(Received 7 June 2025, Accepted 8 August 2025, Scheduled 21 August 2025)

SDF-Net: A Space-Frequency Dynamic Fusion Network for SARATR

Xinlin He, Chao Li*, Kaiming Li, and Ying Luo

Information and Navigation School, Air Force Engineering University, Xi'an 710082, China

ABSTRACT: With the development of deep learning networks, convolutional neural network (CNN) and other techniques provide effective detection methods for synthetic aperture radar automatic target recognition (SAR ATR), and have been widely used. However, due to the objective factors such as complex scene interference inherent in SAR images, the recognition rate of traditional time-domain processing of SAR images is not high enough, which is still a key problem to be solved urgently. To solve this problem, we propose a space-frequency dynamic fusion network (SDF-Net). The network consists of four space-frequency joint processing (SJP) modules connected in series, each comprising convolutional layers and unbiased fast fourier convolution (UFFC) units at different scales to achieve joint feature extraction in the spatial and frequency domains. Building on a four-level series structure, residual paths from the original image features are introduced into the inputs of SJP2, SJP3, and SJP4. Additionally, residual paths from the features output by SJP1 are introduced into the inputs of SJP3 and SJP4, and from SJP2 into the input of SJP4. By incorporating residual paths of features from different stages, the network facilitates cross-stage information interaction, effectively integrating long-distance contextual information. At each fusion node, dynamically generated weights are used for feature fusion, followed by sequential progressive processing through spatial-frequency joint processing, ultimately leading to classification and recognition results. Experimental results on the MSTAR dataset and the FUSAR-Ship1.0 dataset show that compared to traditional methods, this network algorithm achieves a higher recognition rate.

1. INTRODUCTION

Synthetic Aperture Radar (SAR) is an active microwave sensor that emits electromagnetic pulses and coherently processes echoes received at different locations to obtain high-resolution imaging. Synthetic aperture radar enables all-day, all-weather earth observation without being limited by light and climatic conditions, and can even obtain information that it hides through the surface or vegetation. These characteristics make it of great practical value in the fields of military reconnaissance, disaster monitoring, topographic mapping, ocean observation, agriculture, and forestry [1–4]. However, the problems of spot noise, target azimuth sensitivity and complex scene interference inherent in SAR images lead to significant deficiencies in generalization and robustness of traditional methods based on manual features.

In recent years, deep learning has achieved success in the field of target detection [5, 6] and has been introduced into synthetic aperture radar ground target recognition [7–9] to significantly improve the detection performance, so it has been widely used [10–13]. In the field of image processing and computer vision, CNN [14–16] has played an irreplaceable role. Gao et al. [17] proposed an incremental learning method based on strong separability features to solve the problem of forgetting the knowledge learned before. Xu et al. [18] applied a new structural reparameterization method to optimize the learning focus of the network, so that the network can better focus on the key parts of the SAR image. Fukuzaki and Ikehara [19] accelerated the training of large kernel convolution by adjusting the size of the training image and convolution filter to a smaller scale. Wang et al. [20] proposed a two-stage coupled CNN architecture to solve the problem of noise robustness of CNNs. Zang et al. [21] proposed a new layer-wise relevance propagation (LRP) algorithm specifically for understanding the performance of CNN in SAR image target recognition. Marzi et al. [22] proposed a method to classify ten land cover types using a 3-D fully convolutional network (FCN) and 3-D ResNet-50 as the backbone, which was trained from scratch, and multi-phase Sentinel-1 SAR data.

From the above methods, it is evident that CNN and its various evolved algorithms have indeed achieved significant success in image recognition processing. However, it is undeniable that CNN still has many problems, such as sensitivity to the inherent noise of SAR data, limited azimuth robustness and rotation invariance, computational complexity and problems of real-time, insufficient multi-task cooperation and interpretability, and insufficient fusion of local features and global context information. The above limitations are mainly due to the design characteristics of CNN architecture, the complexity of SAR data itself and the diversity of practical application scenarios, which ultimately lead to the low recognition rate of SAR images by network models. Guo et al. [23] proposed a model called multi-level attention network, which effectively combined local structural features and long-distance contextual information to improve the representation ability and recognition accuracy of SAR images. Huang et al. [24] proposed a physically inspired hybrid attention to integrate prior knowledge into the learning process to enhance deep learning models. Li et al. [25] proposed a light attention module called Additive Attention module, which achieves similar performance to SE-ResNet-18 while reducing the computational cost.

* Corresponding author: Chao Li (lchegongda@163.com).

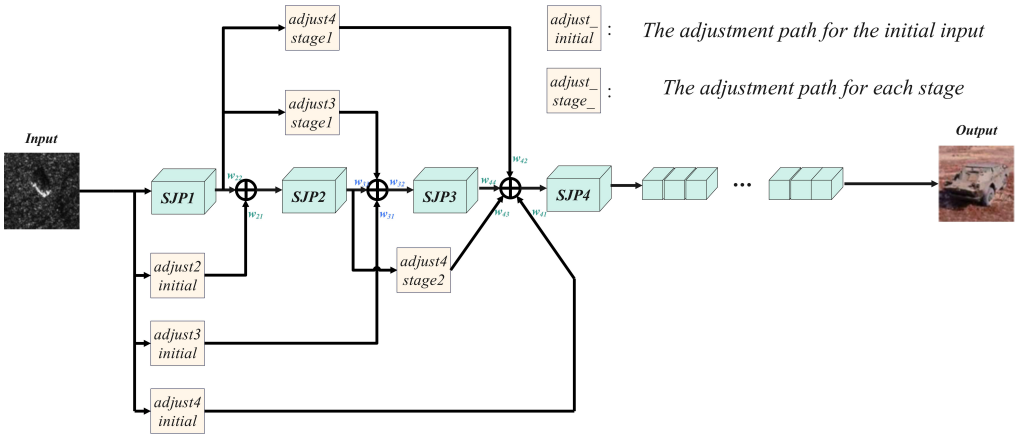

FIGURE 1. The overall structure of the SDF-Net network.

In recent years, fast Fourier transform (FFT) technology has been gradually promoted [26–29]. Inspired by this, we consider that the proper combination of FFT and CNN can provide help for SAR target detection and recognition. Based on the above analysis, to address the low recognition rate of SAR images processed in the time domain within traditional network models, this paper introduces a Space-Frequency Dynamic Fusion Network (SDF-Net) for spatial-frequency dynamic fusion. The network primarily consists of four cascaded spatial-frequency joint processing (SJP) modules and residual paths with dynamic attention mechanisms. Each SJP module includes a traditional time-domain convolutional layer, a frequency domain processing unit called Unbiased Fast Fourier Convolution (UFFC), and two residual paths from the input of the convolutional layer and UFFC, respectively. In the SJP, the residual paths are used to integrate initial information content and retain the original input features; the convolutional layer is used to extract local spatial features, progressively expanding the receptive field; the UFFC, directly connected to the convolutional layer, transforms traditional convolution operations into the frequency domain, capturing global features, thereby improving the single-time-domain processing of traditional networks and enhancing recognition performance. Furthermore, on the basis of a four-level serial structure, each fusion node features distinct residual paths. The residual paths that introduce original image features are fed into the inputs of SJP2, SJP3, and SJP4. Additionally, the residual paths that incorporate the output features from SJP1 are fed into the inputs of SJP3 and SJP4, while the residual paths that incorporate the output features from SJP2 are fed into the input of SJP4. These residual paths primarily use spatial alignment and network expansion techniques to adapt to the main structure. By introducing residual paths that incorporate features from different stages, information can be exchanged across stages, fully integrating long-distance contextual information. Finally, the trained weights are used for fusion, and after four layers of processing, the classifier module produces the final recognition result. The main contributions of this paper are as follows:

(1) The SJP module constructs the joint feature space of frequency and frequency.

(2) The dynamic adaptive fusion path is used to enhance the information interaction capability across stages.

