Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.039846

ARTICLE

RLAT: Lightweight Transformer for High-Resolution Range Profile Sequence Recognition

Xiaodan Wang*, Peng Wang, Yafei Song, Qian Xiang and Jingtai Li

College of Air and Missile Defense, Air Force Engineering University, Xi'an, 710051, China

*Corresponding Author: Xiaodan Wang. Email: afeu_wang@163.com

Received: 20 February 2023 Accepted: 20 April 2023 Published: 26 January 2024

ABSTRACT

High-resolution range profile (HRRP) automatic recognition has been widely applied to military and civilian domains. Present HRRP recognition methods have difficulty extracting deep and global information about the HRRP sequence, which performs poorly in real scenes due to the ambient noise, variant targets, and limited data. Moreover, most existing methods improve the recognition performance by stacking a large number of modules, but ignore the lightweight of methods, resulting in over-parameterization and complex computational effort, which will be challenging to meet the deployment and application on edge devices. To tackle the above problems, this paper proposes an HRRP sequence recognition method based on a lightweight Transformer named RLAT, which consists of rotary position encoding, local-aggregated attention unit (LAU), and lightweight feedforward neural network (LW-FFN). Rotary position encoding is utilized to embed the relative position information for the HRRP sequence. Local aggregation attention unit can effectively aggregate and extract local features by local group linear transformation, and then the self-attention mechanism is adopted for perception and enhancement of global information. Thereby, the enhanced features are extracted by lightweight FFN. In addition, this paper adopts Label Smoothing regularization to add noise to the sample labels, which can improve the generalization performance of the method. Finally, the effectiveness of the proposed method in real scenes is verified based on the MSTAR dataset, a real-world dataset for radar target recognition. Experimental results show that the proposed method achieves superior recognition performance compared to other remarkable methods and achieves significant generalization performance and robustness under variant sample and limited sample conditions. RLAT achieved an accuracy of 99.86% on the MSTAR standard dataset and 99.73% on the MSTAR variant dataset. In particular, it achieves an accuracy of 95.83% with only 274 training samples. Furthermore, the proposed method is more lightweight, with 90.90% reduction in the number of parameters and 96.70% reduction in the computation compared to the Vanilla Transformer, which facilitates deployment in edge devices.

KEYWORDS

Transformer; lightweight model; HRRP sequence recognition; MSTAR; limited data

1 Introduction

Radar Automatic Target Recognition (RATR) has been widely applied to military and civilian domains. Currently, radar high-resolution range profiles (HRRP) are commonly used in RATR due to the advantages of easy acquisition, convenient processing, and small storage space [1]. HRRP is the vector sum of target echoes along the radar line of sight direction, containing rich information on target structure characteristics and scattering point distribution, which has significant applications in the target recognition domain. Therefore, HRRP has been widely used in recognizing aircraft [2,3], ships [4], ballistic missiles [5,6], and military vehicles [7,8].

When the radar detects a moving target, it will move relative to the target to obtain the echo information at several azimuth angles, then HRRPs of consecutive azimuth angles constitute the HRRP sequence [9]. The dynamic temporal features of the target can be extracted effectively with the correlation in an HRRP sequence being modeled for efficient target recognition. Since the dimension of the HRRP sequence is large and there is a large number of noisy regions hidden in HRRP, HRRP sequence recognition is a peculiar class of multivariate time series classification problem. However, existing methods for the HRRP recognition method have limited perception ability of global information and weak representation ability of target information [7], which leads to susceptibility to the noise region and poor recognition performance in real scenes. Therefore, feature enhancement and feature extraction of global information are pivotal issues to improve the recognition performance of HRRP sequences further.

Traditional HRRP sequence recognition methods rely on the manual extraction of features with high discriminability. For example, Timothy et al. [10] proposed a recognition method based on Hidden Markov Model (HMM), which extracted six power spectrum features from high-resolution (HRR) radar signal amplitude *vs.* target distance profiles using HMM. Du et al. [11] proposed a recognition method based on a double-distribution composite statistical model based on the dominant scattering in the range cell of the scattering center model. The range units are divided into three statistical types based on the number of dominant scattering points in the scattering center model's range units. The echoes of different types of range units are modeled as corresponding distribution forms to accomplish the recognition task. Molchanov et al. [12] proposed a recognition method based on micro-Doppler bicoherence features, which extracts the cepstrum coefficients from the micro-Doppler contributions in radar echoes, then calculates the classification features using bicoherence estimation. However, the manually extracted features are susceptible to the influence of subjective factors and have limited recognition performance because of the weak extraction of representative features.

To overcome the limitations of manual feature extraction, machine learning is introduced into HRRP recognition. Lei et al. [13] proposed a Support Vector Machine (SVM) based recognition method, which defines different classifier confidence levels based on the distance between classifiers given by the confusion matrix, and then integrates the support vector machine values and posterior probabilities into the basic probability assignment to achieve a support vector machine and evidence theory combined with the recognition method. Wang et al. [5] proposed an extreme learning autoencoder (1D ELM-LRF-AE) network based on one-dimensional local perceptual domains for meaningful representation learning of HRRP local structures to achieve efficient representation learning and recognition. The above method overcomes the negative effects of subjective factors of researchers and achieves more effective automatic feature extraction but ignores the correlation and temporal information among HRRPs, which causes significant information loss. Besides, the machine learning method is weak in extracting deep features. Thus, the recognition performance still needs to be improved.

Along with the development of deep learning, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are massively applied to HRRP recognition. For example, Xiang et al. [14] proposed a recognition method based on one-dimensional CNN (1D-CNN), which extracts the effective target structure information in HRRP by 1D-CNN and introduces aggregation-perception-recalibration for feature enhancement. Though CNN can effectively extract the local correlation of HRRP sequences, it ignores the temporal information between HRRP sequences. There are limitations to global feature extraction of long sequences since the size of the convolutional kernel limits CNN. In particular, the HRRP sequences in real scenes contain much noisy information, and the local information will harm the generalization of the model due to the influence of noise. Du et al. [15] proposed a recognition method based on a Region-factorized recurrent attentional network with deep clustering, which utilizes the time dependence of recurrent neural network (RNN) in HRRP samples. The clustering mechanism is used to find information regions automatically, weighting the different recognition contributions of the hidden states at each time step. However, RNNs lose important target features when extracting long-range information for long sequences due to the memory loss problem. The essential original information may be lost when the network is stacked deeply. To further enhance the extraction of global information, Pan et al. [16] proposed a recognition method based on CNN-Bi-RNN with Attention Mechanism, which uses convolutional neural networks to obtain a richer embedding representation, and then uses RNN based on Attention Mechanism to extract temporal information, which can use local and global temporal features more effectively, and still maintain high recognition performance for limited samples. With the emergence of the Transformer framework, the long-range information of sequences is modeled by the self-attention mechanism, which adaptively assigns different weights to sequences by calculating the correlation between sequences, paying more attention to the important information of the target region and effectively extracting the global information of sequences. Zhang et al. [17] proposed a recognition method based on a feature-guided Transformer, which effectively enhances the extraction of global information by adding manual features in the attention module and guiding the model to focus on range units with more scattered information, and reduces the dependence on the model on the number of samples. However, the selection of manual features is influenced by human subjective factors and needs further optimization. Diao et al. [18] proposed a recognition method based on Position Embedding-Free Transformer for Radar HRRP Target Recognition, which extracts multiscale information with different weights by combining multiscale convolution with a self-attention mechanism; thus more information and distinguishable features are extracted for recognition. The introduction of multiscale convolution before the self-attention mechanism enables more efficient pre-extraction of multiscale features, but causes a greater computational effort. Although Transformer can achieve better recognition performance, the huge number of model parameters and over-dependence on samples limit its application in edge devices and real scenes.

To achieve a more lightweight and robust recognition method, which facilitates the deployment of real scenes and edge devices. This paper proposes a lightweight HRRP sequence recognition method based on RLAT, which consists of rotary position encoding, local-aggregated attention unit (LAU), and lightweight feedforward neural network (LW-FFN). The method proposed utilizes rotated position encoding to embed relative position information more efficiently. Then, this paper proposes a lightweight local-aggregated attention unit (LAU) to perform local feature aggregation and global perception operations on high-dimensional HRRP sequence data. Feature aggregation can suppress the adverse effects of noise regions, get richer local feature representation, and reduce the number of parameters effectively. Thereby, by putting the aggregated low-dimensional features into the self-attention mechanism, the self-attention mechanism can achieve global information perception and

enhancement, which extracts the long-range correlations of HRRP sequences in time and space domains, effectively enhances the extraction ability of important information in the target region and gets highly distinguishable deep temporal features in HRRP sequences. Besides, the information loss problem of deep networks is also solved through residual connection. Moreover, feature extraction is achieved by LW-FFN, which dramatically reduces the number of parameters compared with the traditional FFN. Finally, Label Smoothing is utilized to introduce label noise to avoid over-reliance of the model on limited training data and enhance the generalization of the proposed method in real scenes. Experiments on MSTAR datasets show that the proposed method improves the recognition performance significantly by effectively reducing the number of parameters and achieves better robustness in both variant targets and limited training data experiments.

The main contribution of this paper is as follows:

(1) Considering the generalization performance and the lightweight of the method, this paper proposes a novel method named RLAT. RLAT consists of Rotary position encoding, LAU, and LW-FFN, which greatly reduces the number of parameters and computational effort by utilizing lightweight modules. In addition, RLAT can represent the relative position information more effectively and deepen the model depth dynamically, which can extract more essential and abstract features.
(2) To alleviate the reliance on training samples and eliminate the undesirable effect of causing redundant information in HRRP sequences. Label smoothing regularization is adopted to add label noise, which can enhance the tolerance to training loss and the generalization of the method.
(3) This paper validates the effectiveness and generalization of the proposed method on real-world datasets, including for variant targets and limited sample conditions; the results illustrate that RLAT has remarkable recognition performance and generalization performance for variant samples and limited samples. Besides, the performance of various position encoding methods and significant hyperparameters in the HRRP sequence task are also explored.

This paper is organized as follows. Section 2 introduces the overall framework of the proposed method and describes the critical detail parts of the method. Section 3 first introduces the construction method of the MSTAR sequence dataset, then verifies the proposed method's effectiveness in real scenes, including variant targets, limited data, and the impact of hyperparameter experiments. Furthermore, comparison experiments verify the proposed method to be more lightweight. Section 4 summarizes the work of this paper and presents the work objectives for the future.

2 Proposed Method

This section presents the overall framework of the proposed method, then introduces and analyzes the principles and details of the essential modules. The overall structure of RLAT is shown in Fig. 1.

HRRP sequence recognition is a particular class of multivariate time series classification problem with high dimensionality, redundant noise information, and limited data. Therefore, suppressing the adverse effects of redundant information in noisy regions and extracting deep valid information and high separability features are essential to improve the performance of HRRP sequence recognition. In addition, existing methods for HRRP sequence recognition extract deep abstract features by stacking a large number of modules, mostly ignoring the problem of lightweight, which is unfavorable for application to real scenes and deployment to edge devices.

