SAR Target Classification Based on Multiscale Attention Super-Class Network

Di Wang, Yongping Song, Junnan Huang, Daoxiang An, *Member, IEEE*, and Leping Chen, *Student Member, IEEE*

***Abstract*—The convolutional neural network (CNN) is widely used in synthetic aperture radar (SAR) target recognition, but conventional CNN mainly adopts a single-scale convolutional kernel, resulting in losing part of the feature information of targets and does not pay enough attention to significant features. On the other hand, conventional CNN approaches only assign fine-class labels to SAR targets, ignoring the high-level semantics information of similar categories, which reduces the feature differences between categories and the generalization ability of the model. Therefore, this article proposes a multiscale attention super-class CNN (MSA-SCNN) for SAR target classification. First, MSA-SCNN combines multiscale feature fusion with the attention module to improve the integrity of SAR target feature representation. The attention module includes channel and spatial attention modules, which realize the weighted enhancement of different scale features. Additionally, MSA-SCNN introduces super-class labels to increase the feature difference between categories. The classification stage consists of a fine-class branch and a super-class branch, and the features trained on the super-class branch are fused to the fine-class branch to improve the network's fine classification ability. Experiments on the moving and stationary target acquisition and recognition dataset and the FUSAR-Ship dataset show that the proposed MSA-SCNN outperforms many current existing state-of-the-art methods.**

***Index Terms*—Convolutional neural network (CNN), multiscale attention, super-class labels, synthetic aperture radar (SAR) target classification.**

I. Introduction

SYNTHETIC aperture radar (SAR) has the ability to obtain high-resolution images all-day and all-weather. Compared with other imaging methods such as optical and infrared, SAR can acquire target information covered by clouds and vegetation. Nowadays, SAR is widely used in numerous fields, e.g., military reconnaissance and geological exploration [1]. In most SAR applications, automatic target recognition (ATR) plays an important research and application value in the field of military reconnaissance, so it has received considerable attention [2].

Manuscript received 22 June 2022; revised 30 August 2022; accepted 6 September 2022. Date of publication 15 September 2022; date of current version 24 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62271492, Grant 62101566, and Grant 62101562 and in part by the Natural Science Fund for Distinguished Young Scholars of Hunan Province under Grant 2022JJ10062. *(Corresponding author: Daoxiang An.)*

The authors are with the College of Electronic Science, National University of Defense Technology, Changsha 410073, China (e-mail: wdnudt@163.com; sypopqjkl@163.com; huangjunnan16@nudt.edu.cn; daoxiangan@nudt.edu.cn; gfkdclp@126.com).

Digital Object Identifier 10.1109/JSTARS.2022.3206901

Generally, a standard SAR ATR system usually consists of three stages: detection, discrimination, and classification recognition [3]. In the detection stage, the region of interest (ROI) is prescreened according to the local grayscale statistics in the SAR images [4]. The discrimination stage significantly removes the false alarm clutter and reserves the real targets [5]. Finally, the target features are extracted, and a classifier is designed to classify the SAR targets.

SAR target classification is one of the vital stages of SAR ATR processing. Generally, the classification methods are divided into two categories: template-based methods [6] and model-based methods [7]. The template-based method needs to build a template database, and the extracted SAR target features will be best matched with the template database during the recognition stage. The classification accuracy is related to the manually designed template, which takes a lot of time to build the template database [8]. The model-based method adopts three-dimensional (3-D) modeling and electromagnetic calculations to simulate SAR images, and iteratively adjusts the model in the process of predicting SAR target chips. Based on these two mainstream methods, many SAR target recognition algorithms have been proposed in recent years, such as principal component analysis [9], linear discriminant analysis [10], support vector machine [11], adaptive boosting [12], conditionally Gaussian model (CGM) [13], and iterative graph thickening [14]. These methods usually need to extract specific features from SAR images and predesign complex target recognition algorithms, which brings tremendous challenges to practical applications.

With the rapid development of deep learning [15], convolutional neural networks (CNNs) have been applied to various computer vision tasks such as image classification [16], object detection [17], semantic segmentation [18], etc., and it has achieved superior performance. CNN directly extracts low-level and high-level features from raw images through convolutional and pooling layers, providing an effective solution for SAR target classification. Many novel works using deep neural networks have proven to be powerful tools for SAR target classification [19], [20], [21].

Most CNN methods only use a single-scale convolutional kernel, resulting in some feature representation loss of the SAR targets. Ai et al. [22] proposed a novel CNN model based on multikernel-size feature fusion (MKSFF-CNN), which uses convolutional kernels of different sizes to extract the multiscale deep features, and then, MKSFF-CNN concatenates the features extracted by the convolutional layers of different dimensions to

achieve the finest classification. Although MKSFF-CNN improves the SAR target feature representation completeness, two important problems should be solved for SAR target classification. First, Multiscale features have information redundancy, and the network needs to automatically focus on important features and suppress unnecessary features to increase the representation power of multiscale features. Second, since SAR images are sensitive to observation conditions, the network needs to introduce prior knowledge to increase the difference of multiscale features between different categories, thereby improving the generalization ability of the model.

In order to increase the feature difference between categories, Zhang et al. [23] designed a two-stream deep network and introduced SAR domain knowledge such as target azimuth and phase into the CNN to assist in classification. For SAR classification tasks, existing methods mainly introduce prior knowledge and fuse features at the network input, ignoring the category labels that actually guide the classification at the network output. These methods only have one kind of fine-class label, and the misclassifications between any two classes are treated equally. In contrast, when humans create categories, nonparallel semantic relations are established between each category [24], and categories with communal features belong to a super-class label, which can assist in fine classification. For example, there are many different types of tanks, they all belong to a super-class label—Tanks. When classifying a tank of an unknown type, it should tend to be classified under the tank super-class label rather than identified as the armored vehicle or another class label. Therefore, super-class labels can improve the generalization ability of the model to unknown classes.

To sum up, in order to solve the problem that the SAR target features extracted by most networks are incomplete and have information redundancy, lack of attention to important features, and small feature differences between categories, this article proposes a multiscale attention super-class CNN (MSA-SCNN) for SAR target classification. In the extraction stage, MSA-SCNN combines multiscale feature fusion with the attention module to improve the integrity of SAR target feature representation. The attention module includes channel and spatial attention modules, which focus on important features and suppress unnecessary features. In the classification stage, MSA-SCNN has a super-class branch and a fine-class branch, which are corresponding to the super-class and fine-class labels. The super-class branch focuses on the communal features of SAR categories so that the feature difference between super-classes increases, whereas the fine-class branch focuses on more refined category features. Finally, features extracted by the super-class branch are fused to the fine class branch to assist in fine classification. The main contributions of this article could be summarized as follows.

1) The multiscale features of SAR targets are analyzed and a new network structure—MSA-SCNN is proposed, which uses convolutional kernels of different sizes to extract feature information of different scales, and fuses these features at each layer. It greatly improves the integrity of the SAR target feature representation.
2) MSA-SCNN uses spatial and channel attention for multiscale feature weighted enhancement, so that the network focuses on target features, suppresses the background clutter, and avoids information redundancy.
3) MSA-SCNN introduces the prior knowledge of super-class labels into CNN and creates two classification branches. The high-level semantic features trained on the super-class branch are fused to the fine-class branch for final classification. The assistance of super-class labels increases the feature difference between the categories, and further improves the generalization ability of the model.

The rest of this article is organized as follows. Section II introduces SAR multiscale features and super-class labels. Section III elucidates the details of the MSA-SCNN model structure and training strategies. Section IV is the experimental part, where comparative experiments and ablation studies are designed and performed. The classification performance is validated on the moving and stationary target acquisition and recognition (MSTAR) dataset and the FUSAR-Ship dataset with a detailed evaluation. Finally, Section V concludes the article.

II. Background Knowledge of MSA-SCNN

In this section, the multiscale features of SAR targets and super-class labels will be discussed in detail. Meanwhile, the division of super-class and fine-class labels will be given. MSA-SCNN uses two methods to improve the integrity of the SAR target feature representation and the generalization ability of the model.

A. Multiscale Features of SAR Targets

Generally speaking, the deeper the network is, the stronger the representation and nonlinear fitting ability it will have. VGG [16] uses stacked small-size convolution kernels to increase the network depth, whereas ResNet [25] introduces skip connections to alleviate the gradient vanishing problem of deep networks, and the network is further deepened. However, the existing SAR target datasets are generally small, and the application of deeper networks can easily lead to overfitting. Therefore, apart from deepening the networks, other methods should be considered to improve the ability of target feature extraction.

GoogLeNet [26] proposed by Szegedy won the championship in the ImageNet large-scale visual recognition challenge competition that year. This model uses a parallel network structure to obtain multichannel features through different convolution branches. The size of the convolutional kernel of each branch is different, which means that the input of the same layer has multiple receptive fields of different sizes, so it can extract multiscale features from images.

Inspired by this, convolutional kernels of different sizes are used to extract features from SAR images. Fig. 1 displays the results of the feature maps after activation function binary processing to facilitate observation and analysis. The usage of 3×3 kernels highlights the local features of the SAR images and divides the target feature maps into multiple small regions, and the holes in the feature maps make the global contour features inconspicuous. As the size of kernels increases, the global contour features of the target are gradually enhanced. But as

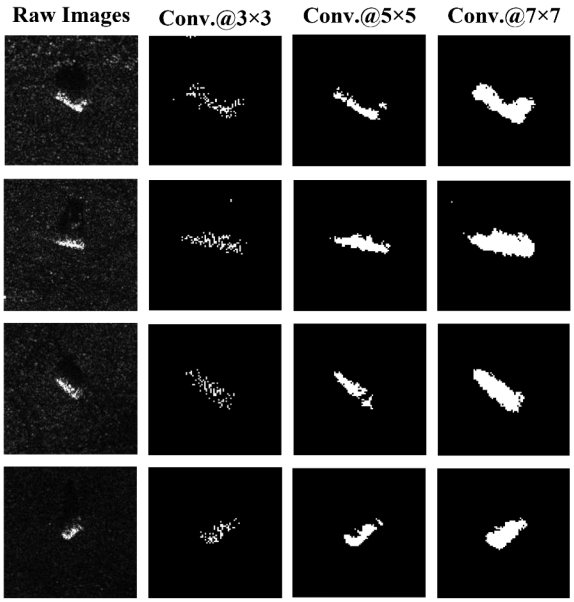

Fig. 1. Multiscale feature maps of SAR targets under different convolution kernels. The first column is the raw SAR images.