The structure of this paper is summarized as follows. Section 2 introduces the network structure of SDF-Net in detail. Section 3 introduces the data and analyzes the experimental results to verify the feasibility and superiority of the network. Finally, Section 4 summarizes this paper.

2. METHODOLOGY

The overall structure of the SDF-Net network proposed in this paper is shown in Fig. 1. A four-level progressive feature extraction structure is adopted, and each stage includes three parallel paths: main convolution path, residual path, and frequency domain convolution path. After preprocessing, the input image passes through four feature extraction stages in sequence, ultimately completing classification via adaptive pooling and fully connected layers. The network's depth and width grow exponentially, with the number of channels expanding from 64 to 512, and the spatial dimensions downsampling from 224×224 to 14×14, forming a hierarchical feature pyramid.

2.1. Dynamic CBAM Module

The overall architecture of the module is shown in Fig. 2, which has a dual attentional synergy mechanism: channel attention and spatial attention. For the input feature map: channel attention uses global average pooling and a fully connected layer to learn channel weights to enhance the response of important feature channels and solve the "what" problem:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(:,:,i,j) \tag{1}$$

$$w_c = \sigma\left(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z_c)\right) \tag{2}$$

$$x_c = x \odot w_c \tag{3}$$

where $x \in R^{B \times C \times H \times W}$ denotes the input features; z_c denotes the average pooling result; H and W denote weight; W_1 and W_2 denote learnable parameters; σ denotes the Sigmoid function; $\odot$ denotes the channel-by-channel multiplication; x denotes the output features; and x_c denotes channel-by-channel multiplication output.

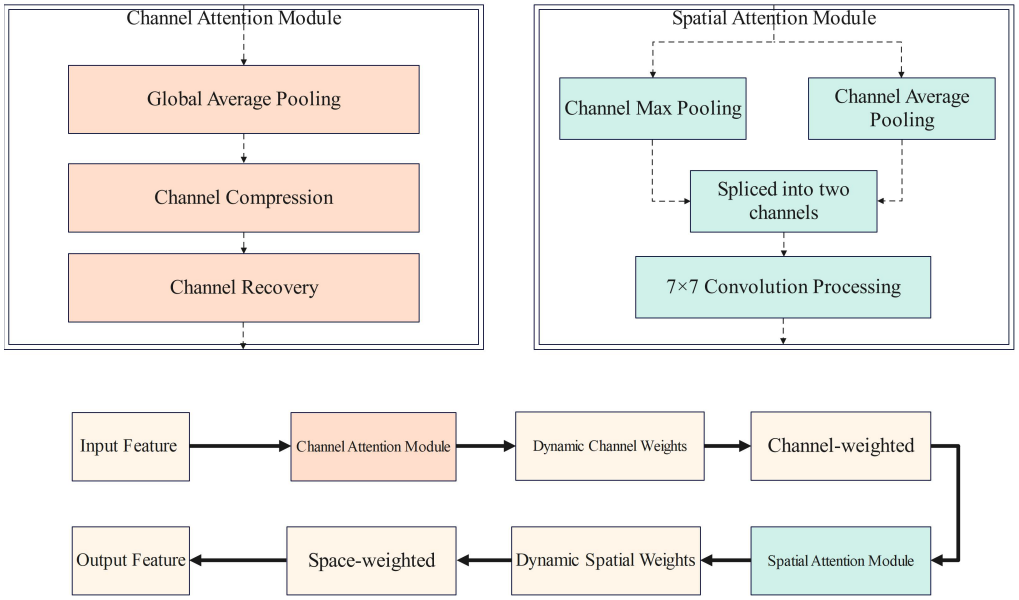

FIGURE 2. Schematic diagram of dynamic CBAM module.

Spatial attention utilizes the features of max pooling and average pooling, adopts 7×7 reflection padding convolution to learn spatial weight map, focuses on key areas, and solves the "where" problem:

$$z_s^{\max} = \max_c x_c \tag{4}$$

$$z_s^{avg} = \frac{1}{C}\sum_{c=1}^{C} x_c \tag{5}$$

$$w_s = \sigma\left(K_{7\times7} * [z_s^{\max}; z_s^{avg}]\right) \tag{6}$$

$$x_{out} = x_c \odot w_s \tag{7}$$

where $z_s^{\max}$ denotes the maximum pooling result, z_s^{avg} the average pooling result, $K_{7\times7}$ the 7×7 reflection padding convolution kernel, and [;] the channel concatenation operation.

Channel attention: compresses global spatial information to the channel dimension, learns the importance of the channel through the bottleneck structure (compression, activation, recovery), and then outputs the channel weight matrix.

Spatial attention: the channel dimension information is fused, and the large receptive field convolution (7×7) is used to capture the spatial relationship, then the spatial weight matrix is output.

The dynamic characteristics are as follows: the weight can be generated in real time according to the input content, and different samples or different positions have adaptive weights. Dual attention coordination conforms to the human visual mechanism. The channel attention first selects the important feature channels, and the spatial attention then focuses on the key spatial areas, forming a "channel-space" cascade attention mechanism. At the same time, it pays attention to "which features are important" (channel dimension) and "where is important" (space dimension).

Key advantages include:

(1) Using 1×1 convolution instead of full connection layer, 1×1 convolution is equivalent to the implementation of full connection layer but more suitable for convolution architecture.

(2) Reflective filling reduces the loss of boundary information and realizes the memory reuse of shared feature maps.

2.2. Space-Frequency Joint Processing (SJP) Module

2.2.1. Main Convolution Path

This path uses multi-scale convolution to extract local features in space, with the overall architecture shown in Fig. 3. The idea behind designing multi-scale convolutions lies in three aspects: First, it can achieve progressive receptive fields, aligning with the cognitive rule of moving from local to global. Second, through multi-scale convolutional kernels at deeper layers, it balances details and context while ensuring the complexity and readability of the network. Third, it employs channel doubling technology, gradually increasing from 64 to 512, maintaining an exponential increase in feature representation capability. In this way, the balance between the depth and width of the network keeps the utilization of graphics processing unit (GPU) at a high level, while ensuring that multi-scale convolution does

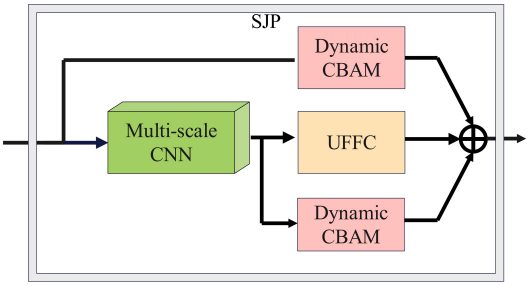

FIGURE 3. Schematic diagram of the SJP module.

TABLE 1. Key data flow changes.

Phase	Input Dimensions	Output Dimensions	Number of Channels	Core Operations	Purpose
1	224×224	112×112	64	5×5 convolution + three-way integration	Capture the underlying texture features
2	112×112	32×32	128	5×5 convolution + dynamic fusion + mixed downsampling	Establish intermediate semantic associations
3	32×32	16×16	256	$5 \times 5 + 3 \times 3$ multi-scale convolution combination	Combine global patterns with local details
4	16×16	8×8	512	$3 \times 3 + 5 \times 5$ convolution inverse order kernel mixing	Deep feature processing with refinement and integration
Classification	8×8	1×1	$1024 \rightarrow N$	global pooling + fully connected	Classification

not have too much impact on the complexity of parameters, clinching the efficiency of training. The key data flow changes are shown in Table 1.

Specific application:

$$\begin{aligned} &\text{Stage 1}: \quad x = W_{5\times5} * x_{in} \\ &\text{Stage 2}: \quad x = W_{5\times5} * x_{in} \\ &\text{Stage 3}: \quad x_1 = W_{5\times5} * x_{in}, \quad x_2 = W_{3\times3} * x_1 \\ &\text{Stage 4}: \quad x_1 = W_{3\times3} * x_{in}, \quad x_2 = W_{5\times5} * x_1 \end{aligned} \tag{8}$$

where x_{in} denotes the input feature, x, x_1, x_2 the convolution output result, $*$ the convolution operation, $W_{3\times3}$ the 3×3 convolution kernel, and $W_{5\times5}$ the 5×5 convolution kernel.