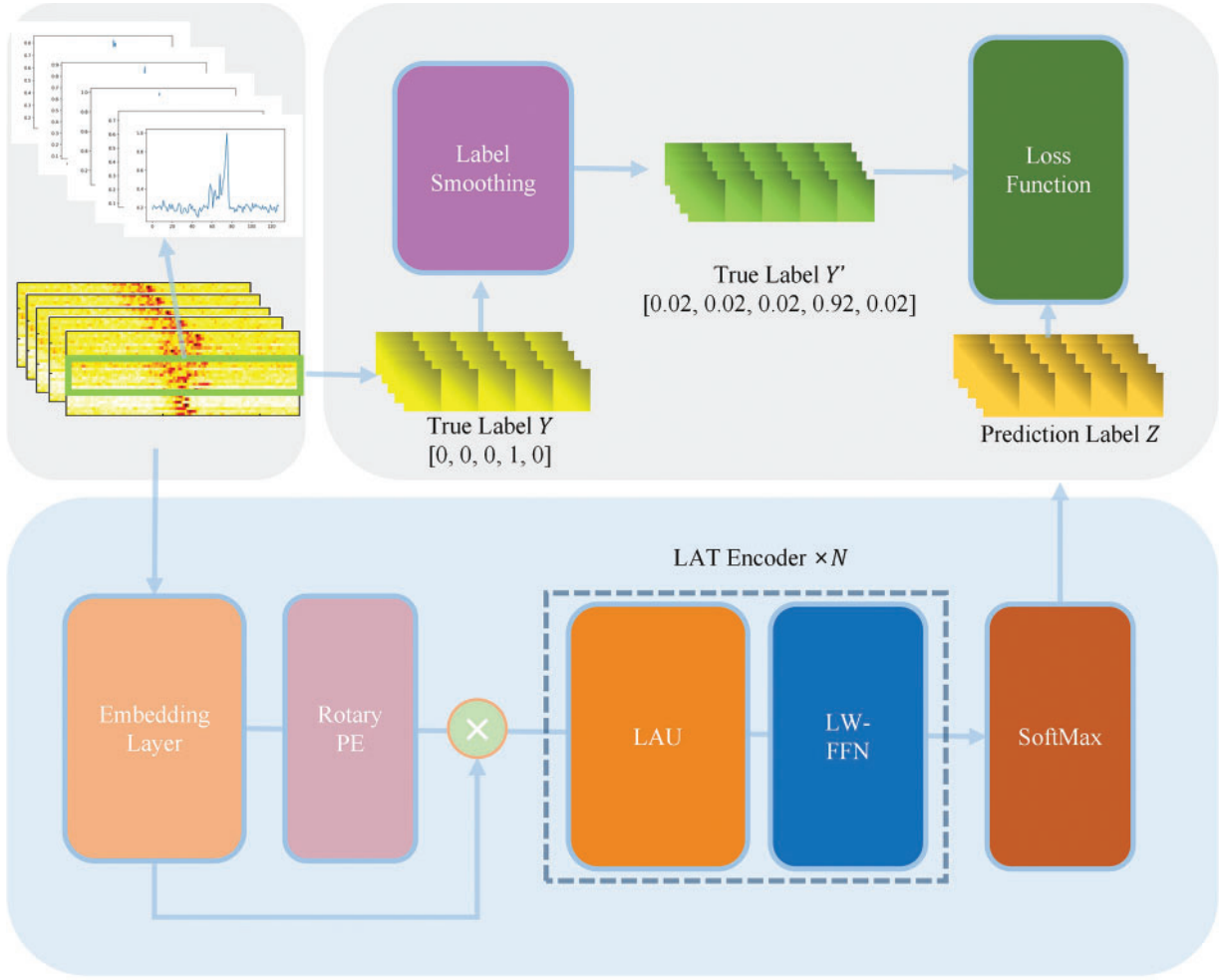


Figure 1: Illustration of the structure of the proposed RLAT

This paper proposes an HRRP sequence recognition method based on RLAT, which consists of rotary position encoding, local-aggregated attention unit (LAU), and lightweight feedforward neural network (LW-FFN). The feature extraction part of RLAT consists of stacked LAT Blocks, which are different from the Encoders of the traditional Transformer. LAT blocks can dynamically adjust the model depth using the adaptive model scaling mechanism. Consequently, the model depth can be dynamically scaled to make the model depth more adaptable to different feature extraction stages, effectively decreasing the number of parameters. Finally, SoftMax is utilized to calculate the probability of each target category achieving recognition.

As shown in Fig. 1, LAT Block is mainly composed of LAU and LW-FFN, where LAU is mainly used for feature enhancement, and LW-FFN is used for feature extraction. LAT Block has significant feature enhancement and feature extraction capabilities and dramatically reduces the number of parameters and computational effort through lightweight methods. To enhance the feature representation of HRRP, traditional deep learning methods introduce richer features by first raising the dimensionality. However, HRRP sequences contain a large amount of redundant noise, and raising the dimensionality often introduces more redundant features, which not only

introduces a massive number of parameters but reduces the recognition ability of the model. The traditional group linear transformation uses the channel shuffle mechanism to enhance the global information extraction ability, but this paper discards the channel shuffle mechanism and inputs the aggregated low-dimensional features into the self-attention mechanism, which can complete the global information perception and enhancement. However, this paper discards the channel shuffle mechanism and inputs the aggregated low-dimensional features into the self-attention mechanism, which can perform global information perception and enhancement to obtain the deep temporal features with high distinguishability in HRRP sequences.

The training process of RLAT is shown in Algorithm 1.

Algorithm 1: Training process of the RLAT

Input: $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N_{data}}$, a dataset of HRRP sequence. θ, initial parameters. N_{epoch} is the number of training iterations. ε is the hyperparameter of Label Smoothing. $\eta \in (0, \infty)$.

Output: $\hat{\theta}$, the trained parameters. The label of the target.

for $iter = 1, 2, \ldots, N_{epoch}$ **do**

  for $n = 1, 2, \ldots, N_{data}$ **do**

    $\boldsymbol{q}(\boldsymbol{\theta}) \leftarrow \mathrm{RLAT}(\boldsymbol{x}_n, \boldsymbol{y}_n | \boldsymbol{\theta})$

    $\mathrm{loss}(\boldsymbol{\theta}) = -\sum_{i=1}^{K} p_i \log q_i(\boldsymbol{\theta})$ where p_i is the probability of the true value, and q_i is the probability of the predicted value. i is the category label of the target.

    Label Smoothing: $Loss(\boldsymbol{\theta}) = \begin{cases} (1-\varepsilon) \cdot \mathrm{loss}(\boldsymbol{\theta}) & i = y \\ \varepsilon \cdot \mathrm{loss}(\boldsymbol{\theta}) & i \neq y \end{cases}$

    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \cdot \nabla Loss(\boldsymbol{\theta})$

  end

end

return $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, the label of the target

2.1 Rotary Position Encoding

Convolutional neural networks and recurrent neural networks get the position information by processing the time series continuously. In contrast, Transformer is a network based on the self-attention mechanism, which is insensitive to position information and needs to add position coding to provide position information for time series. Currently, the commonly used position encoding mainly includes absolute position encoding and relative position encoding. Among them, absolute position encoding is simple to implement, and the number of parameters and computation is smaller. However, the encoded absolute position information is too simplified, which limits the representation of position information, resulting in poor performance in the recognition task. Literature [19] showed that the relevance of sequence data with closer positions is more substantial, thereby adding relative position information is more beneficial to extract the relevance of sequence information. To promote the utilization of relative position information between time series, literature [20] introduced relative position information between sequences in the attention mechanism. Although the performance is improved, the implementation process is complex, and the number of parameters and computations is larger. To achieve a more lightweight relative position encoding, this paper adopts rotary position encoding to add position information to HRRP sequences.

Assuming that the HRRP sequence is $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, where N denotes the length of the sequence. The traditional absolute position encoding adds the position information after the embedding layer, which is calculated as

$$f(\boldsymbol{x}_i, i) = W(\boldsymbol{x}_i + \boldsymbol{p}_i), \tag{1}$$

where i is the position of $\boldsymbol{x}_i$, and $\boldsymbol{p}_i \in \mathbb{R}^d$ is a trainable d-dimension position vector that depends on $\boldsymbol{x}_i$. Absolute position encoding adds mutually independent position information to the sequence data at each position, requiring a large amount of data training to perform better. Since HRRP sequences are usually non-cooperative target data with limited data, absolute position encoding will not be appropriate for solving the HRRP sequence recognition problem.

Relative position encoding is to add relative position information to the self-attention mechanism, which is calculated as

$$\begin{cases} f_q(\boldsymbol{x}_m) = \boldsymbol{W}_q \boldsymbol{x}_m \\ f_k(\boldsymbol{x}_n, n) = \boldsymbol{W}_k(\boldsymbol{x}_n + \boldsymbol{p}_c^k) \\ f_v(\boldsymbol{x}_n, n) = \boldsymbol{W}_v(\boldsymbol{x}_n + \boldsymbol{p}_c^v), \end{cases} \tag{2}$$

where $f_q(\cdot)$, $f_k(\cdot)$ and $f_v(\cdot)$ denote the functions that compute the query, key, and value of the self-attention mechanism, respectively, which are used to calculate the correlation between series data, q, k, v represent query, key, and value of the self-attention mechanism, respectively. $\boldsymbol{p}_c^k, \boldsymbol{p}_c^v \in \mathbb{R}^d$ are the relative position vectors, k and v represent key and value of the self-attention mechanism, respectively, the relative position vectors depend on the relative distance of positions, $c = \text{clip}(m - n, r_{\min}, r_{\max})$ represents the relative distance between the positions m and n, as the distance between the relative positions increases, the correlation between the data decreases, and the limit range of c is set as $r_{\max}$, beyond which the correlation is considered consistent. $\boldsymbol{x}_m, \boldsymbol{x}_n$ are the HRRPs of position and n, respectively. $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ are the learnable parameter matrixes of the self-attention mechanism, respectively.

In pursuit of lightweight relative position encoding, relative position encoding is implemented in the form of absolute position encoding. After adding to the position matrix, the process of the absolute position matrix in the self-attention mechanism is calculated as

$$\boldsymbol{q}_m^{\mathrm{T}} \boldsymbol{k}_n = \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{p}_n + \boldsymbol{p}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{p}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{p}_n, \tag{3}$$

The core concept of relative position encoding is to replace the absolute position vector $\boldsymbol{p}_n$ embedded in the third and fourth terms with the relative position vector $\boldsymbol{p}_{m-n}$, and to replace the third and fourth terms with two trainable vectors $\boldsymbol{u}^{\mathrm{T}}$ and $\boldsymbol{v}^{\mathrm{T}}$. Moreover, the parameter matrixes are replaced by $\boldsymbol{W}_k$ and $\boldsymbol{W}_k'$ respectively to distinguish between content and position. So the method of calculating weights by the self-attention mechanism in the relative position encoding becomes

$$\boldsymbol{q}_m^{\mathrm{T}} \boldsymbol{k}_n = \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k' \boldsymbol{p}_{m-n} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{x}_n + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{W}_q^{\mathrm{T}} \boldsymbol{W}_k' \boldsymbol{p}_{m-n}. \tag{4}$$

To fully use the relative position information in the HRRP sequence to extract the deep temporal information. This paper adopts rotary position encoding to improve the relative position encoding, which is a multiplicative encoding to achieve the relative position encoding utilizing absolute position encoding. The position encoding is calculated as

$$g(\boldsymbol{x}_m, m) = \boldsymbol{R}_{\Theta,m}^d \boldsymbol{W} \boldsymbol{x}_m, \tag{5}$$

where the rotary matrix is

$$\boldsymbol{R}^{d}_{\Theta,m}=\begin{bmatrix}\cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0\\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0\\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & \sin m\theta_{d/2}\\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2}\end{bmatrix}, \tag{6}$$

where $\Theta=\left\{\theta_i=\dfrac{1}{10000^{2(i-1)d}}, i\in[1,2,\ldots,d/2]\right\}$, the calculation process of rotary position encoding in the self-attention mechanism is

$$\boldsymbol{q}_m^{\mathrm{T}}\boldsymbol{k}_n=(\boldsymbol{R}^{d}_{\Theta,m}\boldsymbol{W}_q\boldsymbol{x}_m)^{\mathrm{T}}(\boldsymbol{R}^{d}_{\Theta,n}\boldsymbol{W}_k\boldsymbol{x}_n)=\boldsymbol{x}^T\boldsymbol{W}_q\boldsymbol{R}^{d}_{\Theta,n-m}\boldsymbol{W}_k x_n. \tag{7}$$

The rotary position encoding utilizes a rotary matrix $\boldsymbol{R}^{d}_{\Theta,n-m}$ to introduce relative position information. The rotary position encoding avoids more learnable parameter matrices, which introduces relative position information in a more lightweight way.

2.2 Local-Aggregated Attention Unit

HRRP sequence is a particular class of multivariate time series containing rich temporal and structural information, which is widely used in RATR. However, operations such as adding windows during target detection make redundant information hidden in HRRP, which will adversely impact feature extraction and confuse effective target features. Augmenting attention to important regions of HRRP by using the attention mechanism can effectively suppress the undesirable effects of noisy regions [16] and improve the effectiveness of feature extraction. To enhance the feature enhancement and extraction of HRRP sequences, the local-aggregated attention unit is proposed. Unlike most methods that perform high-dimensional mapping of the input information, LAU downscales the input HRRP sequences, aggregating the features of local information by local group linear transformations. Then the aggregated low-dimensional features are globally perceived by the self-attention mechanism, which effectively enhances the global information extraction ability. Since the HRRP sequence contains a large amount of redundant noisy information, the shallow high-dimensional mapping will confuse the noisy information and the target information. Instead, this paper adopts multilayer local group linear transformations to perform local feature aggregation by first ascending and then descending the local features. Thus, the low-dimensional aggregated features are more easily processed by the self-attention mechanism, and the local feature aggregation can effectively improve the ability of the self-attention mechanism to focus on global information.

