LW-CMDANet

Lang, Ping; Fu, Xiongjun; Feng, Cheng; Dong, Jian; Qin, Rui; Martorella, Marco

DOI:
10.1109/JSTARS.2022.3195074

LW-CMDANet: A Novel Attention Network for SAR Automatic Target Recognition

Ping Lang, *Graduate Student Member, IEEE*, Xiongjun Fu, *Senior Member, IEEE*, Cheng Feng, Jian Dong, Rui Qin, and Marco Martorella, *Fellow, IEEE*

***Abstract*—Deep-learning-based synthetic aperture radar automatic target recognition (SAR-ATR) plays a significant role in the military and civilian fields. However, data limitation and large computational cost are still severe challenges in the actual application of SAR-ATR. To improve the performance of the convolutional neural network (CNN) model with limited data samples in SAR-ATR, this article proposes a novel multidomain feature subspace fusion representation learning method, i.e., a lightweight cascaded multidomain attention network, namely, LW-CMDANet. First, we design a four-layer CNN model to perform hierarchical feature representation learning via the hinge loss function, which can efficiently alleviate the overfitting problem of the CNN model by a nongreedy training style with a small dataset. Then, a cascaded multidomain attention module, based on discrete cosine transform and discrete wavelet transform, is embedded into the previous CNN to further complete the class-specific feature extraction from both the frequency and wavelet transform domains of the input feature maps. Thus, the multidomain attention can enhance the feature extraction ability of previous nongreedy learning manner, to effectively improve the recognition accuracy of the CNN model. Experimental results on small SAR datasets show that our proposed method can achieve better or competitive performance than that of many current existing state-of-the-art methods in terms of recognition accuracy and computational cost.**

***Index Terms*—Discrete cosine transform (DCT), multidomain attention, synthetic aperture radar automatic target recognition (SAR-ATR), wavelet transform.**

I. Introduction

OPERATING in all weather, day-and-night, and high-resolution imaging, a synthetic aperture radar (SAR) plays an increasingly important role in the fields of military and civilian applications, such as surveillance and reconnaissance tasks. Synthetic aperture radar automatic target recognition (SAR-ATR) is one of the significant SAR imagery interpretation tasks [1], which is used to predict the specific category of detected targets (such as military vehicles [2], [3] and terrains [4]) through obtained SAR imagery data with computer processing technology. In recent years, with the rapid development of the deep learning (DL) technique, SAR-ATR has achieved a great success. However, the heavy dependence on the large-scale dataset and the large computational cost of the DL model are still main challenges when DL-based SAR-ATR methods are applied in practical scenarios. The main reasons are as follows: 1) the scatter characteristics of the SAR target are highly sensitive to imaging conditions, such as different azimuth and pose angles of the target; 2) it is obviously expensive and time consuming to acquire and annotate a large number of SAR target images; and 3) the good DL model usually has a large number of parameters and needs large computational cost to be trained before the model reaches convergence.

To address the above issues of SAR-ATR, many researchers have proposed some state-of-the-art (SOTA) methods in recent years, which mainly contain three categories: data augmentation based (e.g., simulated SAR images [5], [6], data augmentation techniques [7], and extra SAR data sample generalization [8], [9]), transfer learning or prior knowledge based (e.g., pretrained techniques [10], [11] and related SAR imaging prior information [5], [6], [17]–[20]), and the fine-grained model structure design or novel learning techniques based (e.g., M-Net [22], multifeature fusion learning [29], [30], [32], and metalearning [33]–[35]).

As for the data augmentation technique, the authors augmented the SAR dataset by simulated SAR images in [5], [6], and [33]. In [7], three image processing techniques (i.e., translation, speckle noising, and pose synthesis) were proposed to augment the SAR dataset. Huang et al. [8] and Song et al. [9] proposed the deep Q-learning and the adversarial autoencoder model to generate extra SAR data samples to enhance the generalization of the DL model, respectively.

As for the transfer learning, the authors in [5], [6], and [33] proposed that the DL model could first learn the physical-related features from the simulated SAR images and then transferred learned prior knowledge into the real SAR images recognition task to improve the generalization ability of the DL model. Huang et al. [10] and Ying et al. [11] first performed the pretraining technique on unlabeled SAR images or optical images to acquire prior knowledge and then transferred learned knowledge

Manuscript received 1 May 2022; revised 22 June 2022 and 20 July 2022; accepted 24 July 2022. Date of publication 29 July 2022; date of current version 22 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61571043 and in part by the 111 Project of China under Grant B14010. *(Corresponding author: Xiongjun Fu.)*

Ping Lang, Xiongjun Fu, and Cheng Feng are with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: langping911220@bit.edu.cn; fuxiongjun@bit.edu.cn; dafeng7@126.com).

Jian Dong is with the School of Electronics and Optics, Army Engineering University, Shijiazhuang 050003, China (e-mail: radarvincent@sina.com).

Rui Qin is with the School of Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: qinrui@cqupt.edu.cn).

Marco Martorella is with the Department of Information Engineering, University of Pisa, 56126 Pisa, Italy (e-mail: m.martorella@iet.unipi.it).

Digital Object Identifier 10.1109/JSTARS.2022.3195074

to the real SAR-ATR task. Moreover, the domain knowledge of SAR imaging (such as range/azimuth angle information [12], [16], attributed scattering center features [17]–[21], and multiscale rotation invariant Haar-like features of SAR images [30]) and the extracted features by the DL model are fused to effectively and efficiently alleviate the overfitting at the limited data scenarios.

As for the fine-grained model structure design or the novel learning technique, many novel models are proposed to improve the features extraction ability of the DL models, such as memory network [22], model compression technique [23], A-ConvNet [24], hybrid inference network [25], novel loss functions [26]–[28], multichannel parallel topology [29], and multiscale prototypical network [31]. The pose angle marginalization learning and the target aspect angle sharing learning between source and target domains were also proposed to improve the recognition performance of the DL model in [13]–[15]. In addition, multiview feature fusion learning [32], metalearning methods [33]–[35], contrastive learning [36], [37], and semisupervised or self-supervised learning [15], [38] were also studied to address the few-sample problem in SAR-ATR tasks.

Although these methods have made a great progress in aforementioned SAR-ATR, they may still exist some challenges in actual SAR-ATR scenarios. The data-augmentation-based methods directly increase the number of training data samples, which can obviously improve the generalization of the DL model. The training process, however, needs more computational cost to train the model before reaching the convergence. The transfer-learning- or prior-knowledge-based methods can provide the interpretability of the DL process to some extent, since the prior knowledge of SAR imaging has been embedded into the learning process. However, the requirement of a large pretrained SAR dataset for the transfer learning is a severe challenge. In addition, it is difficult to extract the complete prior knowledge of the SAR images due to the complexity of the SAR imaging, especially at limited data scenarios. As for the fine-grained model structure design or novel learning-technique-based methods, this type of methods can directly improve the recognition performance of the DL model. However, these recognition methods are complex, which is difficult to design an effective recognition model in a short time. For example, the model compression technique usually needs many experiments to determine the optimal scale of model pruning. Metalearning methods need to construct a large-scale metadataset to effectively train the model.

Inspired by the attention mechanism (AM) [44], [45] and the wavelet transform [51], we propose a novel end-to-end multidomain feature subspace fusion representation learning network to directly improve the recognition accuracy and reduce computational cost at limited data scenarios in this article. More specifically, an end-to-end lightweight network architecture based on a cascade multidomain attention (i.e., LW-CMDANet) is proposed, which is a four-layer lightweight nongreedy HL-convolutional neural network (CNN) model with the hinge loss function based on our previous work [39]. The HL-CNN model can perform hierarchical feature representation learning via the hinge loss function in a nongreedy manner. Thus, it can efficiently alleviate the overfitting problem by a nongreedy training style with the limited dataset. More importantly, a novel cascaded multidomain attention module, based on the discrete cosine transform (DCT) and the discrete wavelet transform (DWT), is proposed to be embedded into the HL-CNN architecture to further complete the class-specific feature extraction from both the frequency and wavelet transform domains of the input feature maps during the training process.

The multifrequency spectrum features and the multiresolution spectrum features of input feature maps obtained by the DCT and the DWT, respectively, can increase the number of feature subspaces of the input feature maps, which can provide the CNN model (i.e., HL-CNN) with higher probability to extract effective generalized (i.e., class-specific) features. The more the generalized features extracted, the better the generalization of the model. In this way, the model can learn more different-level features from the small dataset. In other words, this way can adaptively increase the number of feature subspaces of the feature maps by embedding a cascaded multidomain attention module during the training process, instead of directly augmenting the training data samples, such as [5], [6], and [33].

Moreover, multiresolution spectrum decomposition via the DWT can reduce the size of the feature maps by downsampling. In this way, multidomain feature subspaces (spatial features performed by the convolution operation, DCT, and DWT, frequency features performed by the DCT, and wavelet transform features performed by the DWT) can enrich the feature learning space of the HL-CNN to further improve the feature extraction capacity with small data samples. At the same time, the multidomain feature maps can effectively compensate for the feature extraction deficiency caused by the nongreedy learning of the HL-CNN. In addition, a depthwise separable convolution block [40] is adopted to replace the traditional convolution to reduce the computational burden. The overview of the LW-CMDANet is shown in Fig. 1.

The main contributions and novelties of this article can be summarized as follows.

1) We propose a novel multidomain feature subspace fusion representation learning architecture, which can adaptively fuse spatial features, frequency features, and wavelet transform features to improve the generalized feature extraction capacity of the HL-CNN to enhance the recognition accuracy with small samples in SAR-ATR scenarios.
2) A lightweight nongreedy HL-CNN is developed to improve the generalization performance of the deep CNN and reduce the computational cost.
3) A novel multidomain attention module based on the DCT and the DWT is proposed to perform the frequency transform and the waveform transform of input feature maps, which can increase extra two feature representation learning subspaces of input feature maps, i.e., frequency and wavelet transform spectrum subspaces, respectively. In this way, the model can adaptively improve multidomain feature subspace fusion representation learning in an end-to-end manner to enhance the SAR-ATR performance.

The rest of this article is organized as follows. Section II briefly introduces related works. The methodology of our proposed method is presented in Section III. Section IV describes the experiment and result analyses, as well as discussion. Finally, Section V concludes this article.

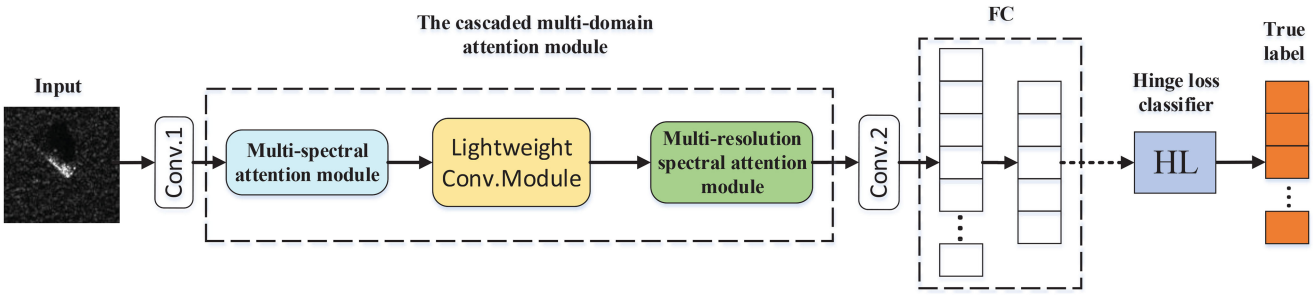


Fig. 1. Overview of the proposed LW-CMDANet architecture.