2.2.2. Unbiased Fast Fourier Convolution (UFFC)

UFFC module is the core of the network. The overall idea is to transform the traditional convolution operation to the frequency domain to capture the global feature dependence. As shown in Fig. 4, the specific steps are: First, the input image is filled to prevent edge effects (boundary artifacts) after FFT and ensure the integrity of frequency domain convolution. Then, the input image and convolution kernel are respectively FFT:

$$F(x) = \text{RFFT2D}(x_{padded}) \tag{9}$$

$$F(W) = \text{RFFT2D}(W_{conv}) \tag{10}$$

where x_{padded} denotes the result of the fill, W_{conv} the convolution kernel, and $F(x)F(W)$ the results of the Fourier transform for the filling results and the convolution kernel, respectively.

Then, element-by-element multiplication is performed in the frequency domain, and the result is inverted FFT. Finally, after cropping and filling, unbiased correction and weighted fusion operation are performed to ensure that the output size is consistent with the input:

$$y = \sum_{c=1}^{C} \mathcal{F}(x) \cdot \mathcal{F}(W) \tag{11}$$

$$y_{freq} = \text{IRFFT2D}(y) \tag{12}$$

where $\cdot$ denotes the element-by-element multiplication of complex numbers, y the result of multiplication of complex numbers, and y_{freq} the real result obtained by inverse Fourier transform.

The advantage of the module is that the $O(n \log n)$ complexity of FFT is more efficient than the $O(n^2)$ complexity of spatial convolution, which breaks through the local receptive field limit and efficiently deals with large receptive fields. In specific applications, when processing 512×512 images, it is faster than the 3×3 convolution and has stronger robustness.

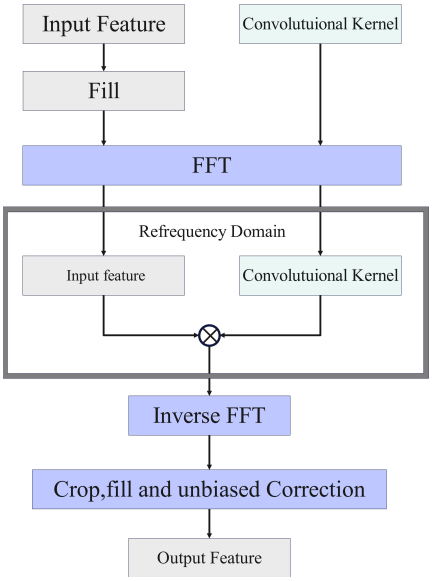

FIGURE 4. Schematic diagram of UFFC module.

2.2.3. Residual Path Containing Dynamic Convolutional Block Attention Module (CBAM)

In each SJP module, there is a residual path that includes the residual from the original input of this stage after Dynamic-CBAM processing. This path primarily uses fixed 3×3 kernels to maintain compatibility with mainstream ResNet architectures. The parameter setting is fixed at stride $= 2$, ensuring that spatial downsampling is synchronized with the main path. In this way, the residual path contributes a significant portion

of the gradient flow, reflecting the rationality of the module design. Additionally, through the Dynamic-CBAM mechanism, selective learning of certain features is emphasized.

2.3. Dynamic Weight Adaptive Feature Fusion

2.3.1. Dynamic Adaptation

In each adjustment path, the following methods are mainly used for network adaptation:

1. Space alignment: Keep feature map integrity through AdaptiveAvgPool to avoid interpolation distortion:

$$x_{adj} = \text{AdaptiveAvgPool2d}(x_{in}, S_{target}) \tag{13}$$

2. Channel expansion: 1×1 convolution is used to realize cross-stage feature dimensioning. The parameter amount is only 1/9 of the standard convolution, reducing the training occupancy of video memory, reducing the training time and improving the efficiency:

$$x_{adj} = W_{1\times1} \cdot x_{adj} \left(W_{1\times1} \in \mathbb{R}^{C_{out}\times C_{in}}\right) \tag{14}$$

where S_{target} denotes the input target size, and x_{adj} denotes the adaptive output result.

3. Batch normalization-rectified linear unit (BN-ReLU): Prevent feature degradation, improve fusion effect, and effectively balance stability and nonlinear expression ability.

2.3.2. Dynamic Weight Fusion

At each different fusion node, the input paths from different stages are fused with emphasis through dynamic weight allocation obtained by continuous training:

$$w = \text{softmax}\left([w_1, w_2, ..., w_N]\right)$$
$$(\text{Stage 2}: N = 2,\ \text{Stage 3}: N = 3,\ \text{Stage 4}: N = 4) \tag{15}$$

$$x_{fused} = \sum_{i=1}^{N} w_i \cdot x_{adj,i} \tag{16}$$

where w_N denotes the weight, $x_{adj,i}$ each adaptive residual path, and x_{fused} the output result after weighted summation.

Experimental results show that during training, the higher-level stage automatically learns to pay more attention to recent features (such as the feature weight of SJP3 in SJP4 reaching 0.43), while the system automatically increases the weight of early features (from 0.15 to 0.31). The fusion weights, as learnable parameters, participate in backpropagation, with the gradient contribution of the main path far exceeding that of the fusion path, indirectly demonstrating the rationality of the dynamic weight design.

2.4. Classifier Module

The global average pooling is used to replace the full connection layer, reducing the number of parameters by 95%. At the same time, Dropout is placed before the last Linear to prevent overfitting. In addition, Dropout can also improve the robustness and accuracy of the model and finally get accurate classification results:

$$\begin{aligned} h &= \text{ReLU}\left(W_{fc1} \cdot z + b_{fc1}\right) \\ y &= W_{fc2} \cdot \text{Dropout}(h) + b_{fc2} \end{aligned} \tag{17}$$

where W_{fc1} and W_{fc2} denote the full connection weight, and b_{fc1} and b_{fc2} denote the bias.

3. EXPERIMENT

3.1. Dataset

3.1.1. Introduction to the MSTAR Dataset

The MSTAR (Moving and Stationary Target Acquisition and Recognition) dataset was jointly developed by the U.S. DARPA (Defense Advanced Research Projects Agency) and AFRL (Air Force Research Laboratory). It is one of the most influential benchmark datasets in the field of SAR (Synthetic Aperture Radar) image target recognition. The dataset is collected using high-resolution panchromatic SAR sensors operating at the X-band (HH polarization), with a resolution of $0.3\,\text{m} \times 0.3\,\text{m}$ and target slice sizes of 128×128 pixels. The target categories and data are divided accordingly. The target types include 10 military targets, such as 2S1 self-propelled howitzer, BRDM-2 armored reconnaissance vehicle, BTR-60 armored personnel carrier, and T-72 tank. Additionally, it provides large SAR images containing natural scenes (such as forests and buildings), supporting target detection and recognition tasks. The data is divided into standard operating conditions (SOCs) and extended operating conditions (EOCs). Under SOC, the training and test sets have the same target models and configurations, differing only in elevation and azimuth angles; under EOC, the test set has significant differences in imaging angles, target configurations, or models, increasing the difficulty of recognition. Fig. 5 shows a comparison of SAR images and optical images for the 10 targets in the MSTAR dataset.