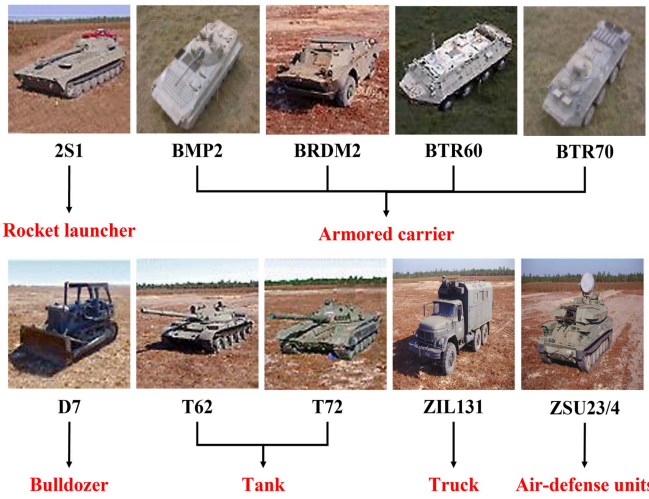

Fig. 2. Super-class label division under the MSTAR dataset, black represents the original fine-class labels, and red represents the super-class labels.

the hole area of the target feature map decreases, the local detail features are lost by degrees. This phenomenon also reveals that the SAR targets have multiscale features. However, the number of multiscale feature channels without any processing is large, which leads to information redundancy. Therefore, this article further extracts important features and suppresses unnecessary features through attention mechanisms and super-class labels.

B. SAR Domain Super-Class Labels

SAR targets contain lots of prior knowledge in the SAR domain, such as azimuth angle, phase information, etc. [23]. Unlike optical images, SAR images are sensitive to the azimuth angle. Under the side-view imaging mode of SAR, some areas will lose echo signals due to the blocking of the target itself, and the attribute scattering center (ASC) that reflects the target structure also changes with the azimuth angle [26]. The SAR phase also contains additional information. However, the final focused image will have phase errors due to motion errors and terrain factors. The introduction of these two prior knowledge can improve the recognition rate [27], but it is easily affected by various external factors and becomes unstable.

The above-mentioned prior knowledge in the SAR domain will be affected by the signal-to-noise ratio of SAR images, and the extraction results of prior information may be contaminated. For the SAR image classification problem, the SAR target category labels are also prior knowledge. While recognizing unknown targets, our humans use prior knowledge to determine the high-level super-class labels of objects first and then identify them finely [24]. Inspired by human recognition of objects, super-class labels are higher level semantic divisions of classes with similar features. The conception of super-class was originally applied to the dataset with unbalanced class distribution [28], which helps minority classes benefit from abundant samples under the same super-class. It is interesting to notice that the SAR targets can also abstract the super-class labels, which are more stable than other prior information, because when the SAR observation parameters and the scene change, the prior information of azimuth angle and phase may change, but the super-class labels will not change. In this way, the super-class labels can be stably used for SAR target recognition ignoring the influence of external factors.

Taking the MSTAR dataset as an example, Fig. 2 shows the super-class labels divided by ten categories of targets, in which the black classes under each optical image represent original fine-class labels, and the red classes represent super-class labels. BMP2, BRDM2, BTR60, and BTR70 are all armored vehicles, so they belong to a super-class label, whereas T62 and T72 belong to different models of tanks, so they are divided into a super-class label. The rest of the vehicles have no common characteristics and belong to their labels. MSA-SCNN introduces super-class labels to guide the network to filter multiscale features so that the feature difference between categories increases and the generalization ability of the model is improved.

III. MSA-SCNN Classification Method

In this section, a novel SAR classification method MSA-SCNN is proposed. This approach combines multiscale feature fusion with the attention module to improve the integrity of SAR target feature representation. Meanwhile, it introduces super-class labels to increase the multiscale feature difference between categories. The basic structure of the MSA-SCNN model will be described. Then, the configuration of training implementation will be given.

A. Structure of MSA-SCNN

The basic structure of the proposed MSA-SCNN is shown in Fig. 3, which adopts multichannel parallel convolutional layers for feature extraction, and each convolutional layer uses convolutional kernels of different sizes, where n is the number of channels, and the convolutional kernel size is $Sn \times Sn$. MSA-SCNN extracts multiscale features and fuses them after

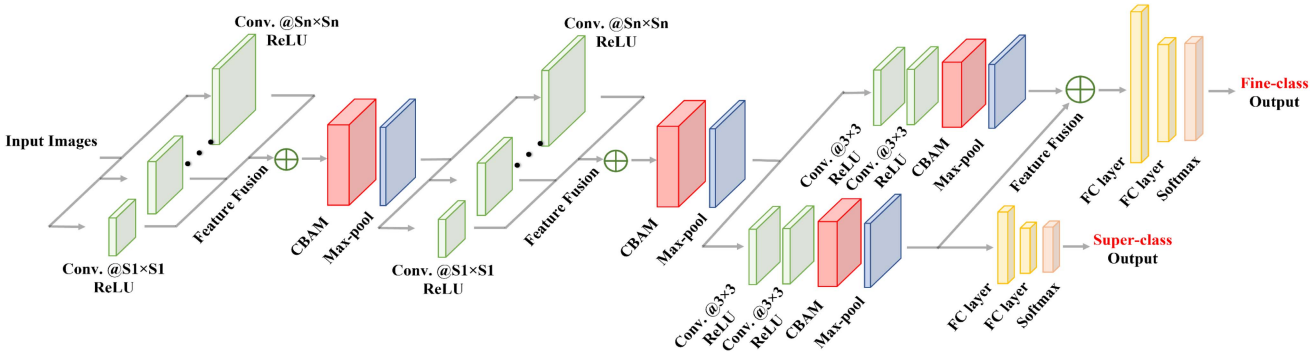

Fig. 3. Structure of MSA-SCN for SAR target classification. Blocks of different colors represent network layers of different structures. Green blocks: Convolution kernels of different sizes and nonlinear activations. Red blocks: Convolutional block attention module (CBAM). Blue blocks: Max-pooling layers. Yellow blocks: Fully connected layers. Orange blocks: Softmax classifier. $S1, S2, \ldots, Sn$: The different sizes of the convolutional kernel.

each convolutional layer to ensure the integrity of the SAR targets. In addition, MSA-SCNN uses attention modules after convolutional layers, which focus on important features, suppress unnecessary features, and avoid information redundancy.

In the classification stage, MSA-SCNN is divided into the super-class branch and the fine-class branch. The subsequent convolutional layer adopts 3×3 small-sized kernels to extract deeper high-level semantic features. Then, the features extracted by the super-class branch are fused into the fine-class branch, which increases the feature difference between categories and assists fine classification. Finally, the softmax classifier assigns a posterior probability to each target category. The outputs of the two branches correspond to fine-class labels and super-class labels. And the total loss of MSA-SCNN is the sum of the loss weights of the two branches. The details of those layers and training operations are described in the rest of this section.

B. Multiscale Convolution and Pooling

The convolutional layer is the core block in our network, and it can automatically extract the multiscale features from the input SAR images. The small-size and large-size kernels can extract local feature information and global contour feature information, respectively. So MSA-SCNN uses parallel multiscale convolution kernels for feature extraction and fusion.

Each convolutional layer has n convolution kernels of different sizes $(S_1 \times S_1, S_2 \times S_2, \ldots, S_n \times S_n)$. Let $a_i^{(l-1)}$ be the ith feature map in the $l-1$ convolutional layer in the proposed network, $a_{j\_n}^{(l)}$ is the jth output feature map of the nth convolutional channel in this layer. Suppose that $w_{ij\_n}^{(l)}$ denote the convolutional kernel operating the ith input feature map to the jth output feature map in the nth convolutional channel, and $b_{j\_n}^{(l)}$ is the jth bias. The forward propagation process in the convolutional layer can be expressed as

$$z_{j\_n}^{(l)} = \sum_i a_i^{(l-1)} * w_{ij\_n}^{(l)} + b_{j\_n}^{(l)} \tag{1}$$

$$a_{j\_n}^{(l)} = \sigma\left(z_{j\_n}^{(l)}\right) \tag{2}$$

where $z_{j\_n}^{(l)}$ denotes the convolutional result before nonlinear activation. The symbol $*$ denotes the convolutional operation, and $\sigma(\cdot)$ is the nonlinear activation function. MSA-SCNN adopts the rectified linear unit (ReLU) [29] function as the nonlinear function, which avoids problems of gradient explosion and disappearance. In addition, the calculation of ReLU is easy and efficient.

After multiscale convolution, the network will fuse these multiscale features. To ensure that feature information of each scale is not lost, MSA-SCNN concatenates these multiscale features. Let $f_n^{(l)}$ be the output feature map of the nth channel in the lth convolutional layer and $F^{(l)}$ is the fused feature of the lth convolutional layer, the feature fusion can be expressed as

$$F^{(l)} = \left[f_1^{(l)}, f_2^{(l)}, \ldots, f_n^{(l)}\right]. \tag{3}$$

Since the 1×1 kernel is unhelpful in increasing the receptive field, the too-large kernel will repeatedly extract feature information when it slides on the SAR image with a small stride [22]. Therefore, in the proposed MSA-SCNN, the number of convolutional channels is four, and the sizes of kernels are $S_1 = 3, S_2 = 5, S_3 = 7, S_4 = 9$.The multiscale features extracted by them are complementary to the fusion process and can represent the SAR target features better.