II. Related Works

A. AM in CNNs

Inspired by the human brain visual processing system [41], the AM can effectively improve the information processing efficiency. The AM can adaptively concentrate on important input information and neglect or less focus on other input information. In recent years, the AM has achieved a great success in DL-based computer vision (CV) [44] and natural language processing [42], mainly including spatial attention [43], channel attention [44], frequency channel attention [45], mixture attention of spatial and channel [46], nonlocal attention [47], class attention, and temporal attention [50].

Wang et al. [43] developed a spatial attention to enhance the significant spatial feature extraction of the DL model in the image classification tasks. Hu et al. [44] proposed a channel attention, i.e., squeeze-and-excitation network (SENet), to extract channelwise feature maps by squeeze [i.e., global average pooling (GAP)] and excitation (i.e., feature learning by multilayer perceptron) operations. The SENet can adaptively recalibrate the channelwise feature representation to weight the channel relationship, which can further bring a significant improvement in recognition performance with only slight additional computational cost. From a different perspective, based on the work of [44], Qin et al. [45] proposed a frequency channel attention block embedded in the CNN, i.e., FCANet, to efficiently extract the frequency-domain features of the channel feature map by the DCT. The experimental results show that the FCANet could improve by 1.8% in terms of the top-one accuracy on ImageNet, compared with the SENet.

Woo et al. [46] developed a mixture attention block, namely, convolutional block attention module (CBAM), which combined channel attention and spatial attention to comprehensively extract the effective input feature maps. In order to reduce the dependence on external information, Wang et al. [47] proposed a nonlocal AM in the CNN architecture to compute the response at a certain position as the weighted sum of all the location features. In recent two years, the self-attention-based transformer structure [42] has achieved a great success in the CV domain [48], [49]. In addition, Yuan et al. [50] proposed a class-specific attention module in the CNN architecture applied in image segmentation.

However, current existing attention modules aforementioned are usually effective in the large-scale dataset scenarios. Moreover, these attentions usually focus on a single feature subspace, e.g., the SENet focuses on the channel feature subspace and the FCANet performs on the frequency feature subspace. In the practical small dataset scenario, these attention-based methods maybe not work well (e.g., the limited feature learning capacity) due to the degradation of the available feature learning space during the training process, which is verified in Section IV (i.e., experiments and results).

B. Wavelet Transform in CNNs

The wavelet-transform-based multiresolution spectrum analysis of the image is good at extracting scale-invariant features [51], which is potential to embed the wavelet transform into the CNN model to effectively capture the spectral and spatial features simultaneously with an end-to-end architecture. Fujieda et al. [52] proposed a novel wavelet CNN model to efficiently perform the multiresolution spectrum and spatial feature extraction of the input image simultaneously during the training process. In order to perform a better tradeoff between the size of receptive field and the computational cost, Liu et al. [53] developed a multilevel wavelet CNN model to increase the receptive field of the convolutional filters and reduce the resolution of the feature maps. In addition, our previous works in [54] proposed a trainable wavelet soft threshold denoising module into the CNN to perform the noisy SAR image target recognition.

Inspired by the attention-based methods and the wavelet decomposition, we proposed a novel multidomain feature subspace fusion representation learning network, i.e., LW-CMDANet, to perform the multidomain feature subspace learning with the small dataset. More specifically, the multidomain feature subspaces of our proposed method contain the spatial feature subspace by the convolutional filters, the DCT, and the DWT, the multispectrum feature subspace by the DCT, and the wavelet decomposition subspace by the DWT. Therefore, the multidomain feature subspace fusion representation learning of the input SAR image can be achieved via an end-to-end CNN model during the training process. Our proposed method can address the problem of the current existing attention-based methods at the aforementioned limited data scenarios.

III. Methodology

In this section, we present the methodology of our proposed method (LW-CMDANet) in detail. First, the problem formulation and the methodology of SAR-ATR are presented. Then, the cascaded multidomain attention module is introduced, including

multispectrum attention and multiresolution spectrum attention. The depthwise separable convolution block and the functional model description of the LW-CMDANet are introduced. Finally, the proposed network model is presented.

A. Problem Formulation and Methodology

In this article, we consider that the SAR-ATR is performed at the limited data scenarios, which is the actual situation in military and civilian application domains. More importantly, it is often difficult to collect a large amount of data due to the military or commercial secrets and SAR imaging characteristics, such as the sensitivity of the observation angle and the complex scatter characteristics. Thus, the small feature space of limited data usually leads to overfitting of the DL model. In addition, it is difficult to train a DL model with a large number of parameters in a short time (i.e., large computational complexity). Aiming to address above problems, we propose a novel end-to-end DL-based SAR-ATR model (i.e., LW-CMDANet).

More specifically, in order to improve the generalization ability of the LW-CMDANet or alleviate the overfitting and reduce the computational cost, we mainly make contributions in the following three aspects (i.e., dataset preprocessing, network design, and model training style).

1) In order to maximally reduce the influence of the land clutter, we slice every sample of the MSTAR dataset into the size of 40×40 centered on the target;
2) We design a lightweight CNN model based on the depthwise separable convolution and the multidomain AM. The parameters of the model can be greatly reduced by the depthwise separable convolution operation. At the same time, the multidomain feature subspace (i.e., spatial, frequency, and wavelet transform domains) fusion representation learning of the input SAR images can be simultaneously performed to improve recognition accuracy during the training process.
3) We adopt the hinge loss function to perform nongreedy training to alleviate the overfitting of the DL model.

B. Cascaded Multidomain Attention

1) From DCT to Multispectrum Attention: Hu et al. [44] proposed a channel attention, i.e., SE module, which consists of *squeeze* and *excitation* operations. Suppose that $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$ is the input feature map of a convolutional layer, and H, W, and C are the height, the width, and the number of channels of the feature map, respectively. $\mathbf{X}$ first performs the *squeeze* operation (i.e., GAP) to generate a channelwise descriptor $\mathbf{z} \in \mathbf{R}^C$ by aggregating the feature maps across the spatial dimensions, i.e., $H \times W$, for each channel of $\mathbf{X}$. Then, the *excitation* operation is used to reduce a set of channelwise weights by a self-gating mechanism with a sigmoid activation function. Therefore, the SE AM is given by

$$att = \text{sigmoid}(f_c(\text{GAP}(X))) \tag{1}$$

where $att \in \mathbf{R}^C$ is the attention vector, i.e., weight vector, sigmoid is the sigmoid activation function, used to generate a scalar ranging from 0 to 1, $f_c(.)$ is a feature mapping operation, such as a fully connected (FC) layer, and GAP is the GAP operation. The weight vector is applied to the corresponding feature map of $\mathbf{X}$ with a channelwise multiplication operation, which yields the output of the SE module by

$$\mathbf{Y}_{:,:,:,i} = att_i \mathbf{X}_{:,:,:,i}, \text{ s.t. } i \in \{0, 1, \ldots, C-1\} \tag{2}$$

where $\mathbf{Y}$ is the output of the SE module, att_i is the ith element of the weight vector, and $\mathbf{X}_{:,:,:,i}$ is the ith channel feature of the input $\mathbf{X}$.

According to the detailed theoretical analysis of [45], the GAP operation of the SE module is a special case of the 2-D DCT, i.e., the low-frequency component of the 2-D DWT is proportional to GAP. The 2-D DCT of an image $\mathbf{X} \in \mathbf{R}^{H\times W}$ can be written as

$$f^{2d}_{h.w} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x^{2d}_{i,j} \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right) \cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \text{ s.t. } h \in \{0, 1, \ldots, H-1\}, w \in \{0, 1, \ldots, W-1\} \tag{3}$$

where $f^{2d} \in \mathbf{R}^{H\times W}$ is the 2-D DCT frequency spectrum of an image x and H and W are the height and the width of x, respectively. A 2-D DCT example of an SAR image is shown in Fig. 2.

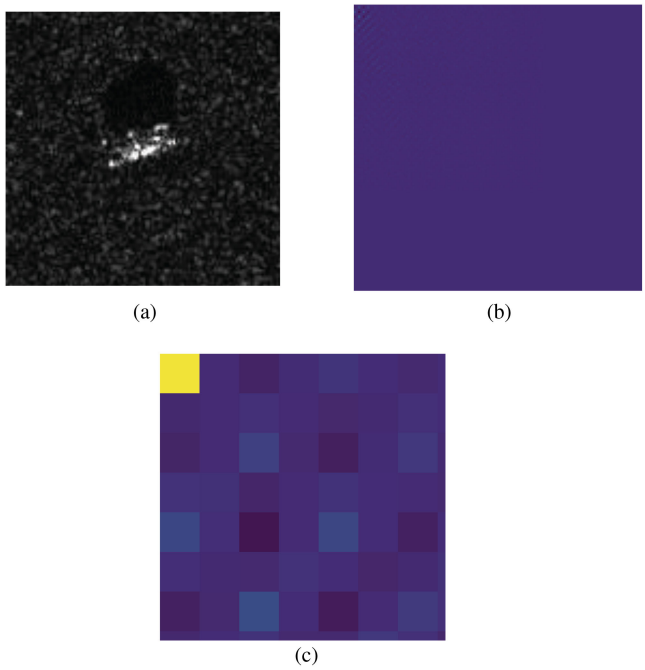


Fig. 2. 2-D DCT of the SAR image. (a) Original. (b) DCT (main components at the upper-left corner). (c) Local zoom of the upper-left corner of (b).

When $h = 0, w = 0$, i.e., $f_{0.0}^{2d}$ is the lowest frequency component, which can be written as

$$f_{0.0}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} \cos\left(\frac{\pi 0}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi 0}{W}\left(j+\frac{1}{2}\right)\right)$$
$$= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d}$$
$$= \text{GAP}(x^{2d})HW. \tag{4}$$

From (4), it can be seen that $f_{0.0}^{2d}$ is proportional to the GAP operation, i.e., the GAP is only a special case of the frequency components of the 2-D DCT. Therefore, it is prospective to incorporate other frequency components to extend the feature subspaces of the existing SE module.

According to (3), the inverse 2-D DCT can be written as

$$x_{i.j}^{2d} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f_{h,w}^{2d} \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)$$
$$\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right)$$
$$\text{s.t. } i \in \{0,1,\ldots,H-1\}, j \in \{0,1,\ldots,W-1\}. \tag{5}$$

For simplicity, we use A to represent the basis case of the inverse 2-D DCT in (5) by

$$A_{h.w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right). \tag{6}$$

Then, an image $\mathbf{X}$ can be written as

$$X = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(f_{0,0}^{2d}A_{0,0}^{i,j} + f_{0,1}^{2^{\sim}d}A_{0,1}^{i,j} + \cdots$$
$$+ f_{H-1,W-1}^{2d}A_{H-1,W-1}^{i,j}). \tag{7}$$

According to (4), (7) can also be written as

$$X = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(\text{GAP}(X)HW)A_{0,0}^{i,j} + f_{0,1}^{2d}A_{0,1}^{i,j} + \cdots$$
$$+ f_{H-1,W-1}^{2d}A_{H-1,W-1}^{i,j}). \tag{8}$$

According to (1) and (8), it is natural to see that the existing SE channel attention can be generalized to produce a novel multispectrum attention by incorporating the multiple frequency components of the 2D DCT.

