the six most challenging classes from the OpenSARUrban dataset, including dense residential building, general residential areas, high-rise buildings, single buildings, industrial storage areas, and vegetation. The RGB, SAR, amplitude, and phase representations for these six classes are shown in Fig. 10.

We tested the recognition performance of the four networks on the six categories of the OpenSARUrban dataset under the same experimental conditions. As shown in Fig. 10, the SAR images in this dataset exhibit strong similarity across six categories and have a wide range of target variations, making the recognition task more difficult. During this experiment, the OpenSARUrban dataset was split into a training set consisting of 2737 images and a test set consisting of 817 images. Fig. 11 displays the confusion matrix of the MsCvFA-CNN for the recognition of the OpenSARUrban dataset. The highest recognition accuracy achieved is 93.02%. The recognition accuracies for the general residential area and vegetation categories reached 100%. It is worth noting that the general residential area category has high similarity with the other four categories, yet it still achieves a recognition accuracy of 100%, demonstrating the significant advantage of MsCvFA-CNN in capturing detailed features. Fig. 12 presents the accuracy of the compare networks on the OpenSARUrban dataset. It is evident that MsCvFA-CNN achieves superior recognition accuracy compared to the other networks across all categories. Furthermore, the accuracy of the compare networks at different epochs is shown in Fig. 13. MsCvFA-CNN achieves its highest recognition accuracy at around 60 epochs, and it requires fewer epochs compared to the other networks.

In conclusion, whether it is the commonly used MSTAR dataset or the challenging OpenSARUrban dataset, the MsCvFA-CNN network can effectively utilize both amplitude and phase information for the recognition of different targets or scenes. It improves the accuracy while reducing the training epochs.



| Actual Label \ Predict Label | Dense residential | General residential | High-rise buildings | Single building | Industrial storage | Vegetation | Precision |
|---|---|---|---|---|---|---|---|
| Dense residential | 111 | 25 | 0 | 0 | 0 | 0 | 136<br>81.62% |
| General residential | 0 | 133 | 0 | 0 | 0 | 0 | 133<br>100.00% |
| High-rise buildings | 0 | 2 | 134 | 5 | 0 | 0 | 141<br>95.04% |
| Single building | 1 | 3 | 0 | 114 | 18 | 0 | 136<br>83.82% |
| Industrial storage | 0 | 1 | 1 | 1 | 136 | 0 | 139<br>97.84% |
| Vegetation | 0 | 0 | 0 | 0 | 0 | 132 | 132<br>100.00% |
| Precision | 112<br>99.11% | 164<br>81.10% | 135<br>99.26% | 120<br>95.00% | 154<br>88.31% | 132<br>100.00% | 817<br>93.02% |

Fig. 11. Confusion matrix of MsCvFA-CNN obtained in the SOC experiment on the OpenSARUrban dataset.

### F. Ablation Experiment

In this section, a series of comprehensive ablation experiments were conducted on different network modules based on the SOC-10 dataset to validate the superiority of MsCvFA-CNN. The motivation of this section's experiment is to investigate how MEFM, CAM, and Cv-CNN affect the performance of the network. MEFM and CAM are eliminated, respectively, from MsCvFA-CNN, and feature visualization is used to test the effects of different modules in SAR ATR. The network structure of the Cv-CNN without MEFM and CAM is shown in Fig. 14. In addition, real-valued networks with the same architecture as MsCvFA-CNN are constructed to examine the impact of imaginary part information on SAR image target recognition.

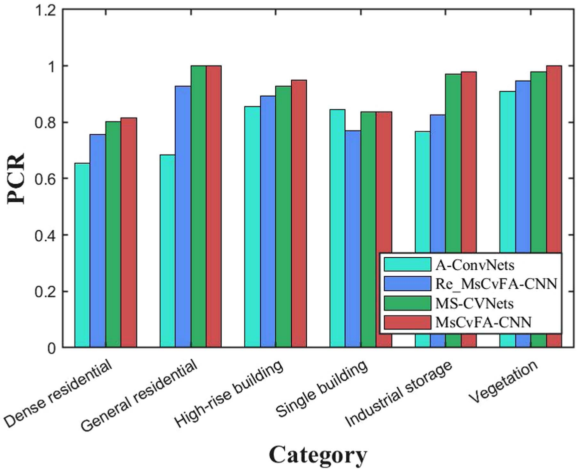


Fig. 12. Performance of the four networks in accurately recognizing targets across six distinct scenarios.