3.1.2. Introduction to the FUSAR-Ship1.0 Dataset

FUSAR-Ship1.0 is a SAR ship detection dataset constructed by Chinese research teams and released in 2019, aiming to advance the study of SAR image ship target detection algorithms. The data primarily comes from SAR images acquired by the domestic Gaofen-3 satellite and TerraSAR-X satellite, covering various imaging modes and polarization techniques. The images include high-resolution SAR images (resolution 1 m–3 m), with scenes spanning complex environments such as ports, open seas, and near-shore areas. Targets include over ten types of ships, including cargo ships, oil tankers, and fishing boats. Rectangular bounding boxes are used to label ship positions, and metadata such as target categories and sizes are provided. Some versions support rotated bounding box labeling (e.g., SSDD+), to reduce background interference and estimate ship orientation. A comparison of the SAR images and optical images of targets in the FUSAR-Ship1.0 dataset is shown in Fig. 6.

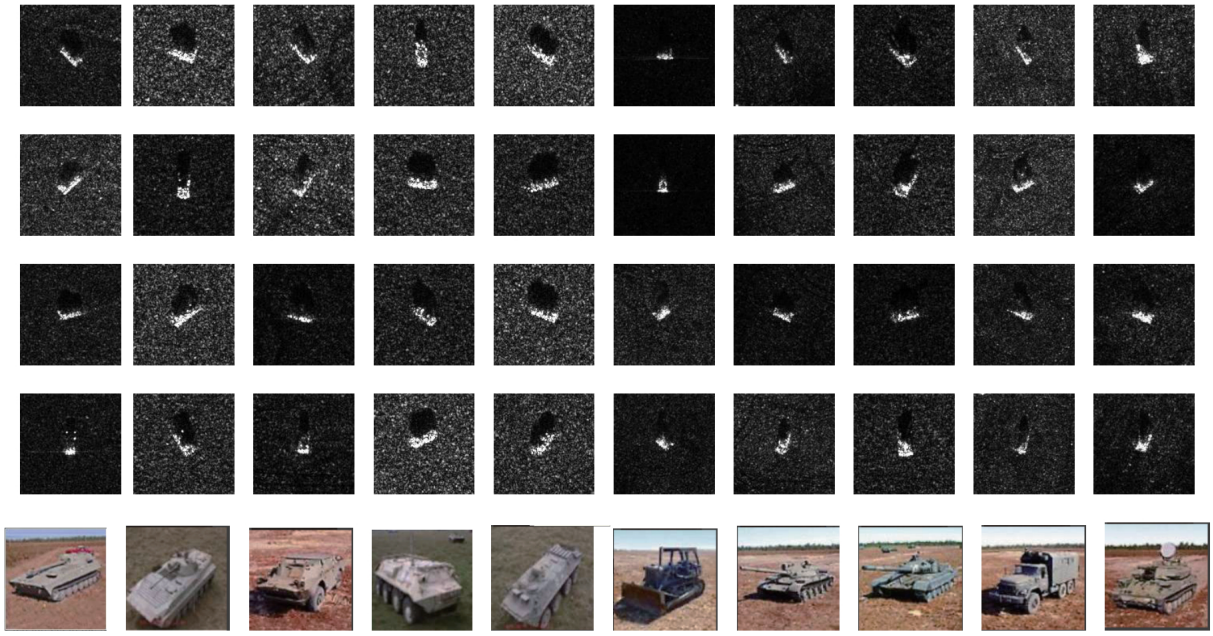

FIGURE 5. Comparison of SAR and optical images of targets in the MSTAR dataset.

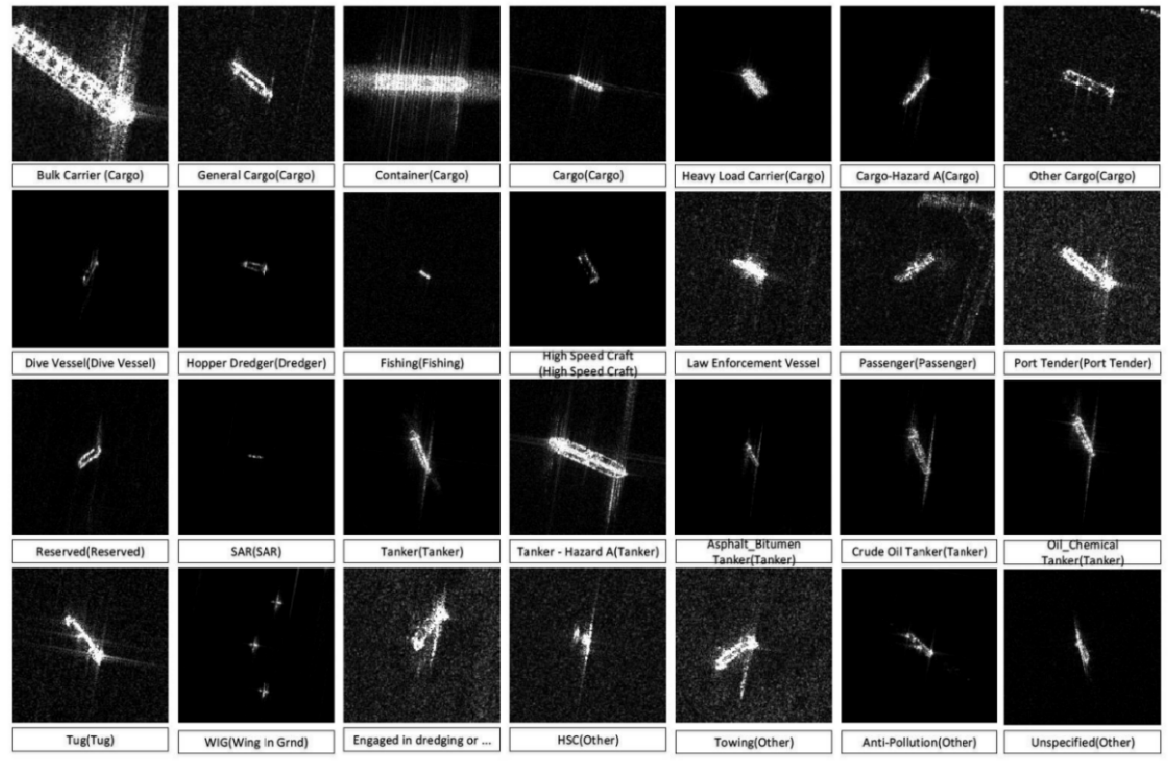

FIGURE 6. Comparison of SAR and optical images of targets in the FUSAR-Ship1.0 dataset.

3.2. Experimental Setup

PyTorch open source library was used to implement the proposed solution in Python 3.10.16. Training was supported by an NVIDIA GeForce RTX 4060 Laptop GPU and AMD Ryzen9 7940HX with Radeon Graphics (2.40 GHz) with CUDA toolkit 11.8. In the experiments of this paper, data preprocessing was performed on both the MSTAR dataset and FUSAR-Ship1.0 dataset. The image size was adjusted to 128×128 pixels to ensure that all input images have a consistent size, meeting the model's requirements for input dimensions. Training was conducted for 100 epochs, with each channel standardized to a mean of 0.5 and a standard deviation of 0.5.