More concretely, the feature map $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$ is first divided into n subparts along with the channel dimension. We denote these parts as $[X^0, X^1, \ldots, X^{n-1}]$, $X^i \in \mathbf{R}^{H\times W\times C'}$, and $C' = \frac{C}{n}$. For each part, a suitable corresponding 2-D DCT frequency component is assigned, which can be written as

$$f_{u,v}^{i} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{:,h,w}^{i}A_{h,w}^{u,v}$$
$$\text{s.t. } i \in \{0,1,\ldots,n-1\} \tag{9}$$

where $f_{u,v}^{i} \in \mathbf{R}^{C'}$ is the frequency component corresponding to X^i, whose index is $[u, v]$. The whole multispectrum attention

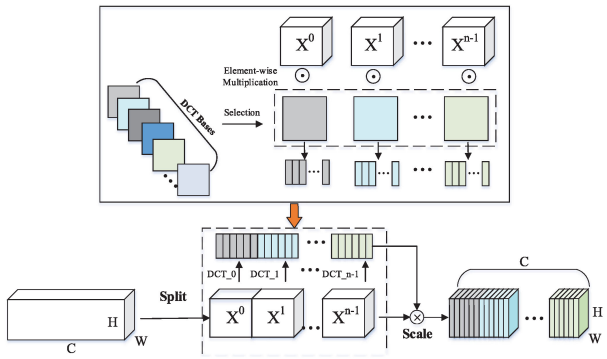


Fig. 3. Overview of the multispectrum attention module.

vector can be concatenated as

$$F = cat([f^0, f^1, \ldots, f^{n-1}]) \tag{10}$$

where $F \in \mathbf{R}^C$. Therefore, according to (1), the multispectrum attention structure can be written as

$$att_{\text{MulSpectral}} = \text{sigmoid}(fc(F)). \tag{11}$$

We can see from (10) and (11) that the multispectrum attention can incorporate extra frequency components into the feature map $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$, instead of containing only the lowest frequency component, such as the GAP operation. The overall illustration of the multispectrum attention is shown in Fig. 3.

2) From the DWT to Multiresolution Spectrum Attention: In order to fully exploit the multiscale decomposition features of the input feature maps, we proposed a novel multiresolution spectrum attention module to perform the multiresolution analysis based on the DWT [51].

An image can be decomposed into four subband images by the 2-D DWT with four convolutional filters, i.e., low-pass filter f_{LL} and high-pass filters f_{LH}, f_{HL}, and f_{HH}. We take Haar wavelet as an example, the four convolutional filters are defined as

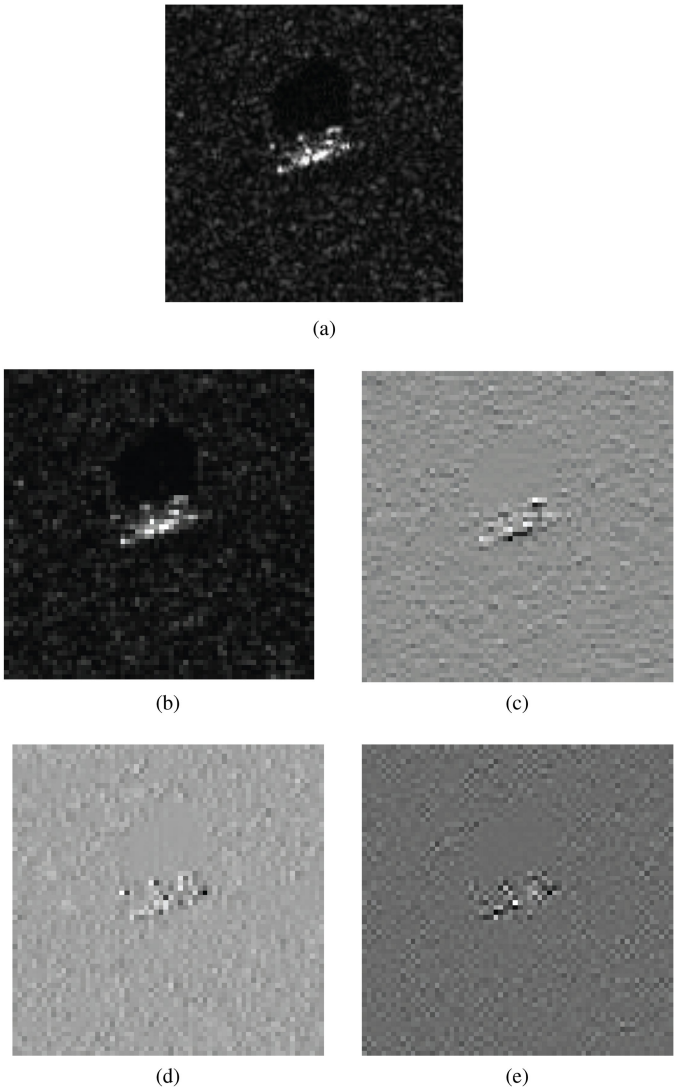
$$f_{LL} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, f_{LH} = \begin{bmatrix}-1 & -1\\ 1 & 1\end{bmatrix}$$
$$f_{HL} = \begin{bmatrix}-1 & 1\\ -1 & 1\end{bmatrix}, f_{HH} = \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}. \tag{12}$$

These filters are orthogonal to each other. Given an image x, the four subband components of x by the decomposition of the 2-D DWT are defined as

$$x_{LL} = (f_{LL} \otimes x)\downarrow_2,\ x_{LH} = (f_{LH} \otimes x)\downarrow_2$$
$$x_{HL} = (f_{HL} \otimes x)\downarrow_2,\ x_{HH} = (f_{HH} \otimes x)\downarrow_2 \tag{13}$$

where $\otimes$ represents the convolutional operation, $\downarrow_2$ represents a downsampling operation with a factor of 2. More concretely, the (i, j)th value of x_{LL}, x_{LH}, x_{HL}, and x_{HH} of the 2-D Haar DWT can be mathematically expressed as

$$x_{LL}(i,j) = x(2i-1, 2j-1) + x(2i-1, 2j)$$
$$+ x(2i, 2j-1) + x(2i, 2j)$$
$$x_{LH}(i,j) = -x(2i-1, 2j-1) - x(2i-1, 2j)$$
$$+ x(2i, 2j-1) + x(2i, 2j)$$



(a) (b) (c) (d) (e)

Fig. 4. DWT of the SAR image. (a) Original. (b) x_{LL}. (c) x_{LH}. (d) x_{HL}. (e) x_{HH}.

$$\begin{aligned} x_{HL}(i,j) = &-x(2i-1,2j-1)+x(2i-1,2j) \\ &-x(2i,2j-1)+x(2i,2j) \\ x_{HH}(i,j) = &\,x(2i-1,2j-1)-x(2i-1,2j) \\ &-x(2i,2j-1)+x(2i,2j). \end{aligned} \tag{14}$$

We take the 2-D DWT of an SAR image as an example; the four decomposition components are shown in Fig. 4.

Inspired by the channel attention [44], we extend the channel attention to the multiresolution spectrum attention in this article, which can efficiently extract features from the 2-D DWT domain. In general, the high-frequency component of an image, i.e., x_{HH}, is the noisy component. In order to alleviate the negative impact of the noisy component on SAR-ATR tasks, we quit the x_{HH} component of the feature maps and consider x_{LL}, x_{LH}, and x_{HL} during the training process. More concretely, the input feature maps of the DWT, denoted as $\mathbf{F} \in \mathbf{R}^{H\times W\times C}$, are first divided into three subparts along with the channel dimension. We denote these parts as $[F^0, F^1, F^2]$, $F^i \in \mathbf{R}^{H\times W\times C'}$, $C' = [\frac{C}{3}]$ ($[x]$ represents that the maximum integer is not more than x). Since "3" is an odd number, the number of channels in the features maps is 2^n, i.e., an even number. Therefore, we there make a specific treatment to assign the number of channels: the number of channels of F^0 and F^1 is the maximum integer not exceeding $[\frac{C}{3}]$ and the remainder of channels are assigned to F^2. For example, if the feature map F has 64 channels, F^0 and F^1 have 21 channels, and F^2 has 22 channels. Therefore, a suitable corresponding 2-D DWT frequency component is assigned to each part, which can be written as

$$\begin{aligned} x_{LL} &= F^0 \otimes f_{LL} \\ x_{LH} &= F^1 \otimes f_{LH} \\ x_{HL} &= F^2 \otimes f_{HL} \end{aligned} \tag{15}$$

where x_{LL}, x_{LH}, and x_{HL} are the low-frequency and high-frequency components corresponding to F^0, F^1, and F^2, respectively. Similar to the SE module, we use *squeeze* and *excitation* operations to obtain the attention vector of feature map $\mathbf{F}$ as

$$att_i = \text{sigmoid}(fc(\text{GAP}(F^i))) \tag{16}$$

where $att_i \in \mathbf{R}^{[\frac{C}{3}]}$, sigmoid is the sigmoid activation function to generate a scalar ranging from 0 to 1, $fc(.)$ is a feature mapping operation, such as an FC layer, and GAP is a GAP operation. The whole multiresolution attention vector, also called weight vector, can be concatenated as

$$att_{\text{MulReso}} = cat([att^0, att^1, att^2]) \tag{17}$$

where $att_{\text{MulReso}} \in \mathbf{R}^C$. This weight vector is applied to the corresponding feature maps of $\mathbf{F}$ via a channelwise multiplication operation, which yields the output of the multiresolution spectrum attention module by

$$\mathbf{Y}_{:,:,:,i} = {att_{\text{MulReso}}}^i \times \mathbf{F}_{:,:,:,i}, \text{ s.t. } i \in \{0, 1, \ldots, C-1\} \tag{18}$$

where $\mathbf{Y}$ is the output of the multiresolution spectrum attention module, ${att_{\text{MulReso}}}^i$ is the ith element of the weight vector, and $\mathbf{F}_{:,:,:,i}$ is the ith channel feature of the input $\mathbf{F}$. The overall illustration of the multiresolution spectrum attention is similar to the previous multispectrum attention, as shown in Fig. 3.

C. Depthwise Separable Convolution Block

Differently from a traditional convolution operation, i.e., a one-step operation of both the filtering and feature combinations, a depthwise separable convolution has a two-step operation: a depthwise convolution and a 1×1 pointwise convolution to substantially reduce the computational cost, which is illustrated in Fig. 5 [40].