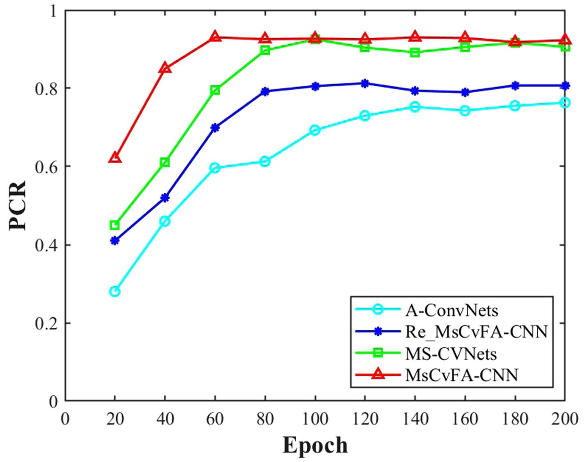


Fig. 13. Performance of the four networks in terms of recognition accuracy at various training epochs on the OpenSARUrban dataset.

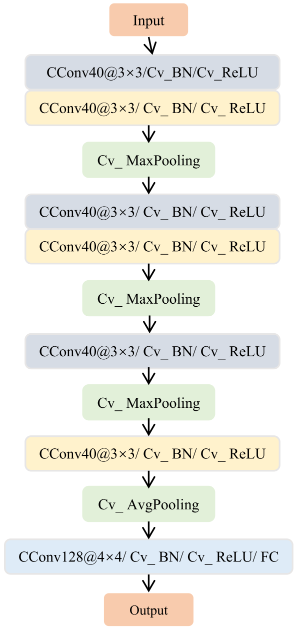
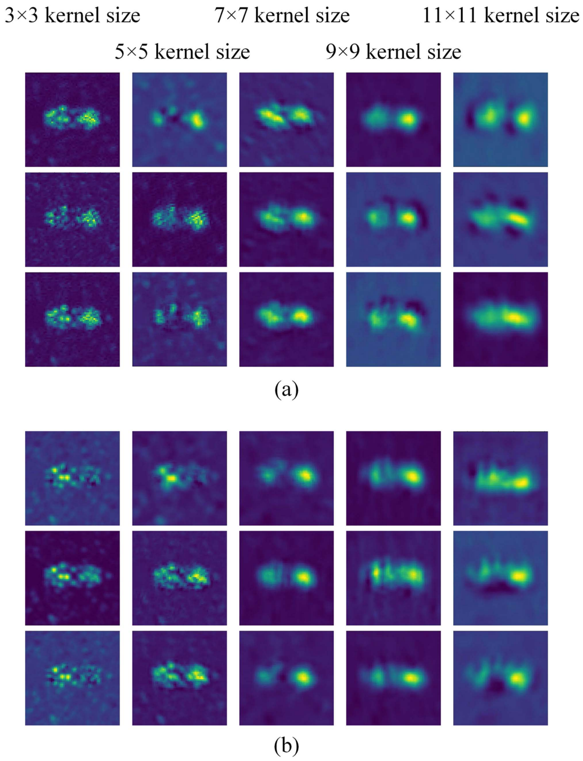

Input
CConv40@3×3/Cv_BN/Cv_ReLU
CConv40@3×3/ Cv_ BN/ Cv_ ReLU
Cv_ MaxPooling
CConv40@3×3/ Cv_ BN/ Cv_ ReLU
CConv40@3×3/ Cv_ BN/ Cv_ ReLU
Cv_ MaxPooling
CConv40@3×3/ Cv_ BN/ Cv_ ReLU
Cv_ MaxPooling
CConv40@3×3/ Cv_ BN/ Cv_ ReLU
Cv_ AvgPooling
CConv128@4×4/ Cv_ BN/ Cv_ ReLU/ FC
Output

Fig. 14. Architecture of single-scale Cv-CNN.



(a)

(b)

Fig. 15. BTR-60 target features extracted using various kernel sizes. (a) Amplitude part characteristics. (b) Phase part characteristics.

