A CNN-Based Method for Sparse SAR Target Classification with Grad-CAM Interpretation

JI Zhongyuan[1,2,3], *ZHANG Jingjing*[1,3*], *LIU Zehao*[4], *LI Guoxu*[5]

1. College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China; 2. College of Criminal Justice, Shandong University of Political Science and Law, Jinan 250014, P. R. China; 3. The Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China; 4. NARI Technology Co., Ltd., Nanjing 211106, P. R. China; 5. China Electronics Technology Corporation Ceyear Technology Co., Ltd., Qingdao 266555, P. R. China

(Received 12 June 2025; revised 10 August 2025; accepted 20 August 2025)

Abstract: In recent years, deeps learning has been widely applied in synthetic aperture radar (SAR) image processing. However, the collection of large-scale labeled SAR images is challenging and costly, and the classification accuracy is often poor when only limited SAR images are available. To address this issue, we propose a novel framework for sparse SAR target classification under few-shot cases, termed the transfer learning-based interpretable lightweight convolutional neural network (TL-IL-CNN). Additionally, we employ enhanced gradient-weighted class activation mapping (Grad-CAM) to mitigate the "black box" effect often associated with deep learning models and to explore the mechanisms by which a CNN classifies various sparse SAR targets. Initially, we apply a novel bidirectional iterative soft thresholding (BiIST) algorithm to generate sparse images of superior quality compared to those produced by traditional matched filtering (MF) techniques. Subsequently, we pretrain multiple shallow CNNs on a simulated SAR image dataset. Using the sparse SAR dataset as input for the CNNs, we assess the efficacy of transfer learning in sparse SAR target classification and suggest the integration of TL-IL-CNN to enhance the classification accuracy further. Finally, Grad-CAM is utilized to provide visual explanations for the predictions made by the classification framework. The experimental results on the MSTAR dataset reveal that the proposed TL-IL-CNN achieves nearly 90% classification accuracy with only 20% of the training data required under standard operating conditions (SOC), surpassing typical deep learning methods such as vision Transformer (ViT) in the context of small samples. Remarkably, it even presents better performance under extended operating conditions (EOC). Furthermore, the application of Grad-CAM elucidates the CNN's differentiation process among various sparse SAR targets. The experiments indicate that the model focuses on the target and the background can differ among target classes. The study contributes to an enhanced understanding of the interpretability of such results and enables us to infer the classification outcomes for each category more accurately.

Key words: sparse synthetic aperture radar; convolutional neural network (CNN); ensemble learning; target classification; SAR interpretation

CLC number: TN925 **Document code:** A **Article ID:** 1005-1120(2025)04-0525-16

0 Introduction

Distinct from conventional radar systems, synthetic aperture radars (SARs) operate effectively under all-day and all-weather conditions and have been deployed in both military and civilian applications[1-2]. Within the realm of SAR image processing, target classification[3-4] has risen in popularity with the advent of deep learning. Deep learning methods surpass manual feature extraction by auto-

*Corresponding author, E-mail address: jingjingzhang@nuaa.edu.cn.

How to cite this article: JI Zhongyuan, ZHANG Jingjing, LIU Zehao, et al. A CNN-based method for sparse SAR target classification with Grad-CAM interpretation[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025, 42(4): 525-540.

http://dx.doi.org/10.16356/j.1005-1120.2025.04.008

matically identifying target features. Schmidhuber[5] provided a comprehensive examination of deep learning methodologies within neural networks, offering insights into their architecture, training processes, and applications across various domains. Alex net[6-7] was developed by Krizhevsky et al. It won the ImageNet challenge with a top-5 error rate of 17.0%, triggering a series of advancements in computer vision tasks such as image classification. The success of deep learning in optical image classification has encouraged researchers to adapt these techniques to SAR image classification, which has led to the creation of specialized deep learning models for SAR images. In 2014, Chen et al.[8] proposed to replace backpropagation with a sparse autoencoder to conduct multi-class SAR image classification tasks. However, the scarcity of large-scale annotated SAR datasets hampers the progression of deep learning in SAR image classification. Consequently, enhancing the classification performance with a limited number of SAR samples has become an important goal in SAR target classification research in recent years.

Specifically, in 2016, Chen et al.[9] further proposed a fully convolutional network (termed A-ConvNets) that helped to avoid the overfitting issue and achieved a 99.13% accuracy rate using all the training samples. And Ding et al.[10] addressed issues related to target translation, speckle noise, and missing poses by proposing three distinct data augmentation techniques. In 2017, Lin et al.[11] introduced the convolutional highway unit (CHU) as a solution to the gradient vanishing problem observed in deep convolutional neural networks (CNNs). Jiang et al.[12] achieved classification accuracy of 96.32% on the MSTAR dataset in 2018, by decomposing each original SAR image into 36 Gabor feature maps at multiple scales and orientations using Gabor filters. These were then used as inputs to a deep CNN[12-13]. In 2019, Wang et al.[14] replaced the rectified linear unit (ReLU) with concatenated ReLU (CReLU) as an activation function to preserve the negative phase information and obtain dual feature maps from the previous layer. They demonstrated that even with only 20% of the MSTAR dataset, a classification accuracy of 88.17% could be achieved. Guo et al.[15] proposed a novel feature learning structure called the compact convolutional autoencoder (CCAE) in 2020, which utilized a dual-channel shared parameter structure to calculate the compactness loss between targets of the same class. They showed that the CCAE could achieve a accuracy of 98.59% by minimizing this compactness loss. Finally, in 2021, Fu et al.[16] introduced the first meta-learning framework for SAR, named MSAR, which learned a favorable initialization and an effective update strategy, thereby validating the applicability of meta-learning to few-shot SAR classification.

In the field of SAR target classification, images recovered through matched filtering (MF) have commonly been used and have achieved considerable success. Nonetheless, MF-based images can be significantly compromised by noise, sidelobes, and clutter. To mitigate these issues, sparse SAR imaging technology has been developed, which enhances the image quality by emphasizing the targets and suppressing background clutter. Sparse recovery algorithms, such as iterative soft thresholding (IST)[17-18] and orthogonal matching pursuit (OMP)[19-20], are typically employed for this enhancement. A novel sparse recovery algorithm, termed BiIST (Bidirectional iterative soft thresholding), was proposed in 2018, building upon the IST method[21-22]. BiIST enhances the target features while maintaining the statistical distribution of the image. Further advancements were made in 2021, when Bi et al.[23] demonstrated that the integration of sparse SAR images with popular object detection methods, such as YOLOv3 and Faster R-CNN, yielded higher accuracy compared to MF-based images. Continuing this trend, in 2022, Deng et al.[24] introduced a new sparse SAR target classification framework named the amplitude-phase CNN (AP-CNN). This framework exploited both the magnitude and phase information from sparse SAR images reconstructed using BiIST. The findings suggested that the combination of sparse SAR images with AP-CNN framework could lead to superior performance[24].