The pooling layer is generally connected after the convolutional layer to reduce the dimension of the feature map, thereby reducing the parameters of the entire network. MSA-SCNN adopts the maximum pooling [30] and calculates the maximum response of the pooling window for output.

C. Multiscale Attention Module

The SAR features after multiscale convolutional have multiple channels, resulting in a lack of attention to significant features and information redundancy. Interestingly, attention not only tells where to focus but also improves the representation of interests. MSA-SCNN adopts the CBAM [31] to focus on important features and suppress unnecessary ones. Fig. 4 shows the structure of CBAM, which includes channel attention and spatial attention modules. Given a multiscale feature $F \in \mathbb{R}^{C \times H \times W}$ as input, average-pooled features and max-pooled features are first generated in each channel. Both features are then forwarded to

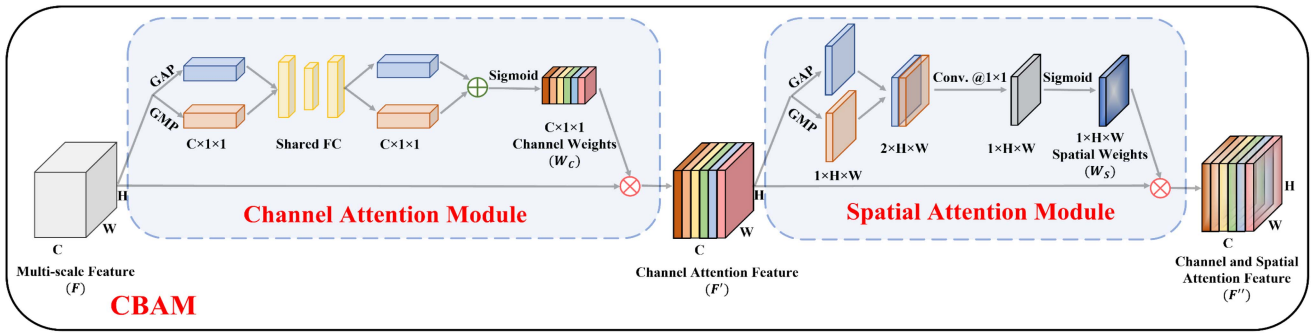

Fig. 4. Structure of CBAM includes a channel attention module and a spatial attention module. GAP: Global average-pooling. GMP: Global max-pooling.

shared fully connected layers and the output features are merged using elementwise summation. Finally, the channel attention weight $W_C \in \mathbb{R}^{C\times1\times1}$ is obtained through the activation function. The channel attention process can be summarized as

$$W_C = \sigma\left(FC\left(\text{AvgPool}\left(F\right)\right) + FC\left(\text{MaxPool}\left(F\right)\right)\right) \quad (4)$$

$$F' = F \otimes W_C \quad (5)$$

where $F'$is the channel attention feature, the symbol $\otimes$ denotes elementwise multiplication, and $\sigma(\cdot)$denotes the sigmoid function.

Different from the channel attention, the spatial attention focuses on the interspatial relationship of multiscale features, which is complementary to the channel attention. Average-pooled features and max-pooled features along the channel axis are first generated. Applying pooling operations along the channel axis is shown to be effective in highlighting informative regions [32]. Then, both features are concatenated and forwarded to a convolution layer to produce spatial attention weight $W_S \in \mathbb{R}^{1\times H\times W}$. The spatial attention process can be summarized as

$$W_S = \sigma\left(\text{Conv.@}1\times1\left(\left[\text{AvgPool}(F'), \text{MaxPool}\left(F'\right)\right]\right)\right) \quad (6)$$

$$F'' = F' \otimes W_S \quad (7)$$

where $F''$is the channel and spatial attention feature, the Conv.@1×1 represent a convolution operation with the filter size of 1×1.

After the introduction of the CBAM, the SAR multiscale features are further filtered on the channel. Meanwhile, in the space, the model more focuses on target features and suppresses the background clutter.

D. Super-Class and Fine-Class Branches

MSA-SCNN introduces the prior knowledge of super-class labels into CNN and creates two classification branches. The structure of the branches is shown in Fig. 5. The outputs of the two branches correspond to the fine-class and super-class labels, respectively. Generally, the fine-class labels are the original labels of the SAR target, and the super-class labels are the high-level classes reassigned for SAR targets. If several fine classes have the same attribute characteristics, they are assigned a super-class label. Let $[C_1, C_2, \ldots, C_m]$be the fine-class labels

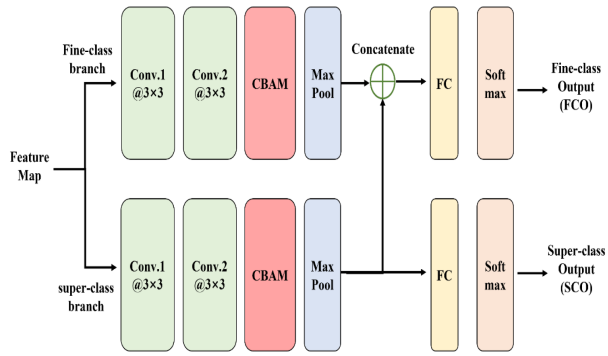

Fig. 5. Super-class and fine-class branches for SAR target classification.

of SAR targets, and $[S_1, S_2, \ldots, S_n]$be the super-class labels, where m and n represent the number of fine-class and super-class labels, Then the number of super-class labels must be less than or equal to the number of fine-class labels (i.e., $n \leq m$).

Due to the small feature size after max pooling, the subsequent convolutional layers use small size 3×3 kernels to extract deep features. The super-class branch extracts the common features of each super-class to increase the difference between classes, whereas the fine-class branch pays more attention to the fine features of different classes. Therefore, the CBAM attention module is also added after the convolutional layer to make the two branches pay different attention.

To increase the feature difference between categories and further improves the generalization ability of the model, the features extracted from the super-class branch are fused into the fine-class branch. Let f_c is the feature extracted by the fine-class branch, and f_s is the feature extracted by the super-class branch. The feature fusion strategy is also concatenating on the feature channel, ensuring that the SAR target feature information is not lost. F, the final feature of the fine-class branch after feature fusion, can be expressed as

$$F = \left[f_c, f_s\right]. \quad (8)$$

The fused feature F will be sent to the fully connected layer and the softmax classifier for final classification.

E. Fully Connected Layer, Dropout, and Softmax

The fully connected layer is essentially equivalent to the spatial transformation of the feature. The feature information will be converted into feature vectors and input into the fully connected layer. Let F_v be the feature vector of the fused feature, the forward propagation process of the fully connected layer is expressed as

$$Z^{(l)} = \sigma\left(F_v \cdot W + B\right) \tag{9}$$

where *W* and *B* are the weights and bias of the fully connected layer, respectively, and $Z^{(l)}$ is the output of the *l*th fully connected layer. Since the fine-class branch fuses the features of the super-class branch, and the length of the feature vector after flattening is longer, the number of neural units in the fully connected layer is also more than that of the super-class branch.

When there are fewer training samples, the deep network will have overfitting problems, resulting in the bad performance of the model on the test set. This article uses the dropout method [33] to suppress this problem effectively. It sets the output of each hidden unit to zero randomly. Since the parameters of the fully connected layer are relatively large, our proposed MSA-SCNN applies the dropout scheme due to complicated parameters in the fully connected layer and sets the dropout probability to 0.5.

The softmax classifier is often adopted for multitarget classification by connecting to the back of the fully connected layer to provide the posterior probability of each category. Let the output of the last fully connected layer be $Z = [z_1, z_2, \ldots, z_C]$, then the corresponding posterior probability for each class can be formulated as

$$p(y_i|Z) = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)} \tag{10}$$

where y_i denotes the *i*th target category and *C* indicates the total number of categories. The output of the softmax classifier is a *C*-dimensional vector, representing the probability of each category. For the fine-class and super-class branches, the softmax output vector dimensions correspond to the number of fine-class labels and super-class labels, respectively. The two branches compute the posterior probabilities for their respective classes.

F. Cost Function and Backpropagation

The cost function of multiclassification is the cross-entropy loss, which is defined as

$$L(w,b) = -\sum_{i=1}^{C} y_i \log p\left(y_i|Z;w,b\right) \tag{11}$$

where *w* and *b* are trainable parameters in the network. In the MSA-SCNN model, there are two classification branches corresponding to two cost functions. Let L_{fc} and L_{sc} be the fine-class and super-class cost functions, and *L* is the total cost function. They are formulated as

$$L_{fc}(w_1,b_1) = -\sum_{i=1}^{m} y_i \log p\left(y_i|Z_{fc};w_1,b_1\right) \tag{12}$$

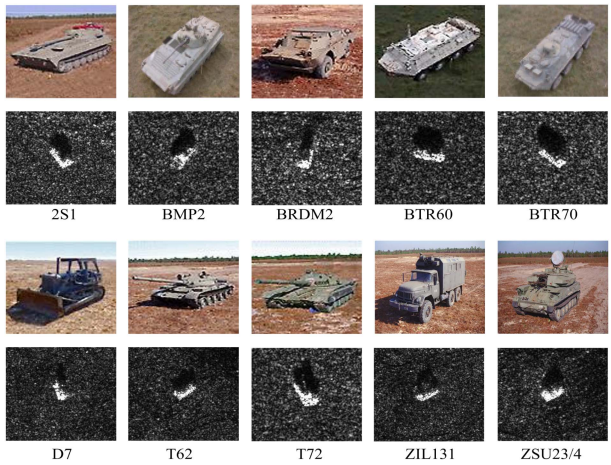

Fig. 6. Optical images and corresponding SAR images of military targets in the MSTAR dataset.

$$L_{sc}(w_2,b_2) = -\sum_{i=1}^{n} y_i \log p\left(y_i|Z_{sc};w_2,b_2\right) \tag{13}$$

$$L(w,b) = \lambda \cdot L_{sc}(w_1,b_1) + (1-\lambda)\cdot L_{fc}(w_2,b_2) \tag{14}$$

where *m* and *n* represent the number of fine-class and super-class labels, and λ is the weight coefficient of the super-class cost function, which is in the range of (0, 1). The total loss is the weighted sum of the super-class loss and the fine-class loss. If the value of λ is too large, the feature extraction will focus more on the super-class labels during training. The λ value of MSA-SCNN in this article is 0.5, considering that the super-class loss and the fine-class loss are equally crucial. Different λ have various impacts on the accuracy, which will illustrate in the experimental section. *w* and *b* can be optimized by continuously minimizing the cost function during the training process, which is favorable to the classification accuracy.