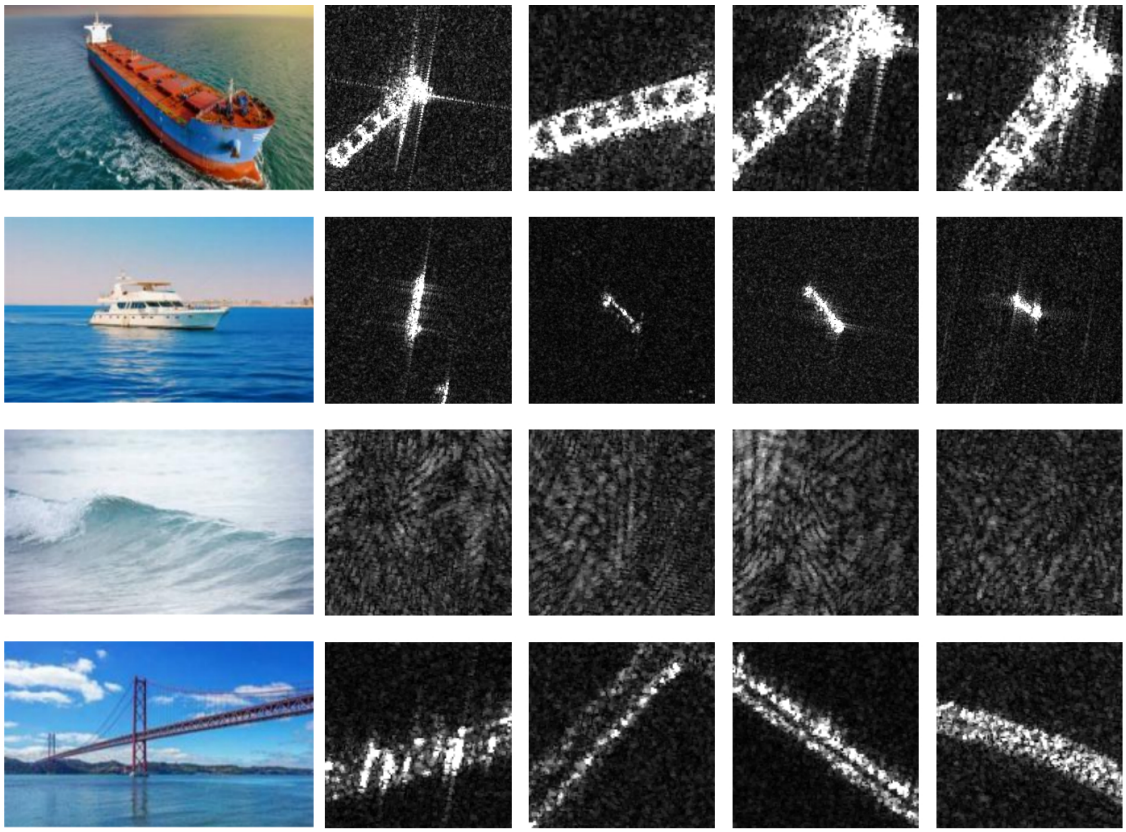

FIGURE 7. Examples of instances in each category of the sub-dataset.

TABLE 2. Detailed distribution of MSTAR data sets at different pitch angles.

Target type	2S1	BMP2	BRDM_2	BTR_60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
Training(17°)	220	220	220	220	220	220	220	220	220	220
Testing(15°)	40	40	40	40	40	40	40	40	40	40

3.3. Detection Results

3.3.1. Results on the MSTAR Dataset

In the experiment, we used SAR images with an elevation angle of 17° as the training set and 15° as the test set, as shown in Table 2. In the comparative tests, the SDF-Net recognition model proposed in this paper was compared with three traditional CNN networks and four advanced methods. As shown in Table 4, the overall recognition rates of traditional networks such as VGG19 [30], ResNet18 [30], and A-ConvNet [31] reached 89.6%, 90.5%, and 92.19%, respectively. With other advanced algorithms, for example, LM-BN-CNN [32] reached 96.44%; CA-MCNN [31] reached 97.81%; proposed CNN-LSTM and proposed CNN [33] reached 98.35% and 98.52%, respectively; and CCAE [34] reached 98.59%.

The results of the recognition rates for ten specific classification targets show that the SDF-Net model proposed in this paper achieved excellent data with a recognition rate of 100% for seven categories: BMP2, BRDM_2, BTR_60, D7, T62, T72, and ZSU_23_4, and good data with a recognition rate of 97.5% and a confusion rate of only 2.5% for the other three categories, which is better than the confusion rate of other methods.

3.3.2. Results on the FUSAR-Ship1.0 Dataset

In this study, we selected a subset of the original dataset and made appropriate modifications, which includes four common categories: Bridge, Bulk Carrier, Ship, and Wave. To effectively evaluate deep learning algorithms, we divided the data into two parts: training set and test set. The number of images for each category is shown in Table 3. Examples of each category are illustrated in Fig. 7. In the comparative experiment, we used the obtained dataset to train different CNN mod-

TABLE 3. Number of images in each category of the sub-dataset.

Category	Training	Test
Bridge	132	61
Bulk Carrier	130	30
Ship	378	66
Wave	312	54

TABLE 4. Recognition performance of each algorithm in MSTAR dataset.

Algorithms	2S1	BMP2	BRDM_2	BTR_60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4	Average
VGG19 [30]	98.9%	64.1%	100.0%	63.1%	81.1%	95.3%	85.7%	92.3%	98.9%	99.6%	89.6%
ResNet18 [30]	98.5%	70.3%	100.0%	75.9%	70.9%	98.2%	98.5%	73.5%	99.6%	99.3%	90.5%
A-ConvNet [31]	91.97%	89.61%	98.18%	96.41%	97.45%	95.26%	95.97%	78.87%	98.91%	100.00%	92.19%
LM-BN-CNN [32]	93.43%	98.30%	94.89%	95.90%	98.47%	99.27%	88.64%	97.77%	97.45%	97.08%	96.44%
CA-MCNN [31]	99.64%	97.27%	99.27%	99.64%	98.98%	99.63%	99.64%	93.85%	100.00%	94.26%	97.81%
Proposed CNN-LSTM [33]	96.00%	92.00%	100.00%	99.00%	96.00%	99.00%	99.00%	100.00%	100.00%	100.00%	98.35%
Proposed CNN [33]	98.00%	92.00%	100.00%	99.00%	97.00%	99.00%	99.00%	100.00%	100.00%	100.00%	98.52%
CCAE [34]	97.44%	97.43%	98.54%	96.92%	99.48%	99.27%	98.53%	100.00%	99.27%	98.90%	98.59%
SDF-Net	97.50%	100.00%	100.00%	100.00%	97.50%	100.00%	100.00%	100.00%	97.50%	100.00%	99.25%

TABLE 5. Recognition performance of each algorithm in FUSAR-Ship1.0 dataset.

Algorithms	VGG16	AlexNet	ResNet18	DenseNet121	Proposed CNN [33]	Proposed CNN-LSTM [33]	ACRM [35]	ACRM-Coreset [35]	CA [36]	SDF-Net
Accuracy	64.45%	66.11%	68.54%	69.44%	69.82%	70.41%	63.2%	71.3%	72.58%	72.99%

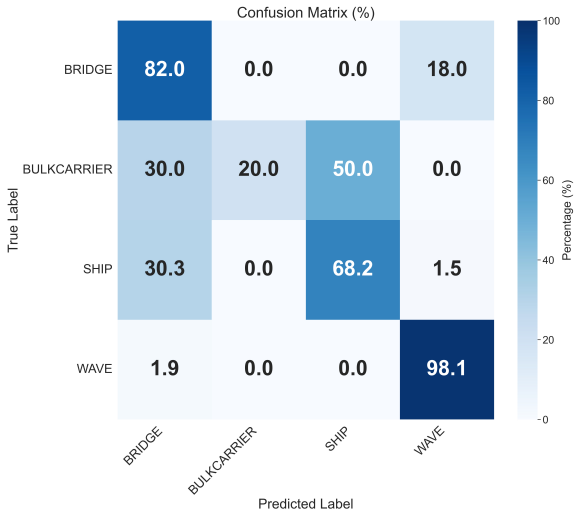

FIGURE 8. The confusion matrix of SDF-Net in sub-dataset test.

els with 100 epochs, and the recognition performance of each algorithm under the FUSAR-Ship1.0 dataset is shown in Table 5. The confusion matrix in Fig. 8 indicates that the proposed method has relatively good classification performance for Bridge and Wave classes, achieving recognition rates of 82% and 98.1%, respectively, with a zero chance of Bulk Carrier misclassifying wave. Objectively speaking, the features of these two classes are more distinct and easier to learn, while Bulk Carrier is prone to confusion and has a lower recognition rate. In terms of overall recognition rate, SDF-Net outperforms traditional CNN architectures, with a smaller training loss. The table provides the accuracy rates of VGG16, AlexNet, ResNet18, and DenseNet121 traditional CNN models, which are 64.45%, 66.11%, 68.54%, and 69.44%, respectively. Compared to SDF-Net's 72.99%, there is a certain gap in the results, which also intuitively reflects the excellent performance, high classification accuracy, and training efficiency of SDF-Net.