The Transformer is a deep neural network model that relies on a self-attention mechanism, it was proposed by Google in 2017[25]. It has emerged as one of the most sophisticated models to handle sequential data, especially suitable for tasks involving natural language processing (NLP). Compared with traditional sequence models, such as the recurrent neural network (RNN)[26] and long short-term memory network (LSTM)[27], the Transformer can handle longer data sequences and model the relationship between any two elements due to the self-attention mechanism. As exploration in the field has progressed, the Transformer has also been adapted for use in computer vision. As the Transformer was initially designed for the processing of natural language, it requires inputs in sequence form. In 2020, Dosovitskiy et al.[28] minimally modified the Transformer to create the vision Transformer (ViT), which dissected an image into patches, assigned these patches to different positions in a sequence, and encoded them. This innovation reduces the reliance on traditional convolution and pooling operations in deep learning networks. Taking a different approach, Liu et al.[29] introduced the swin Transformer, which segmented the input image into multiple levels, with each level being processed by its own Transformer model. The swin Transformer achieves greater efficiency by restricting the self-attention computations to non-overlapping local windows and enabling inter-window connections. This model has shown impressive performance in tasks such as image classification, object detection, and image segmentation.

The Transformer model, recognized for its superior sequence modeling capabilities, holds great potential in SAR target classification and recognition. It can globally model the features within SAR images, leading to enhanced classification accuracy and robustness. While CNNs have long been the foundation of SAR automatic target recognition (ATR), they experience challenges in the case of limited data, such as the inability to expand their width and depth without encountering a bottleneck in feature representation, leading to poor recognition performance. The Transformer and CNN models, as the focal points of current research in SAR target classification, are instrumental in enhancing the accuracy of few-shot SAR target classification[30-32]. To address the overfitting issue in SAR ATR with small sample sizes, Li et al.[33] introduced a non-subsampled Laplacian pyramid decomposition (NSLP)-based ViT model in 2021, named NSLP-ViT. After preprocessing SAR images with NSLP, the model leverages the preprocessed images for ViT network training, effectively mitigating overfitting issues[33]. In 2022, Li et al.[34] further developed a self-attention-based multi-aspect SAR recognition method. This approach eliminates from the sequential dependency of RNNs, which can result in information loss. By discovering correlations within the semantic information in images and employing convolutional autoencoders, this method enhances the noise immunity and reduces the dependency on extensive datasets[34]. Wang et al.[35] proposed a convolutional Transformer (ConvT) tailored to few-shot learning (FSL) in SAR ATR. ConvT constructs hierarchical feature representations and captures the global correlations of local features, thus maintaining the network's capacity for local feature extraction while also considering the global context[35]. Deng et al.[36] combined the ViT architecture with a contrastive learning framework, using a plethora of unlabeled samples for pre-training, followed by fine tuning with a smaller set of labeled data. This approach proved effective for SAR image classification tasks, even with limited labeled samples[36]. To address the scarcity of SAR data, Youk et al.[37] proposed a novel Transformer-based SAR target image translation network. This model was designed to learn a feature space mapping from synthetic to real SAR domains, facilitating the generation of targets at various azimuth angles[37]. Finally, to address the biased SAR data distribution and insufficient model representation in target recognition, Liu et al.[38] introduced a new view semantic Transformation network (VSTNet). This network synthesized samples to enrich the statistical distribution of the training data, demonstrating its robustness and effectiveness in experiments on the MSTAR dataset[39].

In this paper, we introduce a novel model for sparse SAR target classification that leverages transfer learning and ensemble learning. This framework, named TL-IL-CNN (Transfer learning-based interpretable lightweight-CNN), is further examined using gradient-weighted class activation mapping (Grad-CAM) to provide insights into the CNN's decision-making process across various target classes, thereby significantly improving the results' interpretation. To augment the quality of the MF-based SAR dataset, we initially employ the BiIST algorithm, which enhances the image quality and yields a refined sparse SAR dataset. Subsequently, these improved sparse SAR images undergo preprocessing to accommodate varying sizes, before being fed into the TL-IL-CNN. This network is responsible for feature extraction and delivering the final classification outcomes. In the final stage of analysis, we utilize a lightweight model in conjunction with Grad-CAM to elucidate the classification decisions for different sparse SAR target types. In extensive experiments conducted on the MSTAR dataset[38], the TL-IL-CNN framework demonstrates a remarkable classification accuracy of nearly 90% while utilizing only 20% of the training data in a standard operating condition (SOC) scenario. The framework also exhibits superior performance compared to conventional deep learning methods under extended operating conditions (EOCs). The results reveal that the classification framework prioritizes distinct regions of the images when discerning between various sparse SAR targets, offering a deeper understanding of how the framework processes and identifies these targets.

The main contributions of this work can be concluded as follows:

(1) By harnessing the intrinsic advantages of sparse SAR data coupled with transfer learning techniques, this approach significantly enhances the classification precision, thereby setting a new benchmark for future research in face to small sample situation.

(2) A multi-scale integrated migration model is proposed for sparse SAR target classification, and a voting mechanism is used to output test image classification results. This model can improve the classification accuracy and increase the number of correctly predicted target types under the condition of small samples.

(3) Grad-CAM not only enhances the interpretability of the model but also deepens the understanding of its behavior, allowing for a more nuanced analysis of the features that the model "sees" and "considers" important in the SAR image classification tasks.

The remaining sections of this paper are structured as follows. Section 1 briefly describes the principle of the BiIST-based sparse SAR image recovery algorithm. Then, the target classification model of the TL-IL-CNN and the principle of Grad-CAM are introduced in Section 2. Sections 3 and 4 describe the experiments and analysis based on the sparse images. Section 5 concludes our work.

1 Methodology

1.1 BiIST based SAR image enhanceme

Traditional MF techniques are widely used in SAR image formation, yet they often suffer from noise, sidelobes, and clutter interference, which can obscure target features and degrade classification performance. To address these limitations, sparse recovery algorithms have been developed to enhance image quality by promoting sparsity in the target scene. The MF-based[21-22] image serves as the input, and the sparse imaging model can be expressed as

$$\boldsymbol{X}_{\mathrm{MF}} = \boldsymbol{X} + \boldsymbol{N} \tag{1}$$

where $\boldsymbol{X}$ is the area of interest to be recovered and $\boldsymbol{N}$ the difference between $\boldsymbol{X}_{\mathrm{MF}}$ and $\boldsymbol{X}$. Then, $\boldsymbol{X}$ is reconstructed by considering the L_1-norm regularization problem

$$\hat{\boldsymbol{X}} = \min_{\boldsymbol{X}} \left\{ \| \boldsymbol{X}_{\mathrm{MF}} - \boldsymbol{X} \|_{\mathrm{F}}^2 + \beta \| \boldsymbol{X} \|_1 \right\} \tag{2}$$

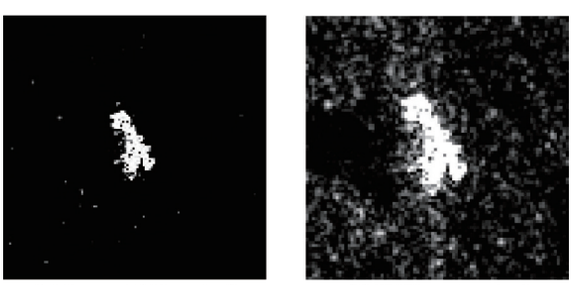
where F denotes the Frobenius norm and β the regularization parameter.