As shown in Fig. 2, LAU consists of local group linear transformations, nonlinear activation, layer normalization, self-attention mechanism, and residual connection. Feature enhancement is effectively performed by local feature aggregation and global perception. The encoded vector $\boldsymbol{I}=g(\boldsymbol{X})$ is used as the input of LAU, which performs local feature aggregation and global feature enhancement, then outputs the enhanced features $\boldsymbol{F}$. In the local feature aggregation stage, assuming that there are total L layers of grouped linear transformations. The first layer is up-dimensioned, and the remaining

$L - \lceil L/2 \rceil$ layers are down-dimensioned. The specific procedure to calculate the number of groups in each layer is

$$n^l = \begin{cases} \min(2^{l-1}, n_{\max}), & 1 \le l \le \lceil L/2 \rceil \\ n^{L-l}, & Otherwise, \end{cases} \tag{8}$$

where n^l is the number of groups of the linear transformation of the l th layer, and $n_{\max}$ is the maximum value of the number of groups of the linear transformation of the group. The upper limit of the group value is set to avoid the number of groups being too large. The linear transformation of each layer is calculated as

$$\boldsymbol{H}^l = \begin{cases} G(\boldsymbol{I}, \boldsymbol{W}^l, \boldsymbol{b}^l, g^l), & l = 1 \\ G(\mathrm{M}(\boldsymbol{I}, \boldsymbol{H}^{l-1}), \boldsymbol{W}^l, \boldsymbol{b}^l, g^l), & Otherwise, \end{cases} \tag{9}$$

where $\mathrm{M}(\cdot)$ denotes the operation, including residual connection, nonlinear activation function GELU, and layer normalization. Local feature aggregation can obtain deep local features via first up-dimensioning and then down-dimensioning the local features. Moreover, the residual connection is used to avoid the loss of original information in each layer of the local group linear transformation in the model deepening. Since the low-dimensional aggregated features are facilitated to be processed by the self-attention mechanism, the local feature aggregation can effectively enhance the ability of the self-attention mechanism to focus on global information. Finally, the low-dimensional aggregated features are used as the input of the self-attention mechanism for the global perception operation.

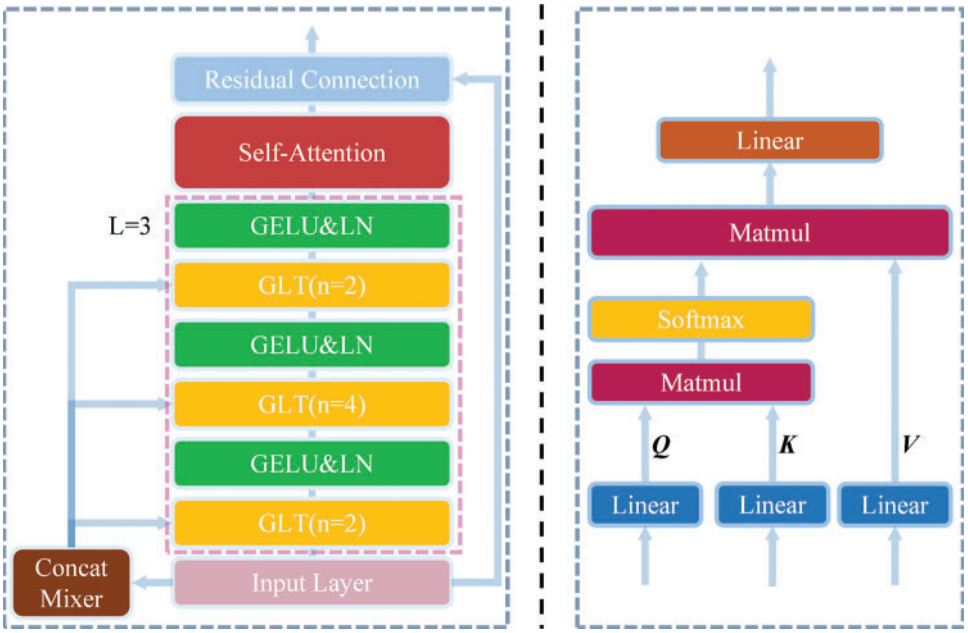


Figure 2: Illustration of the structure of LAU

The self-attention mechanism conducts long-range modeling by calculating correlations between HRRP data, which enhances target features with high discriminability and suppresses the undesirable effects of redundant noisy information, thus effectively performing global enhancement of features after local aggregation. According to the relevant principles of information retrieval, the self-attention mechanism calculates the correlation of sequence data by query vector and key vector to obtain the attention matrix and then calculates the globally enhanced features with the value matrix as

$$\boldsymbol{A} = SA\left(\boldsymbol{H}^{L}\right) = Attention\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{10}$$

where $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{W}_{Q}$ is the query vector, $\boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}_{K}$ is the key vector, and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}_{V}$ is the value vector, $\boldsymbol{W}_{Q}$,$\boldsymbol{W}_{K}$,$\boldsymbol{W}_{V}$ are the learnable parameter matrixes $\boldsymbol{W}_{s}$, respectively.

Finally, to ensure that the dimensionality of the input and output is consistent, the vector needs to be up-dimensioned first after the self-attention mechanism processing. At the same time, to avoid the loss of important features due to the excessive depth of the model, the residual connection is finally added to retain the vital information in the original features, which can be obtained as

$$\boldsymbol{F} = \boldsymbol{I} + \boldsymbol{W}_{s}\boldsymbol{A}, \tag{11}$$

where $\boldsymbol{W}_{s}$ is the parameter matrix for dimensioning the output of the self-attention mechanism.

2.3 Lightweight Feedforward Neural Network

LAU has a deeper network structure and more significant feature enhancement capability than the traditional multi-head attention mechanism. Therefore, this paper uses a lightweight feedforward neural network instead of the traditional feedforward neural network for feature extraction. Assuming that the dimensionality of the input features $\boldsymbol{F}$ is d_m, the traditional FFN (as shown in the left panel of Fig. 3) adopts the way of first up-dimensioning to $4d_m$ and then down-dimensioning to d_m for feature extraction, which has a large number of parameters. Since LAU has a significant effect on feature enhancement, this paper adopts lightweight FFN (shown in the right panel of Fig. 3) for feature extraction, which consists of a fully connected layer, a nonlinear activation function, and a residual connection, and is mainly used for feature extraction. The lightweight FFN is first down-dimensioned to $\frac{d_m}{4}$, then up-dimensioned to d_m. As for lightweight FFN, the total number of parameters is $\frac{d_m^2}{4} \times 2 = \frac{d_m^2}{2}$ for both up- and down-dimensional processes, but the traditional FFN first up-dimensionalizing to $4d_m$ and then down-dimensionalizing to d_m, the total number of parameters is $4d_m^2 \times 2 = 8d_m^2$. Therefore, the lightweight FFN reduces the number of parameters by 16 times compared with the traditional FFN. The computational process of lightweight FFN for feature extraction is

$$\boldsymbol{O}_{m} = Feedforward(\boldsymbol{F}) = \boldsymbol{F} + ((\text{Relu}(\boldsymbol{F}\boldsymbol{W}_{1} + \boldsymbol{b}_{1}))\boldsymbol{W}_{2} + \boldsymbol{b}_{2}), \tag{12}$$

where $\boldsymbol{W}_{1}, \boldsymbol{b}_{1}$ are the parameter matrix and bias vector at dimensionality reduction, respectively, $\boldsymbol{W}_{2}, \boldsymbol{b}_{2}$ are the parameter matrix and bias vector at dimensionality reduction, respectively, and Relu$(\cdot)$ are the nonlinear activation functions.

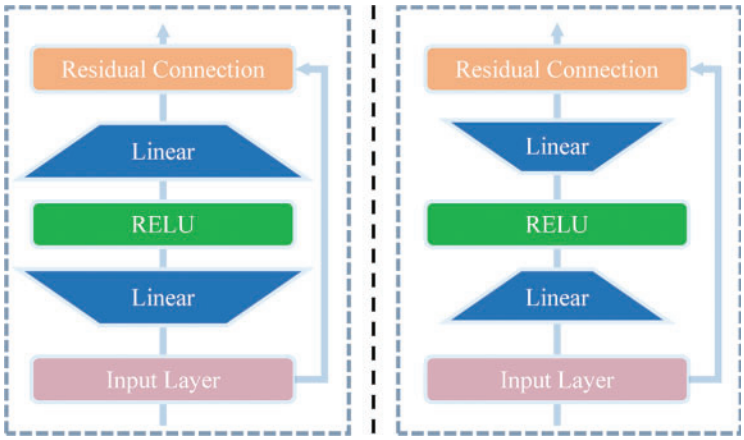


Figure 3: Illustration of the structure of LW-FFN

Following the stacked layer LAT blocks process, finally using SoftMax as the classifier, the output is

$$\boldsymbol{Y} = \text{SoftMax}\left(\boldsymbol{O}_M\right) = \frac{\exp\left(o_i^M\right)}{\sum_{k=1}^{K} o_k^M}, \tag{13}$$

where $\boldsymbol{O}_M$ is the output of the LAT block at the M th layer, $\exp(\cdot)$ denotes the exponential function, and K is the total number of categories.

2.4 Label Smoothing Regularization

Against the background of non-cooperative targets, the current HRRP sequence samples are limited in quantities. Furthermore, the HRRPs in real scenes are in a complex noise environment, and there are still some differences in HRRPs of the same targets, which strongly leads to the overfitting problem in HRRP sequence recognition. To solve the overfitting problem, the Label Smoothing regularization strategy is adopted [21]. Label Smoothing adds label noise to avoid the model over-reliance on limited training samples and enhances the generalization performance of the proposed method for application in real scenes.

When coding the labels of the samples, the probability distribution of the traditional one-hot coding is

$$p_i = \begin{cases} 1 & i = y \\ 0 & i \neq y. \end{cases} \tag{14}$$

To enhance the generalization performance of the model, one-hot coding is modified to soft one-hot coding to add fuzzy noise to the labels, thus reducing the weight of real sample labels in the computational loss. Consequently, the model will not be overly dependent on a limited number of samples, avoiding falling into local optimal solutions, which finally achieves suppression of the overfitting problem. Once Label Smoothing is added, the probability distribution of soft one-hot

labels is

$$p_i = \begin{cases} (1-\varepsilon) & i = y \\ \dfrac{\varepsilon}{K-1} & i \neq y, \end{cases} \tag{15}$$

where K denotes the total number of categories in the task, i denotes the number of categories, and ε is the hyperparameter.

When using the cross-entropy loss function to calculate the loss values between the predicted values and true values, the cross-entropy loss function is calculated as

$$Loss = -\sum_{i=1}^{K} p_i \log q_i, \tag{16}$$

where p_i is the probability of the true value, and q_i is the probability of the predicted value.

The neural network will optimize the model in the direction of low loss value during the training process. However, over-reliance on the training set data will reduce the generalization performance of the recognition task of HRRP sequences in real scenes. The Label Smoothing regularization strategy will avoid overconfidence in the network, slow down the penalty intensity of the loss, and avoid the model falling into the local optimal solution, and its loss function for each category is calculated as

$$Loss_i = \begin{cases} (1-\varepsilon) \cdot Loss & i = y \\ \varepsilon \cdot Loss & i \neq y, \end{cases} \tag{17}$$

where ε is the hyperparameter. In the neural network training process, the optimal prediction probability distribution is obtained when minimizing the cross-entropy loss values of the predicted and true values, as

$$Z_i = \begin{cases} \log \dfrac{(K-1)(1-\varepsilon)}{\varepsilon + \alpha} & i = y \\ \alpha & i \neq y, \end{cases} \tag{18}$$

where K denotes the total number of categories in the task, ε is a hyperparameter, and α is a real number.

As derived from the prediction probability distribution, Label Smoothing regularization can increase the tolerance to the existence of errors between the true values and predicted values. Consequently, Label Smoothing can prevent the model from over-relying on the training set samples, which can prevent the model from falling into local optimal solutions and enhance the generalization performance of the model.

