GLOBAL-LOCAL INFORMATION INTERACTIVE LEARNING NETWORK FOR SAR TARGET RECOGNITION WITH LIMITED SAMPLES

Lei Miao, Haohao Ren, Yue Li, Lin Zou, Xuegang Wang*

University of Electronic Science and Technology of China
School of Information and Communication Engineering
Chengdu, China 611731

ABSTRACT

Deep learning-based synthetic aperture radar (SAR) automatic target recognition (ATR) have shown great potential recently, however, the performance of these methods is subject to the number of annotated samples. In real application scenarios, it tends to acquire limited number of samples due to the acquisition cost, in which case the exising ATR method is susceptible to over-fitting. To achieve SAR target recognition robustly in the case of limited samples, this paper proposes a global-local information interactive learning network. Specifically, we first develop a global-local interactive learning architecture, which is dedicated to extract global-local discriminative feature by interactively integrating the merits of local convolution and sparse self-attention. Then, a hierarchical feature discrimination module is proposed to improve intra-class compactness and inter-class divergence, thereby boosting the recognition performance of the ATR model. Evaluation experiments on the moving and stationary target acquisition and recognition (MSTAR) dataset illustrate that the proposed method is superior to advanced SAR ATR methods under the condition of limited samples.

***Index Terms*—** Synthetic aperture radar (SAR), Automatic target recognition (ATR), Limited samples.

1. INTRODUCTION

Synthetic aperture radar (SAR) is capable of imaging around the clock regardless of weather conditions and object obscuration. Therefore, it plays an important role in military and civilian applications. To effectively interpret the rich information of SAR images, SAR automatic target recognition (ATR) has become a crucial research area in recent years.

As an efficient and powerful machine learning method without manual design, deep learning techniques have been successfully applied to SAR ATR. For example, Chen *et al.* introduced a full convolutional neural network (CNN) for SAR ATR [1]. Pei *et al.* proposed a parallel processing bar depth network model for joint recognition by utilizing SAR images from multiple viewpoints [2]. In [3], a multi-scale convolutional capsule network was designed to achieve robust target recognition in complex SAR application scenarios.

However, existing deep learning-based SAR ATR methods rely largely on the availability of sufficient training samples. Unfortunately, only a limited number of labeled SAR samples are acquired in most real SAR applications due to the acquisition cost. To address this problem, Zheng *et al.* [4] devised a dynamic multi-discriminator architecture based on generative adversarial network (GAN) to generate labeled images. Wang *et al.* introduced a self-consistent augmentation rule to semi-supervised framework, which can fully utilize unlabeled data for network training [5]. Lang *et al.* proposed a multi-domain feature subspace fusion method, which mitigates the overfitting problem caused by the limited training samples [6].

In real-world scenarios, the lack of sufficient training samples presents a significant challenge to extract discriminative target feature from limited SAR images. In order to achieve robust SAR target recognition with a few labeled samples, we propose a global-local information interactive learning network. To be specific, a global-local interactive learning architecture combing local convolution and sparse self-attention is first designed to comprehensively extract global-local discrminative features. Then, we develop a hierarchical feature discrimination module to learn a feature space with both intra-class compactness and inter-class divergence, which can further improve the recognition performance of the proposed network The experimental results on MSTAR dataset demonstrate that our proposed method outperforms exsiting SAR ATR approaches.

2. PROPOSED METHOD

The proposed method is composed of global-local interactive learning architecture are hierarchical feature discrimination module, as depicted in Fig. 1. First, global-local feature is extracted by global-local interactive learning architecture integrating local convolution and sparse self-attention. Af-

Corresponding author: Haohao Ren (email: haohao_ren@uestc.edu.cn).
Thanks to the National Natural Science Foundation of China under Grant 62201124 and 42027805 for funding.

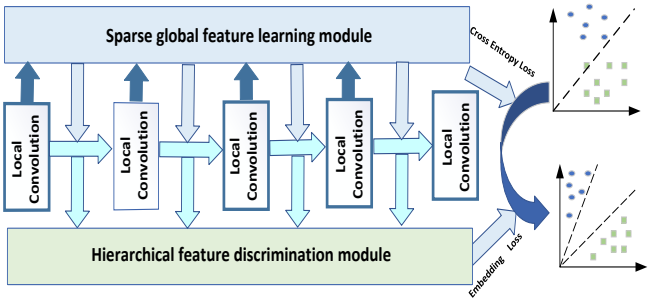

Fig. 1. Overall framework of the proposed approach.