The proposed CNN and LSTM-based CNN methods achieved accuracy rates of 69.82% and 70.41% in [33], while the aspect continual recognition model (ACRM) and Coreset-based ACRM methods demonstrated 63.2% and 71.3% accuracy rates in [35], respectively. The cosine affinity (CA) method developed in [36] attained 72.58% accuracy. Compared with prior approaches, SDF-Net outperformed these methods with significantly higher recognition rates. These comparative results further validate the effectiveness of SDF-Net.

3.4. Ablation Study

This part conducts ablation experiments on the three key parts of the network model to verify the contribution of these modules to improve the performance of the model.

3.4.1. Results of the MSTAR Dataset

Necessity of Dynamic CBAM. This section primarily analyzes the impact of Dynamic CBAM on experimental results. According to the data in Table 6, combining the use of Dynamic CBAM mechanism can improve performance by 1.12%. Data

TABLE 6. Results of ablation experiments on MSTAR dataset.

Number	Dynamic CBAM	UFFC	Residual Path	Accuracy
1	✓	×	×	96.93%
2	×	✓	×	97.87%
3	×	×	✓	97.47%
4	×	✓	✓	98.13%
5	✓	×	✓	97.87%
6	✓	✓	×	98.40%
7	✓	✓	✓	99.25%

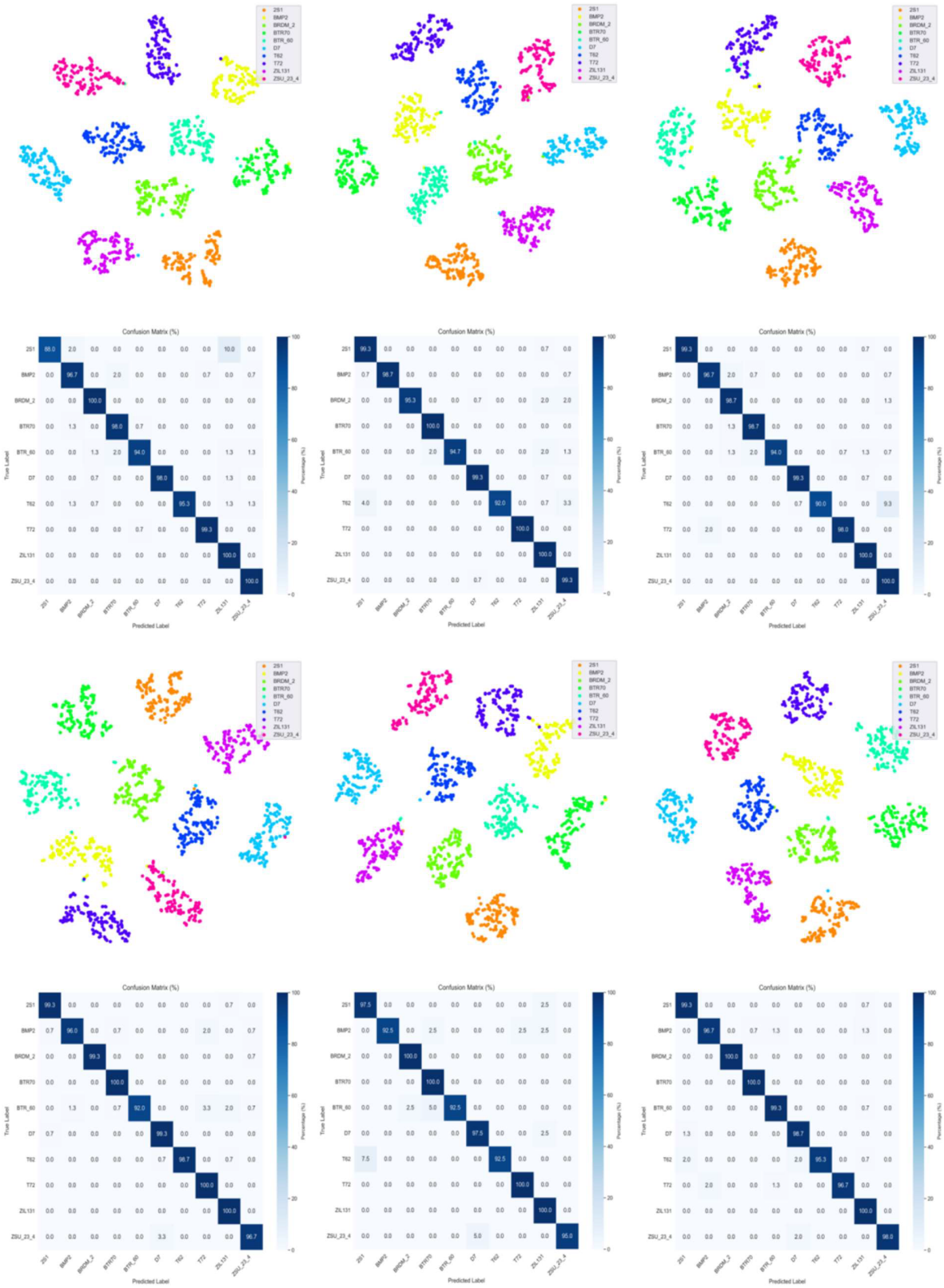

FIGURE 9. Confusion matrix of the ablation experiment on the MSTAR dataset.

in the table, rows 3 and 5, show that performance can be improved by 0.40%. These performance comparisons indicate that using Dynamic CBAM can more effectively adjust feature information into the network model, and dynamic dual attention mechanisms facilitate channel-space collaborative optimization.

Necessity of UFFC. This section mainly analyzes and compares the impact of UFFC on experimental results. Accord-