Although the proposed MSA-SCNN has two branches, the way of training parameters is similar to one-branch approaches. And backpropagation [34] can still be used to compute gradients and update network parameters.

IV. Experiments and Analysis

A. Datasets and Random Cropping

The SAR image datasets used in this article are the MSTAR dataset [35] and the FUSAR-Ship dataset [36]. Both datasets are detailed in the following.

1) MSTAR Dataset: The dataset is acquired by the Sandia National Laboratory, operating at *X*-band with a high resolution of 0.3 m and HH polarization. The MSTAR dataset includes ten different classes of military vehicles (rocket launcher: 2S1; armored carrier: BMP2, BRDM2, BTR60, and BTR70; bulldozer: D7; tank: T62 and T72; truck: ZIL131; and air defense unit: ZSU23/4), which are captured under different conditions, such as aspect angle, depression angle, and serial number. The optical images of the targets and their corresponding SAR images are displayed in Fig. 6. To comprehensively evaluate the classification performance of the proposed MSA-SCNN, the standard

TABLE I
NUMBER OF TRAINING AND TEST IMAGES FOR THE SOC EXPERIMENTAL SETUP

Super-classes	Fine-classes	Number	
		Train (17°)	Test (15°)
Rocket launcher	2S1	299	274
Armored carrier	BMP2	233	196
	BRDM2	298	274
	BTR60	256	195
	BTR70	233	196
Bulldozer	D7	299	274
Tank	T62	299	273
	T72	232	196
Truck	ZIL131	299	274
Air defense unit	ZSU23/4	299	274

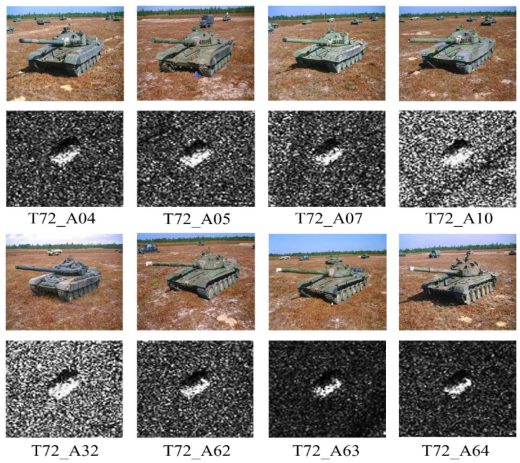

Fig. 7. Optical images and corresponding SAR images of the T72 variants.

operating condition (SOC) and extended operating condition (EOC) [35] were used to test the algorithm.

The SOC refers to that the serial numbers and target configurations in the testing set are the same as those in the training set, but with different aspects and depression angles. Table I lists a summary of the SOC experimental setup, showing that SAR images with the depression angle of 17° and 15° belong to the training set and the testing set, respectively. The proposed MSA-SCNN adopts two kinds of labels (fine-class labels and super-class labels). According to the common attributes of the military vehicles, ten fine-class labels are divided into six super-class labels, named rocket launcher (2S1), truck (ZIL131), tank (T62 and T72), armored carrier (BTR60, BTR70, BRDM2, and BMP2), air defense units (ZSU23/4), and bulldozer (D7).

The EOC is closer to real battlefield situations and this article selects the configuration-variant (EOC-C) and version-variant (EOC-V) datasets. The EOC-C refers to the addition or removal of discrete components on the target, such as removing the fuel barrel on the T72. In addition, the EOC-V refers to target version variation, which means that after some armored vehicles are finalized, they will be upgraded, such as adding the state-of-the-art reactive armor or replacing the main gun with a larger caliber. The EOC dataset contains two BMP2 variants (9566 and c21) and ten T72 variants (812, S7, A04, A05, A07, A10, A32, A62, A63, and A64). Optical images and the corresponding SAR images of the eight T72 targets are shown in Fig. 7. It can be seen that the T72 variants are almost indistinguishable, which brings challenges to SAR target recognition.

A summary of EOC-C and EOC-V for training and testing datasets is listed in Tables II and III. There are four target types (BMP2, BRDM2, BTR70, and T72) for EOC training sets with a depression angle of 17°. The EOC-C test set has two target types (BMP2 and T72) with seven different configuration variations, and EOC-V has one target type (T72) with five version variations. The EOC dataset is of great significance for evaluating the generalization ability of the MSA-SCNN model.

2) FUSAR-Ship Dataset: The dataset is constructed by 126 original Gaofen-3 images, covering a large variety of sea, land, coast, river, and island scenarios. It includes different classes of

TABLE II
NUMBER OF TRAINING IMAGES FOR THE EOC-C AND EOC-V EXPERIMENTAL SETUP (DEPRESSION: 17°)

Super-classes	Fine-classes	Types	Number
Armored carrier	BMP2	9563	233
	BRDM2	E-71	298
	BTR70	c71	233
Tank	T72	132	232

TABLE III
NUMBER OF TEST IMAGES FOR THE EOC-C AND EOC-V EXPERIMENTAL SETUP (DEPRESSION: 15° AND 17°)

Datasets	Super-classes	Fine-classes	Types	Number
EOC-C	Armored carrier	BMP2	9566	428
		BMP2	c21	429
	Tank	T72	812	426
		T72	A04	573
		T72	A05	573
		T72	A07	573
		T72	A10	567
EOC-V	Tank	T72	S7	419
		T72	A32	572
		T72	A62	573
		T72	A63	573
		T72	A64	573

ship chips as well as samples of strong scatterer, bridge, coastal land, islands, sea, and land clutter. In this article, ten categories of samples are used to verify the effectiveness of the proposed MSA-SCNN model, and the SAR images of ten categories are shown in Fig. 8. In addition, a summary of FUSAR-Ship for training and testing datasets is listed in Table IV. Like the MSTAR dataset, the fine-class labels are divided into several super-class labels, named ships (cargo, fishing, tanker, and other

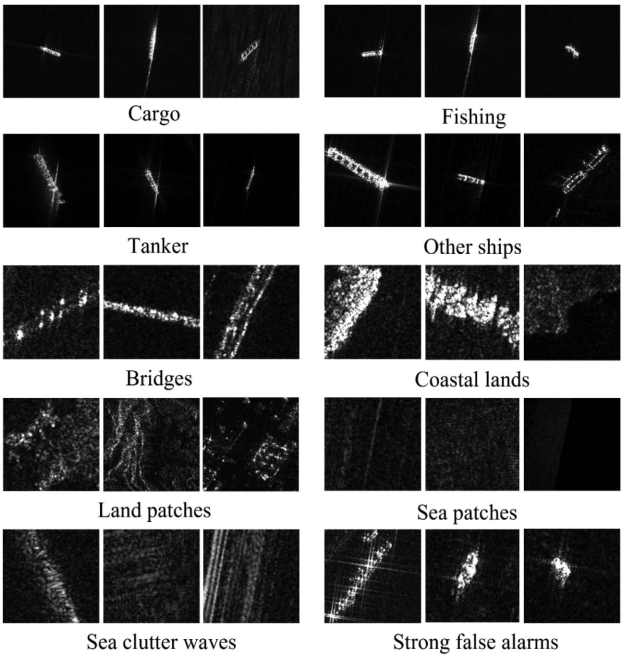

Fig. 8. SAR images of ten categories in the FUSAR-Ship dataset.

TABLE IV
NUMBER OF TRAINING AND TEST IMAGES FOR THE FUSAR-SHIP EXPERIMENTAL SETUP

Super-classes	Fine-classes	Number	
		Train	Test
Ships	Cargo	366	156
	Fishing	248	106
	Tanker	150	64
	Other ships	312	133
Lands	Bridges	1023	438
	Coastal lands	707	303
	Land patches	1137	487
Sea	Sea patches	1250	535
	Sea clutter waves	1378	590
Strong scatterer	Strong false alarms	299	128

ships), lands (bridges, coastal lands, and land patches), sea (sea patches and sea clutter waves), and strong scatterer (strong false alarms).

Both MSTAR and FUSAR-Ship datasets have different sizes of images for categories. To ensure that the SAR image size is the same as the network input size (88×88), this article adopts the random cropping method to process the data uniformly. Ten image slices are cropped for each SAR image as the training set, one of which is the center crop as the raw SAR image dataset, and the rest nine are randomly cropped as an expanded dataset. The randomly cropped SAR target may be incomplete, which is helpful to improve the generalization ability of the network, while avoiding the overfitting problem. Fig. 9 shows the result of randomly cropping one of the SAR images.

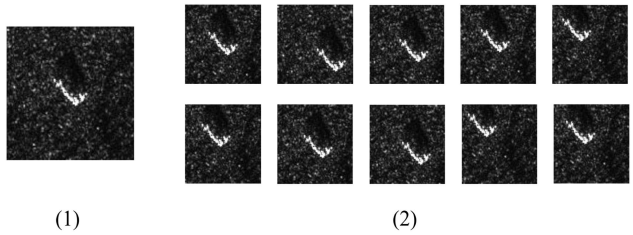

Fig. 9. Schematic diagram of random cropping from the MSTAR dataset. (1) Raw SAR image. (2) Cropped SAR image slices. The first one on the top row is a center-cropped SAR image slice, and the remaining nine are randomly cropped SAR image slices.