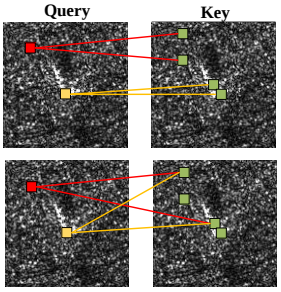

Fig. 2. Sparse and Dense global information.

terwards, the hierarchical feature discrimincation module is leveraged to strength the intra-class compactness and inter-class divergence at each feature layer, so as to promote the discrimination of the ATR model. Finally, a mixed loss combining embedding loss and classification loss is exploited to optimize the whole model. Hereinafter, the details of the proposed method are elaborated.

2.1. Global-local information interactive learning

How to mine rich and non-redundant discrimination information from limited samples is one of the feasible solutions to improve SAR target recognition performance under limited sample conditions. With this idea in mind, this paper proposes a global-local interactive learning architecture to extract rich global-local features, as shown in Fig. 1.

The proposed global-local information interactive learning architecture is capable of extracting rich global-local features in a interactive learning manner. Among them, multiple convolution operations are leveraged to extract local discrimination information, as plotted in Fig. 1. In view of the limited training samples, on the basis of self-attention mechanism [7, 8], a global feature extraction operation based on sparse self-attention mechanism is developed to mine global discrimination information, whose structure is depicted in Fig. 3. In contrast to the conventional global attention that pays excessive attention to global information from other unrelated locations, as shown in Fig.2, the proposed sparse global feature extraction method resorts to the discrete cosine transform-based key query location search strategy to realize sparse global feature relations by selectively weighting the location.

Let $\mathbf{F_1} \in \mathbb{R}^{H\times W\times C}$ be the local feature map extracted from convolution operation. To obtain global discrimination features, a convolution operation with kernel size of 1×1 is first performed along different dimensions, the output features are denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. Then, $\mathbf{Q}$ is divided into m groups along the channel dimension, denoted as $[\mathbf{Q}^0, \mathbf{Q}^1, \ldots, \mathbf{Q}^{m-1}]$, where $\mathbf{Q}^i \in \mathbb{R}^{H\times W\times C'}$ and $C' = \frac{C}{m}$. Two-dimensional discrete cosine transform (2D-DCT) is commonly used for image compression. Therefore, it can serve as an adaptive weighted preprocessing operation. In this paper, multiple frequency components of 2D-DCT are utilized to compress features. To adaptively establish sparse global relationships from other related locations, discrete 2D cosine transform is first performed at different $\mathbf{Q}^i$. Mathematically,

$$\begin{aligned} Freq^i &= \text{2DDCT}^{u_i,v_i}\left(\mathbf{Q}^i\right) \\ &= \sum_{\boldsymbol{h}=0}^{\boldsymbol{H}-1}\sum_{\boldsymbol{w}=0}^{\boldsymbol{W}-1} X_{h,w}^i \mathbf{B}_{h,w}^{\boldsymbol{u}_i,\boldsymbol{v}_i} \\ s.t. i &\in \{0,1,\cdots,n-1\} \end{aligned} \tag{1}$$

where H and W are the height and width of the input X, $\mathbf{B}_{h,w}^{i,j}$ is defined as

$$\begin{aligned} \mathbf{B}_{h,w}^{\boldsymbol{u}_i,\boldsymbol{v}_i} &= \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \\ s.t. h &\in \{0,1,\cdots,H-1\}, w \in \{0,1,\cdots,W-1\} \end{aligned} \tag{2}$$

where $[u_i, v_i]$ refers to the frequency component 2D indices corresponding to $\mathbf{Q}^i$, $Freq^i \in \mathbb{R}^{C'}$.

Next, all $Freq^i$ (i=1,$\cdots$,m) are concatenated in serial way, i.e.,

$$Freq = Concat([Freq^0, Freq^1, \cdots, Freq^{m-1}]) \tag{3}$$

where $Concat(\cdot)$ represents concatenation operation.

Then, a set of adaptive weighting on $\mathbf{Q}$ can be obtained according to the following formula:

$$\mathbf{A} = \sigma(fc(Freq)) \tag{4}$$

where fc represents full connection operation layer, σ is the nonlinear function $sigmoid(\cdot)$.

Finally, on the basis of self-attention, sparse global feature can be obtained according to the following formula:

$$\mathbf{F_g} = softmax(\frac{(\mathbf{A}\odot\mathbf{Q})\mathbf{K}^T}{\sqrt{d_k}})V \odot \mathbf{F_1} \tag{5}$$

where d_k is the dimension of $\mathbf{Q}$. The architecture of sparse global feature extraction is depicted in Fig. 3.