Assuming that a traditional convolution layer takes a feature map $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$ as input and yields a feature map $\mathbf{Y} \in \mathbf{R}^{H'\times W'\times C'}$ as output, where C and C' are the number of input and output channels, respectively. The traditional convolution layer is parameterized by a convolution kernel $\mathbf{K}$ of size $k \times k \times C \times C'$, where k is the spatial square dimension of the convolution kernel and C and C' are identically defined

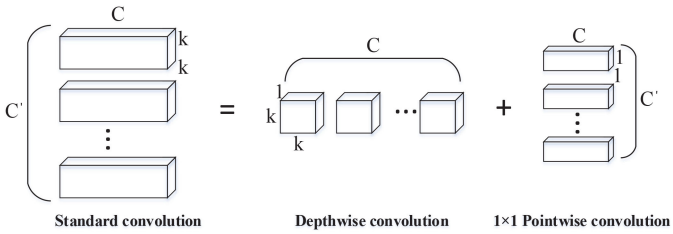


Fig. 5. Relationship between the standard convolution and the depthwise separable convolution.

aforementioned. The output feature map $\mathbf{Y}$ of the traditional convolution is computed by

$$Y_{k,l,c'} = \sum_{i,j,c} K_{i,j,c,c'} X_{k+i-1,l+j-1,c}. \tag{19}$$

According to (19), the computational cost of the traditional convolution depends on the kernel size, the number of the output channels, and the size of the input feature map, which can be calculated by

$$\text{Cost} = k \cdot k \cdot C \cdot C' \cdot H \cdot W. \tag{20}$$

As for the depthwise separable convolution, the first step is the depthwise convolution operation, i.e., each input channel has a filter to perform the channelwise convolution, which can be computed by

$$\acute{Y}_{k,l,c} = \sum_{i,j} \acute{K}_{i,j,c} X_{k+i-1,l+j-1,c} \tag{21}$$

where K' is the depthwise convolution kernel with the size of $k \times k \times C$. The computational cost of this operation can be written as

$$\text{Cost}_{\text{depthwise}} = k \cdot k \cdot C \cdot H \cdot W. \tag{22}$$

Similarly, the computational cost of the 1×1 pointwise convolution can be written as

$$\text{Cost}_{\text{pointwise}} = C \cdot C' \cdot H \cdot W. \tag{23}$$

The sum of the computational cost of depthwise and pointwise convolutions can be written as

$$\text{Cost}_{\text{sum}} = k \cdot k \cdot C \cdot H \cdot W + C \cdot C' \cdot H \cdot W. \tag{24}$$

Compared to the traditional convolution operation, the reduction of the computational cost of the depthwise separable convolution is computed by

$$\begin{aligned} \text{Ratio} &= \frac{k \cdot k \cdot C \cdot H \cdot W + C \cdot C' \cdot H \cdot W}{k \cdot k \cdot C \cdot C' \cdot H \cdot W} \\ &= \frac{1}{C'} + \frac{1}{k^2}. \end{aligned} \tag{25}$$

According to (25), the depthwise separable convolution can dramatically reduce the computational cost. For example, the depthwise separable convolution can achieve eight to nine times less computational cost when using 3×3 convolution kernel, compared with the traditional convolution.

D. Model Description

The proposed LW-CMDANet consists of two traditional convolution blocks, a cascaded multidomain attention module, and two FC layers. The cascaded multidomain attention module includes a multispectrum attention block, a multiresolution spectrum attention block, and a lightweight convolution module (consists of two depthwise separable convolution blocks). In addition, we use the hinge loss function as a classifier. The overview of the LW-CMDANet is shown in Fig. 1.

The model first extracts the low-level spatial features of the input SAR image via the traditional convolution layer, such as contexture and edge features. The multispectrum attention is, then, used to improve frequency-domain feature extraction capability by the 2-D DCT of feature maps. In order to reduce the parameters of the traditional convolution operation, we introduce two lightweight depthwise separable convolution blocks [40] to perform feature extraction, which can be used to replace the traditional convolution layer to alleviate the overfitting problem. However, the 2-D DCT is a global frequency spectral transform, and it is difficult to perform local detailed information analysis. In order to take advantage of the multiresolution detailed information analysis of feature maps, we propose a multiresolution spectrum attention module by the 2-D DWT, followed by the previous lightweight convolution block, to efficiently perform multiresolution spectrum feature extraction. The multiresolution spectrum attention module is followed by a traditional convolution layer and two FC layers to extract high-level features and form a feature vector. In addition, we propose a nongreedy training manner via the hinge loss function as a classifier to alleviate the overfitting.

E. Network Architecture

The network architecture of the LW-CMDANet is illustrated in Fig. 6. More concretely, each standard convolutional layer block contains a 64-filter with 3×3 convolutional kernel, a rectified linear unit (ReLU) activation function, and a batch normalization (BN) layer. In addition, each convolution layer is followed by a 2×2 max-pooling layer. Two FC layers include 128- and 10-D outputs, respectively. The cascaded multidomain attention module will be introduced in detail as follows.

1) Multispectrum Attention Module: The multispectrum attention module consists of multigroup feature maps along with the channel dimension, 2-D DCT-based selection, groupwise 2-D DCT, two FC layers (including 16 and 64 output channels, respectively), a ReLU, and a sigmoid activation, as shown in the multispectrum attention module of Fig. 6. This module can adaptively assign a learning scale (i.e., weight) value to each output channel of the second convolution block. This proportional value is used to weight the importance of the channelwise features, which is ranging from 0 to 1 and controlled by the sigmoid activation.

In order to alleviate the computational cost, we select main four frequency spectrum components to construct the multispectrum attention along with the channel dimension of the feature maps. Therefore, we first split the 64-channel feature maps into four groups (denoted by X^0, X^1, X^2, and X^3,

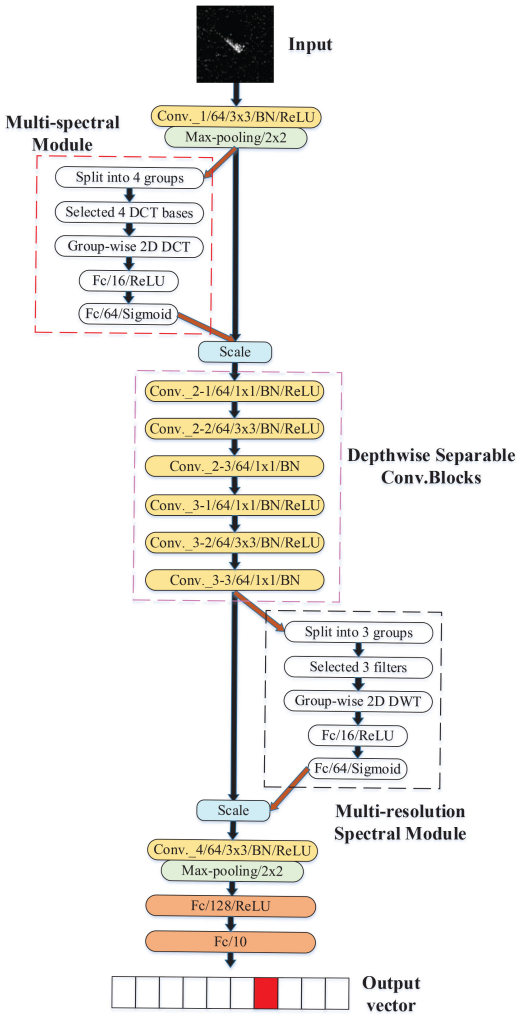


Fig. 6. Network architecture of the LW-CMDANet.

respectively) along with the channel dimension, each of which has 16 channels. From the 2-D DCT of an SAR image, the main frequency components are mainly focused on the upper-left part of the frequency component matrix, as shown in Fig. 2. It is expensive to test all the frequency component combinations to determine the selection of frequency components. Simplistically, we choose four main frequency components closest to the upper-left corner, i.e., the index is $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, which corresponds to the four groups of the feature map, respectively. According to (9), the 2D DCT frequency components of the groupwise feature maps can be written as

$$f_{u,v}^{i} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{:,h,w}^{i} A_{h,w}^{u,v}$$
$$\text{s.t. } [u,v] \in \{0,1\},\ i \in \{0,1,2,3\} \tag{26}$$

where $f_{u,v}^{i} \in \mathbf{R}^{16}$ is the frequency component corresponding to X^i, and H and W are the height and the width of feature maps, respectively.

2) Multiresolution Spectrum Attention Module: Similar to the multispectrum attention module, the multiresolution spectrum attention module consists of multigroup feature maps along with the channel dimension, the wavelet filter selection, groupwise 2-D DWT, two FC layers (including 16 and 64 output channels, respectively), a ReLU, and a sigmoid activation, as shown in the multispectrum attention module of Fig. 6.

In order to reduce the computational burden, we adopt the first-level 2-D DWT, which has four filters to obtain three main frequency components, i.e., x_{LL}, x_{LH}, and x_{HL}, respectively. We first split the last 64-channel feature maps into three groups along with the channel dimension, denoted by X^0, X^1, and X^2 ($[X^0, X^1] \in \mathbf{R}^{21}$ and $X^2 \in \mathbf{R}^{22}$), respectively. We select Haar wavelet base to perform the 2-D DWT of the input feature maps, as illustrated in (12), i.e., f_{LL}, f_{LH}, and f_{HL}. According to (15)–(18), we can obtain the output of the multiresolution spectrum attention module, illustrated as follows:

$$\mathbf{Y}_{:,:,:,i} = att_{\text{MulReso}}{}^{i} \times \mathbf{X}_{:,:,:,i},\ \text{s.t. } i \in \{0,1,\ldots,63\}. \tag{27}$$

3) Depthwise Separable Convolution Block: In order to reduce the model parameters to alleviate the overfitting problem, we adopt two identical depthwise separable convolution blocks between the multispectrum attention module and the multiresolution spectrum attention module, as illustrated in Fig. 6. Each depthwise separable convolution block consists of three convolution layers: the first one contains a 64-filter with 1×1 convolutional kernel, a ReLU activation function, and a BN layer; the second one contains a 64-filter with 3×3 convolutional kernel, a ReLU activation function, and a BN layer; and the third one includes a 64-filter with 1×1 convolutional kernel and a BN layer.

IV. Experiments and Results

In this section, the detailed experimental design and the results analysis are provided and compared. In order to better compare with the existing SOTA methods, we evaluate our proposed method by four small datasets from the MSTAR dataset [3]. The MSTAR dataset is widely used to verify the effectiveness of the existing SAR-ATR methods. The experimental results are compared with some existing SOTA methods, such as A-convNet [24], FCANet [45], SENet [44], and CBAMNet [46]. The preprocessing of the dataset is performed on the Pycharm (2018.2.4 version) soft platform and running on Windows 7 with Intel CPU i5-4460 (3.20 GHz) and RAM (16.0 GB). The network model training and evaluation are carried out on the Pytorch architecture and run on a Linux server with Intel Xecon CPU (2.20 GHz) and NVIDIA Tesla GPU K80 (24.0 GB).

A. Dataset Descriptions

MSTAR is a baseline X-band SAR imagery dataset with a resolution of 0.3×0.3 m, including ten classes of ground targets, such as BMP2 (infantry combat vehicle), BTR70 (armored personnel carrier), T72 (main tank), etc. The samples are shown in Fig. 7. The number of training and testing datasets of MSTAR is shown in Table I. The depression angle of training and testing datasets is 17° and 15°, respectively. The azimuth angle of each class is full of 360° for each class.

TABLE I
NUMBER OF SAMPLES OF THE MSTAR DATASET (SOC)

Dataset	Depression	Azimuth	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU234	Total
Training	17°	360°	299	233	298	256	233	299	299	232	299	299	**2747**
Testing	15°	360°	274	195	274	195	196	274	273	196	274	274	**2425**

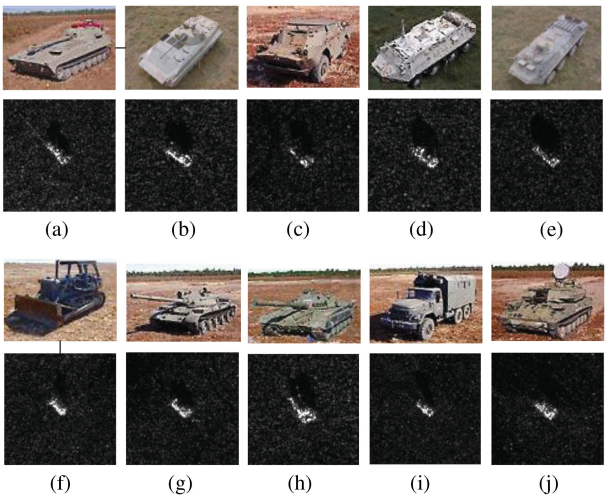


Fig. 7. MSTAR data samples [3], optical images (top), and their corresponding SAR images (bottom). (a) 2S1. (b) BMP2. (c) BRDM2. (d) BTR60. (e) BTR70. (f) D7. (g) T62. (h) T72. (i) ZIL131. (j) ZSU234.