In comparison with MF-based results, sparse

SAR images have less clutter and noise and fewer sidelobes. Hence, the target contour is more obvious. In order to quantify the superiority of BiIST based SAR image reconstruction method on suppressing noise and clutters, we use target-to-background ratio (TBR) to evaluate the merits of sparse images, which is expressed as

$$\mathrm{TBR}(\boldsymbol{X}) \triangleq 20\lg\left(\frac{\max_{(p,q)\in\mathcal{T}}|(\boldsymbol{X})_{(p,q)}|}{(1/N_{\mathcal{B}})\sum_{(p,q\in\mathcal{B})}|(\boldsymbol{X})_{(p,q)}|}\right) \tag{3}$$

where (p,q) denotes the pixel point in the area of interest, $\mathcal{T}$ the target area, and $N_{\mathcal{B}}$ the number of pixels in the background area $\mathcal{B}$. The target area was manually delineated by thresholding and morphological operations to cover the target's physical extent, while the background area was defined as a 50-pixel wide buffer zone surrounding excluding the target region itself. The higher the TBR, the more prominent the target is and the better suppression of noise and clutter. The efficacy of the BiIST algorithm in enhancing the target-to-background ratio (TBR) for targets within the same category is demonstrated in Fig.1. The comparative data presented in Table 1 indicate a substantial increase in the TBR for sparse images post-reconstruction. Notably, the TBR for images reconstructed using the BiIST algorithm surpasses the TNR for those processed with the MF algorithm by over 25 dB. This significant improvement underscores the BiIST algorithm's potential in improving the quality of SAR imagery.



(a) Sparse SAR method (b) MF-based method

Fig.1 SAR image recovered by BiIST and MF based methods

Table 1 Comparison of TBR values of MF-based target and BiIST-based sparse SAR target dB

Target	Sparse method	MF method
1	58.046	30.806
2	59.556	28.627
3	58.281	29.284

1.2 Transfer learning and ensemble learning

The proposed SAR target classification framework is based on the transfer learning and ensemble learning techniques. Transfer learning is a technique that utilizes a pre-trained CNN as the feature extractor and fine tunes the network on a small dataset of SAR images. This process is aimed at learning the features that are specifically relevant to SAR target classification, as shown in Fig.2, where FC stands for full connected layer.

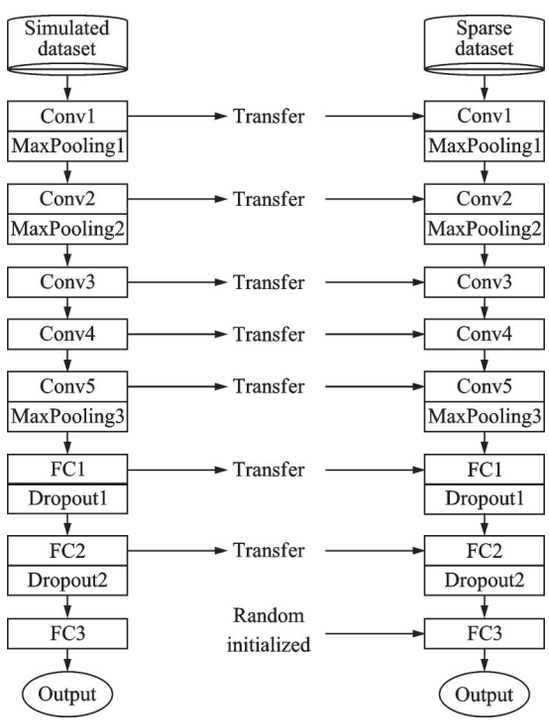


Fig.2 Principle of transfer learning

This approach takes advantage of the powerful features learned by the CNN from a large dataset of simulated SAR images, which comprises seven types of vehicles with three different distributions and includes 3 024 targets in each category. Initially, the simulated SAR dataset is randomly divided into two sets, with 70% for training and 30% for testing. The original learning rate is set at 0.01, and all parameters in the network are randomly initial-

ized to better adapt to the target domain.

Ensemble learning, on the other hand, combines the predictions of multiple CNNs that have been trained on the SAR dataset. This method can enhance the robustness and accuracy of the classification results by leveraging the advantages of different CNNs. Specifically, we employ three different CNN architectures as base models and train them independently. We then merge the predictions of these models using a weighted average approach to obtain the final classification results of the ensemble framework.It is important to note that when training the different CNNs, we use the same training data but with varying input sizes to accommodate targets of different scales.

As illustrated in Fig.3, the proposed TL-IL-CNN framework is structured into three main components. The first component involves preprocessing the input data into two different sizes, specifically 88 pixel×88 pixel and 128 pixel ×128 pixel. As previously mentioned, training on images of varying sizes contributes to the framework's robustness against the size variations encountered during inference. This is particularly beneficial in scenarios where the sizes of the input images are not standardized. Furthermore, by accommodating differently sized inputs, the TL-IL-CNN is able to recognize salient features at multiple scales, thereby enhancing its efficiency and accuracy. The second component is the feature extraction module, which consists of three branches. (1) The first branch is A-ConvNet[9], which is an established architecture. (2) The second branch is a deep network that we developed, featuring five convolutional layers. The first two convolutional layers are each followed by a ReLU activation and a max-pooling layer. Subsequently, there are two additional convolutional layers, and the final layer is the combination of a convolutional layer, ReLU activation, and a max-pooling layer, leading to two fully connected layers. To prevent overfitting, we integrate dropout layers after the first two fully connected layers in the deep network, referred to as CNN-Dropout. (3) The third branch is a lightweight network devised by ourselves, which is significantly shallower than the deep network, comprising three convolutional layers. Each layer is succeeded by a ReLU activation function and a max-pooling layer, culminating in a single fully connected layer. The feature extraction layers highlighted in red within Fig.3 are pretrained on a simulated SAR dataset. This pretraining step aims to accelerate the convergence and enhance the classification accuracy. The final component is the fusion module, which integrates the predictions from the various branches using a voting strategy to produce the ultimate classification outcome.

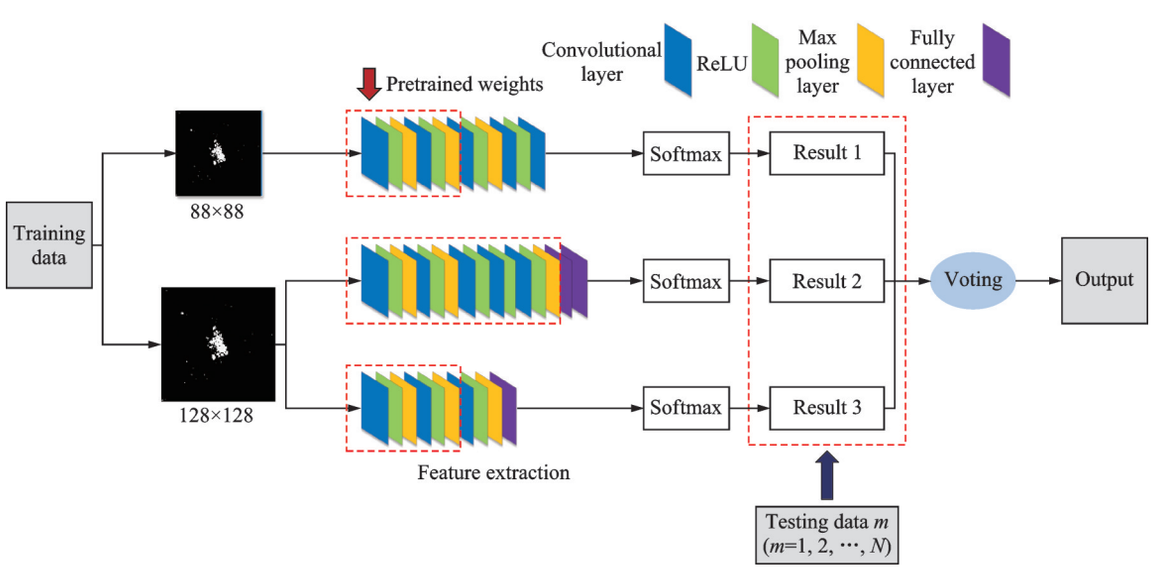


Fig.3 Architecture of TL-IL-CNN

2 Visualization of CNN

Grad-CAM is a technique used to visualize the regions of an input image that are significant for a neural network's prediction. Fig.4 illustrates the core principle: It involves using the gradients of the output class score with respect to the feature maps of the network's last convolutional layer to ascertain the importance of each feature map in making the prediction, where W represents the weight. Grad-CAM begins by forward propagating an input image through the network to generate a class activation map (CAM)[26], which is a 2D representation that emphasizes the image regions most important to the predicted class.