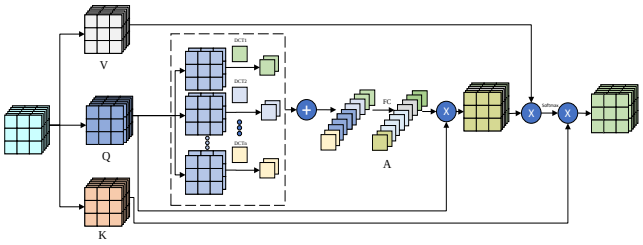

Fig. 3. Sparse self-attention mechanism.

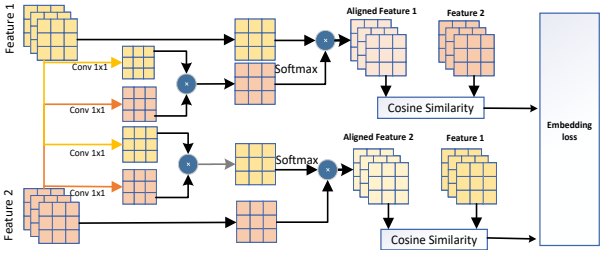

Fig. 4. Feature identification module.

2.2. Hierarchical feature discrimination module

To enhance the discrimination of global-local features, a hierarchical feature discrimination module is developed. Concretely, inspired by contrastive learning, the features of the same targets are pulled closer while the features of different targets are pushed away at each feature layer. Given two features $\mathbf{F}_i$, $\mathbf{F}_j$ extracted from the same or different category of targets, $\{\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i\}$, and $\{\mathbf{q}_j, \mathbf{k}_j, \mathbf{v}_j\}$ are first obtained through the convolution operation with the kernel size of 1×1, respectively. Then, the alignment operation between different features is performed, aiming to focus on the same position for two feature maps. Concretely, Aligning $\mathbf{F}_i$ with $\mathbf{F}_j$, the attention graph of $\mathbf{F}_j$ is first obtained by multiplying the key $\mathbf{k}_i$ and the query $\mathbf{q}_j$, which is then applied to $\mathbf{v}_i$ to finally obtain $\mathbf{s}_{i|j}$. The computation is as follows.

$$\mathbf{s}_{i|j} = \mathbf{v}_i * \text{softmax}\left(\frac{\mathbf{q}_j \mathbf{k}_i^\top}{\sqrt{d}}\right). \tag{6}$$

Similarly, $\mathbf{s}_{j|i}$ can be obtained via $\mathbf{k}_j$, $\mathbf{q}_i$ and $\mathbf{v}_j$, This in turn yields a feature similarity score:

$$\text{sim}(\mathbf{F}_i, \mathbf{F}_j) = \left[(||\mathbf{v}_i||_2)^\top ||\mathbf{s}_{j|i}||_2 + (||\mathbf{v}_j||_2)^\top ||\mathbf{s}_{i|j}||_2\right] \tag{7}$$

where $||*||_2$ is the ℓ_2 normalization operation.

To enhance intra-class compactness and inter-class divergence, according to supervised contrastive loss, the following embedding loss in the hierarchical feature discrimination module is defined to optimize the global-local information interactive learning network, i.e.,

$$\begin{aligned} L_{\text{f}} &= \sum_{i=1}^{2N} \frac{1}{2N_{y_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i\neq j} \cdot \mathbb{1}_{y_i = y_j} \cdot \ell_{ij} \\ \ell_{ij} &= -\log \frac{\exp(sim(\mathbf{F}_i, \mathbf{F}_j))/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i\neq k} \cdot \exp(sim(\mathbf{F}_i, \mathbf{F}_k))/\tau)} \end{aligned} \tag{8}$$

with τ as a scalar temperature parameter, $\mathbb{1}_{condition} \in \{0, 1\}$ as an indicator function whose value is 1 if the condition is satisfied. N_{y_i} as the total number of samples with the same label y_i, The network structure of the proposed hierarchical feature discrimination module is drawn in Fig. 4.

2.3. Network optimization

The cross-entropy loss is adopted for classification, which is defined as cross-entropy loss formula

$$L_c = -\sum_{i=1}^{C}\sum_{j=1}^{N} Y_{x_j,i} \log\left(p\left(y = i\,|x_j\right)\right) \tag{9}$$

where N is the number of samples for i-th category of target, $Y_{x_j,i} = 1$ if the samples x_j belongs to i-th category, otherwise, $Y_{x_j,i} = 0$, $p(y = i|x_j)$ denotes the predicted probability that the classifier gives the sample x_j belonging to category i.

Combing Eq. (8) and Eq. (9), the mixed loss can be expressed as follows:

$$L = L_c + \lambda L_f \tag{10}$$

where λ is a weight to balance the importance of two losses.