In order to better validate the performance of the proposed LW-CMDANet, we construct four types of small training subdatasets, i.e., subset-200, subset-100, subset-50, and subset-20, from the overall training dataset of MSTAR to train the LW-CMDANet in our experiments. The number of these subdatasets for each class is different. For example, there are 200 SAR image samples for each class in subset-200 (a total of 2000 samples for ten classes). In addition, we select 20 samples of each class from the training dataset of MSTAR and a total of 200 samples to construct a validation dataset, which is different from the previous four training subdatasets. The validation dataset of the four experiments is the same. The testing dataset of each class is the testing dataset of the MSTAR, which has not been changed in each experiment. To alleviate the degradation of clutter or speckle noise in SAR-ATR performance, we crop the original SAR image (all training, validation, and testing datasets) with a size of 128×128 into 40×40 centered on the midpoint of the original image.

B. Setting and Training

All the experiments adopt the Adam optimizer [56], and the initial learning rate is 10^{-4}, which is half annealed every 50 training epochs. The total number of training epochs is 200. The batch size is 64. After each training epoch, we use all the testing datasets to test the performance of the trained model. According to the analysis of qualitative and quantitative results, we evaluate the performance of our proposed method through the training loss curve, average test accuracy curve, and feature map visualization compared to some existing SOTA methods.

C. Results

This part has analyzed all the experimental results of our proposed method and existing SOTA methods in detail from quantitative and qualitative experimental results. The performance metrics include recognition accuracy and computational complexity. In addition, the ablation experiments have been conducted to further verify the effectiveness of our proposed method. These experimental results have confirmed the feasibility and efficiency of our proposed method compared to the existing SOTA methods.

1) LW-CMDANet Performance: The training loss curve and the test recognition accuracy curve of the LW-CMDANet on four different MSTAR training subsets are shown in Figs. 8 and 9, respectively. As seen from Fig. 8(a), overfitting appears in the subset-20 experiment, since the validation loss curve fluctuates in the range 0.03–0.04 during all the training process, which is inconsistent with the training loss curve, while the testing accuracy of the subset-20 experiment is only about 52% when the LW-CMDANet reaches convergence, as illustrated in Fig. 9. The reason behind this is that the number of data samples of the subset-20 is extremely small, i.e., 20 samples for each class. There are more than 30 SAR image samples within 0–360° of azimuth angles for each class of the MSTAR dataset. However, the SAR imaging process is very sensitive to the azimuth angle of the target [57]. That is, different angles may produce very different SAR images with a same target, since the electromagnetic scatter features of the target are extremely different at different azimuths. These impact factors include target shape, size, material types, and so on. Therefore, the samples of subset-20 for each class are incomplete, which leads to insufficient feature representation learning in the training process.

With the number of training samples increasing, the training performance is better, i.e., the validation and training loss curves are more constant, which is closed to 0 when the training process reaches convergence. Subset-50 converges at about the 50th training epoch, while subset-100 and subset-200 converge at about 25th and 10th training epoch, as shown in Fig. 8(b) and (c), respectively. The more the SAR samples, the better the test accuracy performance, as shown in Fig. 9. The test accuracy rate of subset-20, subset-50, subset-100, and subset-200 is about 55.34%, 89.93%, 92.15%, and 96.63%, respectively, when the training process reaches convergence.

In addition, in order to more clearly demonstrate the good performance of our proposed method on each target category recognition task, we have tested the recognition performance of the trained LW-CMDANet on four experimental scenarios with the testing dataset of MSTAR. The recognition results are illustrated with confusion matrixes, as shown in Fig. 10.

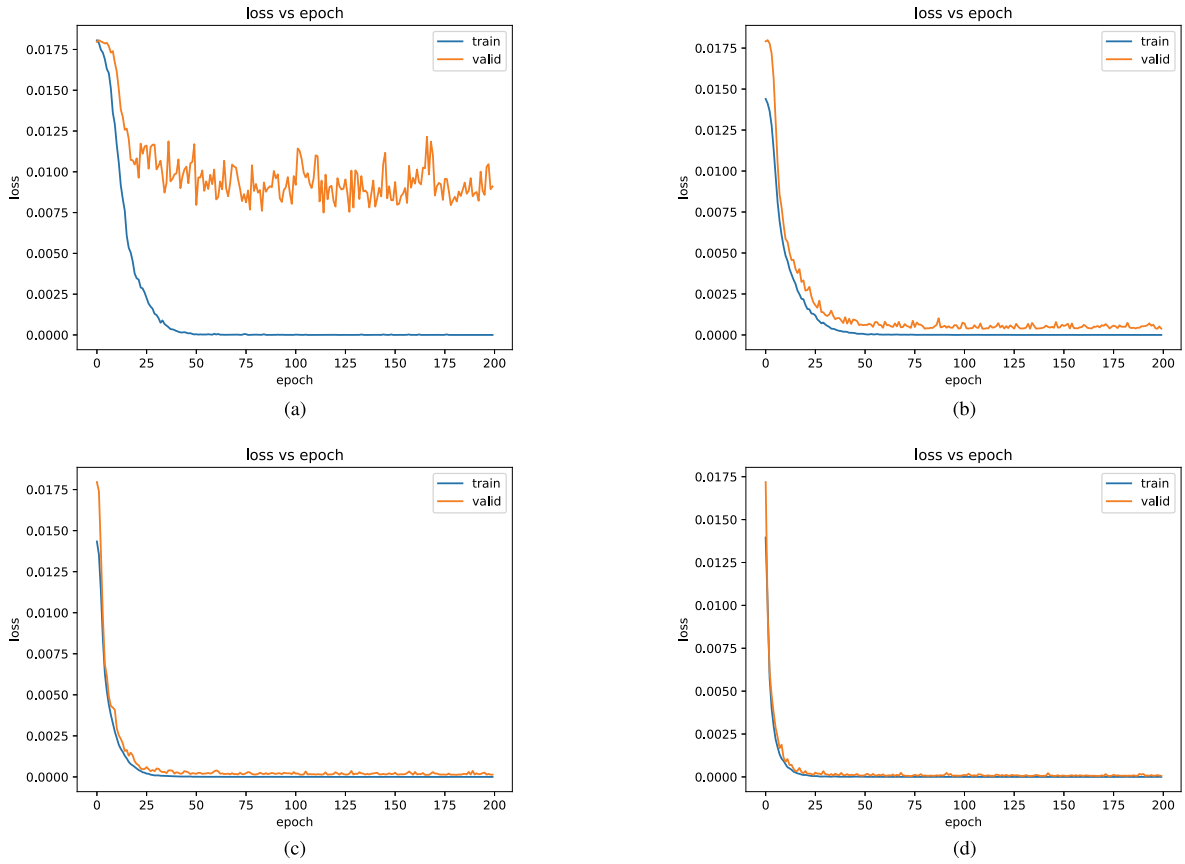


Fig. 8. Loss curves of the LW-CMDANet training on four different MSTAR training subsets. (a) Subset-20. (b) Subset-50. (c) Subset-100. (d) Subset-200.

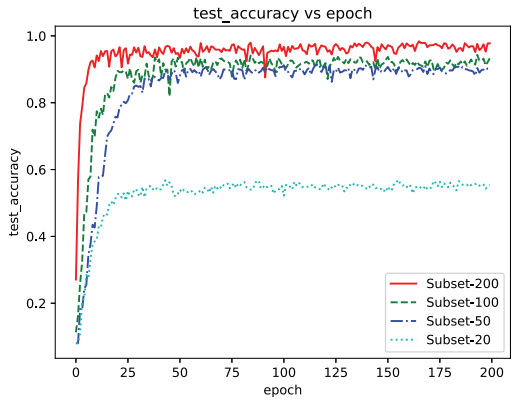


Fig. 9. Test recognition accuracy curves of all four experiments.

The diagonal line of the confusion matrix shows the number of correct recognized samples for each class; others are the wrong samples. It can be seen from Fig. 10 that with more training samples, the performance of the model is higher in terms of recognition accuracy, which is consistent with the result analysis of Fig. 9.

2) Comparison With the Existing SOTA Methods: Besides, we have compared the performance between the proposed LW-CMDANet and the existing SOTA methods in terms of test accuracy, as illustrated in Table II. In order to better compare to the existing SOTA method, most of these compared SOTA methods have a similar architecture, i.e., the same number of convolutional layers. For example, the baseline CNN has four standard convolutional layers and two FC layers. The HL-CNN represents that this CNN model uses the hinge loss function; the MobileNet represents that the middle two layers are depth-wise separable convolution blocks; and the CNN-SENet introduces the SE attention module into the CNN model. Similarly, the attention module of CA, FCA, CBAM, the cascade FCA, and DWT is also embedded into the CNN or A-ConvNet model to construct the CNN-CANet, CNN-FCANet, CNN-CBAMNet, CNN-FCA-DWTNet, and A-CNN-FCA-DWTNet, respectively. In addition, we have also compared with latest methods, such as YOLO-DMCCA [60], SAR-OVSM [64], SAR-VGG-KNN [65], SAR-HOG [66], and SAR-BoVM [67]. As seen from Table II, when the training dataset is subset-200, the test accuracy rate is more than 90% for all the methods except for the CNN-CANet is about 85%.

The test accuracy of the CNN-FCA-DWTNet is higher than that of our proposed LW-CMDANet when the training dataset is subset-200 and subset-100. The reason behind this is that it is insufficient to fully extract hierarchical features via the depth-wise separable convolution operation of the LW-CMDANet, due to the reduction of convolutional parameters compared to standard convolution. If the model has enough data for training, it is more advantageous to extract more effective input features by slightly increasing the model parameters. When the training dataset is subset-50 or subset-20, the accuracy of our proposed LW-CMDANet (i.e., 89.93% and 55.34%, respectively) is higher than that of the CNN-FCA-DWTNet (i.e., 86.16% and 55.04%).