3 Experiment Results and Analysis

3.1 Datasets

The MSTAR dataset is a standard dataset widely used for SAR target recognition [7,22,23]. Its data source is a high-resolution clustered synthetic aperture radar, which operates in the X-band with a resolution of 0.3 m × 0.3 m and HH polarization. The MSTAR dataset includes ten categories of targets, such as T72, BMP2, and BTR70. The data with a pitch angle of 17° in the dataset is used as the training set, and the data with a pitch angle of 15° are used as the test set. The azimuth angles of all targets cover 0∼360°. Dataset 1 includes the original MSTAR dataset training set of 2747 SAR

images, and the test set includes 2348 SAR images. To further test the generalization performance of the model, this paper adds four variant targets of BMP2 (SN-9563), BMP2 (SN-C21), T72 (SN-812), and T72 (SN-S7) to the test set to make up dataset 2. Dataset 2 includes 2747 SAR images in the training set of the original MSTAR dataset and 3203 SAR images in the test set. In this paper, SAR images are converted into HRRP sequences according to the method specified in reference [24], and the MSTAR sequence dataset 2 is composed as shown in Table 1.

Table 1: MSTAR sequence dataset 2

Category of training set	Training set (17°)	Category of test set	Test set (15°)
2S1	2990	2S1	2740
		BMP2 (SN-9566)	1960
BMP2 (SN-9566)	2330	BMP2 (SN-9563)	1950
		BMP2 (SN-C21)	1960
BRDM-2	2980	BRDM-2	2740
BTR70 (SN-C71)	2330	BTR70 (SN-C71)	1960
BTR60	2560	BTR60	1950
D7	2990	D7	2740
T62	2990	T62	2730
		T72 (SN-132)	1960
T72 (SN-132)	2320	T72 (SN-812)	1950
		T72 (SN-S7)	1910
ZIL131	2990	ZIL131	2740
ZSU23/4	2990	ZSU23/4	2740
Total	27470	Total	32030

The conversion steps are as follows: The dataset is first converted into a complex SAR image, and then an Inverse Fast Fourier Transform (IFFT) is carried out in the orientation dimension of the complex SAR image, and the data obtained along the distance dimension is the HRRP complex sequence. Then the HRRP sequence is obtained after modulo the HRRP complex sequence. 100 HRRP samples could be obtained for each complex SAR image, and the average of every 10 HRRPs can be obtained as 10 average HRRP samples. Consequently, the training set of the original MSTAR dataset includes 24,270 HRRP samples, and the test set includes 32,030 HRRP samples.

Assuming that the length of the generated HRRP sequence is $L(L \leq 50)$, the sliding window algorithm for generating the HRRP sequence is shown in Algorithm 2.

Algorithm 2: HRRP sequence generation algorithm

Step 1: The azimuth blocks are arranged in order, and to obtain the same number of HRRP sequences, $L-1$ previous azimuth blocks are added after the 50th block, so that the sliding window data contains a total of $50+L-1$ azimuth blocks;

Step 2: According to the order of azimuth blocks, the first HRRP sequence shall be taken from the 1st to the L block, so that the HRRP sequence with the length of L could be obtained;

Step 3: Slide the sliding window down and repeat Step 2 until the whole block is taken;

(Continued)

Algorithm 2 (continued)

Step 4: Return to the 1st HRRP of the azimuth block, and repeat steps 2 and 3 from the 2nd to the $L+1$ azimuth block in the order of azimuth blocks;

Step 5: Repeat step 4 until 50 azimuth blocks are taken, and then obtain the same number of HRRP sequences as the HRRP data.

Use the sliding window algorithm to process HRRP data according to the steps shown in Fig. 4.

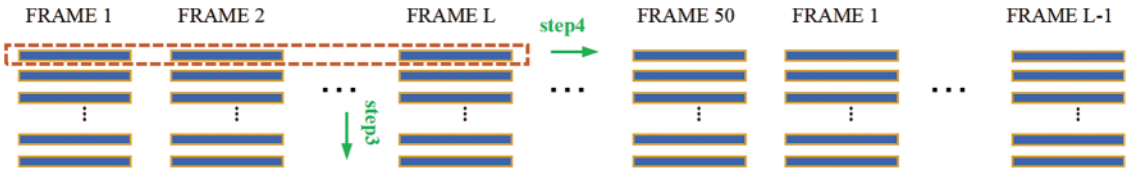


Figure 4: Schematic diagram of HRRP sequence generation

As can be seen from Fig. 4, the azimuth angle of 360° is divided into 50 azimuth blocks, and each block contains 7.2°, in which the sampling interval of each SAR image is 1°, and each SAR image can be processed, and ten average HRRP samples are obtained, so the sampling interval of each average HRRP sample is 0.1°. In previous research, denoised and enhanced samples are used as input for the model, which leads to a poor generalization of the model. In this paper, only the HRRP data are energy normalized. In real scenes, data of individual angles are often missing due to aircraft motion. Therefore, the dataset did not interpolate the missing data in the MSTAR dataset, which makes the data more consistent with the real situation.

After processing by the sliding window method, this paper gets 24,270 HRRP sequence samples in the training set and 23,480 HRRP sequence samples in the test set from dataset 1; get 24,270 HRRP sequence samples in the training set and 32,030 HRRP sequence samples in the test set from dataset 2.

The HRRP sequence samples of some targets are shown in Figs. 5a–5h are the corresponding HRRPs, respectively. It can be seen that MSTAR, as a real-world dataset, sample contains a large amount of noisy redundant information, which causes greater difficulties and challenges for effective feature extraction during recognition.

3.2 Recognition Performance Comparison Experiments

To verify the recognition performance of the proposed methods, seven frequently used baseline methods, LSTM [25], GRU [26], 1D-CNN [14], TCN [27], gMLP [28], XCM [29], and Transformer [30], are selected as comparison methods in this paper. Moreover, the model architecture and parameters for the comparison experiments were designed according to the references to achieve optimal model performance. The recognition performance of each method is verified in MSTAR dataset 1, and the recognition results of the comparison experiments on ten categories of targets are shown in Table 2.

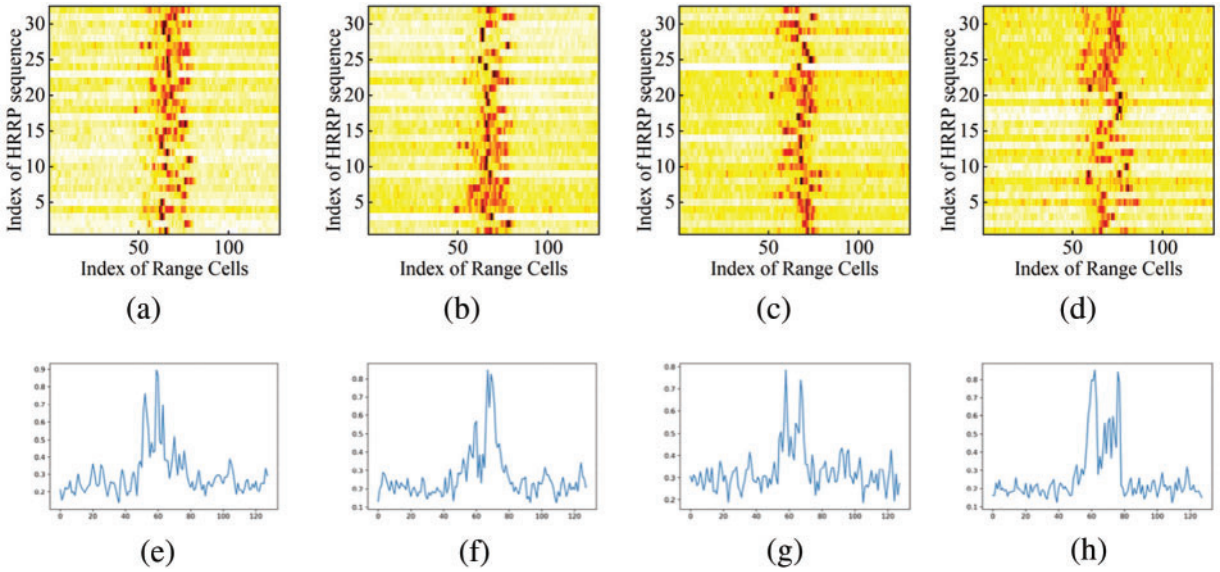


Figure 5: Part of samples of MSTAR datasets

Table 2: Recognition accuracy of compare experiments on dataset 1 (%)

	LSTM	GRU	1D-CNN	TCN	gMLP	XCM	Transformer	Proposed
2S1	89.4 (−10.32)	95.26 (−4.48)	**99.78** (+0.04)	98.46 (−1.28)	66.06 (−33.68)	97.15 (−2.59)	98.56 (−1.18)	99.74
BMP2 (SN-9566)	99.95 (−0.05)	99.80 (−0.20)	99.95 (−0.05)	**100.00** (0.00)	71.94 (−28.06)	99.08 (−0.92)	98.74 (−1.26)	**100.00**
BRDM-2	94.64 (−5.18)	90.29 (−9.53)	98.61 (−1.21)	99.74 (−0.08)	67.99 (−31.83)	99.85 (+0.03)	**99.96** (+0.14)	99.82
BTR70 (SN-C71)	82.50 (−17.5)	93.62 (−6.38)	95.97 (−4.03)	86.68 (−13.32)	81.73 (−18.27)	94.80 (−5.20)	99.69 (−0.31)	**100.00**
BTR60	99.54 (−0.46)	**100.00** (0.00)	99.74 (−0.26)	**100.00** (0.00)	91.69 (−8.31)	99.69 (−0.31)	98.73 (−1.27)	**100.00**
D7	**100.00** (0.00)	**100.00** (0.00)	**100.00** (0.00)	**100.00** (0.00)	97.23 (−2.77)	**100.00** (0.00)	99.96 (−0.04)	**100.00**
T62	95.02 (−4.72)	94.14 (−5.60)	**99.89** (+0.15)	95.75 (−3.99)	89.60 (−10.14)	98.06 (−1.68)	97.85 (−1.89)	99.74
T72 (SN-132)	**100.00** (0.00)	**100.00** (0.00)	**100.00** (0.00)	**100.00** (0.00)	95.51 (−4.49)	**100.00** (0.00)	**100.00** (0.00)	**100.00**
ZIL131	85.15 (−14.3)	91.24 (−8.21)	98.39 (−1.06)	97.29 (−2.16)	71.82 (−27.63)	**99.78** (+0.33)	99.77 (+0.32)	99.45
ZSU23/4	99.74 (−0.26)	99.89 (−0.11)	99.82 (−0.18)	**100.00** (0.00)	96.97 (−3.03)	**100.00** (0.00)	99.96 (−0.04)	**100.00**
Average value	94.48 (−5.38)	96.17 (−3.69)	99.25 (−0.61)	97.94 (−1.92)	82.77 (−17.09)	98.90 (−0.96)	99.32 (−0.54)	**99.86**

As shown in Table 2, the average accuracy of the RLAT proposed in this paper is the highest for ten categories of target recognition, reaching 99.86%, which is 17.09% better than the gMLP, more than 3.69% better than the commonly used recurrent neural networks LSTM and GRU, 0.61%

and 1.92% better than the remarkable performance of convolutional neural networks 1D-CNN and TCN, respectively, 0.96% better than XCM, 0.54% better than Transformer with the same network structure. Besides, the proposed RLAT achieves optimal recognition performance on six targets, TCN achieves optimal performance on five targets, and other methods are less than five, reflecting that the proposed method is more stable than others. The experimental results show that the recognition performance of the Transformer and this work are higher than other methods, which indicates that the long-range modeling information using Transformer can mine the long-range temporal information and represent the features of HRRP sequences more effectively. In addition, RLAT has powerful local feature extraction and global perception capabilities, which can extract the local and global multi-level information between sequences more efficiently than the traditional Transformer, thus achieving the highest recognition performance.

As shown in Fig. 6, the proposed method has a more stable and balanced recognition performance for ten categories of targets, and all other methods have certain recognition shortcomings. In particular, the recognition performance of gMLP and LSTM is very volatile, with a fluctuation range of 31.17% and 17.50%, the fluctuation range of GRU is 9.71%, the fluctuation range of 1D-CNN and TCN is 4.03% and 13.32%, the fluctuation range of XCM is 5.20%, and the fluctuation range of Transformer is 2.15%, respectively. In comparison, the maximum fluctuation range of the proposed method is less than 0.55%. The results illustrate the effectiveness of the RLAT, which can suppress the adverse effects of noisy information and effectively extract highly distinguishable target features.