3. EXPERIMENTAL RESULTS

To assess the effectiveness of the proposed method, this section conducts evaluation experiments on the moving and stationary target acquisition and recognition (MSTAR) dataset. The publicly released MSTAR dataset includes ten categories of ground military targets. SAR images are acquired under various depression angles covering full aspects from 0° to 360°, which was collected by Sandia National Laboratories using a spotlight SAR sensor with 10 GHz X-band. Referring to previous works [4], SAR images captured at 17° depression angle are used as training dataset, while those collected at 15° depression angle are selected for evaluation. In the following experiments, we select $K \in \{20, 40, 80\}$ samples from each class. To avoid experimental contingency, each experiment is repeated for 10 runs to calculate the average recognition rate.

In the following experiments, the hyper-parameter λ in Eq. 10 is set to 0.5. This paper adopts Adam optimizer to optimize the proposed method, and the learning rate is set to 0.01. To illustrate the superiority of the proposed method, six widely used SAR target recognition method, including CNN[4], GAN-CNN[4], MGAN-CNN [4], DNN [9], SSL[5], HDLM[10] are employed as competitors.

3.1. Recognition performance comparisons

To demonstrate the superiority of the proposed method with limited samples, multiple experiments are conducted under different limited sample conditions. Table 1 lists the average recognition rate of each method under different experimental conditions. One can see that the proposed method consistently outperforms all competitors in SAR ATR task with limited samples, and it is worth noting that the proposed method does not use any unlabelled data compared to other methods. When the number of training samples is reduced from 80 to 20, the performance of the proposed method still maintains above 96%, indicating that the proposed method is less affected by the reduction of training samples.

Table 1. Comparison of Performance (%) Under MSTAR

Class	Sample number for each class			
	All	80	40	20
CNN	97.03	93.88	88.35	83.80
DNN	96.50	93.76	87.73	79.39
GAN-CNN	97.53	94.91	90.13	84.39
MGAN-CNN	97.81	94.91	90.82	85.23
SSL	-	98.65	97.11	92.62
HDLM	-	98.80	97.85	95.17
Ours	**-**	**99.34**	**98.23**	**96.00**

3.2. Ablation studies

In this section, we perform ablation experiments to verify the effectiveness of each module of the proposed method, i.e., sparse self-attention (SS) and hierarchical feature discrimination (HFD). 25 samples are randomly selected from each category of target to compose the training set. Experimental results of the ablation study are shown in Table 2. The proposed modules have all contributed to improving recognition performance, especially when these modules work together, achieving the highest recognition accuracy of 97.36%.

Table 2. Ablation experiments

SS	HFD	Accuracy(%)
✗	✗	88.87
✓	✗	96.28
✗	✓	92.58
✓	✓	97.36

4. CONCLUSION

This paper proposes a global-local information interactive learning network to solve the problem of SAR ATR with limited samples. On one hand, the proposed method is capable of extracting rich features by interactively integrating local convolution and sparse global self-attention. On the other hand, the hierarchical feature discrimination module is developed to enhance intra-class compactness and inter-class divergence in feature embedding space, thereby boosting SAR target recognition accuracy with limited samples. Experiments on MSTAR show that the proposed method outperforms advanced SAR ATR methods with limited samples.

5. REFERENCES

[1] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 4806–4817, 2016.

[2] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T.-S. Yeo, "Sar automatic target recognition based on multiview deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 2196–2210, 2018.

[3] H. Ren, X. Yu, L. Zou, Y. Zhou, X. Wang, and L. Bruzzone, "Extended convolutional capsule network with application on sar automatic target recognition," *Signal Processing*, vol. 183, 2021.

[4] C. Zheng, X. Jiang, and X. Liu, "Semi-supervised sar atr via multi-discriminator generative adversarial network," *IEEE Sensors Journal*, vol. 19, pp. 7525–7533, 2019.

[5] C. Wang, J. Shi, Y. Zhou, X. Yang, Z. Zhou, S. Wei, and X. Zhang, "Semisupervised learning-based sar atr via self-consistent augmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 4862–4873, 2021.

[6] P. Lang, X. Fu, C. Feng, J. Dong, R. Qin, and M. Martorella, "Lw-cmdanet: A novel attention network for sar automatic target recognition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[8] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global context networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 6881–6895, 2023.

[9] D. A. E. Morgan, "Deep convolutional neural networks for ATR from SAR imagery," in *Algorithms for Synthetic Aperture Radar Imagery XXII*, vol. 9475, 2015, p. 94750F.

[10] C. Wang, J. Pei, J. Yang, X. Liu, Y. Huang, and D. Mao, "Recognition in label and discrimination in feature: A hierarchically designed lightweight method for limited data in sar atr," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.