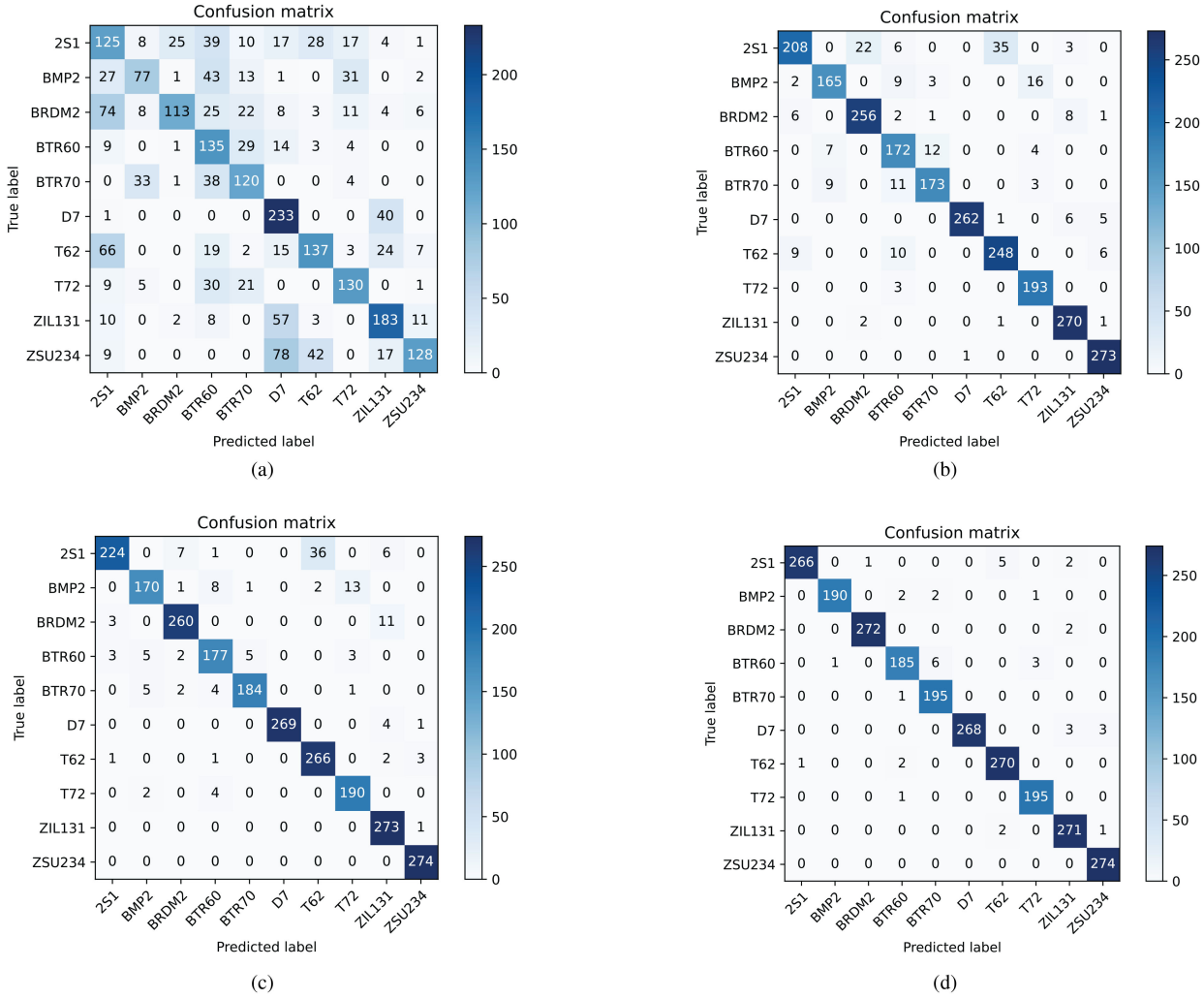


Fig. 10. Confusion matrixes on four experimental scenarios: (a) subset-20, (b) subset-50, (c) subset-100, and (d) subset-200.

TABLE II
TEST ACCURACY PERFORMANCE BETWEEN THE LW-CMDANET AND THE EXISTING SOTA METHODS ON DIFFERENT MSTAR DATA SUBSETS

Model	Subset-200(%)	Subset-100(%)	Subset-50(%)	Subset-20(%)
Baseline CNN	95.86	89.69	83.57	54.20
A-ConvNet [24]	96.58	90.93	82.24	47.89
ResNet [58]	95.97	89.27	83.55	53.67
HL-CNN	96.38	90.82	83.91	53.87
MobileNet [40]	96.91	90.02	83.91	53.87
CNN-SENet [44]	96.63	91.73	86.11	54.95
CNN-FCANet [45]	92.24	89.66	82.05	51.61
CNN-CBAMNet [46]	96.02	90.84	82.25	51.61
CNN-CANet [59]	85.88	73.87	70.41	49.54
YOLO-DMCCA [60]	91.38	86.69	60.56	40.23
SAR-OVSM [64]	93.68	76.35	50.57	45.67
SAR-VGG-KNN [65]	92.67	88.54	65.66	49.89
SAR-HOG [66]	93.26	74.58	48.97	43.26
SAR-BoVW [67]	93.38	78.32	53.65	48.23
A-CNN-FCA-DWTNet	95.94	87.84	73.66	44.20
CNN-FCA-DWTNet	**97.15**	**92.98**	86.16	55.04
LW-CMDANet (proposed)	96.63	92.15	**89.93**	**55.34**
LW-FCANet	96.76	91.88	89.75	54.82
LW-DWTNet	96.36	92.09	89.76	55.31

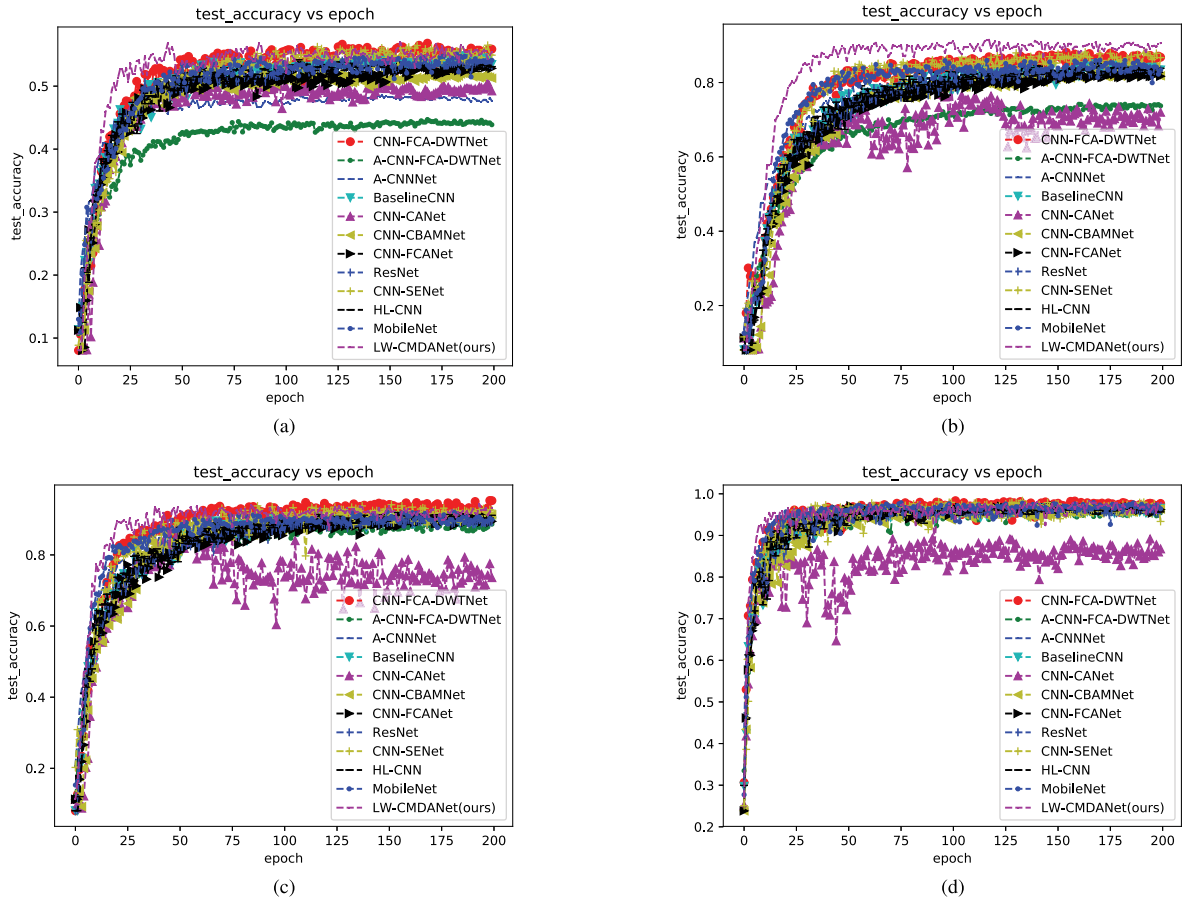


Fig. 11. Convergence curves of our proposed method and comparison SOTA methods on four experiment scenarios: (a) subset-20, (b) subset-50, (c) subset-100, and (d) subset-200.

The reason behind this is that the extracted features by the depthwise separable convolution of the LW-CMDANet are sparse and generalized when the number of data samples is small, which is beneficial to increase the generalization capacity to alleviate the overfitting, while the standard convolution has more parameters to extract more features, which is redundant for the limited data samples to some extent. In addition, compared to [68], the accuracy of our proposed method is 89.93%, which is higher than 88% in [68] in the subset-50 experiment. However, the accuracy of our proposed method is 92.15%, which is lower than 95% in [68] in the subset-100 experiment. The reason of this is that except for 100 labeled samples for each class, [68] has extra unlabeled samples for each class to perform unsupervised learning to assist the supervised learning on labeled samples. Therefore, with the knowledge of unlabeled samples, [68] has higher accuracy on the subset-100 experiment.

In order to further verify the effectiveness of our proposed method, we compare the convergence performance between our proposed method and the existing SOTA methods on the test dataset during the training process. The convergence curves of four kinds of experiments are illustrated in Fig. 11. It can be seen from Fig. 11 that our proposed method converges the fastest on all the experiments. The convergence point is at about 50, 30, 25, and 10 epochs, as shown in Fig. 11(a)–(d), respectively, while more training epochs are needed to reach the convergence point for the comparison existing SOTA methods. Moreover, our proposed method has higher performance in terms of recognition accuracy, especially on the subset-50 experiment, which achieves the highest accuracy rate, i.e., about 89.93% when it reaches convergence. From the quantitative (see Table II) and qualitative (see Fig. 11) experimental results, our proposed method, i.e., LW-CMDANet, has higher performance in terms of SAR target recognition accuracy when the number of data samples is small compared with the existing SOTA methods.

The stability of the extracted features by the model can directly affect the recognition performance. We have verified the feasibility and effectiveness of our proposed method via the visualization of the feature map stability. For simplicity, we take the A-ConvNet and the baseline CNN model as comparative examples. We have done the feature stability comparison experiments between our proposed method and comparison SOTA methods by t-distribute stochastic neighbor embedding (t-SNE) [61], as illustrated in Fig. 12. In order to visualize the experimental results, the input high-dimensional feature maps, extracted by the trained model, are reduced to 2-D by t-SNE. In general, the effective and efficient model has larger interclass difference and higher intraclass aggregation in the feature map space than that of the poor model. It can be seen from the 2-D feature representation space of t-SNE in Fig. 12, with the number of training data samples increasing, the feature maps of different targets are more distinguishable. More specifically, in the subset-20 experiment, it is difficult to observe the interclass