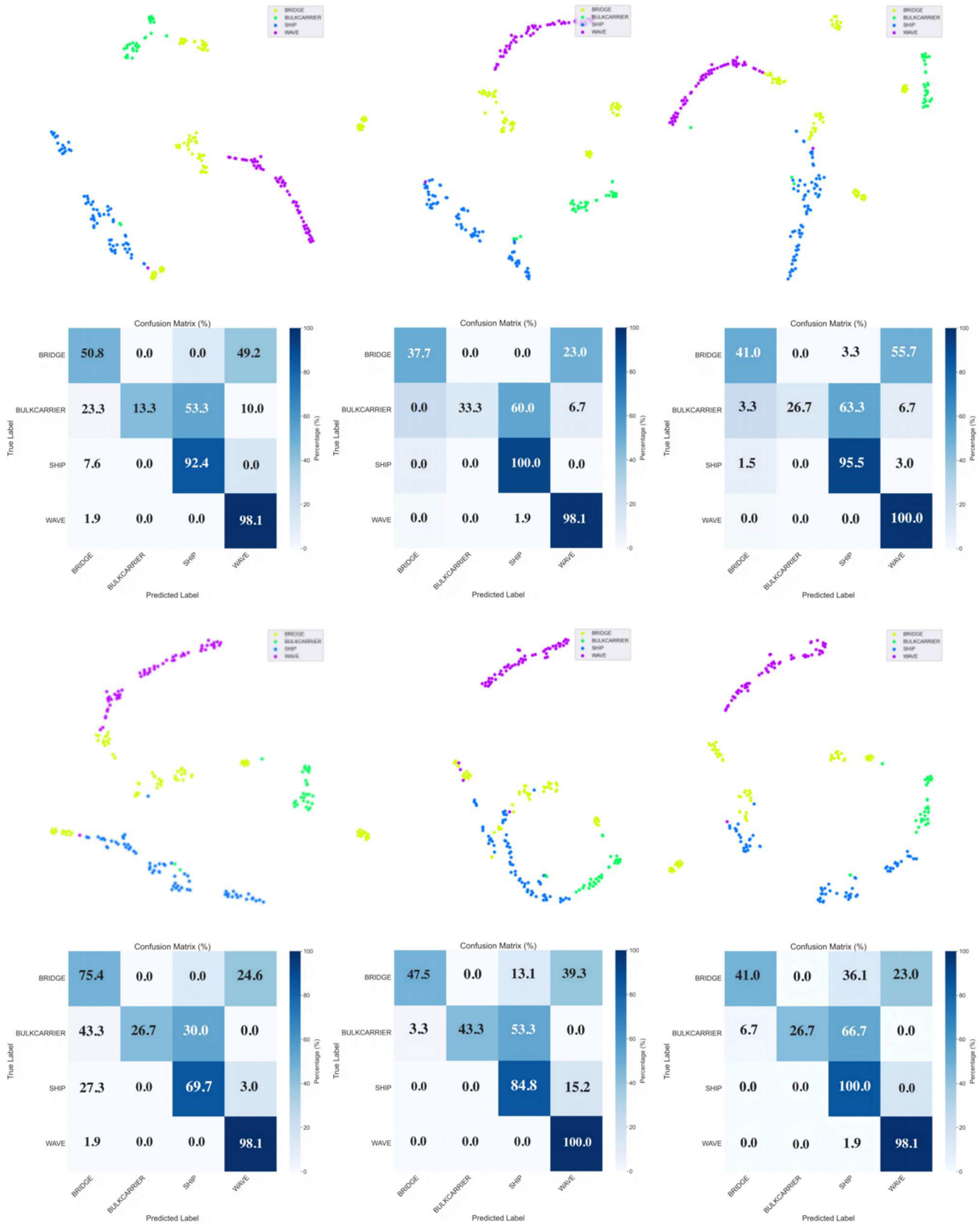

FIGURE 10. Confusion matrix of the ablation experiment on the FUSAR-Ship1.0 dataset.

ing to data in Table 6, the performance of UFFC can be improved by 1.38%. Data in the table, rows 3 and 4, show that the performance can be enhanced by 0.66%. The above performance comparisons indicate that using UFFC for frequency domain information processing and constructing a joint space of spatial-frequency features can more effectively integrate feature information, thereby improving recognition accuracy and efficiency.

Necessity of Residual Path. This section mainly analyzes and compares the impact of Residual Path on experimental results. According to the data in Table 6, combining the use of Residual Path mechanism can improve recognition performance by 0.85%. Data in the table, rows 1 and 5, show that recognition performance can be improved by 0.94%. These performance comparisons indicate that using multiple paths of Residual Path to integrate feature information helps en-

hance recognition performance. Fig. 9 presents the distributed stochastic neighbor embedding (TSNE) visualization and confusion matrix under six conditions. The above experiments demonstrate that each module can improve the accuracy of SAR target detection.

3.4.2. Results of the FUSAR-Ship1.0 Dataset

Necessity of Dynamic CBAM. This section primarily analyzes the impact of Dynamic CBAM on experimental results. According to the data in Table 7, combining the use of Dynamic CBAM mechanism can improve performance by 0.48%. Data in the table, rows 3 and 5, show that performance can be improved by 0.95%. These performance comparisons indicate that using Dynamic CBAM can more effectively adjust feature information into the network model, and dynamic dual attention mechanisms facilitate channel-space collaborative optimization.

TABLE 7. Results of ablation experiments on FUSAR-Ship1.0 dataset.

Number	Dynamic CBAM	UFFC	Residual Path	Accuracy
1	✓	×	×	70.62%
2	×	✓	×	71.76%
3	×	×	✓	71.09%
4	×	✓	✓	72.51%
5	✓	×	✓	72.04%
6	✓	✓	×	72.04%
7	✓	✓	✓	72.99%

Necessity of UFFC. This section mainly analyzes and compares the impact of UFFC on experimental results. According to the data in Table 7, the performance of UFFC can be improved by 0.95%. Data in the table, rows 3 and 4, show that the performance can be improved by 1.42%. These performance comparisons indicate that using UFFC for frequency domain information processing and constructing a joint space of spatial-frequency features can more effectively integrate feature information, thereby enhancing recognition accuracy and efficiency.

Necessity of Residual Path. This section mainly analyzes and compares the impact of Residual Path on experimental results. According to the data in Table 7, combining the use of Residual Path mechanism can improve recognition performance by 0.95%. Data in the table, rows 1 and 5, show that recognition performance can be improved by 1.42%. These performance comparisons indicate that using multiple paths of Residual Path to integrate feature information helps enhance recognition performance. Fig. 10 presents TSNE visualization and confusion matrices under six conditions. The above experiments demonstrate that each module can improve the accuracy of SAR target detection.

3.4.3. Comparative Analysis of Data Sets FUSAR-Ship1.0 and MSTAR

Experimental results demonstrate that the FUSAR-Ship1.0 dataset exhibits lower recognition accuracy than the MSTAR dataset. To address this disparity, we analyze three key factors:

(1) Target Feature Variability: While MSTAR primarily focuses on ground vehicles (e.g., tanks and armored vehicles) with distinct structural patterns, the FUSAR-Ship1.0 dataset contains ship targets sharing high structural similarity. This similarity reduces feature differentiation between categories, complicates feature extraction and classification processes, and ultimately limits performance improvement.

(2) Background Complexity: Unlike MSTAR datasets uniformly textured backgrounds with clear object boundaries, FUSAR-Ship1.0 datasets feature dynamic sea clutter backgrounds with low radiometric contrast. The complex and significant background interference further hinders effective target identification.

(3) Model Attention Mechanism: Conventional spatial attention mechanisms tend to focus on compact central features in MSTAR datasets. However, for FUSAR-Ship1.0's scattered ship components (e.g., superstructures and deck equipment), these mechanisms struggle to effectively target dispersed critical areas. This limitation may explain the suboptimal performance of the model in ship recognition tasks.

4. CONCLUSION

This paper primarily addresses the issues of low recognition rates and poor stability in traditional convolutional neural networks when dealing with synthetic aperture radar images. We propose the SDF-Net architecture, which constructs a spatiotemporal feature space through frequency domain convolution units (UFFC). By leveraging dynamic adaptive feature fusion mechanisms to achieve cross-stage information interaction, adding dynamic dual attention mechanisms to realize channel-space collaborative optimization, and adopting multi-scale mixed convolution kernel design to balance receptive fields and computational efficiency, we enhance classification accuracy and training efficiency. In experiments, we analyze and compare the performance of VGG19, ResNet18, A-ConvNet, and other CNN models on the MSTAR dataset, as well as primarily compare them with VGG16, AlexNet, ResNet18, and DenseNet121 network models on the FUSAR-Ship1.0 dataset. From the experimental results, the SDF-Net architecture model proposed in this paper shows a significant advantage over traditional CNN models and other similar current methods in terms of recognition performance. This paper plays a positive role in generalizing CNN models for SAR target recognition [37].

In addition, it is not easy to obtain large-scale SAR image data sets. How to improve the recognition performance under small sample conditions and the exploration of semi-supervised learning methods will be the future work.

ACKNOWLEDGEMENT

This work is supported in part by the National Natural Science Foundation of China under Grant 62131020 and 62371468.

REFERENCES

[1] Wang, J., Z. Cui, T. Jiang, C. Cao, and Z. Cao, "Lightweight deep neural networks for ship target detection in SAR imagery," *IEEE Transactions on Image Processing*, Vol. 32, 565–579, 2023.

[2] Dong, Y., F. Li, W. Hong, X. Zhou, and H. Ren, "Land cover semantic segmentation of port area with high resolution SAR images based on SegNet," in *2021 SAR in Big Data Era (BIGSARDATA)*, 1–4, Nanjing, China, 2021.

[3] Frey, O., C. L. Werner, and R. Coscione, "Car-borne and UAV-borne mobile mapping of surface displacements with a compact repeat-pass interferometric SAR system at L-band," in *IGARSS 2019 — 2019 IEEE International Geoscience and Remote Sensing Symposium*, 274–277, Yokohama, Japan, Aug. 2019.