To explore the impact of convolution kernels with different sizes on target feature extraction, the features extracted by the complex-valued convolution layer is visualized. The target features extracted by different convolution kernels are displayed in Fig. 15, with (a) representing the amplitude part and (b) representing the phase part. Using a 3 × 3 kernel size preserves the target's global and local texture features, which are distributed across the target, thereby mitigating the effects of background noise. Conversely, the features extracted using a 5 × 5 kernel size are mainly localized, making some features more susceptible to noise interference. The 7 × 7 kernel size captures both the overall characteristics of the target and enhances the global features extracted using the 3 × 3 kernel size. However, there is a conflict between the features extracted using the 9 × 9 and 7 × 7 convolution kernels. The use of an 11 × 11 convolution kernel facilitates the extraction of global features from the target. The results of the experiment show that the feature representation ability of the network improves as the convolution kernel size increases. Meanwhile, in order to achieve the highest recognition accuracy while taking into account the computational resources, combining the accuracy of different convolution kernels in Table VI, we found that using multibranch convolution kernels with sizes of 3 × 3, 7 × 7, and 11 × 11 yields the highest recognition rate of 98.53%. Although some redundancy exists between the amplitude and

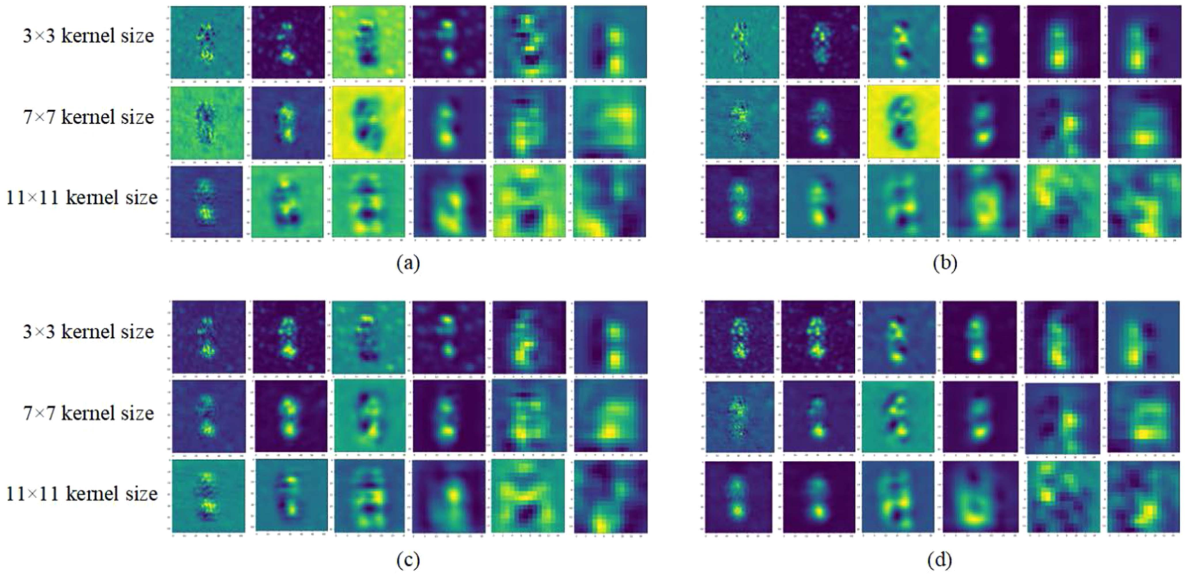


Fig. 16. Attention distribution of CAM on target features. (a) and (b) for the amplitude and phase components of BRT-60 prior to the incorporation of CAM. (c) and (d) for the amplitude and phase components of BRT-60 after the integration of CAM.