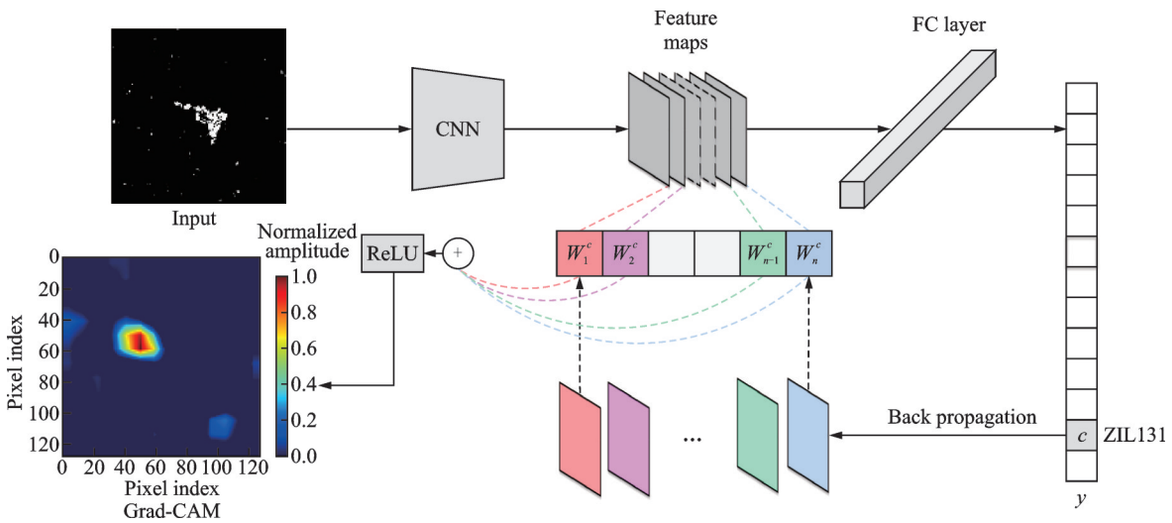


Fig.4 Flowchart of Grad-CAM

For enhanced detail in the localization map, Grad-CAM calculates the gradients of the class score for the output concerning the feature maps of the final convolutional layer. These gradients are then used to weight the feature maps, yielding a coarse localization map that accentuates the image regions most influential to the prediction. Formally, let $f_k(i,j)$ denote the activation of the kth feature map at the spatial location (i,j), and let y_c represent the predicted probability for class c. The Grad-CAM algorithm computes the gradient of y_c with respect to f_k as

$$\alpha_{kc} = 1/Z \cdot \sum_i \sum_j \partial y_c / \partial f_k(i,j) \tag{4}$$

where Z is a normalization factor and the summation is over all spatial locations of f_k. The gradient α_{kc} is then used to weight the feature maps, producing a localization map L_c for class c, which can be written as

$$L_c = \mathrm{ReLU}\left(\sum_k \alpha_{kc} \cdot f_k\right) \tag{5}$$

where ReLU is the rectified linear unit activation function. The localization map generated by Grad-CAM can be displayed as a heatmap superimposed on the input image, highlighting the areas that play a pivotal role in the network's prediction. This visualization technique is instrumental in understanding the decision-making process of deep neural networks and identifying the specific features that they rely on to make predictions.

One of the primary benefits of Grad-CAM is its universal applicability to any CNN, obviating the need for alterations to the existing network architecture. Moreover, Grad-CAM is versatile enough to illuminate not only the primary objects of interest but also the contextual elements, such as the background and adjacent objects, that the network takes into account when classifying an image.

3 Experimental Results

3.1 Dataset

This section introduces the reconstruction of a sparse SAR image dataset featuring military vehicle samples from the MSTAR project[40]. Fig.5 exhibits

the original SAR images for each category, alongside their corresponding optical images. The MSTAR dataset, which is publicly available, is composed of complex image data derived using the MF algorithm. This dataset includes a diverse array of SAR target samples at a resolution of 0.3 m×0.3 m, and it has been extensively employed in SAR target detection, recognition, and classification research. In this dataset, there are ten different classes of military vehicle targets.

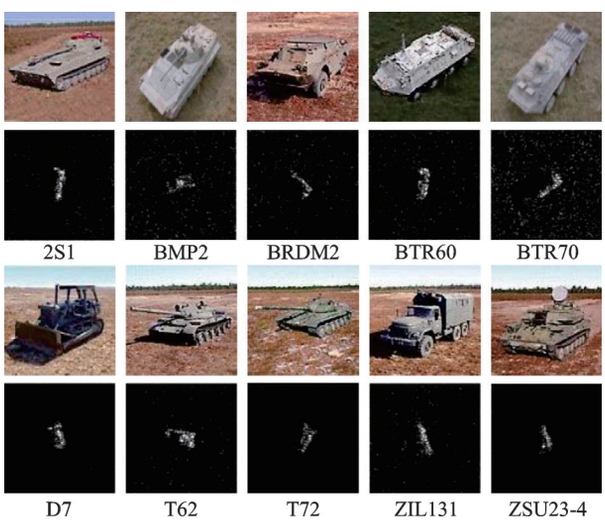


Fig.5 SAR image and corresponding optical image of ten classes of vehicles in MSTAR dataset

The dataset encompasses SAR image segments of stationary ground targets, spanning various categories and aspect angles from 0° to 360°. Within the MSTAR dataset, there are ten types of military vehicles, and Table 2 provides a detailed description of the data. The target segments in MSTAR are divided into SOC and EOC. For the purposes of this study, the BiIST algorithm has been applied to construct a sparse SAR dataset based on the MSTAR samples.

Table 2 Data description for SOC

Class	Serial No.	Training set (Depression 17°)	Testing set (Depression 15°)
2S1	B01	299	274
BMP2	SN9563	233	196
BRDM2	E-71	298	274
BTR60	Kloyt7532	256	195
BTR70	C71	233	196
D7	92v13015	299	274
T62	A51	299	273
T72	SN132	232	196
ZIL131	E12	299	274
ZSU23-4	D08	299	274
Total		2 747	2 426

The simulated dataset[41] was generated using electromagnetic scattering modeling based on physical optics and shooting-and-bouncing-ray techniques, incorporating three distinct ground background distributions (urban, woodland, desert) to simulate realistic SAR conditions. This dataset was selected for pretraining due to: (1) Geometric similarity of vehicle models to MSTAR targets, (2) controlled variation in depression angles matching MSTAR acquisition parameters, and (3) speckle noise characteristics emulating real SAR systems. We configure the learning rate and the number of epochs at 0.05 and 200, respectively. The pretraining phase yields accuracy exceeding 99% on the testing set, thereby establishing a robust foundation for subsequent experiments.