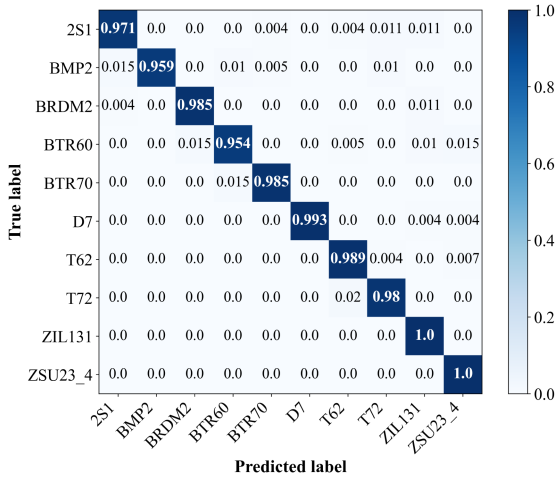

Fig. 10. Confusion matrix of fine-class classification result by MSA-SCNN under SOC experiment (Accuracy rate: 98.31%).

B. Results and Analysis Under SOC

In this experimental setup, the performance of the proposed architecture will be evaluated under the SOC dataset. The super-class and fine-class labels are set according to Table I. And all the training SAR images are randomly cropped to expand the datasets. Each convolutional layer has four channels with dimensions 3 × 3, 5 × 5, 7 × 7, and 9 × 9. Meanwhile, considering that the super-class and fine-class losses are equally significant, the super-class loss weight λ is set to 0.5.

Fig. 10 shows the fine-class classification performance of the proposed MSA-SCNN in the form of the confusion matrix on the SOC experiment. The confusion matrix is the visualization tool used to evaluate the target classification performance, whose rows correspond to the true category labels of the target, and columns represent the predicted category labels of the target. The total accuracy of ten fine-class classifications is calculated to reach 98.31%. From Fig. 10, the diagonal elements are much larger than the confusion matrix in other positions, which means MSA-SCNN has high classification accuracy for each target type in the SOC experiment. For the ZIL131 and ZSU23/4 categories, the accuracy even achieves 100%.

MSA-SCNN also has classifier output in the super-class branch. Fig. 11 shows the super-class classification performance in the form of a confusion matrix on the SOC experiment. It can

TABLE V
CLASSIFICATION ACCURACY OF VARIOUS TARGETS IN DIFFERENT METHODS UNDER SOC EXPERIMENT (%)

Method	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU23/4	Total
MSRIHL-CNN [21]	85.77	82.56	96.35	96.41	98.98	98.18	91.21	100	98.54	93.80	94.14
VDCNN [37]	96.30	98.35	98.70	93.10	96.60	99.45	97.95	98.95	99.25	99.45	97.81
VGG-Net [16]	95.71	95.63	97.81	94.87	98.32	99.10	97.93	97.81	99.64	99.71	97.75
Res-Net [25]	93.07	99.49	99.27	91.10	89.80	98.54	90.11	90.31	97.81	100	95.38
ViT-B/16 [38]	95.27	96.41	98.46	96.41	91.84	100	99.17	98.95	100	100	97.98
Swin-T [39]	96.91	92.31	95.25	96.92	100	100	98.91	99.49	99.27	100	98.06
CA-MCNN [40]	99.64	97.27	99.27	99.64	98.98	99.63	99.64	93.85	100	94.16	97.81
MKSFF-CNN [22]	93.80	94.36	97.45	98.46	99.49	99.27	95.24	100	99.27	97.81	97.44
MSA-SCNN	**97.08**	**95.92**	**98.54**	**95.39**	**98.47**	**99.27**	**98.90**	**97.96**	**100**	**100**	**98.31**

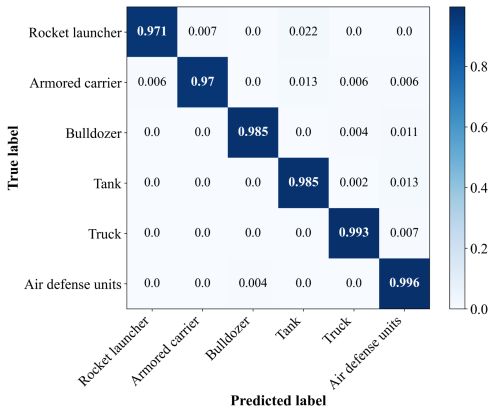

Fig. 11. Confusion matrix of super-class classification result by MSA-SCNN under SOC experiment (Accuracy rate: 98.02%).

be seen that the classification accuracy under six super-class labels reaches 98.02%, which indicates that the features learned by the MSA-SCNN model can distinguish super-classes.

In order to comprehensively validate the superiority of the proposed MSA-SCNN, a series of commonly used SAR target classification methods are compared with the proposed MSA-SCNN. These methods include MSRIHL-CNN [21], VDCNN [37], VGG-Net [16], Res-Net [25], ViT-B/16 [38], Swin-T [39], CA-MCNN [40], and MKSFF-CNN [22]. MSRIHL-CNN optimally fuses the deep features extracted by CNN and the local edge features extracted by Haar-like template. VDCNN is a multiview deep neural network that fuses SAR image features from multiple views for the same target layer by layer. VGG-Net and Res-Net are two commonly used deep neural network models. ViT and Swin-T are current existing state-of-the-art methods for image classification. CA-MCNN adopts the ASC model to extract SAR target component information and then fuses it with the deep features of CNN to improve the recognition accuracy. MKSFF-CNN uses convolutional kernels of different sizes to extract the multikernel-size deep features of the SAR target, and then, these features are fused in an optimal way to acquire the lowest loss.

Table V displays the classification accuracy of these methods under the SOC experiment. MSRIHL-CN, VDCNN, and CA-MCNN introduce the prior knowledge in the SAR domain, which improves the completeness of SAR feature representation. However, the way to acquire prior knowledge is very complicated, and the prior information is easily changed by external factors. VGG and Res-Net increase the network depth to extract the deep features of the SAR target and improve the classification accuracy. However, they use fixed-size convolutional kernels, which lose part of the scale features of SAR images. MKSFF-CNN uses convolutional kernels of different sizes to extract the multiscale features, but does not pay enough attention to significant features and has information redundancy. ViT and Swin-T replace the backbone network from CNN to transformer structure, and the accuracy is improved.

The MSA-SCNN proposed in this article focuses on important features and suppresses unnecessary features due to combining multiscale feature fusion with the attention module. Meanwhile, the prior knowledge of super-class labels is introduced to increase the multiscale feature difference between categories. The final accuracy rate reaches 98.31%, which is higher than other methods. In particular, the 2S1, T62, ZIL131, and ZSU23/4 categories have higher accuracy, and it is found that these categories belong to different super-class labels, indicating that the model has indeed learned the different features of super-classes. When fusing features into the fine-class branch, the differences between the categories are increased, and the recognition accuracy is also improved.

To more intuitively explain the effectiveness of MSA-SCNN, the raw SAR test images and the output vectors of the fully connected layers of MSA-SCNN are mapped to a 2-D Euclidean space by the t-distributed stochastic neighbor embedding (t-SNE) [41] algorithm. The t-SNE is a powerful dimensionality reduction algorithm that can help us study the distribution characteristics of high-dimensional data in low-dimensional space.

Fig. 12 illustrates the input SAR images and the fine-class classification output of Res-Net, MKSFF-CNN, and MSA-SCNN. It can be observed that the visualization results of the raw samples are mixed and difficult to classify. The outputs of Res-Net and MKSFF-CNN have been significantly improved, but the feature distribution is uneven, which is easy to misidentify. However, after being processed by MSA-SCNN, the samples with the same class label became closer, and the feature distribution distances between categories are farther, so they are easier to be recognized. Fig. 13 illustrates the super-class classification

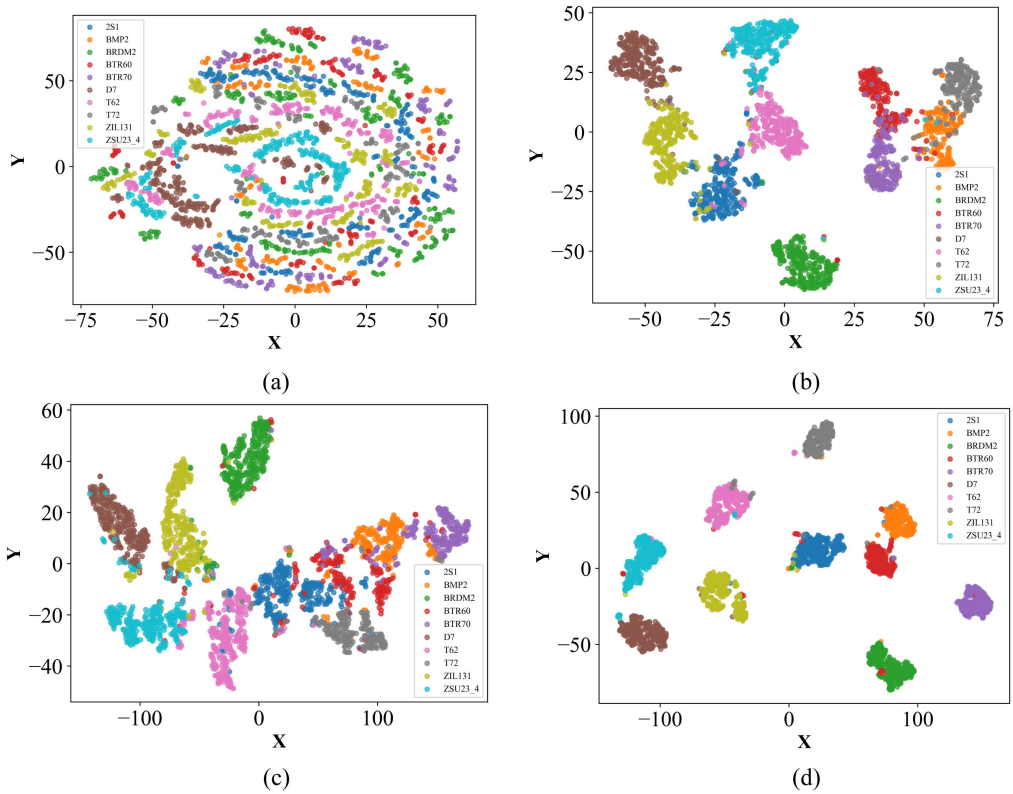

Fig. 12. Visualization of fine-class classification result. Points with the same color belong to the same target class. (a) Input of SAR images under the fine-class labels. (b) Output of Res-Net. (c) Output of MKSFF-CNN. (d) Output of the fine-class branch in MSA-SCNN.