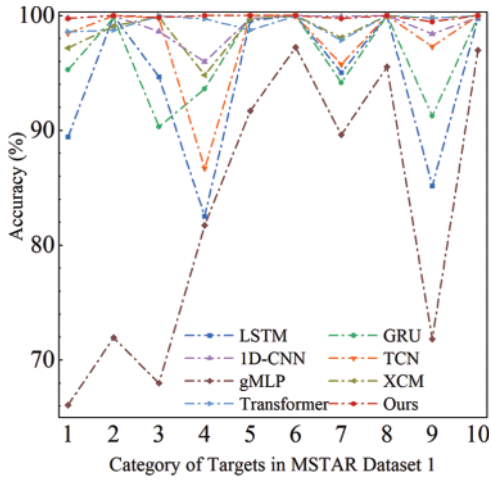


Figure 6: Accuracy of 10 targets in MSTAR dataset 1; the numbers on the x-axis represent ten targets in dataset 1, respectively

3.3 Robustness Comparison Experiments

For real-world application scenes, HRRP data usually come from non-cooperative targets, which usually contain variant versions, resulting in the shape configurations of the variant targets being different from those of the original targets. The recognition performance of the variant targets is an essential factor in measuring the method's robustness. The robustness of the proposed method on the variant dataset is verified by setting up comparison experiments on the dataset MSTAR dataset 2. Dataset 2 is unchanged compared with the training set of Dataset 1, but about 36% of variant samples

are added in the test set, which is mainly distributed on the BMP2 and T72 targets, constituting an unbalanced dataset simultaneously. The experimental results are shown in Table 3.

Table 3: Recognition accuracy of compare experiments on dataset 2 (%)

	LSTM	GRU	1D-CNN	TCN	gMLP	XCM	Transformer	Proposed
2S1	84.49 (−13.47)	83.36 (−14.60)	**99.31** (+1.35)	97.55 (−0.41)	32.66 (−65.30)	99.23 (+1.27)	89.51 (−8.45)	97.96
BMP2 (SN-9566)	98.14 (−1.79)	89.05 (−10.88)	96.70 (−3.23)	94.57 (−5.36)	46.20 (−53.73)	94.63 (−5.30)	99.58 (−0.35)	**99.93**
BRDM-2	95.51 (−4.13)	93.94 (−5.70)	99.45 (−0.19)	98.80 (−0.84)	69.20 (−30.44)	98.28 (−1.36)	99.41 (−0.23)	**99.64**
BTR70 (SN-C71)	77.24 (−22.76)	93.32 (−6.68)	94.74 (−5.26)	88.42 (−11.58)	70.51 (−29.49)	97.40 (−2.60)	96.94 (−3.06)	**100.00**
BTR60	96.46 (−2.77)	94.72 (−4.51)	98.36 (−0.87)	**100.00** (+0.77)	68.31 (−30.92)	**100.00** (+0.77)	97.59 (−1.64)	99.23
D7	99.85 (−0.15)	99.64 (−0.36)	**100.00** (0.00)	**100.00** (0.00)	86.57 (−13.43)	**100.00** (0.00)	**100.00** (0.00)	**100.00**
T62	91.98 (−7.95)	95.86 (−4.07)	99.19 (−0.74)	92.67 (−7.26)	54.95 (−44.98)	98.35 (−1.58)	99.00 (−0.93)	**99.93**
T72 (SN-132)	97.85 (−2.15)	99.85 (−0.15)	92.51 (−7.49)	97.44 (−2.56)	79.05 (−20.95)	**100.00** (0.00)	99.98 (−0.02)	**100.00**
ZIL131	86.57 (−13.43)	66.31 (−33.69)	99.01 (−0.99)	98.72 (−1.28)	48.72 (−51.28)	97.08 (−2.92)	99.41 (−0.59)	**100.00**
ZSU23/4	**100.00** (0.00)	**100.00** (0.00)	99.93 (−0.07)	**100.00** (0.00)	92.85 (−7.15)	**100.00** (0.00)	98.88 (−1.12)	**100.00**
Average value	94.11 (−5.62)	92.03 (−7.70)	97.35 (−2.38)	96.78 (−2.95)	64.22 (−35.51)	98.25 (−1.48)	98.31 (−1.42)	**99.73**

As shown in Table 3, the recognition accuracy of each method decreases due to the increased variant samples and the higher generalization performance required for the model. Nevertheless, the proposed method still achieves the highest average recognition accuracy of 99.73%, which is only 0.13% lower compared to dataset 1. gMLP decreases by 18.55%, LSTM, and GRU by 0.37% and 4.14%, 1D-CNN, and TCN by 1.90% and 1.16%, respectively, and Transformer by 1.01%. At the same time, RLAT is 35.51% better than the gMLP, more than 5.62% better than the commonly used recurrent neural networks LSTM and GRU, 2.38% and 2.95% better than the remarkable performance of convolutional neural networks 1D-CNN and TCN, respectively, 1.48% better than XCM, 1.42% better than Transformer with the same network structure. In addition, the proposed RLAT achieves optimal recognition performance on eight targets, while all other compared methods are less than 4, which is even superior to dataset 1. RLAT has long-range modeling capabilities and dynamically deepens the model depth by LAU, enabling the extraction of more essential and abstract features. Therefore, RLAT shows more remarkable stability on the variant dataset, which has a stronger generalization performance than other methods.

As shown in Fig. 7, the recognition performance of RLAT for the variant dataset is more stable and balanced, with a maximum fluctuation range of only 2.04%. In comparison, the fluctuation ranges of the comparison methods LSTM and GRU are 22.76% and 16.64%, 1D-CNN and TCN are 5.37% and 10.49%, Transformer is 7.49%, XCM is 11.58%, and gMLP is 60.19%, respectively. The results illustrate that the global temporal features extracted by RLAT are more robust, stable, and

distinguishable, as well as Label Smoothing can avoid over-reliance on training samples and further improve the generalization performance to variant samples.

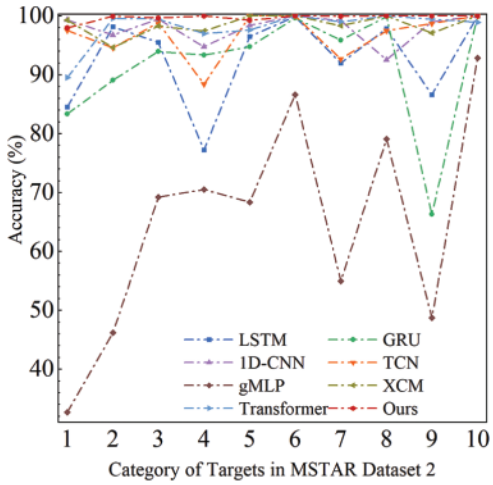


Figure 7: Accuracy of 10 targets in MSTAR dataset 2; the numbers on the x-axis represent 10 targets in dataset 1, respectively

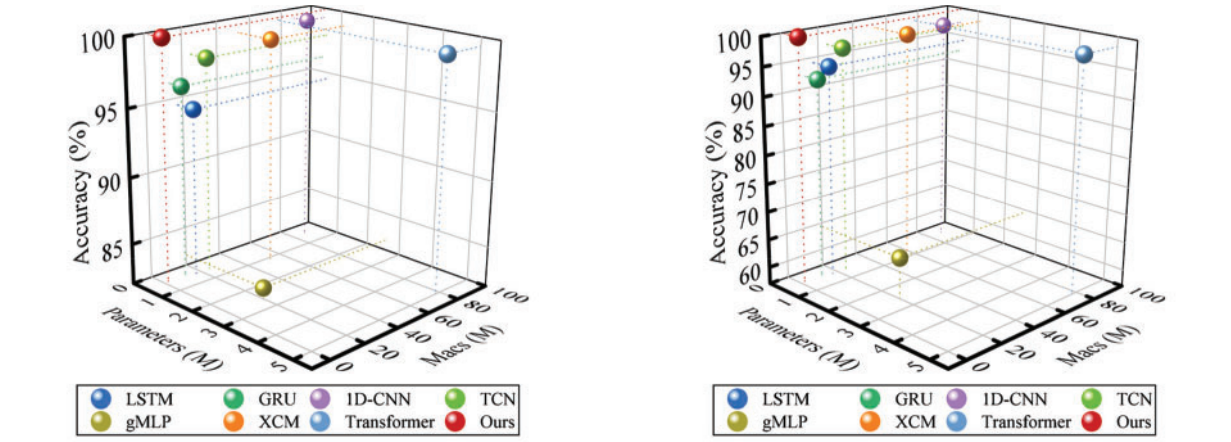
3.4 Lightweight Comparison Experiments

RLAT achieves remarkable recognition performance in both the MSTAR standard dataset D 1 and the variant dataset D 2. To verify the lightweight of the proposed method, the number of parameters and the computational effort for comparing the various methods are shown in Table 4.

Table 4: Recognition accuracy of lightweight compare experiments (%)

	LSTM	GRU	1D-CNN	TCN	gMLP	XCM	Transformer	Proposed
Params (M)	0.53	0.40	0.43	0.51	2.55	1.06	4.98	0.43
Macs (M)	17.07	12.80	88.52	25.19	20.51	53.13	77.76	2.56

As shown in Table 4, the proposed RLAT achieves significant lightweight in terms of the number of parameters and computation, with 90.90% reduction in the number of parameters and 96.70% reduction in the computation compared to the Vanilla Transformer. Since RLAT uses the LAU module, the number of parameters and computations is significantly reduced while ensuring recognition performance. Excluding GRU, the number of parameters of RLAT is smaller than other comparable models, and the computation of the proposed method is smaller than other comparable models. In particular, the results show that the computation of 1D-CNN and XCM is severely increased due to the introduction of convolutional neural networks. As shown in Fig. 8, RLAT achieves better recognition performance under the premise of a more lightweight network structure, which illustrates that RLAT is more favorable for edge devices and real-world application deployment. As shown in Fig. 8, the relationship between accuracy, number of parameters and computation can be more intuitively obtained. The experimental results show that RLAT achieves better recognition performance with a smaller number of parameters and computations.



(a) Lightweight comparison experiments of D 1 (b) Lightweight comparison experiments of D 2

Figure 8: The result of lightweight comparison experiments

3.5 Limited Sample Comparison Experiments

HRRP sequence recognition under limited samples is one of the significant challenges currently. Recognition performance for limited samples is verified for the sequence length of HRRP sequences and the number of training samples. The limited sequence length can improve the real-time performance of the recognition, which achieves the target recognition several times earlier according to the demand. Then, the limited training samples can effectively verify the generalization performance of the model under the non-cooperative target conditions, which can effectively reduce the training time. At the same time, limited samples mean less target information, which has higher expectations on the feature extraction ability of the model. Comparison experiments are set up to verify the recognition performance of the proposed method under the limited sample condition using limited-length HRRP sequence samples and limited training samples. HRRP sequence generation algorithm is used for the generation of datasets, the length of the HRRP sequence is set as $\{1, 2, 4, 8, 16\}$, and the rate of training samples is set as $\{1\%, 2\%, 5\%, 10\%, 50\%, 90\%\}$, respectively. The following abbreviates MSTAR dataset 1 as D 1 and MSTAR dataset 2 as D 2. The severe experimental conditions are set to verify the recognition performance of the proposed method under the limited sample conditions, and the experimental results are shown in Tables 5 and 6.