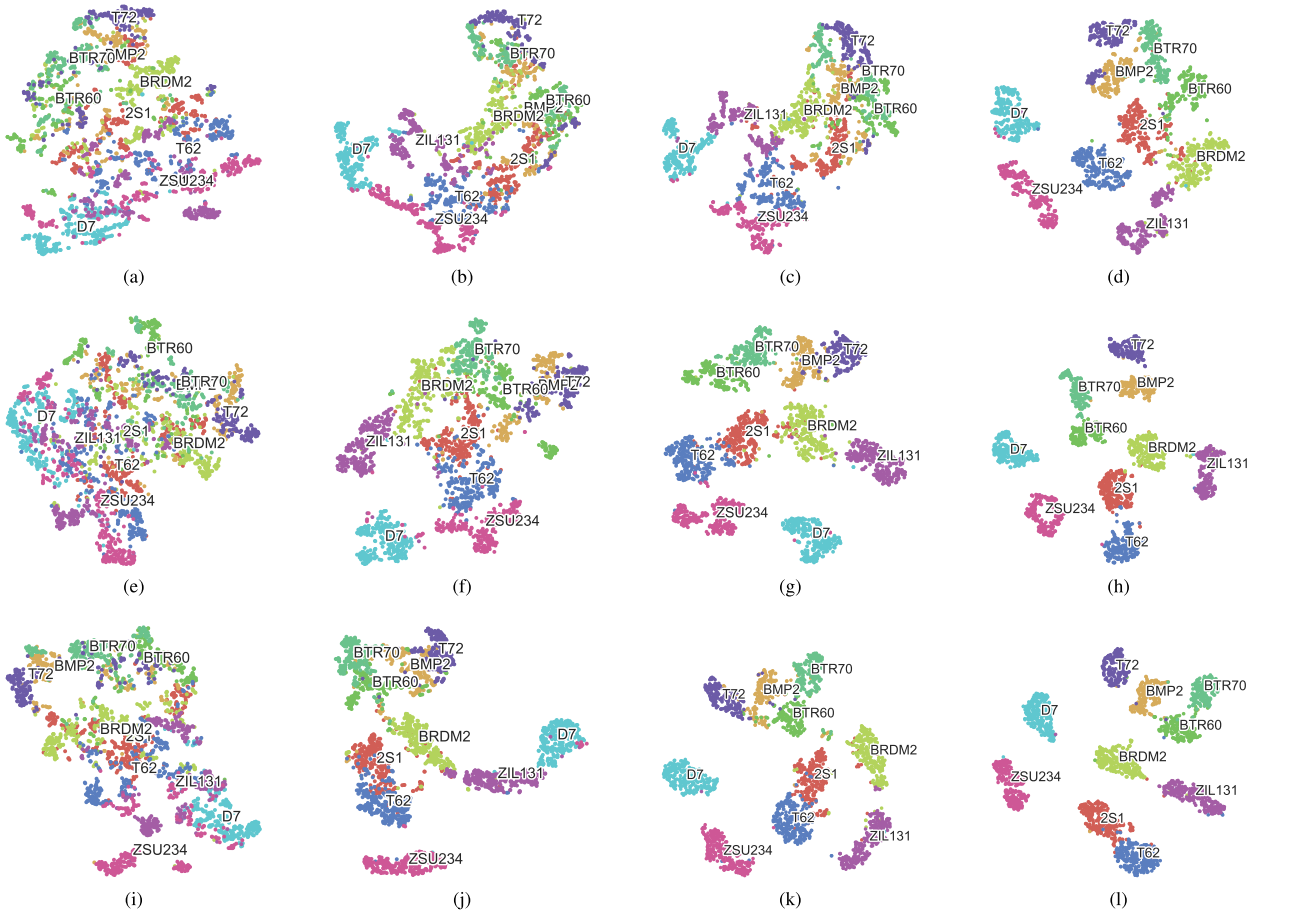


Fig. 12. Feature map comparison at the 2-D t-SNE space. (a)–(d) are the feature maps extracted by the pretrained baseline CNN. (e)–(h) are the feature maps extracted by the pretrained A-ConvNet. (i–l) are the feature maps extracted by pretrained our proposed method, i.e., LW-CMDANet. From the first column to the last column represents the experiments of subset-20, subset-50, subset-100, and subset-200, respectively. The different colors of circle points represent different types of SAR targets.

difference, which presents a large degree of confusion on the recognition results, as shown in Fig. 12(a), (e), and (i). As for the subset-200 scenario, the interclass difference between the different targets is obvious, and there is a high aggregation degree in the intraclass, as illustrated in Fig. 12(d), (h), and (l). Compared to the feature maps extracted by the baseline CNN and the A-ConvNet in the four experiments, our proposed method, i.e., LW-CMDANet, has higher performance in terms of interclass feature map separation and intraclass feature map aggregation. Therefore, our proposed method has effective and efficient recognition performance, which is consistent with the quantitative results, as shown in Table II.

3) Computational Complexity: In addition to recognition accuracy, the computational complexity is also a key factor for SAR-ATR, which is determined by the learnable parameters, data size, other nonlearnable parameters, and so on. Taking the subset-20 experiment as an example, we have compared the training time and the testing (i.e., inference) time between our proposed method and comparison SOTA methods, as shown in Figs. 13 and 14 [64], respectively. The testing time is regarded as the computational cost when only one SAR image from the testing dataset is used as the input of the trained model. It can be seen from Figs. 13 and 14 that the models with the cascaded multispectrum and multiresolution spectrum attention

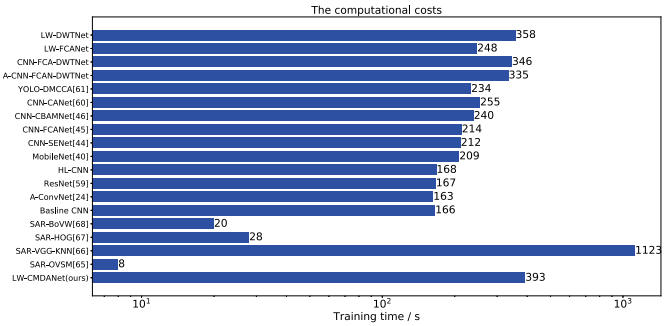


Fig. 13. Training time on the subset-20 experiment.

module need more computational cost except for SAR-VGG-KNN, such as LW-CMDANet, CNN-FCA-DWTNet, and A-CNN-FCA-DWTNet. Except for the learnable parameters, there are many nonlearnable operations in the cascade multispectrum and multiresolution spectrum attention module, such as the 2-D DCT and 2-D DWT operations of the feature map. Compared to some other SOTA models, e.g., baseline CNN and ResNet, these operations need a little bit larger computational cost. Although the training time of our proposed LW-CMDANet is 393 s (less than 7 min) for all 200 epochs, the testing time is just about

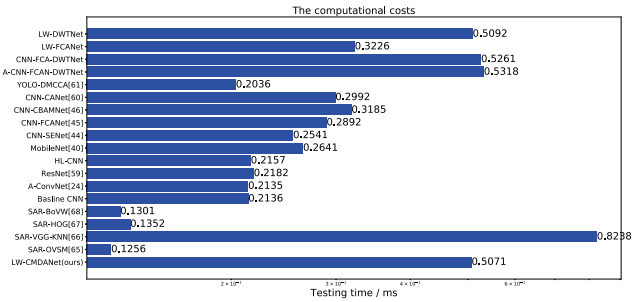


Fig. 14. Testing time on the subset-20 experiment.

0.5071×10^{-3} s. As seen from Figs. 13 and 14, the training time or testing time of YOLO-DMCCA, SAR-HOG, SAR-BovW, and SAR-OVSM is obviously lower than that of LW-CMDANet, such as the training time of SAR-OVSM is only 8 s. Since the YOLO-DMCCA model has pretrained in the large-scale dataset to obtain the prior knowledge, which can accelerate the training speed in the following SAR-ATR task. The SAR-HOG, SAR-BovW, and SAR-OVSM are not DL-based methods, which do not have a large number of parameters to train during the training process. In addition, the features of these methods are manually extracted via the feature extractor, such as Gabor filter in SAR-BovW. However, the recognition accuracy of these methods is limited than that of LW-CMDANet, as shown in Table II. In addition, the training time of [68] is 2.35 s per training epoch, while our proposed method is 1.97 s. The experimental result analysis of the computational complexity has demonstrated that our proposed method has better or competitive performance compared to some existing SOTA methods.

4) Ablation Experiments: In order to further verify the effectiveness of our proposed method, we have implemented the ablation experiments. Compared with our proposed method, i.e., LW-CMDANet, the two comparison models are designed: one has only a multispectrum attention module, i.e., LW-FCANet, while the other has only a multiresolution spectrum attention module, i.e., LW-DWTNet. The experimental results of test accuracy and the computational complexity of ablation experiments are shown in the last two rows of Table II and Figs. 13 and 14. It can be seen from Table II that the test accuracy of our proposed method, LW-FCANet, and LW-DWTNet is 96.63%, 96.76%, and 96.36% in the subset-200 experiment, respectively. The LW-FCANet is slightly higher than our proposed method. One reason is that when the training dataset is relatively sufficient, the DWT may cause information loss, because the high-frequency component of the DWT, i.e., x_{HH}, has been excited in our proposed method, while the test accuracy of our proposed method is higher than LW-FCANet and LW-DWTNet in all the subset-100, subset-50, and subset-20 experiments. The reason is that when the training data are small, the multidomain features can improve the generalization of the model. When compared to the MobileNet [40] (without any attention module), our proposed method has a higher test accuracy, i.e., 92.15%, 89.93%, and 55.34% in subset-100, subset-50, and subset-20 experiments, respectively, while the MobileNet is 90.02%, 83.91%, and 53.87%, respectively. As seen from Fig. 13, the training time of our proposed method (393 s) is larger than that of two ablation experiments (248 and 358 s, respectively), since our proposed method is more complex, which includes two attention modules. The inference time, i.e., testing time of an input image, of our proposed method is 0.5071×10^{-3} s, which is larger than 0.3226×10^{-3} of LW-FCANet and smaller than 0.5092×10^{-3} of LW-DWTNet, as shown in Fig. 14. These experimental results have demonstrated that the computational cost of our proposed method is better or more competitive.

D. Discussions

Our proposed method can effectively alleviate the overfitting and computational cost problems at limited data scenarios, which benefits from the following four aspects, i.e., data preprocessing, cascaded multispectrum and multiresolution spectrum attention module, depthwise separable convolution, and nongreedy learning strategy. First, we slice all the input SAR images into the size of 40×40 as data preprocessing, which can efficiently alleviate the degradation of clutter or speckle in the recognition performance. Owing to the input interference, such as noise, the performance of the deep neural network may be severely deteriorated.

Second, the extracted features are more sparse and abstractive, the generalization of the model is better, and the model is more beneficial to alleviate the overfitting problem [62]. Multidomain feature subspace fusion representation learning, performed by the convolution operation, cascaded multispectrum (i.e., 2-D DCT), and multiresolution spectrum (i.e., 2-D DWT) attention module, is effective and efficient, which can contribute to completely extract features of the input image from spatial, frequency, and wavelet transform domains at the same time via an end-to-end model. The multidomain feature subspace fusion can greatly enrich the feature representation space of the input image. In this way, the feature extraction of the proposed method can be performed from the spatial domain, frequency domain, and wavelet transform simultaneously. These extracted feature maps have higher degree of sparsity and generalization compared to only the original spatial feature space. Therefore, multidomain feature subspace fusion representation learning can efficiently alleviate the overfitting problem and improve the recognition accuracy in the case of limited data.

Third, the more the parameters of the model, the easier the overfitting appears in the model . In order to reduce the number of parameters of the proposed method, we introduce the depthwise separable convolution operation to replace the standard convolution to reduce the eight to nine times of parameters (as explained in Section III-C). Therefore, the extracted features of the model are sparse and less redundant, which can alleviate the overfitting to some extent.

Finally, we adopt a nongreedy training strategy to replace the standard greedy training (i.e., cross-entropy loss function) method. More concretely, we use the hinge loss function to replace the traditional cross-entropy loss function to perform nongreedy learning. This strategy is effective to address the overfitting problem, as illustrated in our experimental results.

However, our proposed method has main four limitations.

1) We have sliced the original SAR image size of 128×128 into 40×40 to alleviate the bad influence of background

clutter and noise in the SAR-ATR task. Therefore, our proposed method is effective only for spotlight SAR images, since this type of image has high signal-to-clutter or noise ratio, high resolution, and small imaging background. Therefore, its application is limited.