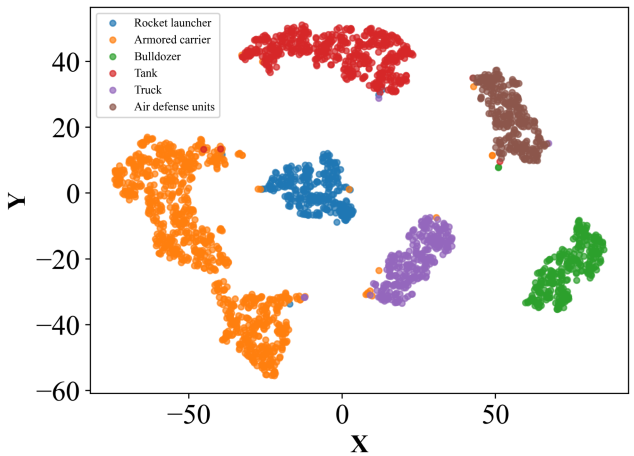

Fig. 13. Visualization of super-class classification result of MSA-SCNN. Points with the same color belong to the same target class.

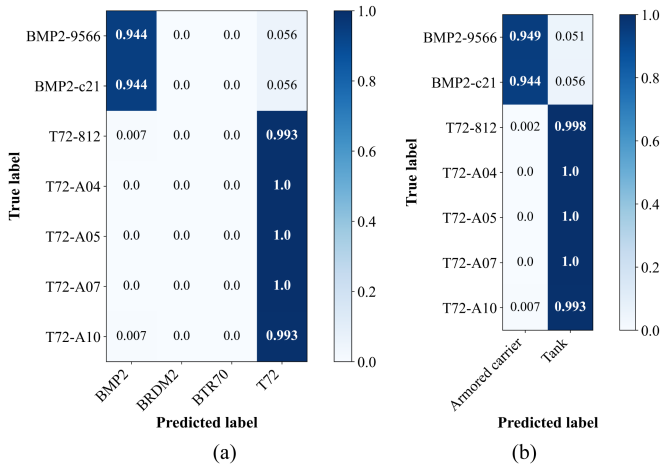

Fig. 14. Confusion matrix of classification result by MSA-SCNN under EOC-C experiment. (a) Fine-class classification result. Accuracy rate: 98.46%. (b) Super-class classification result. Accuracy rate: 98.57%.

output of MSA-SCNN. The introduction of super-class labels also makes the super-class samples more clearly separated from each other.

C. Results and Analysis Under EOC

In the EOC experiment, all training SAR images of EOC are also randomly cropped to expand the datasets. Each convolutional layer has four channels with dimensions 3×3, 5×5, 7×7, and 9×9. Meanwhile, considering that the super-class and fine-class losses are equally significant, the super-class loss weight λ is set to 0.5.

Figs. 14 and 15 show the fine-class and super-class classification performance of the proposed MSA-SCNN in the form of the confusion matrix on the EOC-C and EOC-V experiments. It can be seen from the figures that both the fine-class accuracy and the super-class accuracy are around 97%. Especially for the T72 variants A04, A32, and A62, the accuracy has reached 100%.

TABLE VI
CLASSIFICATION ACCURACY OF VARIOUS TARGETS IN DIFFERENT METHODS UNDER EOC-C EXPERIMENT (%)

Method	BMP2-9566	BMP2-c21	T72-812	T72-A04	T72-A05	T72-A07	T72-A10	Total
CGM [13]	85.95	89.95	78.87	70.16	87.61	76.79	82.01	81.22
Res-Net [25]	83.88	86.71	98.30	99.68	99.43	98.48	96.47	95.24
VDCNN [37]	92.55	95.75	98.75	91.57	99.15	90.52	96.90	95.45
ViT-B/16 [38]	93.46	95.80	98.35	95.11	96.68	91.27	94.18	94.87
Swin-T [39]	90.65	93.24	97.42	96.34	98.08	95.81	97.88	95.85
MSA-SCNN	**94.39**	**94.41**	**99.30**	**100**	**100**	**100**	**99.29**	**98.46**

TABLE VII
CLASSIFICATION ACCURACY OF VARIOUS TARGETS IN DIFFERENT METHODS UNDER EOC-V EXPERIMENT (%)

Method	T72-S7	T72-A32	T72-A62	T72-A63	T72-A64	Total
CGM [13]	85.92	83.39	77.31	71.20	68.94	76.86
Res-Net [25]	98.09	98.75	96.14	95.68	93.63	96.37
VDCNN [37]	94.72	99.62	96.87	93.12	92.95	95.46
ViT-B/16 [38]	98.60	97.73	90.23	92.67	99.28	95.50
Swin-T [39]	97.90	99.48	96.86	93.19	96.90	96.86
MSA-SCNN	**98.81**	**100**	**100**	**100**	**99.13**	**99.63**

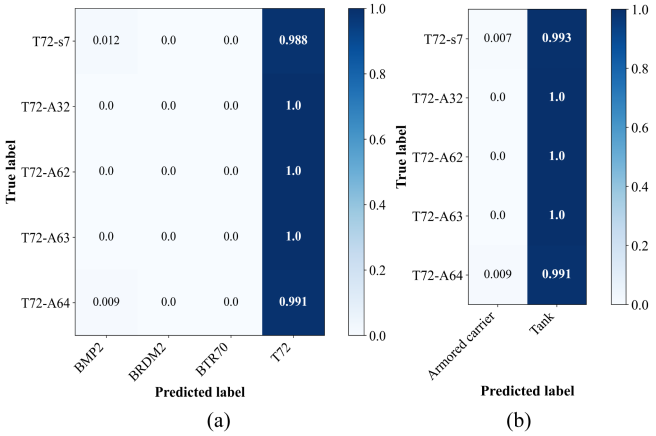

Fig. 15. Confusion matrix of classification result by MSA-SCNN under EOC-V experiment. (a) Fine-class classification result. Accuracy rate: 99.63%. (b) Super-class classification result. Accuracy rate: 99.70%.

Despite the targets having lots of variants, after introducing the super-class labels and attention module, the common features of the same super-class can be extracted so that the target can be well recognized. These results also substantiate that the proposed network can adapt to the target classification of different types and has a good generalization ability.

To sufficiently verify the superiority of the proposed MSA-SCNN under the EOC experiment, this section compares the MSA-SCNN with other methods under the EOC experiment. Tables VI and VII display the EOC-C and EOC-V classification accuracy of different methods. It can be seen that the CGM has lower accuracy than the deep learning methods and cannot effectively extract the target features. Although Res-Net increases the network depth, it cannot fully represent SAR target features only by fixed-size convolutional kernels. The performance of ViT and Swin-T on EOC is not as good as that on SOC, and the accuracy is similar to the CNN methods. VDCNN fuses multiview SAR target feature information, but it cannot effectively extract common features from variant targets, resulting in no significant improvement in classification accuracy. The MSA-SCNN combines multiscale feature fusion with the attention module to get a more complete SAR target feature representation and introduces super-class labels for different category types. Therefore, the network learns the common features of the super-class and assists in the final fine classification. And the MSA-SCNN can adapt to different types and configuration variants of the target, with 98.46% accuracy on EOC-C and 99.63% accuracy on EOC-V. Finally, these results also indicate that the proposed MSA-SCNN outperforms other methods.

D. Results and Analysis Under FUSAR-Ship

The FUSAR-Ship dataset has a variety of complex categories, including ships, lands, sea clutter waves, strong scatterers, etc., which brings huge challenges to SAR image classification. Meanwhile, it can also verify the effectiveness of the proposed method MSA-SCNN. Figs. 16 and 17 show the fine-class and super-class classification performance of the proposed MSA-SCNN in the form of the confusion matrix. It can be seen from the figures that the fine-class accuracy and the super-class accuracy are 94.05% and 98.44%, respectively.

Similar to the previous experiments, this section compares MSA-SCNN with other methods, and the accuracy of each category is recorded in Table VIII. In terms of total accuracy, the proposed MSA-SCNN is more than 10% higher than other CNN methods. ViT and Swin-T, the two optimal image classification methods, also have higher accuracy than the conventional CNN methods. In terms of each category, the cargo, fishing, and tanker categories have lower accuracy than the others in all methods, whereas MSA-SCNN introduces super-class labels,

TABLE VIII
CLASSIFICATION ACCURACY OF VARIOUS CATEGORIES IN DIFFERENT METHODS UNDER FUSAR-SHIP EXPERIMENT (%)

Method	Bridges	Cargo	Coastal lands	Fishing	Land patches	Other ships	Sea clutter waves	Sea patches	Strong false alarms	Tanker	Total
VGG-Net [16]	89.50	53.21	78.88	37.74	77.00	45.86	95.76	95.51	80.47	50.00	81.67
Res-Net [25]	64.38	69.87	77.23	34.91	80.49	91.73	98.47	97.94	96.88	87.5	83.71
ViT-B/16 [38]	98.40	69.23	98.35	58.49	97.54	81.20	99.15	98.69	93.75	65.62	93.78
Swin-T [39]	98.63	79.49	95.71	44.34	96.71	66.16	98.64	97.76	92.97	70.31	92.55
MKSFF-CNN [22]	91.78	66.67	89.77	50.94	89.53	82.71	98.31	98.69	76.56	78.12	89.59
MSA-SCNN	**97.26**	**79.49**	**93.07**	**60.38**	**96.30**	**87.97**	**98.81**	**98.69**	**92.19**	**84.38**	**94.05**

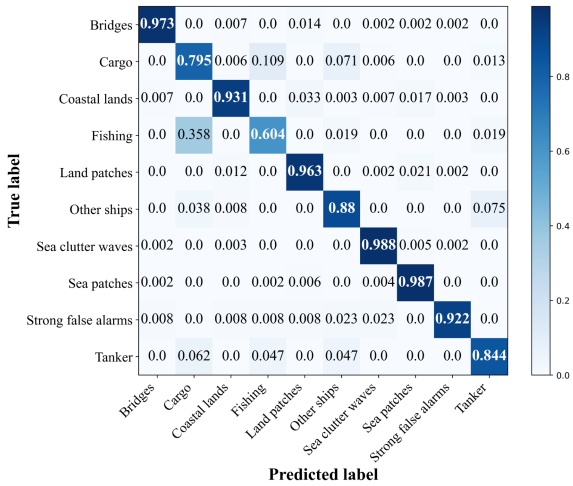

Fig. 16. Confusion matrix of fine-class classification result by MSA-SCNN under FUSAR-Ship dataset (Accuracy rate: 94.05%).