TABLE VI
Results From the Ablation Experiment

| Cv-CNN | | MEFM | | CAM | | Size of convolution kernels | Accuracy (%) | The Number of epoch |
|---|---|---|---|---|---|---|---|---|
| Yes | No | Yes | No | Yes | No | | | |
| ✓ | | | ✓ | | ✓ | 3×3 | 93.37 | 192 |
| ✓ | | | ✓ | ✓ | | 3×3 | **96.58** | **115** |
| ✓ | | ✓ | | | ✓ | 3×3 and 5×5 and 7×7 | 97.10 | 203 |
| ✓ | | ✓ | | | ✓ | 3×3 and 5×5 and 9×9 | 97.58 | 212 |
| ✓ | | ✓ | | | ✓ | 3×3 and 5×5 and 11×11 | 97.59 | 217 |
| ✓ | | ✓ | | | ✓ | 3×3 and 7×7 and 9×9 | 98.11 | 241 |
| ✓ | | ✓ | | | ✓ | 3×3and 7×7 and 11×11 | **98.53** | 272 |
| ✓ | | ✓ | | | ✓ | 3×3 and 9×9 and 11×11 | 98.25 | 295 |
| ✓ | | ✓ | | | ✓ | 5×5 and 7×7and 9×9 | 96.99 | 299 |
| ✓ | | ✓ | | | ✓ | 5×5and 7×7and 11×11 | 97.67 | 317 |
| | | ✓ | | | ✓ | 7×7and 9×9and 11×11 | 96.13 | 330 |
| ✓ | | ✓ | | ✓ | | 3×3and 7×7and 11×11 | **99.84** | **158** |
| | ✓ | | ✓ | | ✓ | 3×3 | 91.14 | 167 |
| | ✓ | | ✓ | ✓ | | 3×3 | 93.20 | 95 |
| | ✓ | ✓ | | | ✓ | 3×3and 7×7and 11×11 | 94.87 | 183 |
| | ✓ | ✓ | | ✓ | | 3×3and 7×7and 11×11 | 96.12 | 125 |

The bold values represent the maximum values among the numbers compared.

phase of target features obtained through convolutional kernels, it has been observed that the phase feature is more localized to a specific point on the target. This suggests that employing Cv-CNN can enhance the model's accuracy.

To further validate the role of the CAM in the complex-valued feature extraction stage. The visualization of the feature extraction for the amplitude and phase part of BTR_60 before and after merging CAM is presented in Fig. 16. Fig. 16(a) and (b) shows the amplitude and phase features of the target extracted by three branches respectively without CAM, as shown in each column of Fig. 16. With the convolution depth increases, the target features extracted by the convolution kernels gradually transform from spatial features such as contours to abstract local detail features. However, as the extracted features become increasingly abstract, nearly one-third of the features are not related to the target itself, resulting in a waste of computational resources. After adding CAM, as shown in (c) and (d), compared with (a) and (b), the convolutional layer with CAM pays more attention to the target itself when extracting target features, thus concentrating computational resources on useful target features and greatly reducing computation time. In addition, after adding the first layer of attention mechanism, the feature of the subsequent convolution layer is more focused on the vicinity of the target rather than the background part. Since the CAM is specifically designed for complex-valued SAR images, applying CAM to the model can simultaneously improve the network's attention to important amplitude and phase features of the target. Furthermore, as shown in the third part of Table VI, adding CAM significantly reduces the epoch required for the Cv-CNN to reach the highest recognition accuracy. Therefore, the CAM can enhance the precision and speed of MsCvFA-CNN.

Subsequently, after determining the optimal convolution kernel sizes for MEFM and the impact of CAM on the complex-valued network, we investigated the influence of imaginary components on network performance. Experimental results revealed that for networks without MEFM and CAM, the addition of imaginary components led to a 2.23% improvement in recognition accuracy. In networks with MEFM and CAM, the incorporation of imaginary components resulted in a 3.72% increase in recognition accuracy. This underscores the crucial role of imaginary components in enhancing the accuracy of SAR image target recognition.

In summary, our study highlights the critical role of selecting suitable convolution kernel sizes for accurate and robust target feature extraction. Moreover, we have demonstrated that employing multibranch convolution kernels with varying sizes and complex-valued attention mechanisms can enhance the precision of SAR target recognition while minimizing training time.

## V. Conclusion

The network proposed in this article is specifically designed for complex-valued SAR images that contain phase information. The novel complex-valued attention module enhances the focus on important amplitude and phase features of the targets. The addition of CAM enhances both the accuracy of the network and reduces the number of epochs required for achieving high accuracy by almost half. In addition, the MEFM utilizes a parallel topology structure with multiple kernel sizes for feature extraction, coupled with CAM in the feature extraction and fusion stage, which allows the model to concentrate on crucial features, thereby enhancing its ability to express target features. The findings from our analysis on the complex-valued MSTAR dataset highlight an enhancement of 8.7% in network accuracy through the integration of CAM and MEFM. Furthermore, favorable recognition results are achieved on the large-scale scene dataset OpenSARUrban. The experiments confirm that MsCvFA-CNN outperforms other traditional CNN and Cv-CNN models in terms of recognition performance. The potential of MsCvFA-CNN in SAR ATR is highly promising.

## References

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

[2] D. E. Dudgeon and R. T. Lacoss, "An overview of automatic target recognition," *Lincoln Lab. J.*, vol. 6, no. 1, pp. 3–10, 1993.