Table 5: Recognition accuracy for limited sequence length (%)

	D 1					D 2				
Methods	1	2	4	8	16	1	2	4	8	16
LSTM	58.83 (−10.38)	77.30 (−6.98)	86.76 (−5.61)	90.63 (−4.44)	92.82 (−5.60)	55.69 (−8.71)	71.40 (−5.75)	82.26 (−4.29)	88.68 (−3.30)	89.74 (−6.16)
GRU	64.00 (−5.21)	78.51 (−5.77)	87.07 (−5.30)	91.78 (−3.29)	95.14 (−3.28)	57.59 (−6.81)	72.97 (−4.18)	83.90 (−2.65)	91.31 (−0.67)	92.96 (−2.94)
1D-CNN	65.52 (−3.69)	77.26 (−7.02)	86.44 (−5.93)	92.65 (−2.42)	96.51 (−1.91)	61.39 (−3.01)	71.75 (−5.40)	80.81 (−5.74)	88.66 (−3.32)	93.81 (−2.09)
TCN	68.30 (−0.91)	82.54 (−1.74)	88.85 (−3.52)	92.47 (−2.60)	95.19 (−3.23)	62.37 (−2.03)	76.64 (−0.51)	84.78 (−1.77)	90.87 (−1.11)	94.23 (−1.67)

(Continued)

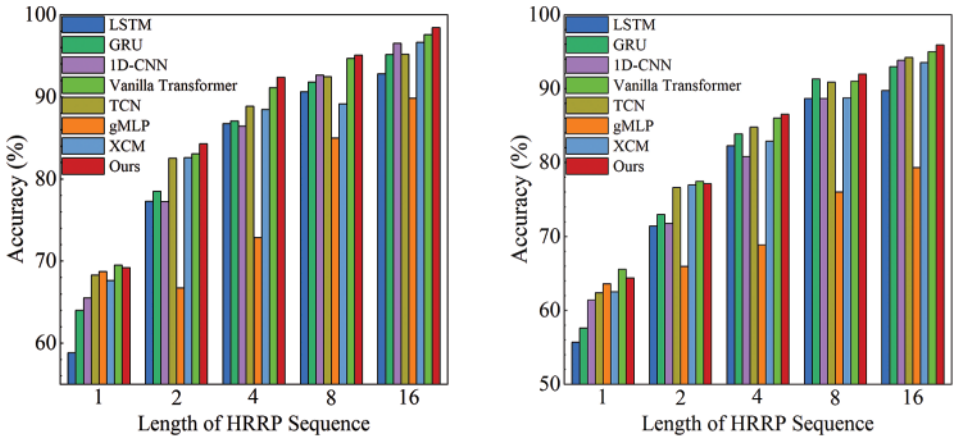
Table 5 (continued)

	D 1					D 2				
Methods	1	2	4	8	16	1	2	4	8	16
gMLP	68.71 (−0.50)	66.71 (−17.57)	72.86 (−19.51)	85.00 (−10.07)	89.83 (−8.59)	63.58 (−0.82)	65.96 (−11.19)	68.87 (−17.68)	76.01 (−15.97)	79.34 (−16.56)
XCM	67.63 (−1.58)	82.59 (−1.69)	88.47 (−3.90)	89.14 (−5.93)	96.64 (−1.78)	62.52 (−1.88)	76.99 (−0.16)	82.90 (−3.65)	88.73 (−3.25)	93.51 (−2.39)
Transformer	**69.51** (+0.30)	83.06 (−1.22)	91.14 (−1.23)	94.70 (−0.37)	97.58 (−0.84)	**65.54** (+1.14)	**77.44** (+0.28)	86.02 (−0.53)	91.01 (−0.97)	95.02 (−0.88)
Proposed	69.21	**84.28**	**92.37**	**95.07**	**98.42**	64.40	77.15	**86.55**	**91.98**	**95.90**

Table 6: Recognition accuracy for limited training data (%)

	D 1						D 2					
Methods	1%	2%	5%	10%	50%	90%	1%	2%	5%	10%	50%	90%
LSTM	56.54 (−39.29)	60.00 (−39.31)	63.05 (−36.34)	91.31 (−8.11)	93.87 (−5.85)	94.25 (−5.54)	45.96 (−47.4)	50.33 (−46.95)	57.78 (−41.11)	87.12 (−11.87)	93.20 (−6.25)	94.05 (−5.66)
GRU	60.03 (−35.80)	61.51 (−37.80)	62.18 (−37.21)	82.77 (−16.65)	93.57 (−6.15)	94.86 (−4.93)	49.35 (−44.01)	52.62 (−44.66)	53.46 (−45.43)	77.53 (−21.46)	90.32 (−9.13)	91.20 (−8.51)
1D-CNN	93.41 (−2.42)	96.80 (−2.51)	97.40 (−1.99)	98.48 (−0.94)	98.94 (−0.78)	99.25 (−0.54)	90.94 (−2.42)	96.07 (−1.21)	96.36 (−2.53)	96.56 (−2.43)	97.14 (−2.31)	97.34 (−2.37)
TCN	87.95 (−7.88)	90.59 (−8.72)	95.63 (−3.76)	96.01 (−3.41)	96.33 (−3.39)	97.50 (−2.29)	86.23 (−7.13)	89.06 (−8.22)	94.96 (−3.93)	95.50 (−3.49)	95.82 (−3.63)	96.01 (−3.70)
gMLP	19.65 (−76.18)	23.05 (−76.26)	37.39 (−62)	51.26 (−48.16)	75.43 (−24.29)	77.45 (−22.34)	17.61 (−75.75)	20.54 (−76.74)	35.53 (−63.36)	48.40 (−50.59)	60.98 (−38.47)	63.06 (−36.65)
XCM	88.03 (−7.80)	92.72 (−6.59)	96.00 (−3.39)	98.89 (−0.53)	98.78 (−0.94)	98.87 (−0.92)	87.37 (−5.99)	92.00 (−5.28)	94.67 (−4.22)	96.60 (−2.39)	97.45 (−2.00)	97.64 (−2.07)
Transformer	80.42 (−15.41)	92.14 (−7.17)	95.59 (−3.80)	95.61 (−3.81)	98.36 (−1.36)	98.65 (−1.14)	75.21 (−18.15)	89.88 (−7.40)	93.97 (−4.92)	94.08 (−4.91)	96.66 (−2.79)	97.81 (−1.90)
Proposed	**95.83**	**99.31**	**99.39**	**99.42**	**99.72**	**99.79**	**93.36**	**97.28**	**98.89**	**98.99**	**99.45**	**99.71**

As shown in Table 5 and Fig. 9, the recognition accuracy of HRRP sequences increases with sequence length. The proposed method performs more remarkably than other comparative experiments on most short sequences. The recognition performance of RLAT outperforms the methods except for the Transformer in both the standard dataset D 1 and the variant dataset D 2, which illustrates that the Transformer-based methods utilize the long-range modeling capability to effectively extract valid target information from short HRRP sequences and reduce the adverse effects of noisy redundant information. Since Transformer has a more complex model structure than RLAT and sequences with lengths 1 and 2 contain less information, the recognition performance is slightly higher for D 1 with sequence length 1 and for D 2 dataset with sequence lengths 1 and 2. As the sequence length increases, RLAT can perform feature selection more effectively and discard the adverse effects of redundant information hidden in HRRP sequences, which can achieve more significant recognition performance than Transformer. To present the results of the comparison experiments more visually, the comparison experiments with limited sequence length are shown in Fig. 9.



(a) Recognition accuracy of D 1 (b) Recognition accuracy of D 2

Figure 9: Recognition accuracy for limited sequence length of different methods

The Transformer relies on a large number of training samples to improve the recognition performance, which severely constrains the application of the Transformer in the field of HRRP recognition. In contrast, the feature enhancement and feature extraction capabilities of the RLAT are more remarkable and can effectively improve the recognition performance for limited training samples. To verify the recognition performance of the proposed method under the condition of limited training samples, the sequence length is set as 32, and the number of training samples is kept at {1%, 2%, 5%, 10%, 50%, 90%}, respectively, where 1% of the original training set contains only 274 training samples, and the recognition performance is shown in Table 6.

As shown in Table 6 and Fig. 10, the recognition accuracy of HRRP sequences increases with the increase of training samples. RLAT achieves more remarkable recognition performance than other methods on the standard dataset D 1 and the variant dataset D 2 with limited training samples. In particular, RLAT achieves 95.83% accuracy on the MSTAR standard dataset D 1 when the training set is only 1% of the original training set. Compared to Transformer, the accuracy is improved by 15.41%, compared to the gMLP, improved by 76.18%, compared to LSTM and GRU, improved by more than 35.80%, compared to TCN, 1D-CNN, and XCM, improved by more than 2.42%. RLAT achieves 93.36% accuracy in variant dataset D 2 when the training set is only 1% of the original training set. Compared to Transformer, the accuracy is improved by 18.15%, compared to the gMLP, improved by 75.75%, compared to LSTM and GRU, improved by more than 44.01%, compared to TCN, 1D-CNN, and XCM, more than improved by 2.42%. RLAT can extract valid information from limited samples more efficiently, while Transformer exhibits severe sample dependence and performs poorly in limited sample experiments. In addition, the vulnerability of gMLP in limited sample recognition tasks is also reflected by its near failure at low sample amounts.

RLAT can achieve remarkable recognition when the training sample is only 274, while other methods are more dependent on the number of samples and seriously affect the training performance when the training sample plummets. The results indicate that RLAT has more outstanding generalization under the limited sample condition and can more effectively recognize HRRP sequences under non-cooperative targets. Since RLAT utilizes LAU for feature enhancement, which can extract local and global multi-level features and dynamically adjust the model depth, making feature extraction more effective. At the same time, Label Smoothing can reduce the dependence of RLAT on training samples

and enhance the generalization performance, so that RLAT can still efficiently recognize variant targets under limited sample conditions. To present the results of the comparison experiments more visually, the comparison experiments with limited training data are shown in Fig. 10.

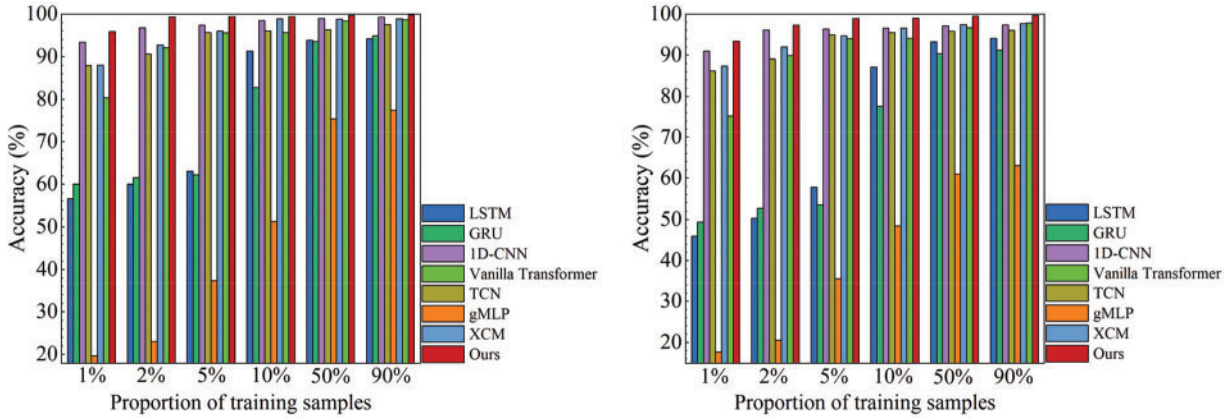


Figure 10: Recognition accuracy for limited training data of different methods

3.6 Position Encoding Comparison Experiments

RLAT is a network based on the self-attention mechanism, which is insensitive to the position information of HRRP sequences. Hence, position encoding is necessary to add position information to HRRP sequences for more efficient time-series feature extraction. Currently, commonly used position encoding mainly includes absolute position encoding and relative position encoding. Absolute position encoding LAPE [31] and relative position encoding T5 [19], XLNET [32] and DEBERTa [20], and Rotary Position Encoding (RoPE) [33] are selected for comparison experiments to verify the different validity for HRRP sequence recognition. The experimental results are shown in Table 7.

Table 7: Recognition accuracy of different positional encoding methods (%)

	D 1					D 2				
Methods	T5	XLNET	DEBERTa	LAPE	RoPE	T5	XLNET	DEBERTa	LAPE	RoPE
RLAT	99.64	99.67	99.74	99.39	**99.86**	98.39	98.05	98.05	98.50	**99.73**

As shown in Table 7, for standard dataset D 1, RoPE achieves 99.86% recognition accuracy, which is more than 0.12% better than other relative position encoding methods and 0.45% better than absolute position coding, and the recognition accuracy of relative position encoding is slightly higher than that of absolute position encoding. For variant dataset D 2, RoPE achieves 99.73% recognition accuracy, which is more than 1.34% better than other relative position encoding methods, and 1.23% better than the absolute position encoding method. However, the recognition accuracy of absolute position encoding is slightly higher than relative position encoding.

To analyze the recognition performance of various position encoding methods more visually, Fig. 11 shows the recognition performance of 5 position encoding methods for the standard dataset D 1.