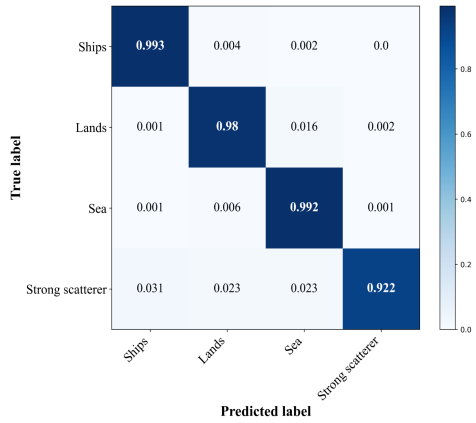

Fig. 17. Confusion matrix of super-class classification result by MSA-SCNN under FUSAR-Ship dataset (Accuracy rate: 98.44%).

which increases the feature difference between categories and makes them easier to distinguish.

All the experiments carried out have manifested that the proposed MSA-SCNN has a good recognition capability in both the MSTAR and FUSAR-Ship datasets, and clearly verify the superiority of the proposed framework.

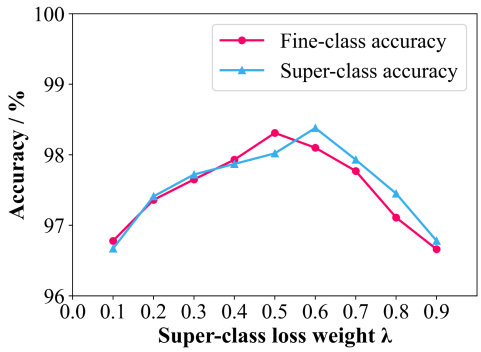

Fig. 18. Line graph of classification accuracy of different super-class loss weights under SOC experiment.

E. Analysis of Super-Class Loss Weight λ

This section gives the classification performance of the proposed MSA-SCNN model under different super-class loss weights λ. The experiments are carried out under the SOC dataset, with ten fine-class and six super-class labels. All conditions follow the settings in Table I and only change the different super-class loss weights λ. Fig. 18 clearly shows the change in fine-class and super-class accuracies with a line graph.

It can be seen from the figure that with the increase of λ, the super-class and fine-class accuracies both increase first and then decrease. When λ is 0.5, the super-class and fine-class losses are considered equally important, and the fine-class accuracy reaches the maximum value of 98.31%. However, the super-class accuracy is not the maximum at this time. When λ is 0.6, the super-class accuracy reaches the maximum value of 98.38%. The model trained with a larger λ pays more attention to the super-class features. When the λ is too large, the fine-class branch has little effect on the network, and the learned features are not enough to distinguish fine classes. In this way, the total loss of the network is too large, and the accuracy of super-class and fine-class both decreases.

When λ approaches 0, the network hardly pays attention to the features extracted by the super-class branch, which reduces the final feature difference of the target. So the super-class and fine-class accuracy are both low. These experiments also certify the importance of introducing super-class labels. In practical applications, the value of λ can be flexibly changed as needed.

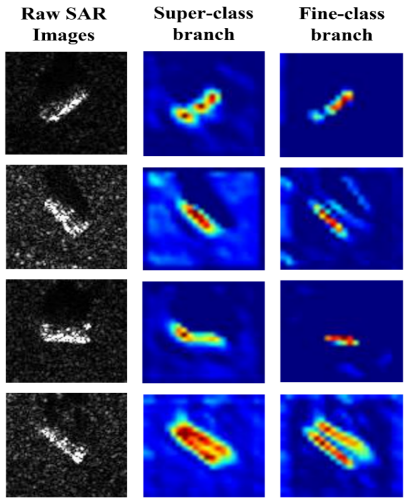

Fig. 19. CBAM visualization results of super-class and fine-class branches of MSA-SCNN.

F. Attention Visualization

To analyze the effect of the attention module CBAM, we apply the Grad-CAM [42] to the super-class and fine-class branches using SAR images from the SOC dataset. Grad-CAM is a recently proposed visualization method that calculates the importance of the spatial locations in convolutional layers. By observing the regions that MSA-SCNN has considered important for predicting a class, we attempt to look at how this network is making good use of multiscale features.

The best visualizations are often obtained after the deepest convolutional layer in the network, and localizations get progressively worse at shallower layers [42]. Therefore, this article selects the last CBAM layer features of the super-class and fine-class branches in MSA-SCNN to generate visualization results. Fig. 19 illustrates the CBAM visualization results of different branches of MSA-SCNN, and Grad-CAM results clearly show areas of interest. It can be seen that different branches pay different attention to SAR targets. The super-class branch focuses on the global contour features of the SAR target, whereas the fine-class branch pays more attention to the local detail features. Therefore, the two branches make full use of the multiscale features of the SAR target. In addition, the introduction of the attention module makes the model focus on the target features, and the background clutter features of the SAR image are suppressed. Finally, MSA-SCNN fuses the features of the two branches to further increase the multiscale feature difference between categories and improve the generalization ability of the model, which is well proved by the EOC and the FUSAR-Ship experimental results.

G. Classification Accuracy Evaluation Under Small-Size Training Datasets

In order to show that the proposed MSA-SCNN still has good performance even with a small sample size, a series of comparative experiments are carried out under the incomplete SOC datasets. A certain proportion of samples are randomly selected for each category in SOC datasets for training. Fig. 20 shows the classification accuracy of each method under different numbers of training samples with the line graph.

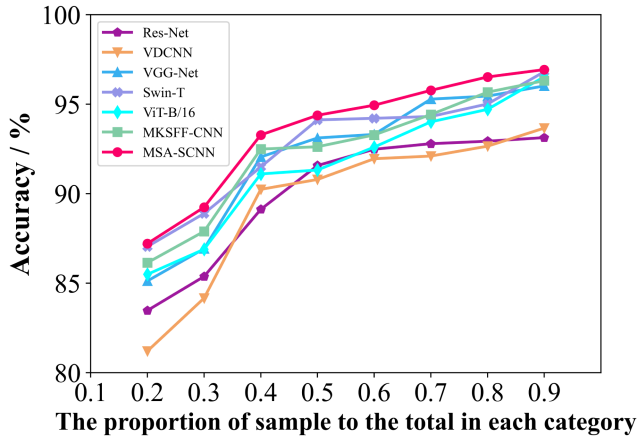

Fig. 20. Classification accuracy of different methods under a small sample size in the SOC datasets.

TABLE IX
NUMBER OF PARAMETERS, MODEL SIZE, AND COMPUTATION TIME SPENT BY EACH METHOD

Method	Number of parameters /10^6	Model size /MB	Computation time per batch/s
VGG-Net [16]	17.82	203	0.225
Res-Net [25]	38.21	436	0.274
VDCNN [37]	15.61	85	0.116
ViT-B/16 [38]	85.81	327	0.118
Swin-T [39]	27.53	105	0.112
MKSFF-CNN [22]	71.73	348	0.095
MSA-SCNN	**6.91**	**27**	**0.107**

It can be seen from Fig. 20 that no matter how the proportion of samples changes, the classification accuracy of MSA-SCNN is always higher than that of other methods. Even if the proportion of the sample is 0.2 in each category, the classification accuracy of MSA-SCNN can reach 87.21%. This can be explained by the introduction of the attention module and super-class labels, which makes MSA-SCNN easier to focus on important features and increase the feature differences between categories with small samples. Meanwhile, MSA-SCNN avoids the problem of overfitting and improves the generalization ability of the model.

H. Model Size and Computing Efficiency Evaluation

The computation time is an important indicator of the efficiency of the classification method. All experiments were run under the same computing station, which is composed of an Intel Core i7-7700K CPU with 4.20 GHz frequency, a 32.0 GB memory, and an Nvidia GeForce RTX 3090 GPU with 24.0 GB memory. The number of parameters, model size, and the computation time per batch including 64 SAR images for all models are recorded in Table IX.

TABLE X
RESULTS OF THE ABLATION EXPERIMENT

Multi-scale feature fusion		Attention module CBAM		Super-class branch		Size of convolutional kernels	Accuracy (%)
Yes	No	Yes	No	Yes	No		
	✓		✓		✓	3×3	94.32
	✓	✓			✓	3×3	96.13
	✓		✓	✓		3×3	96.48
	✓	✓		✓		3×3	97.13
✓		✓		✓		3×3 and 5×5	97.56
✓		✓		✓		3×3 and 7×7	97.44
✓		✓		✓		5×5 and 7×7	97.26
✓		✓		✓		5×5 and 9×9	97.13
✓		✓		✓		3×3, 5×5, and 7×7	98.26
✓		✓		✓		3×3, 5×5, and 9×9	97.93
✓		✓		✓		5×5, 7×7, and 9×9	97.46
✓		✓		✓		**3×3, 5×5, 7×7, and 9×9**	**98.31**

In terms of the model size and parameters, MSA-SCNN is much smaller than that of other models. This is because the designed network structure is simple and the number of network layers is small. Compared to MSA-SCNN, VGG-Net, Res-Net, and MKSFF-CNN have higher structural complexity, so the model has more parameters. The ViT model using the transformer structure has the largest number of parameters. Swin-T uses the tiny version, so its model parameters are less than ViT. MKSFF-CNN fuses the multiscale features of each layer into the final fully connected layer, resulting in a sharp increase in the parameters of the fully connected layer. Compared to MKSFF-CNN, the proposed MSA-SCNN introduces an attention module to avoid the redundancy of multiscale feature information and greatly reduce the parameters of the fully connected layer. In addition, in terms of computation time, MSA-SCNN has a shorter inference time than VGG-Net and Res-Net with deeper networks. MSKFF-CNN has a parallel multiscale feature extraction network, so the computation time is equivalent to MSA-SCNN. In general, the proposed MSA-SCNN model has a simpler structure and shorter computation time than other methods.