[3] L. Novak, G. Owirka, and A. Weaver, "Automatic target recognition using enhanced resolution SAR data," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 35, no. 1, pp. 157–175, Jan. 1999.

[4] S. Du, J. J. Mallorqui, and F. Zhao, "ACE-OT: Polarimetric SAR data-based amplitude contrast enhancement algorithm for offset tracking applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[5] E. R. Keydel, S. W. Lee, and J. T. Moore, "MSTAR extended operating conditions: A tutorial," in *Proc. Algorithms Synthetic Aperture Radar Imagery III*, 1996, vol. 2757, pp. 228–242.

[6] A. Masjedi, M. J. V. Zoej, and Y. Maghsoudi, "Classification of polarimetric SAR images based on modeling contextual information and using texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 932–943, Feb. 2016.

[7] D. Xiang, T. Tang, S. Quan, D. Guan, and Y. Su, "Adaptive superpixel generation for SAR images with linear feature clustering and edge constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3873–3889, Jun. 2019.

[8] X.-M. Li, Y. Sun, and Q. Zhang, "Extraction of sea ice cover by Sentinel-1 SAR based on support vector machine with unsupervised generation of training data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3040–3053, Apr. 2021.

[9] A. Richardson, D. G. Goodenough, H. Chen, B. Moa, G. Hobart, and W. Myrvold, "Unsupervised nonparametric classification of polarimetric SAR data using the K-nearest neighbor graph," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 1867–1870.

[10] E. Santi et al., "On the use of COSMO-SkyMed X-band SAR for estimating snow water equivalent in alpine areas: A retrieval approach based on machine learning and snow models," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[11] H. Lang, G. Yang, C. Li, and J. Xu, "Multisource heterogeneous transfer learning via feature augmentation for ship classification in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[12] J. Liu, M. Xing, H. Yu, and G. Sun, "EFTL: Complex convolutional networks with electromagnetic feature transfer learning for SAR target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[13] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Inform.*, vol. 14, no. 7, pp. 3235–3243, Jul. 2018.

[14] T. Jiang, Z. Cui, Z. Zhou, and Z. Cao, "Data augmentation with Gabor filter in deep convolutional neural networks for SAR target recognition," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 689–692.

[15] H. Shi, X. U. Yuelei, Z. Yang, L. I. Shuai, and L. I. Yueyun, "Target recognition method based on deep belief network," *J. Comput. Appl.*, vol. 34, no. 11, pp. 3314–3317, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1096–1105.

[17] J. Kong and F. Zhang, "SAR target recognition with generative adversarial network (GAN)-based data augmentation," in *Proc. Int. Conf. Adv. Infocomm Technol.*, 2021, pp. 215–218.

[18] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[19] H. Bi, G. Bi, B. Zhang, and W. Hong, "Complex-image-based sparse SAR imaging and its equivalence," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5006–5014, Sep. 2018.

[20] Y. Zhao, X.-F. Yuan, M. Zhang, and H. Chen, "Radar scattering from the composite ship-ocean scene: Facet-based asymptotical model and specular reflection weighted model," *IEEE Trans. Antennas Propag.*, vol. 62, no. 9, pp. 4810–4815, Sep. 2014.

[21] C. Li, J. Liu, and C. Duan, "Imaging characteristic for large elliptical orbit SAR," in *Proc. Asia-Pac. Conf. Synthetic Aperture Radar.*, 2019, pp. 1–5.

[22] P. K. Sanyal, D. M. Zasada, R. P. Perry, and D. W. Winters, "Using shaped phase-thresholds for detecting moving targets in multiple-channel SAR," in *Proc. IEEE Radar Conf.*, 2009, pp. 1–5.

[23] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, “Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” *Lect. Notes Comput. Sci.*, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[25] K. El-Darymli, P. Mcguire, E. W. Gill, D. Power, and C. Moloney, “Characterization and statistical modeling of phase in single-channel synthetic aperture radar imagery,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 3, pp. 2071–2092, Jul. 2015.

[26] M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, “A mathematical motivation for complex-valued convolutional networks,” *Neural Computation*, vol. 28, no. 5, pp. 815–825, 2016.

[27] M. Wilmanski, C. Kreucher, and A. Hero, “Complex input convolutional neural networks for wide angle SAR ATR,” in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2016, pp. 1037–1041.

[28] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, “Complex-valued convolutional neural network and its application in polarimetric SAR image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.