[4] Lang, H., G. Yang, C. Li, and J. Xu, "Multisource heterogeneous transfer learning via feature augmentation for ship classification in SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, 1–14, 2022.

[5] Zhang, T., S. Tian, S. E. Abhadiomhen, Z. Xu, X.-j. Shen, J. Wang, and C. Wang, "Low-rank representation based model for SAR image denoising and edge-preserving," in *2023 SAR in Big Data Era (BIGSARDATA)*, 1–4, Beijing, China, 2023.

[6] Li, C., J. Ni, Y. Luo, D. Wang, and Q. Zhang, "A dual-branch spatial-frequency domain fusion method with cross attention for SAR image target recognition," *Remote Sensing*, Vol. 17, No. 14, 2378, 2025.

[7] Zhang, Z., D. Wu, D. Zhu, and Y. Zhang, "A multichannel SAR ground moving target detection algorithm based on subdomain adaptive residual network," *IEEE Geoscience and Remote Sensing Letters*, Vol. 20, 1–5, 2023.

[8] Geng, Z., Y. Xu, B.-N. Wang, X. Yu, D.-Y. Zhu, and G. Zhang, "Target recognition in SAR images by deep learning with training data augmentation," *Sensors*, Vol. 23, No. 2, 941, 2023.

[9] Chen, J., N. Amjad, and W. Yang, "SAR and multispectral image fusion using multibranch CNN and cross domain learning for local climate zone classification," in *2024 21st International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 484–489, Murree, Pakistan, 2024.

[10] Pan, X., W. Wang, L. Wu, and N. Li, "Improved moving target imaging method for a multichannel HRWS SAR system," *IEEE Geoscience and Remote Sensing Letters*, Vol. 20, 1–5, 2023.

[11] Yang, Y., Y. Wan, and Q. Wan, "A 2D-PRM-based atmospheric phase correction method in GB-SAR interferometry application," *IEEE Sensors Letters*, Vol. 7, No. 6, 1–4, 2023.

[12] Liu, Y., M. Lin, Y. Mo, and Q. Wang, "SAR — Optical image matching using self-supervised detection and a transformer-CNN-based network," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 1–5, 2024.

[13] Yuan, S., Z. Yu, C. Li, and S. Wang, "A novel SAR sidelobe suppression method based on CNN," *IEEE Geoscience and Remote Sensing Letters*, Vol. 18, No. 1, 132–136, 2021.

[14] Zhao, C., X. Fu, and J. Dong, "CGA-Det: A CNN-GNN-based oriented SAR ship detector for complex scenes," *IEEE Geoscience and Remote Sensing Letters*, Vol. 22, 1–5, 2025.

[15] Hao, Y., J. Wu, Y. Yao, and Y. Guo, "A robust anchor-free detection method for SAR ship targets with lightweight CNN," *IEEE Transactions on Instrumentation and Measurement*, Vol. 74, 1–19, 2025.

[16] Zhang, S., Q. Cheng, D. Chen, and H. Zhang, "Image target recognition model of multi-channel structure convolutional neural network training automatic encoder," *IEEE Access*, Vol. 8, 113 090–113 103, 2020.

[17] Gao, F., L. Kong, R. Lang, J. Sun, J. Wang, A. Hussain, and H. Zhou, "SAR target incremental recognition based on features with strong separability," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, 1–13, 2024.

[18] Xu, T., P. Xiao, and H. Wang, "MobileShuffle: An efficient CNN architecture for spaceborne SAR scene classification," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 1–5, 2024.

[19] Fukuzaki, S. and M. Ikehara, "Faster training of large kernel convolutions on smaller spatial scales," *IEEE Access*, Vol. 12, 161 312–161 328, 2024.

[20] Wang, J., T. Zheng, P. Lei, and X. Bai, "Ground target classification in noisy SAR images using convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 11, No. 11, 4180–4192, 2018.

[21] Zang, B., L. Ding, Z. Feng, M. Zhu, T. Lei, M. Xing, and X. Zhou, "CNN-LRP: Understanding convolutional neural networks performance for target recognition in SAR images," *Sensors*, Vol. 21, No. 13, 4536, 2021.

[22] Marzi, D., J. I. S. Jara, and P. Gamba, "A 3-D fully convolutional network approach for land cover mapping using multitemporal sentinel-1 SAR data," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 1–5, 2024.

[23] Guo, Y., Z. Zeng, M. Jin, J. Sun, Z. Meng, and W. Hong, "Multilevel attention networks for synthetic aperture radar automatic target recognition," *IEEE Geoscience and Remote Sensing Letters*, Vol. 21, 1–5, 2024.

[24] Huang, Z., C. Wu, X. Yao, Z. Zhao, X. Huang, and J. Han, "Physics inspired hybrid attention for SAR target recognition," *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 207, 164–174, 2024.

[25] Li, X., Q. Xu, X. Chen, and C. Li, "Additive attention for CNN-based classification," in *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*, 55–59, Takamatsu, Japan, 2021.

[26] Lingyun, G., E. Popov, and D. Ge, "Spectral network combining fourier transformation and deep learning for remote sensing object detection," in *2022 International Conference on Electrical Engineering and Photonics (EExPolytech)*, 99–102, St. Petersburg, Russian Federation, 2022.

[27] Khatavkar, S. A. and N. B. Sambre, "DeepPatchNet: A FFT-CNN-based fusion approach for resilient phase correction in underwater image reconstruction," in *2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)*, 1–7, Bengaluru, India, 2024.

[28] Meng, Y., J. Wu, S. Xiang, J. Wang, J. Hou, Z. Lin, and C. Yang, "A high-throughput and flexible CNN accelerator based on mixed-radix FFT method," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 72, No. 2, 816–829, 2025.

[29] Shi, H., G. Cao, Y. Zhang, Z. Ge, Y. Liu, and D. Yang, "F3Net: Fast Fourier filter network for hyperspectral image classification," *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, 1–18, 2023.

[30] Ruan, X., L. Wang, J. Guo, D. Zhu, and C. Hu, "CNN-based SAR automatic target recognition using SAR raw data," in *2021 CIE International Conference on Radar (Radar)*, 1405–1408, Haikou, Hainan, China, 2021.

[31] Li, Y., L. Du, and D. Wei, "Multiscale CNN based on component analysis for SAR ATR," *IEEE Transactions on Geoscience and*

Remote Sensing, Vol. 60, 1–12, 2021.

[32] Zhou, F., L. Wang, X. Bai, and Y. Hui, "SAR ATR of ground vehicles based on LM-BN-CNN," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 12, 7282–7293, 2018.

[33] Toumi, A., J.-C. Cexus, A. Khenchaf, and M. Abid, "A combined CNN-LSTM network for ship classification on SAR images," *Sensors*, Vol. 24, No. 24, 7954, 2024.

[34] Wang, K., Q. Qiao, G. Zhang, and Y. Xu, "Few-shot SAR target recognition based on deep kernel learning," *IEEE Access*, Vol. 10, 89 534–89 544, 2022.

[35] Chen, H., C. Du, J. Zhu, and D. Guo, "Target-aspect domain continual learning for SAR target recognition," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 63, 1–14, 2025.

[36] Xie, N., M. Xiong, F. Wei, T. Zhang, Z. Yang, and W. Yu, "CA-LOSS: A cosine affinity loss for imbalanced SAR ship classification," in *IGARSS 2024 — 2024 IEEE International Geoscience and Remote Sensing Symposium*, 9070–9074, Athens, Greece, 2024.

[37] Lu, L., G. Zhang, Y. Nie, J. Liu, Y. fang, G. Zhang, and Y. Wu, "Application of improved CNN in SAR image noise reduction," *Journal of Physics: Conference Series*, Vol. 1792, No. 1, 012053, 2021.