I. Ablation Experiment

The ablation experiments are designed under the SOC datasets to illustrate the superiority of MSA-SCNN objectively and comprehensively. The MSA-SCNN is divided into three parts: multiscale feature fusion, attention module CBAM, and super-class branch. Table X records the impact of each part on the SAR target classification performance.

According to Table X, all the multiscale feature fusion, attention module, and super-class branch methods are beneficial for enhancing classification accuracy. First, only adding the super-class branch improves the classification accuracy by about 2%, reaching 96.48%, demonstrating the effectiveness of super-class labels in the MSA-SCNN. And only adding the attention module also improves the accuracy. Then, after applying convolutional kernels with different sizes, the feature representation of SAR targets is more complete and the classification accuracy is better. When two convolutional kernels of different sizes are combined and used for feature extraction, the classification accuracy obtained increases by about 3% to reach 97%. When three convolutional kernels of different sizes are combined, the accuracy is further improved. Finally, when convolutional kernels with sizes of 3×3, 5×5, 7×7, and 9×9 are combined, such as the proposed MSA-SCNN, the classification accuracy obtained increases by about 4% to reach 98.31%. In addition, the ablation experiment also proves that it is reasonable for the proposed MSA-CNN to choose the convolutional kernels with sizes of 3×3, 5×5, 7×7, and 9×9 for feature extraction.

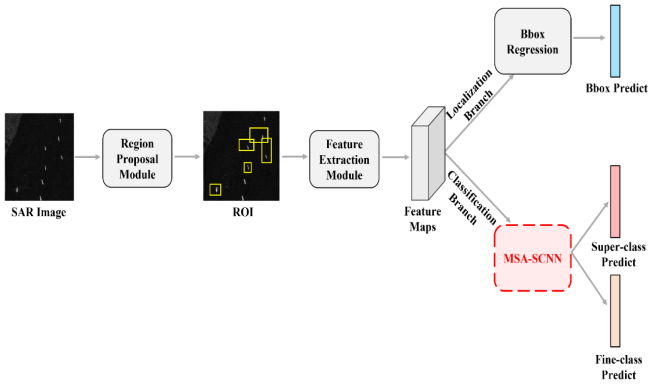

Fig. 21. MSA-SCNN model combined with two-stage detector for SAR target detection.

J. Combine With Detection Networks

The proposed MSA-SCNN model can be combined with the detector to achieve the task of SAR target detection. Object detection methods are usually divided into two-stage detectors and one-stage detectors. In two-stage detectors, e.g., Faster-RCNN [43] and FPN [44], the ROIs are generated by the region proposal module in the first stage. Then, the features of these proposals are processed by two branches of bounding box regression and classification. Since the bounding box regression and classification of the two-stage detector are separated, the proposed MSA-SCNN can replace the original classification branch. Fig. 21 shows the combination of MSA-SCNN and the two-stage detector.

In one-stage detectors, e.g., single shot detector (SSD) [45], you only look once (YOLO) [46], the network directly predicts locations and class labels of the potential object at several feature maps without ROI proposals. Therefore, if MSA-SCNN is to be combined with the single-stage detector, a separate branch needs to be added, and this work will be studied in the future.

V. CONCLUSION

This article proposes a novel network called MSA-SCNN for SAR target classification. First, this method combines multiscale feature fusion with the attention module, so that the network focuses on target features, suppresses the background clutter,

and avoids information redundancy. Second, MSA-SCNN introduces super-class labels to extract the common features of the super-classes, which are fused into the fine-class branch to increase the feature difference between the categories. Finally, experiments on the MSTAR and FUSAR-Ship datasets show that MSA-SCNN can achieve better classification performance than the traditional CNN methods. Especially in EOC and FUSAR-Ship experiments, the generalization ability of MSA-SCNN is stronger. Future work will include research on other prior knowledge of SAR images and MSA-SCNN combined with the one-stage detector.

References

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

[2] A. K. Mishra and B. Mulgrew, "Automatic target recognition," in *Encyclopedia of Aerospace Engineering*. Hoboken, NJ, USA: Wiley, 2010, pp. 1–8.

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

[4] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009.

[5] G. Gao, "An improved scheme for target discrimination in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 277–294, Jan. 2011.

[6] L. M. Novak, G. J. Owirka, W. S. Brower, and A. L. Weaver, "The automatic target-recognition system in SAIP," *Lincoln Lab. J.*, vol. 10, no. 2, pp. 187–202, 1997.

[7] K. Ikeuchi, M. D. Wheeler, T. Yamazaki, and T. Shakunaga, "Model-based SAR ATR system," *Proc. SPIE*, vol. 2757, pp. 376–387, Jun. 1996.

[8] K. El-Darymli, E. W. Gill, P. Mcguire, D. Power, and C. Moloney, "Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review," *IEEE Access*, vol. 4, pp. 6014–6058, 2016.

[9] A. K. Mishra and B. Mulgrew, "Bistatic SAR ATR using PCA-based features," *Proc. SPIE*, vol. 6234, pp. 244–252, 2006.

[10] X. Liu, Y. Huang, J. Pei, and J. Yang, "Sample discriminant analysis for SAR ATR," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2120–2124, Dec. 2014.

[11] Q. Zhao and J. C. Principe, "Support vector machines for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 643–654, Apr. 2001.

[12] Y. Sun, Z. Liu, S. Todorovic, and J. Li, "Adaptive boosting for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 1, pp. 112–125, Jan. 2007.

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

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

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representation*, 2015, pp. 1–14.

[17] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[18] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retrieval*, vol. 7, no. 2, pp. 87–93, 2018.

[19] S. A. Wangner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2861–2872, Dec. 2016.

[20] 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.

[21] J. Ai, R. Tian, Q. Luo, J. Jin, and B. Tang, "Multi-scale rotation-invariant Haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10070–10087, Dec. 2019.

[22] J. Ai, Y. Mao, Q. Luo, L. Jia, and M. Xing, "SAR target classification using the multikernel-size feature fusion-based convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2021, Art. no. 5214313, doi: 10.1109/TGRS.2021.3106915.

[23] L. Zhang et al., "Domain knowledge powered two-stream deep network for few-shot SAR vehicle recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5215315, doi: 10.1109/TGRS.2021.3116349.

[24] K. Li, N. Y. Wang, Y. Yang, and G. Wang, "SGNet: A super-class guided network for image classification and object detection," in *Proc. Conf. Robots Vis.*, Burnaby, BC, Canada, 2021, pp. 127–134.

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

[26] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[27] B. Ding, G. Wen, X. Huang, C. Ma, and X. Yang, "Target recognition in synthetic aperture radar images via matching of attributed scattering centers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3334–3347, Jul. 2017.

[28] Y. Zhou, Q. Hu, and Y. Wang, "Deep super-class learning for long-tail distributed image classification," *Pattern Recognit.*, vol. 80, no. 1, pp. 118–128, 2018.

[29] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[30] Y. Zhou and R. Chellappa, "Computation of optical flow using a neural network," in *Proc. IEEE Int. Conf. Neural Netw.*, 1988, pp. 71–78.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.

[32] K. Nikos and Z. Sergey, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent.*, Paris, France, 2017.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] J. Bouvrie, "Notes on convolutional neural networks," Center Biol. Comput. Learn., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2006, pp. 38–44.

[35] T. D. Ross, S. W. Worrell, V. J. Velten, J. C. Mossing, and M. L. Bryant, "Standard SAR ATR evaluation experiments using the MSTAR public release data set," *Proc. SPIE*, vol. 3370, pp. 566–573, Sep. 1998.

[36] X. Hou, W. Ao, Q. Song, J. Lai, H. Wang, and F. Xu, "FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition," *Sci. China (Inf. Sci.)*, vol. 63, no. 4, pp. 40–58, 2020.

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

[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[40] Y. Li, L. Du, and D. Wei, "Multiscale CNN based on component analysis for SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5211212, doi: 10.1109/TGRS.2021.3100137.

[41] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 1532–4435, 2008.

[42] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 91–99, Jun. 2015.

[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[45] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "Y ou only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

Di Wang received the B.S. degree in electronic information science and technology from the Hefei University of Technology, Hefei, China, in 2021. He is currently working toward the M.S. degree in information and communication engineering from the National University of Defense Technology, Changsha, China.

His research interests include SAR image interpretation, radar targets detection, and artificial intelligence.

Yongping Song received the B.S. degree in electronic engineering, and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2012, 2014, and 2019, respectively.

He was a Lecturer with the Air Force Early Warning Academy until 2021. He is currently an Assistant Researcher with the National University of Defense Technology. His research interests include MIMO radar image formation, radar targets detection, and radar antijamming.

Junnan Huang received the B.S. degree in electronic engineering in 2020 from the National University of Defense Technology, Changsha, China, where he is currently working toward the M.S. degree in electronic information.

His research interests include SAR image processing and SAR image interpretation.

Daoxiang An (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2011, respectively.

He is currently an Associate Professor with the National University of Defense Technology. His current research interests include ultra-wideband SAR, circular SAR, video SAR imaging, SAR interferometry, SAR-GMTI, and SAR image interpretation.

Leping Chen (Student Member, IEEE) received the B.S. degree in electronic engineering and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2011, 2014, and 2018, respectively.

He is currently a Lecturer with the National University of Defense Technology. His current research interests include circular synthetic aperture radar image formation and high-resolution SAR image formation.