[29] J. Guan, J. Liu, P. Feng, and W. Wang, “Multiscale deep neural network with two-stage loss for SAR target recognition with small training set,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[30] Z. Zeng, J. Sun, Z. Han, and W. Hong, “SAR automatic target recognition method based on multi-stream complex-valued networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[31] F. Gao, W. Shi, J. Wang, A. Hussain, and H. Zhou, “A semi-supervised synthetic aperture radar (SAR) image recognition algorithm based on an attention mechanism and bias-variance decomposition,” *IEEE Access*, vol. 7, pp. 108617–108632, 2019.

[32] Y. Li, L. Du, and D. Wei, “Multiscale CNN based on component analysis for SAR ATR,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[33] N. N. Aizenberg, Y. L. Ivas’Kiv, and D. A. Pospelov, “A certain generalization of threshold functions,” *Dokrady Akademii Nauk SSSR*, vol. 196, no. 11, pp. 1287–1290, 1971.

[34] E. Abdelmaksoud, A. Hassen, N. Hassan, and M. Hesham, “Convolutional neural network for Arabic speech recognition,” *Egyptian J. Lang. Eng.*, vol. 8, pp. 27–38, 2020.

[35] M. M. Lau, K. H. Lim, and A. A. Gopalai, “Malaysia traffic sign recognition with convolutional neural network,” in *Proc. Int. Conf. Dig. Signal Process*, 2015, pp. 1006–1010.

[36] W. A. Gardner, “Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique,” *Signal Process.*, vol. 6, no. 2, pp. 113–133, 1984.

[37] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[38] F. Nielsen and R. Nock, “Entropies and cross-entropies of exponential families,” in *Proc. Int. Conf. Image Process.*, 2010, pp. 3621–3624.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[40] D. Hong et al., “More diverse means better: Multimodal deep learning meets remote-sensing imagery classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[41] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, “Interleaved group convolutions for deep neural networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4383–4392.

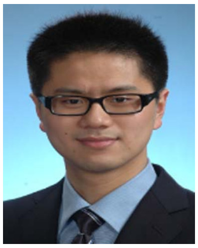
[42] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, “Target classification using the deep convolutional networks for SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[43] S. Feng, K. Ji, F. Wang, L. Zhang, X. Ma, and G. Kuang, “Electromagnetic scattering feature (ESF) module embedded network based on ASC model for robust and interpretable SAR ATR,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[44] J. Zhao, Z. Zhang, W. Yao, M. Datcu, H. Xiong, and W. Yu, “OpenSARUrban: A Sentinel-1 SAR image dataset for urban interpretation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 187–203, 2020.

**Xiaoqian Zhou** received the B.Eng. degree in communication engineering from the Shandong University of Technology, Zibo, China, in 2020. She is currently working toward the M.Eng. degree in information and communication engineering with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China.

Her research interests include machine learning and deep learning, with applications to remote sensing.



**Cai Luo** (Senior Member, IEEE) received the B.Eng. degree in electrical information engineering from Wuhan University, Wuhan, China, in 2006, the M.Sc. degree in electrical and electronic engineering from the University of Sheffield, Sheffield, U.K., in 2008, and the Ph.D. degree in electronic and computer engineering, robotics, and telecommunication from the University of Genoa, Genoa, Italy, in 2012.

He is currently an Assistant Professor in the China University of Petroleum (East China), Qingdao, China. From 2012 to 2013, he was a Marie Curie Fellow with the Technical Research Centre of Finland, Espoo, Finland. From 2013 to 2015, he was an Assistant Researcher with the Department of Automation, Tsinghua University, Beijing, China. His research interests include UAV biomimetic design, dynamic and control of robotic system.

**Peng Ren** (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in computer science from the University of York, York, U.K.

He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. His research interests include remote sensing and machine learning.

Dr. Ren was a recipient of the K. M. Scott Prize from the University of York in 2011 and the Eduardo Caianiello Best Student Paper Award at the 18th International Conference on Image Analysis and Processing in 2015, as one co-author.

**Bin Zhang** received the B.Eng. degree in electronic engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2003, and the M.Eng. degree in test and measurement technology and instrumentation from the China Academy of Launch Vehicle Technology, Beijing, China, in 2006.

He is the Head of the Product R&D Department, Beijing Research Institute of Telemetry, Beijing, China. His research focuses on the field of microwave remote sensing.