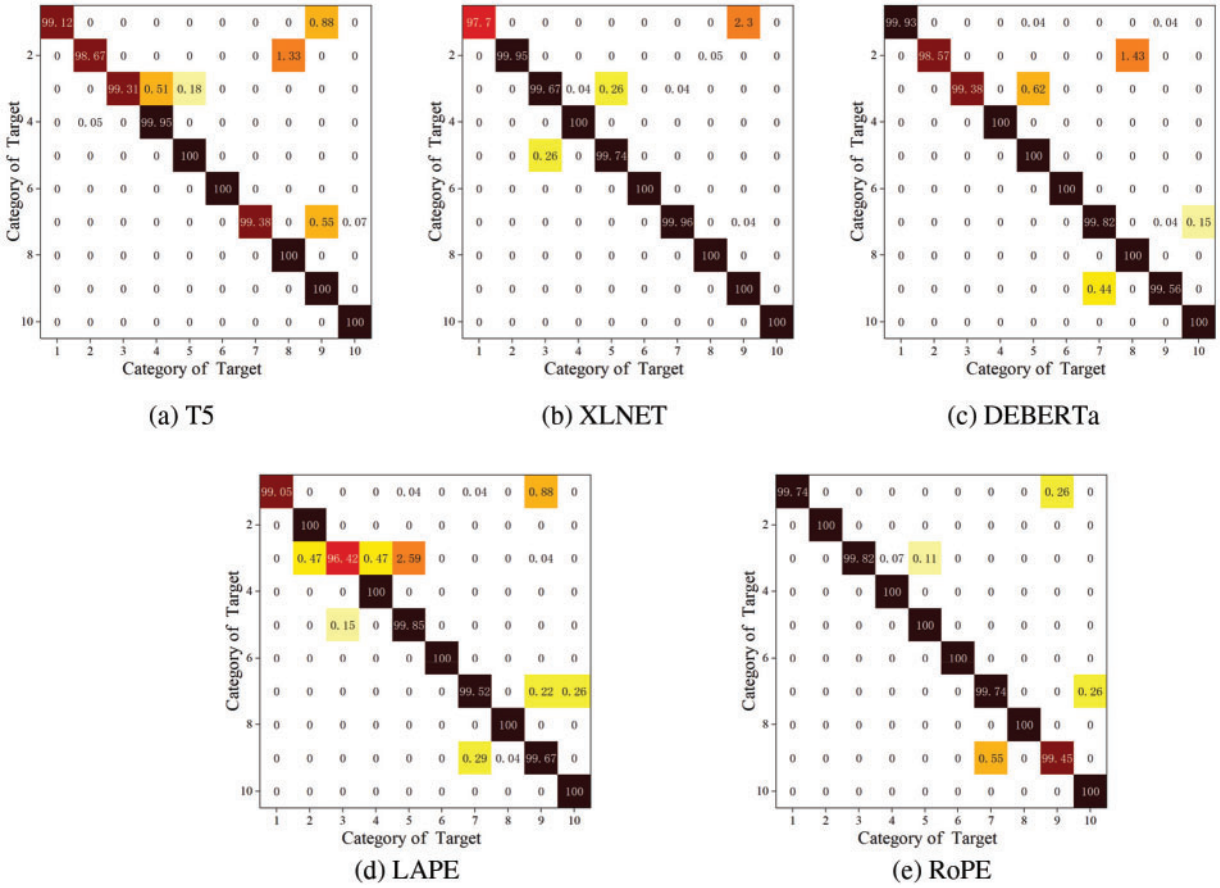
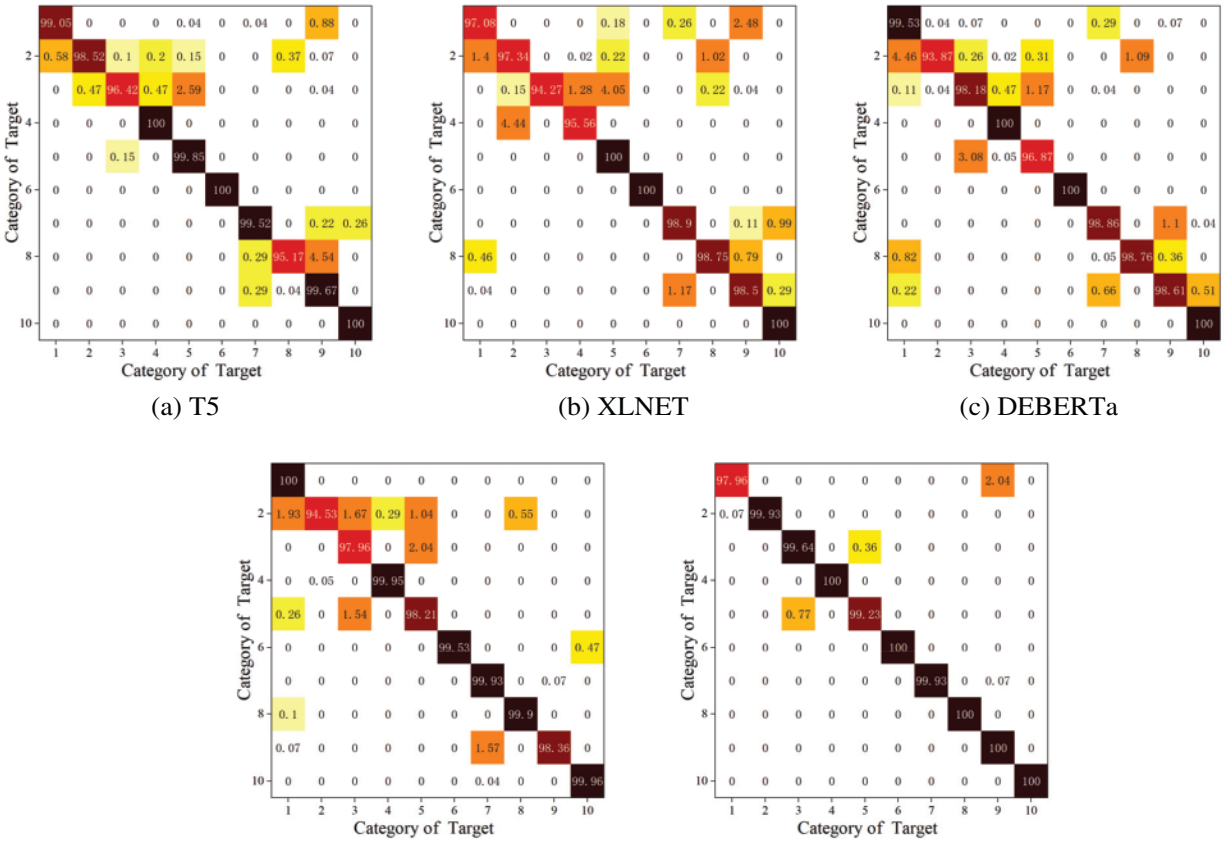


Figure 11: Confusion matrix for different positional encoding methods on D 1

As shown in Fig. 11, for the MSTAR standard dataset D 1, which contains ten categories of military targets, RoPE achieves optimal recognition accuracy for eight categories of targets with a maximum fluctuation in the accuracy of 0.55%, XLNET achieves optimal performance for six categories of targets with a maximum fluctuation of 2.30%, T5 achieves optimal performance for five categories of targets with a maximum fluctuation of 1.33%, DEBERTa achieves optimal performance for five types of targets with a maximum fluctuation of 1.43%, and LAPE achieves optimal performance for five types of targets with a maximum fluctuation of 3.58%. The results show that the recognition performance of RoPE is significantly better than other methods because RoPE has the advantages of both absolute position encoding and relative position encoding, which is more conducive to extracting temporal features. Meanwhile, relative position encoding not only has a higher accuracy than the absolute position encoding method but also has a more balanced recognition performance. Since relative position encoding can extract the relative information between HRRP sequences more effectively, it is beneficial to extract the temporal correlation.

As shown in Fig. 12, for the MSTAR variant dataset D 2, variant targets were added to the test set. RoPE achieved optimal recognition performance for eight categories of targets with a maximum fluctuation in the accuracy of 2.04%, XLNET achieved optimal performance for three categories of targets with a maximum fluctuation of 5.73%, T5 achieved optimal performance for three categories of targets with a maximum fluctuation of 3.58%, DEBERTa achieved optimal performance for

three categories of targets with a maximum fluctuation of 6.13%, and LAPE achieves the optimal performance for three categories of targets with a maximum fluctuation of 6.13%. The results show that the recognition performance of RoPE is significantly better than other methods. Furthermore, the average recognition accuracy of absolute position encoding is higher than other relative position encoding methods. Since the variant dataset has 36% more variant samples, which requires a higher generalization of the recognition method, relative position encoding introduces more parameters in the self-attention mechanism, leading to an overfitting problem in the recognition of variant samples. As RoPE has the advantages of both absolute relative position encoding and relative position encoding, which is more conducive to the extraction of temporal features and more robust to variant samples.



(a) T5 (b) XLNET (c) DEBERTa

(d) LAPE (e) RoPE

Figure 12: Confusion matrix for different positional encoding methods on D 2

3.7 Effect of Significant Hyperparameters

Hyperparameters play a crucial role in deep learning models. For RLAT, the three hyperparameters of LAU mapping dimension E, the number of stacked layers of LAT M, and the maximum depth of LAU are essential to the model performance, among which the width of a single LAU can be effectively controlled by E, and the number of stacked layers of LAT and the maximum depth of LAU can affect the feature extraction ability of the model in terms of model depth. Meanwhile, three hyperparameters are coupled with each other, so the three hyperparameters are combined to verify their effects on the model. Set the range of mapping dimension $Embedding = \{64, 128, 256, 512\}$, the

range of stacking layers $M = \{1, 2, 3, 4, 5, 6, 7, 8\}$, and the L L AU maximum depth $L = \{2, 4, 6, 8, 10\}$, where Figs. 13a–13c show the experimental results in MSTAR standard dataset D 1 and Figs. 13d–13f show the experimental results in MSTAR variant dataset D 2.