2) Since our proposed method is a multidomain feature subspace fusion representation learning, which may not be well in robustness with respect to disturbed input image.
3) The choice of frequency components in the frequency spectrum attention module is only four specific frequency components, which maybe degrade the recognition performance on other datasets. Therefore, the optimal choice strategy of frequency components after the DCT of feature maps should be studied.
4) The DWT and the DCT of feature maps need larger computational cost compared to some existing attention module.

Therefore, it needs more computational time during training and testing processes than some SOTA methods, such as CNN-CANet [59] and CNN-CBAMNet [46]. Therefore, our proposed method maybe is limited in some practical real-time scenarios. According to above limitations, we will go further study our proposed method in our future work.

V. Conclusion

This article proposed an alternative end-to-end lightweight network based on a cascade multidomain attention (i.e., LW-CMDANet) to improve the recognition performance of the DL model in the limited data sample scenarios. Our proposed method made full use of the advantage of the multidomain feature subspace fusion representation learning method and the lightweight CNN design to improve the feature extraction capacity of the DL model. These extracted features were sparse and generalized, which can effectively and efficiently alleviate the overfitting and computational cost problems of the deep-CNN-based model. The qualitative and quantitative experimental results on the MSTAR dataset demonstrated that our proposed method has better or competitive performance compared to the existing SOTA methods. Our proposed method has a bright application prospect in the practical SAR-ATR field. However, there are still some issues that need to be improved, such as the optimal choice strategy of frequency components after the DCT of feature maps and the acceleration of the DWT of feature maps. In addition, our proposed method was verified only at the standard MSTAR dataset. We will verify the performance of our proposed method at more complex SAR imaging conditions in the future work.

References

[1] D. Blacknell and H. Griffiths, *Radar Automatic Target Recognition (ATR) and Non-Cooperative Target Recognition (NCTR)*. London, U.K.: Inst. Eng. Technol., 2013.

[2] I. M. Gorovyi and D. S. Sharapov, "Comparative analysis of convolutional neural networks and support vector machines for automatic target recognition," in *Proc. IEEE Microw. Radar Remote Sens. Symp.*, 2017, pp. 63–66.

[3] E. R. Keydel, S. Lee, and J. Moore, "MSTAR extended operating conditions: A tutorial," *Proc. SPIE*, vol. 2757, no. 1, pp. 228–242, Jun. 1996, doi: 10.1117/12.242059.

[4] Y. Zhou, H. P. Wang, F. Xu, and Y. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016, doi: 10.1109/LGRS.2016.2618840.

[5] Q. He, L. Zhao, K. Ji, and G. Kuang, "SAR target recognition based on task-driven domain adaptation using simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4019205, doi: 10.1109/LGRS.2021.3116707.

[6] S. Du, J. Hong, Y. Wang, K. Xing, and T. Qiu, "Physical-related feature extraction from simulated SAR image based on the adversarial encoding network for data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4016705, doi: 10.1109/LGRS.2021.3100642.

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

[8] K. Huang, W. Nie, and N. Luo, "Fully polarized SAR imagery classification based on deep reinforcement learning method using multiple polarimetric features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 3719–3730, Oct. 2019, doi: 10.1109/JSTARS.2019.2913445.

[9] Q. Song, F. Xu, and Y. Jin, "SAR image representation learning with adversarial autoencoder networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, 2019, pp. 9498–9501, doi: 10.1109/IGARSS.2019.8898922.

[10] Z. L. Huang, Z. X. Pan, and B. Lei, "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data," *Remote Sens.*, vol. 9, 2017, Art. no. 907.

[11] Z. Ying et al., "TAI-SARNET: Deep transferred atrous-inception CNN for small samples SAR ATR," *Sensors*, vol. 20, no. 6, 2020, Art. no. 1724.

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

[13] J. Oh, G. Y. Youm, and M. Kim, "SPAM-Net: A. CNN-based SAR target recognition network with pose angle marginalization learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 701–714, Feb. 2021, doi: 10.1109/TCSVT.2020.2987346.

[14] Y. Sun, Y. Wang, H. Liu, N. Wang, and J. Wang, "SAR target recognition with limited training data based on angular rotation generative network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1928–1932, Nov. 2020, doi: 10.1109/LGRS.2019.2958379.

[15] Z. Wen, Z. Liu, S. Zhang, and Q. Pan, "Rotation awareness based self-supervised learning for SAR target recognition with limited training samples," *IEEE Trans. Image Process.*, vol. 30, pp. 7266–7279, 2021, doi: 10.1109/TIP.2021.3104179.

[16] G. Dong and H. Liu, "Global receptive-based neural network for target recognition in SAR images," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1954–1967, Apr. 2021, doi: 10.1109/TCYB.2019.2952400.

[17] J. Zhang, M. Xing, and Y. Xie, "FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2174–2187, Mar. 2021, doi: 10.1109/TGRS.2020.3003264.

[18] S. Feng, K. Ji, L. Zhang, X. Ma, and G. Kuang, "SAR target classification based on integration of ASC parts model and deep learning algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10213–10225, 2021, doi: 10.1109/JSTARS.2021.3116979.

[19] J. Zhang, M. Xing, G.-C. Sun, and Z. Bao, "Integrating the reconstructed scattering center feature maps with deep CNN feature maps for automatic SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4009605, doi: 10.1109/LGRS.2021.3054747.

[20] J. Liu, M. Xing, H. Yu, and G. Sun, "EFTL: Complex convolutional networks with electromagnetic feature transfer learning for SAR target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5209811, doi: 10.1109/TGRS.2021.3083261.

[21] C. Wang, X. Liu, J. Pei, Y. Huang, Y. Zhang, and J. Yang, "Multiview attention CNN-LSTM network for SAR automatic target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12504–12513, Nov. 2021, doi: 10.1109/JSTARS.2021.3130582.

[22] R. H. Shang, J. M. Wang, L. C. Jiao, R. Stolkin, B. Hou, and Y. Li, "SAR targets classification based on deep memory convolution neural networks and transfer parameters," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2834–2846, Aug. 2018.

[23] C. Zhong, X. Mu, X. He, J. Wang, and M. Zhu, "SAR target image classification based on transfer learning and model compression," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 412–416, Mar. 2019, doi: 10.1109/LGRS.2018.2876378.

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

[25] L. Wang, X. Bai, C. Gong, and F. Zhou, "Hybrid inference network for few-shot SAR automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9257–9269, Nov. 2021, doi: 10.1109/TGRS.2021.3051024.

[26] X. Zhou, T. Tang, Y. Cui, L. Zhang, and G. Kuang, "Novel loss function in CNN for small sample target recognition in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4018305, doi: 10.1109/LGRS.2021.3109342.

[27] C. Wang et al., "Label noise modeling and correction via loss curve fitting for SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5216210, doi: 10.1109/TGRS.2021.3121397.

[28] J. Guan, J. Liu, P. Feng, and W. Wang, "Multiscale deep neural network with two-stage loss for SAR target recognition with small training set," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4011405, doi: 10.1109/LGRS.2021.3064578.

[29] 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, 2021, Art. no. 5214313, doi: 10.1109/TGRS.2021.3106915.

[30] 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, doi: 10.1109/TGRS.2019.2931308.

[31] S. Wang, Y. Wang, H. Liu, and Y. Sun, "Attribute-guided multi-scale prototypical network for few-shot SAR target classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12224–12245, Nov. 2021, doi: 10.1109/JSTARS.2021.3126688.

[32] J. F. Pei, Y. L. Huang, W. B. 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.

[33] K. Wang, G. Zhang, Y. Xu, and H. Leung, "SAR target recognition based on probabilistic meta-learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 682–686, Apr. 2021, doi: 10.1109/LGRS.2020.2983988.

[34] L. Li, J. Liu, L. Su, C. Ma, B. Li, and Y. Yu, "A novel graph meta-learning method for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4015705, doi: 10.1109/LGRS.2021.3097130.

[35] K. Fu, T. Zhang, Y. Zhang, Z. Wang, and X. Sun, "Few-shot SAR target classification via meta-learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 2000314, doi: 10.1109/TGRS.2021.3058249.

[36] Y. Zhai et al., "Weakly contrastive learning via batch instance discrimination and feature clustering for small sample SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5204317, doi: 10.1109/TGRS.2021.3066195.

[37] C. Wang, H. Gu, and W. Su, "SAR image classification using contrastive learning and pseudo-labels with limited data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4012505, doi: 10.1109/LGRS.2021.3069224.

[38] C. Wang et al., "Semisupervised learning-based SAR ATR via self-consistent augmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4862–4873, Jun. 2021, doi: 10.1109/TGRS.2020.3013968.

[39] R. Qin et al., "A semi-greedy neural network CAE-HL-CNN for SAR target recognition with limited training data," *Int. J. Remote Sens.*, vol. 41, no. 20, pp. 7889–7911, 2020, doi: 10.1080/01431161.2020.1766149.

[40] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[41] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 11, 1254–1259, Nov. 1998, doi: 10.1109/34.730558.

[42] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[43] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458, doi: 10.1109/CVPR.2017.683.

[44] J. Hu, L. Shen, and S. Albanie, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[45] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 763–772.

[46] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision*. New York, NY, USA: Springer, 2018.

[47] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[48] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Comput. Surveys*, 2022, doi: 10.1145/3530811.

[49] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.

[50] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2020, *arXiv:1909.11065*.

[51] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[52] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," 2018, *arXiv:1805.08620*.

[53] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level wavelet convolutional neural networks," IEEE Access, vol. 7, pp. 74973–74985, 2019, doi: 10.1109/ACCESS.2019.2921451.

[54] R. Qin, X. Fu, J. Chang, and P. Lang, "Multilevel wavelet-SRNet for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4009005, doi: 10.1109/LGRS.2021.3050891.

[55] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[56] K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.

[57] 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, doi: 10.1109/MGRS.2013.2248301.

[58] 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.

[59] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.

[60] M. Amrani, A. Bey, and A. Amamra, "New SAR target recognition based on YOLO and very deep multi-canonical correlation analysis," *Int. J. Remote Sens.*, Aug. 2021, doi: 10.1080/01431161.2021.1953719.

[61] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Stochastic neighbor embedding with Gaussian and student-t distributions: Tutorial and survey," 2020, *arXiv:2009.10301*.

[62] S. Wiedemann, K. -R. Mller, and W. Samek, "Compact and computationally efficient representation of deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 772–785, Mar. 2020, doi: 10.1109/TNNLS.2019.2910073.

[63] P. Lang et al., "RRSARNet: A novel network for radar radio sources adaptive recognition," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11483–11498, Nov. 2021, doi: 10.1109/TVT.2021.3104824.

[64] M. Amrani, F. Jiang, Y. Xu, S. Liu, and S. Zhang, "SAR-oriented visual saliency model and directed acyclic graph support vector metric based target classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3794–3810, Oct. 2018, doi: 10.1109/JSTARS.2018.2866684.

[65] M. Amrani and F. Jiang, "Deep feature extraction and combination for synthetic aperture radar target classification," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042616, doi: 10.1117/1.JRS.11.042616.

[66] M. Amrani, K. Yang, D. Zhao, X. Fan, and F. Jiang, "An efficient feature selection for SAR target classification," in *Adv. in Multimedia Inf. Process.* Cham, Switzerland: Springer, 2017, pp. 68–78.

[67] M. Amrani, S. Chaib, I. Omara, and F. Jiang, "Bag-of-visual-words based feature extraction for SAR target classification," in *Proc. 9th Int. Conf. Digit. Image Process.*, 2017, pp. 327–332.

[68] Z. Yue et al., "A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition," *Cogn. Comput.*, vol. 13, pp. 795–806, 2021.