3.2 Evaluation of transfer learning on sparse SAR target classification

To illustrate the impact of transfer learning on sparse SAR target classification, we take the lightweight network branch as a case study. We quantitatively assess the effect of varying the number of pretrained layers, specifically transferring 3, 5, and 7 layers, with the reconstructed MSTAR dataset serving as the input. Fig.6 showcases the classification performance on a sparse SAR dataset with different volumes of training samples. The results indi-

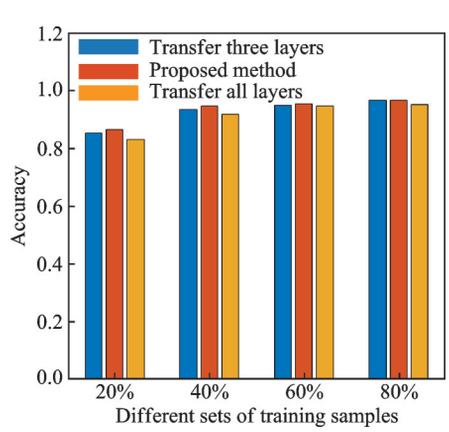


Fig.6 Comparative results based on transferring different numbers of pretrained layers

cate that, with a sufficient number of training samples, the various transfer learning approaches yield comparable classification accuracy. However, when the number of training samples is limited to 20%, the proposed transfer learning method demonstrates superior performance, outperforming the other two methods by 3.51% and 1.37%, respectively.

3.3 Experiments based on MF dataset

In this section, we evaluate the classification performance of the proposed TL-IL-CNN network using the original MSTAR dataset under SOC. For this purpose, samples at 17° are utilized as the training data, while samples at 15° are employed for validation.

A series of experiments is conducted with varying training sample sizes of specifically 20%, 40%, 60%, and 80%. As detailed in Tables 3 and 4, the TL-IL-CNN network achieves a classification accuracy of 85.04%. Notably, the classification performance for targets across seven categories shows an improvement, with the models for 2S1 and ZIL131 vehicles demonstrating a significant increase, 11 additional targets are correctly predicted for each class. On the ZIL131 class, in particular, the model

Table 3 Confusion matrix of single lightweight CNN with transfer learning via 20% training samples under SOC

Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU23-4	Average
2S1	201	5	4	7	9	0	27	0	21	0	
BMP2	2	127	3	21	3	2	5	28	0	5	
BRDM2	9	23	219	1	6	0	1	4	11	0	
BTR60	1	5	8	159	11	1	3	6	0	1	
BTR70	1	12	2	11	170	0	0	0	0	0	
D7	0	0	0	0	0	266	0	0	5	3	
T62	3	0	0	1	0	2	251	3	4	9	
T72	2	3	1	5	0	0	33	149	0	3	
ZIL131	0	0	0	0	0	2	7	0	261	4	
ZSU23-4	0	0	0	0	0	14	0	0	4	256	
Accuracy/%	73.3	64.80	79.93	81.54	86.73	97.08	91.94	76.02	95.26	93.43	84.87

Table 4 Confusion matrix of TL-IL-CNN ensemble via 20% training samples under SOC

Class	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU	Average
2S1	212 (↑11)	2	9	5	6	0	16	2	22	0	
BMP2	13	131 (↑14)	2	14	8	2	2	18	0	4	
BRDM2	6	26	214 (↓5)	2	0	1	1	22	0		
BTR60	1	2	17	161 (↑2)	6	1	0	1	3	3	
BTR70	8	23	4	6	153 (↓17)	0	0	0	2	0	
D7	0	0	0	0	0	268 (↑2)	0	0	3	3	
T62	1	0	0	0	0	2	244 (↓7)	4	17	5	
T72	13	2	1	3	1	0	22	147 (↓2)	4	3	
ZIL131	0	0	0	1	0	0	0	0	272 (↑11)	1	
ZSU	0	0	0	0	0	8	1	0	4	261 (↑5)	
Accuracy/%	77.37	66.84	78.10	82.56	78.06	97.81	89.38	75.00	99.27	95.26	85.04

achieves accuracy exceeding 99%, underscoring the potentiality of TL- IL-CNN for effective SAR target classification in scenarios with limited training data.

3.4 Comparison between TL-IL-CNN and Transformer

The advent of the Transformer architecture has reduced the dominance of CNN in image processing tasks, including image classification, segmentation, and detection. ViT has particularly excelled in image classification, offering a high-performance alternative that is also resource-efficient. However, the efficacy of the ViT for SAR image classification remains untested.

This section compares the performance of the ViT and the proposed TL-IL-CNN framework in the few-shot setting, i.e., using 20% of sparse SAR image data. According to the results compiled in Table 5, the ViT, despite its impressive capabilities in processing optical images, is outperformed by the TL-IL-CNN in the context of limited sample sizes. The average accuracy of the TL-IL-CNN surpasses that of the ViT by 8.45%. Furthermore, in nine out of ten categories, the TL-IL-CNN achieves higher accuracy than the ViT, demonstrating its suitability for small sample SAR image classification tasks.

Table 5 Classification accuracy of each category via 20% training samples under SOC %

Model	2S1	BMP2	BRDM2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU23-4	Average
ViT	83.45	77.67	85.43	80.87	72.73	76.29	83.47	83.51	78.84	89.51	80.75
TL-IL-CNN	89.42	86.22	80.66	86.15	78.57	98.18	88.28	88.27	95.26	95.99	89.20

3.5 Experiments based on TL-IL-CNN and sparse SAR dataset under SOC

In this set of experiments, we train the three models independently using subsets of the available data, specifically 20%, 40%, 60%, and 80%, to evaluate their performance in SAR image classification against conventional deep learning approaches. As depicted in Fig.7, the TL-IL-CNN model excels, achieving nearly 90% classification accuracy with only 20% of the training data. Its performance surpasses that of A-ConvNet[9] and AP-CNN[24] by 3.46% and 7.20%, respectively. Moreover, with abundant training data, the TL-IL-CNN consistently demonstrates higher or comparable accuracy, demonstrating the method's robustness and efficacy.

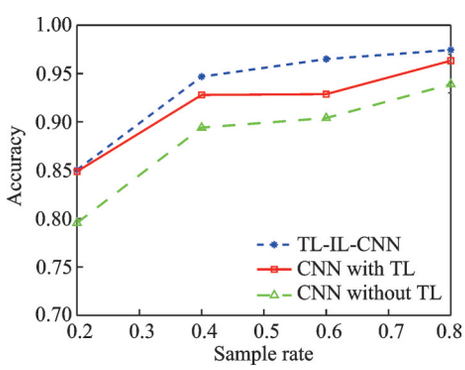


Fig.7 Experimental results of different SOTA (state-of-the-art) methods under SOC

To determine the upper bound of the TL-IL-CNN's performance, we undertake experiments that utilize the full dataset to train the model from scratch. The goal is to pinpoint the maximum accuracy that can be achieved. Fig.8 illustrates the outcomes of these tests. Although the proposed method does not surpass the model trained with the whole training set, it achieves slightly lower performance when using only 20% of the training data.

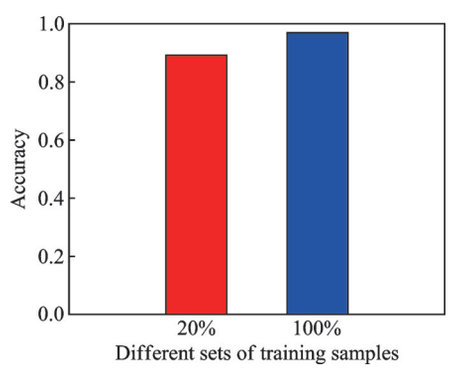


Fig.8 Classification accuracy under 20% samples with transfer learning and 100% samples with random initialization

3.6 Experiments based on TL-IL-CNN and sparse SAR dataset under EOC

Unlike the SOC, the training and testing sets exhibit considerable differences in depression angle in the EOC, as outlined in Table 6. Specifically, targets are captured at 17° for training and 30° for testing. Moreover, discrepancies in target labels present an additional challenge; for instance, the T72 is designated as SN132 in the training set and as A64 in the testing set, complicating the classification of different target categories.

Table 6 Data description for EOC

Class	Serial No.	Training set (Depression 17°)	Testing set (Depression 15°)
2S1	B01	299	288
BRDM2	E-71	298	287
T72	SN132/A64	299	288
ZSU23-4	D08	299	288
Total		1 195	1 151

In this challenging context, we randomly select 20% of the sparse SAR images under the EOC to train the TL-IL-CNN. Despite the pronounced disparity between the training and testing data, the results displayed in Table 7 are promising. The TL-IL-CNN secures overall accuracy of 78.54%, surpassing the performance of standalone lightweight models by margins of 1.74%, 4.95%, and 0.61%, respectively. Remarkably, for targets such as the ZSU23-4, the accuracy achieved is as high as 93.40%.

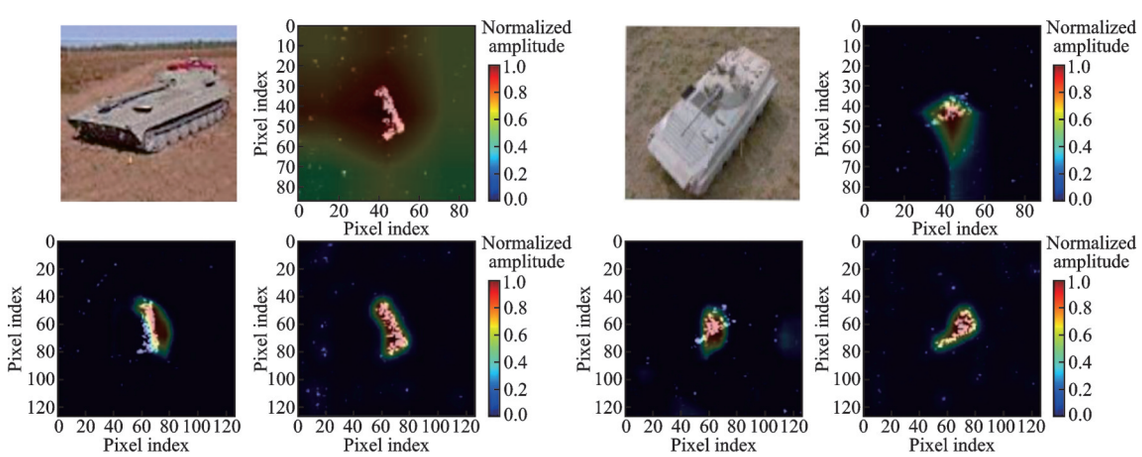
Table 7 Confusion matrix of TL-IL-CNN based on sparse SAR image dataset under EOC

Class	2S1	BRDM2	T72	ZSU23-4	Average
2S1	204	37	33	14	
BRDM2	27	228	3	29	
T72	12	21	203	52	
ZSU23-4	11	6	2	269	
Accuracy/%	70.83	79.44	70.49	93.40	78.54

Multiple experiments based on a sparse SAR dataset are conducted to demonstrate the performance of the proposed TL-IL-CNN. It can be found that, under both the SOC and EOC, the TL-IL-CNN performs better than the typical lightweight models, indicating that the proposed framework can improve the final classification accuracy under limited samples.

4 SAR Target Interpretation Based on Grad-CAM

In the experiments, we utilize Grad-CAM to interpret sparse SAR images, focusing on whether the models consider the target or the background. To ensure a robust analysis and prevent the data insufficiency from influencing the results, 80% of the available samples are used for training. Grad-CAM visualizations of the last convolutional layer activations are then generated for each target class. The visualizations in Fig.9 reveal that, across most categories, the models demonstrate strong recognition capabilities. Despite some models occasionally diverting their focus away from the actual targets, the



(a) 2S1 (b) BMP2

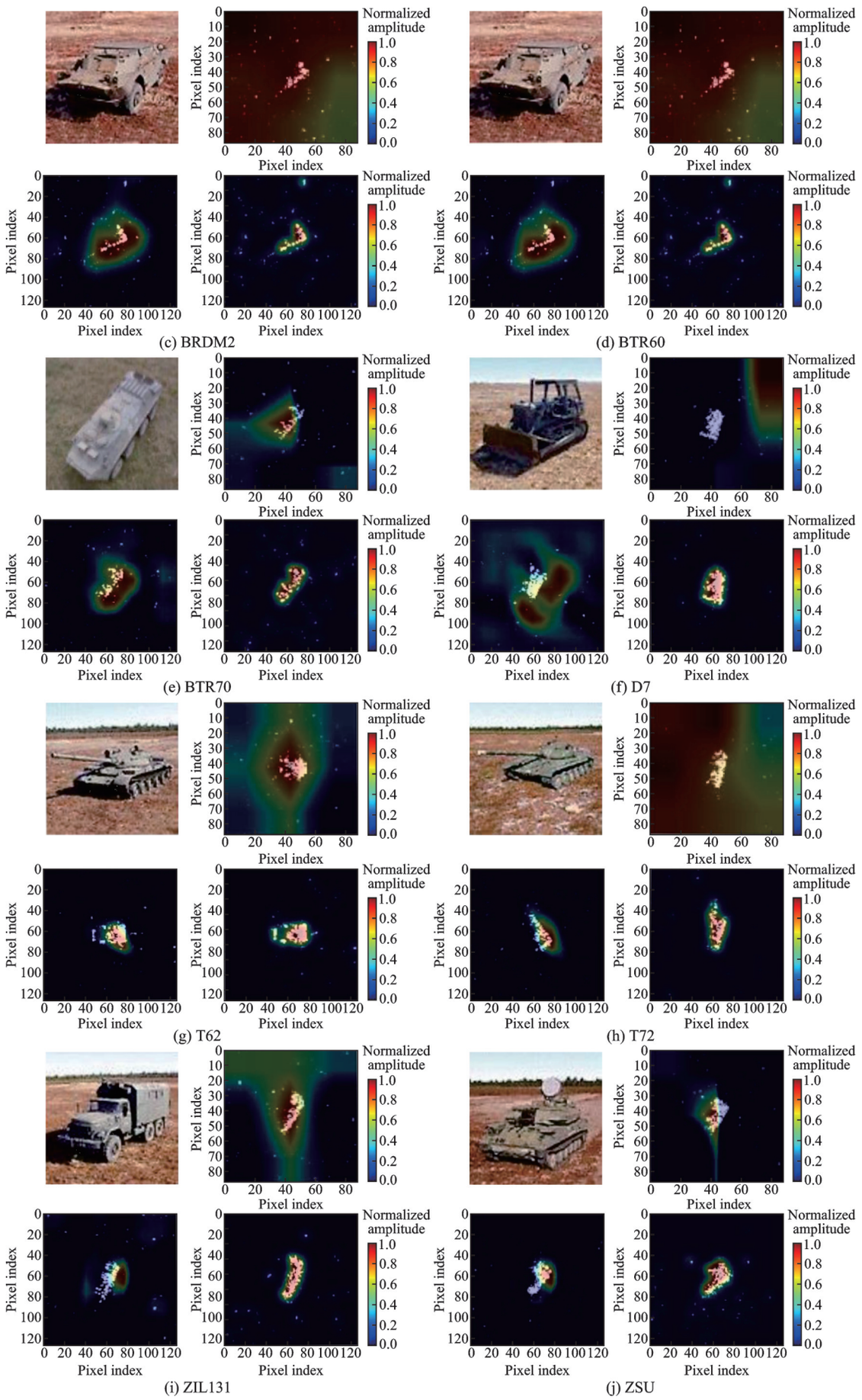


Fig.9 Grad-CAM of different kinds of sparse SAR target and corresponding optical images

collaborative strength of ensemble learning compensates for this. The ensemble enables other models to maintain their target focus, collectively contributing to the framework's impressive overall classification performance. For instance, as shown in Fig.9, the model predominantly focuses on the main body structure of BMP2 and T72 tanks (e.g., turret and hull) due to their distinctive metallic components that produce strong scattering signatures. Conversely, for BRDM2 amphibious vehicles and ZSU23-4 anti-aircraft systems, significant attention is given to the surrounding background context, likely because their lower-profile designs blend with terrain features in SAR imagery, making contextual cues like shadow patterns and ground texture critical for differentiation. This divergence in focus correlates with target characteristics: Heavily armored vehicles (BMP2/T72) exhibit dominant scattering centers, while low-profile targets (BRDM2/ZSU23-4) require environment-context integration.

To quantitatively support the observation of differential background focus, we introduce a target-to-background activation ratio (TBAR) metric defined as

$$\mathrm{TBAR}=\frac{\mu_{\mathrm{target}}}{\mu_{\mathrm{background}}} \tag{6}$$

where μ_{target} and $\mu_{\mathrm{background}}$ denote the mean activation intensity within the manually delineated target area and background area. Analysis of three representative classes reveals significant variation: BMP2 (TBAR=8.7±1.2), BRDM2 (TBAR=2.1±0.8), and ZSU23-4 (TBAR=1.9±0.7). This quantitatively confirms that heavily armored targets (BMP2) predominantly activate target regions while low-profile vehicles (BRDM2/ZSU23-4) exhibit comparable activation in background areas.

In future work, we plan to systematically analyze the learned features to understand the underlying mechanisms driving the model's decision making. Additionally, we aim to explore the synergy of Grad-CAM with other visualization methodologies, such as saliency maps and occlusion sensitivity maps, to construct a more holistic picture of the classification process. Our future endeavors will also include evaluating the transferability of the learned features across different SAR datasets and application scenarios.

5 Conclusions

We propose a novel framework for sparse SAR target classification that is based on transfer learning and ensemble learning, named TL-IL-CNN. First, the BiIST algorithm is used to reconstruct the MSTAR dataset, enhancing its performance. Subsequently, sparse SAR images of various sizes are used to form the training data set. The TL-IL-CNN framework is then introduced to classify targets and to output the final results for each testing image using a voting module. The experimental results demonstrate that the TL-IL-CNN outperforms a single lightweight model in both SOC and EOC. Notably, with only 20% of the training samples under the SOC, the TL-IL-CNN still achieves a classification accuracy of 89.20%, with some classes displaying accuracy of over 95%. Additionally, we employ Grad-CAM to visualize the CNN model and interpret the classification results. The experiments reveal that, through the use of ensemble learning, the framework compensates for individual models that occasionally focus on incorrect areas. This ensures that other models within the ensemble maintain their focus on the correct targets, resulting in excellent overall classification performance. In future research, we will continue to explore SAR target interpretation to deepen our understanding of how the framework learns and recognizes sparse SAR targets.

References

[1] CURLANDER J C, MCDONOUGH R N. Synthetic aperture radar: Systems and signal processing[M]. New York, USA: Wiley, 1991.

[2] HENDERSON F M, LEWIS A J. Principle and application of imaging radar[M]. New York, USA: John Wiley and Sons, 1998.

[3] DUGEON D E, LACOSS R T. An overview of automatic target recognition[J]. Lincoln Laboratory J, 1993, 6(1): 3-10.

[4] KREITHEN D E, HALVERSEN S D, OWIRKA G J. Discriminating targets from clutter[J]. Lincoln Laboratory Journal, 1993, 6(1): 25-51.

[5] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61: 85-117.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[8] CHEN S, WANG H. SAR target recognition based on deep learning[C]//Proceedings of 2014 International Conference on Data Science and Advanced Analytics (DSAA). Shanghai, China: IEEE, 2014: 541-547.

[9] CHEN S, WANG H, XU F, et al. Target classification using the deep convolutional networks for SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4806-4817.

[10] DING J, CHEN B, LIU H, et al. Convolutional neural network with data augmentation for SAR target recognition[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(3): 364-368.

[11] LIN Z, JI K, KANG M, et al. Deep convolutional highway unit network for SAR target classification with limited labeled training data[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(7): 1091-1095.

[12] JIANG T, CUI Z, ZHOU Z, et al. Data augmentation with Gabor filter in deep convolutional neural networks for SAR target recognition[C]//Proceedings of 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia, Spain: IEEE, 2018: 689-692.

[13] YANG J, PENG Z. SAR target recognition based on spectrum feature of optimal Gabor transform[C]//Proceedings of 2013 International Conference on Communications, Circuits and Systems (ICCCAS). Chengdu, China: IEEE, 2013: 230-234.

[14] WANG Z, XU X. Efficient deep convolutional neural networks using CReLU for ATR with limited SAR images[J]. The Journal of Engineering, 2019, 2019(21): 7615-7618.

[15] GUO J, WANG L, ZHU D, et al. Compact convolutional autoencoder for SAR target recognition[J]. IET Radar, Sonar & Navigation, 2020, 14(7): 967-972.

[16] FU K, ZHANG T, ZHANG Y, et al. Few-shot SAR target classification via metalearning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 2000314.

[17] DAUBECHIES I, DEFRISE M, DE MOL C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint[J]. Communications on Pure and Applied Mathematics, 2004, 57(11): 1413-1457.

[18] BI H, BI G. Performance analysis of iterative soft thresholding algorithm for L_1 regularization based sparse SAR imaging[C]//Proceedings of 2019 IEEE Radar Conference (RadarConf). Boston, MA, USA: IEEE, 2019: 1-6.

[19] PATI Y C, REZAIIFAR R, KRISHNAPRASAD P S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition[C]//Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers. Pacific Grove, CA, USA: IEEE, 1993: 40-44.

[20] DONOHO D L, TSAIG Y, DRORI I, et al. Sparse solution of underdetermined systems of linear equations by stage wise orthogonal matching pursuit[J]. IEEE Transactions on Information Theory, 2012, 58(2): 1094-1121.

[21] BI H, BI G, ZHANG B, et al. A novel iterative thresholding algorithm for complex image based sparse SAR imaging[C]//Proceedings of the 12th European Conference on Synthetic Aperture Radar. Aachen, Germany: VDE, 2018: 1-5.

[22] BI H, BI G. A novel iterative soft thresholding algorithm for L_1 regularization based SAR image enhancement[J]. Science China Information Sciences, 2019, 62(4): 49303.

[23] BI H, DENG J, YANG T, et al. CNN-based target detection and classification when sparse SAR image dataset is available[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 6815-6826.

[24] DENG J, BI H, ZHANG J, et al. Amplitude-phase CNN-based SAR target classification via complex-valued sparse image[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 5214-5221.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 14: 1-11.

[26] LIPTON Z, BERKOWITZ J, ELKAN C, et al. A critical review of recurrent neural networks for sequence learning[EB/OL]. (2015-05-29). https://arxiv.org/abs/1506.00019.

[27] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[28] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22). https://arxiv.org/abs/2010.11929.

[29] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.

[30] TAI Y, TAN Y, XIONG S, et al. Few-shot transfer learning for SAR image classification without extra SAR samples[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 2240-2253.

[31] ZHANG Y, LU D, QIU X, et al. Scattering-point topology for few-shot ship classification in SAR images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 10326-10343.

[32] LIU X, WU Y, LIANG W, et al. High resolution SAR image classification using global-local network structure based on vision transformer and CNN[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4505405.

[33] LI S, LANG P, FU X, et al. Automatic target recognition of SAR images based on transformer[C]//Proceedings of 2021 CIE International Conference on Radar (Radar). Haikou, Hainan, China: IEEE, 2021: 938-941.

[34] LI S, PAN Z, HU Y. Multi-aspect convolutional-transformer network for SAR automatic target recognition[J]. Remote Sensing, 2022, 14(16): 3924.

[35] WANG C, HUANG Y, LIU X, et al. Global in local: A convolutional transformer for SAR ATR FSL[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4509605.

[36] DENG J, ZHU Y, ZHANG S, et al. SAR image recognition using ViT network and contrastive learning framework with unlabeled samples[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 21: 4000205.

[37] YOUK G, KIM M. Transformer-based synthetic-to-measured SAR image translation via learning of representational features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5206318.

[38] LIU Z, WU F, WEN Z, et al. View-semantic transformer with enhancing diversity for sparse-view SAR target recognition[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5211610.

[39] HU H, CUI Z, ZHOU Z, et al. SAR-UT: A synthetic-to-measured SAR image translation network based on transformer[C]//Proceedings of 2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena, CA, USA: IEEE, 2023: 6243-6246.

[40] DIEMUNSCH J, WISSINGER J. Moving and stationary target acquisition and recognition model-based automatic target recognition: Search technology for a robust ATR[J]. Proceedings of SPIE, 1998, 3370: 481-492.

[41] MALMGREN-HANSEN D, KUSK A, DALL J, et al. Improving SAR automatic target recognition models with transfer learning from simulated data[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(9): 1484-1488.

Acknowledgements This work was supported in part by the National Natural Science Foundation (Nos.62271248, 62401256), in part by the Natural Science Foundation of Jiangsu Province (Nos.BK20230090, BK20241384), and in part by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of China (No.KLSMNR-K202303).

Authors

The first author Mr. **JI Zhongyuan** received his B.S. degree in criminal science and technology from Shandong Police College, Jinan, China, in 2010, and M.S. degree from People's Public Security University of China (PPSUC), Beijing, China, in 2014. From 2010 to 2017, he worked in Beijing Railway Public Security Bureau, China. Since 2017, he has been working in the Shandong University of Political Science and Law, China. He engaged in teaching and judicial expertise of document, trace and electronic data. Since 2022, he has been studying in Nanjing University of Aeronautics and Astronautics (NUAA) for a Ph.D. degree, majoring in electronic and information engineering. He majors in the image processing and object recognition and detection.

The corresponding author Dr. **ZHANG Jingjing** received his B.E. degree in electronic information engineering from the University of Science and Technology of China (USTC), Anhui, China, in 2009, and a Doctor of Engineering degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2017. From 2017 to 2021, he worked in School of Information Science and Technology, Fudan University, China. Since 2021, he has been with College of

Electronics and Information Engineering, Nanjing University of Aeronautics and Astronautics, China, as an associate professor. His research interests include the design, modeling and calibration of polarimetric SAR systems, SAR imaging, and polarimetric and polarimetric interferometric SAR signal processing and applications.

Author contributions Dr. **ZHANG Jingjing** conceived the research framework and theoretical foundations, while Mr. **JI Zhongyuan**, Mr. **LIU Zehao** and Mr. **LI Guoxu** designed methodologies, conducted analyses, and drafted the manuscript. Mr. **JI Zhongyuan** led revisions with technical refinements, supported by Dr. **ZHANG Jingjing**'s theoretical insights. Both authors jointly validated results, resolved discrepancies, and approved the final version. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

基于卷积神经网络的稀疏SAR目标梯度加权类激活映射分类方法

姬忠远[1,2,3], 张晶晶[1,3], 刘泽昊[4], 李国旭[5]

(1.南京航空航天大学电子信息工程学院,南京 211106,中国;2.山东政法学院刑事司法学院,济南 250014,中国;3.南京航空航天大学雷达成像与微波光子学教育部重点实验室,南京 211106,中国;4.南瑞科技股份有限公司,南京 211106,中国;5.中电科思仪科技股份有限公司,青岛 266555,中国)

摘要:近年来,深度学习在合成孔径雷达(Synthetic aperture radar, SAR)图像处理中得到了广泛的应用。然而,大规模标记SAR图像的采集具有挑战性和成本高的特点,当SAR图像有限时,分类精度往往很差。为了解决这个问题,本文提出了一种新的稀疏SAR目标分类框架,称为基于转移学习的可解释轻量化卷积神经网络(Transfer learning-based interpretable lightweight convolutional neural network, TL-IL-CNN)。此外,本文采用增强梯度加权类激活映射(Gradient-weighted class activation mapping, Grad-CAM)来缓解深度学习模型中的"黑箱"效应,并探索CNN对各种稀疏SAR目标进行分类的机制。首先,采用一种新的双向迭代软阈值(Bidirectional iterative soft thresholding, BiIST)算法来生成比传统匹配滤波(Matched filtering, MF)方法更高质量的稀疏图像。然后,在模拟的SAR图像数据集上预训练多个浅层CNN。利用稀疏SAR数据集作为CNN的输入,评估了迁移学习在稀疏SAR目标分类中的有效性,并提出了TL-IL-CNN的融合方法,以进一步提高分类精度。在MSTAR数据集上的实验结果表明,所提出的TL-IL-CNN在标准操作条件(Standard operating conditions, SOC)下仅需20%的训练数据就可以达到近90%的分类准确率,在小样本情况下超过了典型的深度学习方法,如ViT (Vision Transformer)。它甚至在扩展操作条件(Extended operating conditions, EOC)下表现出更好的性能。此外,Grad-CAM的应用阐明了CNN在各种稀疏SAR目标之间的分化过程。实验结果表明,该模型以目标为中心,不同目标类别的背景会有所不同。本研究有助于加深对此类结果可解释性的理解,以能够更准确地推断每个类别的分类结果。

关键词:稀疏合成孔径雷达;卷积神经网络;集成学习;目标分类;SAR解译