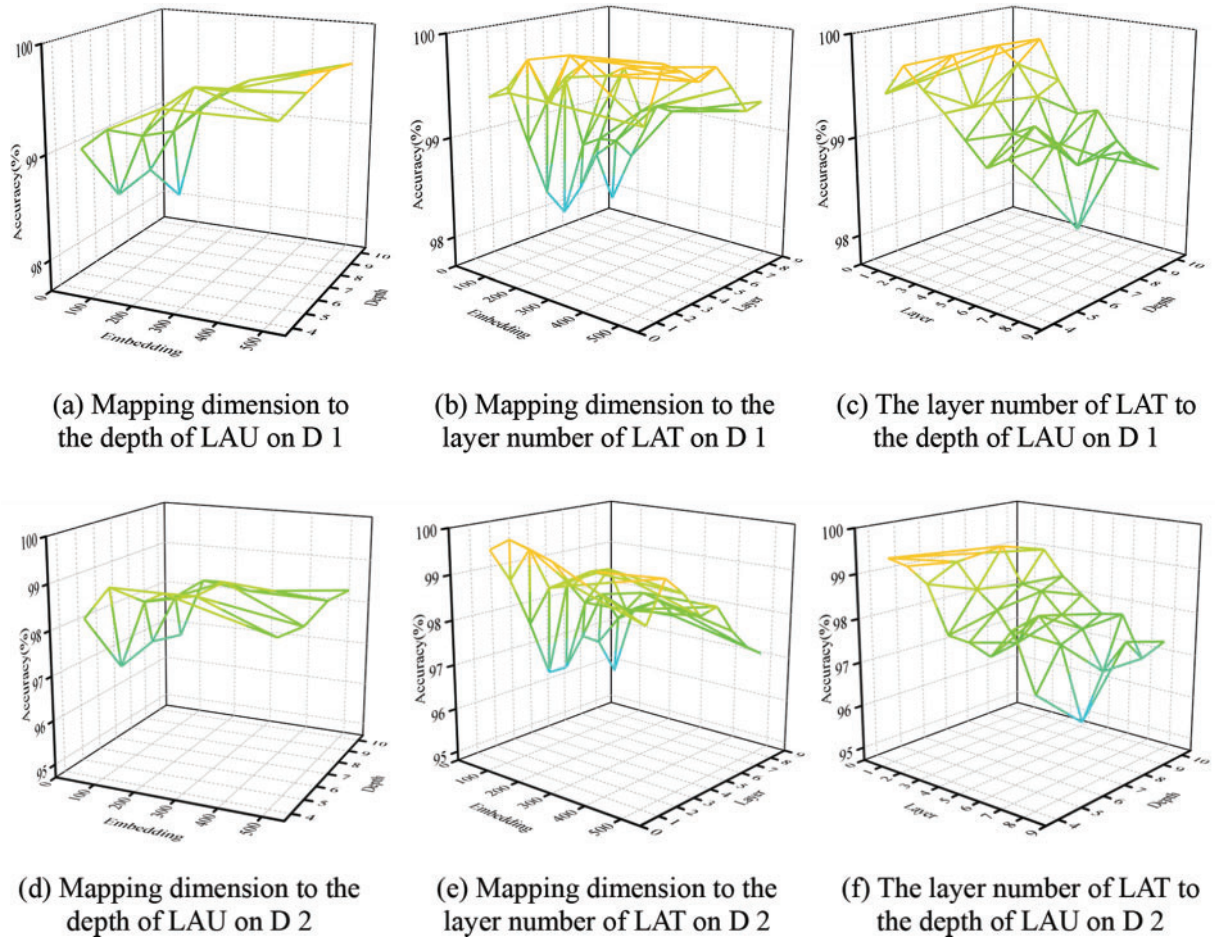


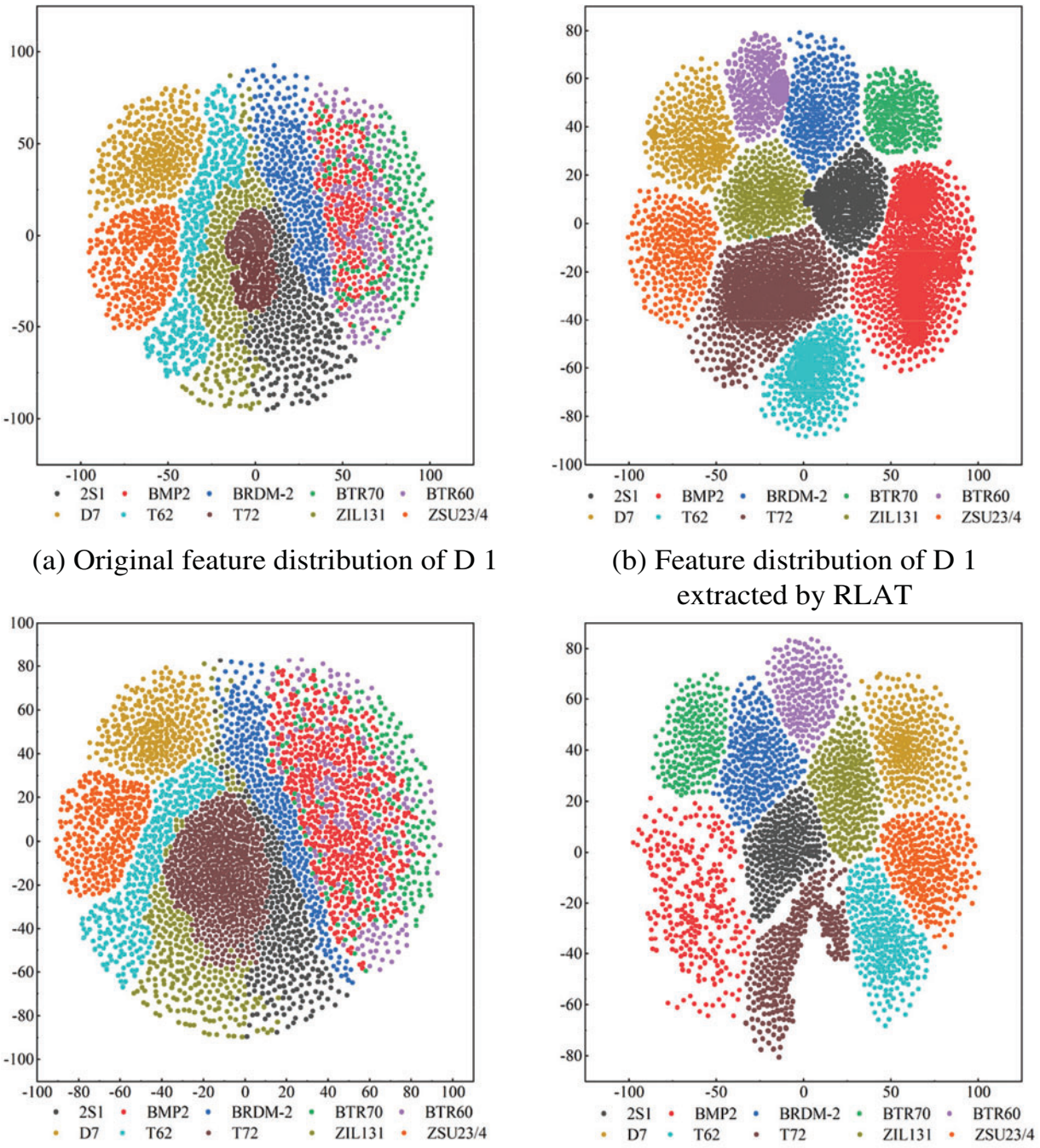
Figure 13: The influence of essential hyperparameters

For the MSTAR standard dataset D 1, as shown in Fig. 13a, the recognition performance of the proposed method increases with the increase of and L. Since the width and depth of LAU increase with the increase of E and L, which can extract richer and more abstract features. As shown in Fig. 13b, the recognition performance of the proposed method shows an increasing trend with the increase of E and shows an increasing and then decreases trend with the increase of M, which shows that too heavy stacking of LATs will harm the recognition performance instead. Too deep models will lead to a sharp increase in the number of parameters, resulting in a severe overfitting problem of the model. As shown in Fig. 13c, the recognition performance of the proposed method tends to increase and then decrease with the increase of M, and shows a slow growth trend with the increase of L. It can be concluded that and L mainly affects the width and depth of LAUs, so they show a positive correlation. While M controls the depth of the whole model, it shows a trend of first increasing and then decreasing. When $M = 2$, the recognition performance is the best. Since the HRRP sequence contains a large number of noisy regions, too deep models can lead to increased overfitting problems.

For the MSTAR variant dataset D 2, the generalization performance of the model is considered to be more challenging due to the addition of a large number of variant samples. As shown in Fig. 13d, the recognition performance of the proposed method tends to increase and then decrease with the increase of E and , because the width of LAU and depth of LAT increase with the increase of E and L, which can extract richer and more abstract features. However, the overfitting problem will be more significant on the variant dataset with increased parameters. As shown in Fig. 13e, the recognition performance of the proposed method shows a trend of increasing and then decreasing with the increase of E and M. The appropriate model width and depth are beneficial to enhance the feature extraction. However, as the width and depth of the model increase, it will lead to a dramatic increase in the number of parameters and aggravate the overfitting problem on the variant dataset. As shown in Fig. 13f, the recognition performance of the proposed method shows a trend of increasing and then decreasing with the increase of M and L. It can be concluded that, unlike the standard dataset, there are certain differences between the training set samples and test set samples in the variant dataset, which are highly susceptible to overfitting problems. Therefore, when the width and depth of the model increase, it shows a trend of first increasing and then decreasing. The recognition performance is best when $E = 128, M = 2, L = 6$. Models that are too wide and deep will extract a large amount of redundant noise information, which can lead to the aggravation of overfitting problems. The research on the variant dataset is the practical demands. Considering the complex real-world circumstances, researching variant datasets is more relevant and necessary. We will investigate the influence of different scales of variant datasets in future work.

3.8 The Visualization of Feature Extraction

To verify the effectiveness of the proposed method for feature extraction, the features are visualized using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm, which reduces the features to 2 dimensions. Figs. 14a–14b show the original feature distribution and the feature distribution extracted by RLAT for dataset D 1, respectively, and Figs. 14c–14d shows the original feature distribution and the feature distribution extracted by RLAT for dataset D 2, respectively. Fig. 14a shows the distribution of the original features of the ten categories of targets with a high degree of sample overlap and poor discrimination. After feature extraction by RLAT, the feature distribution of Fig. 14b is highly distinguishable, with small intra-class distance and large inter-class distance, which is significantly distinguishable. The distribution distinguishability of the original features of Fig. 14c is poorer compared with Fig. 14a because many variant samples are added, leading to more serious sample confusion and increasing classification difficulty. Fig. 14d has a higher distinction of feature distribution with a significant improvement compared to Fig. 14c, which can effectively achieve the recognition task. Both the feature distributions in Figs. 14b and 14d achieve significant distinguishability compared to the original distribution, illustrating the effectiveness of RLAT feature extraction. Since LAU can deepen the model depth dynamically, it is conducive to extracting the essential abstract features of HRRP sequences and can achieve a more effective temporal feature representation.



(a) Original feature distribution of D 1

(b) Feature distribution of D 1 extracted by RLAT

(c) Original feature distribution of D 2

(d) Feature distribution of D 2 extracted by RLAT

Figure 14: Visualization of RLAT feature extraction

4 Conclusions

This paper explores the application of Transformer in HRRP sequence recognition, and proposes a lightweight Transformer-based HRRP sequence recognition method called RLAT, which utilizes a more lightweight rotary position encoding, local-aggregated attention units, lightweight feedforward neural networks, and Label Smoothing to outperform other baseline methods in real scenes significantly. Besides, RLAT effectively reduces the number of parameters and computation of the model, which helps the application and deployment in edge devices. This paper also explores the recognition performance of the proposed method under variant targets and limited samples, and verifies that the generalization performance of the proposed method is significantly better than other methods. Finally, this paper further investigates the effect of position encoding on recognition performance and the effect of essential hyperparameters of the proposed method, which shows that RoPE can represent the relative position information between temporal features more effectively than other position encoding methods, and the hyperparameters have a significant impact on the recognition performance of RLAT, especially the number of stacked layers of LAT. Future work will further improve the lightweight level of the model, improve the recognition performance of the model under limited samples and variant targets, and extend the proposed method to the research work on open-set recognition of HRRP.

Acknowledgement: The authors would like to thank the editors and reviewers for their review and recommendations.

Funding Statement: This work was supported by the National Natural Science Foundation of China (Grant Numbers 61876189, 61703426, 61273275); the Young Talent Fund of University Association for Science and Technology in Shaanxi, China (Grant Number 20190108); and the Innovation Talent Supporting Project of Shaanxi, China (Grant Number 2020KJXX-065).

Author Contributions: Conceptualization, W. X. and W. P.; methodology, W. X. and W. P.; software, X. Q. and L. J.; validation, W. P., X. Q. and L. J.; formal analysis, W. X. and W. P.; resources, W. P. and L. J.; writing—original draft preparation, W. P.; writing—review and editing, X. Q.; supervision, W. X. and S. Y.; project administration, W. X. and S. Y.; funding acquisition, W. X. and S. Y. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The MSTAR datasets we used are a publicly available dataset, which can be downloaded in the Air Force Moving and Stationary Target Recognition Database at https://www.sdms.afrl.af.mil.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Chen, L. Du and L. Y. Liao. "Survey of radar HRRP target recognition based on parametric statistical model," *Journal of Radars*, vol. 11, no. 6, pp. 1020–1047, 2022.

[2] J. Chen, L. Du, G. B. Guo, L. W. Yin and D. Wei, "Target-attentional CNN for radar automatic target recognition with HRRP," *Signal Processing*, vol. 196, pp. 108497, 2022.

[3] X. D. Liu, L. Wang and X. R. Bai, "End-to-end radar HRRP target recognition based on integrated denoising and recognition network," *Remote Sensing*, vol. 14, no. 20, pp. 5254, 2022.

[4] C. L. Lin, T. P. Chen, K. C. Fan, H. Y. Cheng and C. H. Chuang, "Radar high-resolution range profile ship recognition using two-channel convolutional neural networks concatenated with bidirectional long short-term memory," *Remote Sensing*, vol. 13, no. 7, pp. 1259, 2021.

[5] X. D. Wang, R. Li, J. Wang, L. Lei and Y. F. Song, "One-dimension hierarchical local receptive domains based extreme learning machine for radar target HRRP recognition," *Neurocomputing*, vol. 418, pp. 314–325, 2020.
[6] Q. Xiang, X. D. Wang, J. Lai, Y. F. Song, R. Li *et al.,* "Multiscale group-fusion convolutional neural network for high-resolution range profile target recognition," *IET Radar, Sonar & Navigation*, vol. 16, no. 12, pp. 1997–2016, 2022.
[7] X. D. Wang, P. Wang, Y. F. Song and J. T. Li, "Recognition of HRRP sequence based on TCN with attention and elastic net regularization," in *Proc. of 2022 Int. Conf. on Image Processing, Computer Vision and Machine Learning (ICICML)*, Xi'an, China, pp. 346–351, 2022.
[8] Y. F. Zhang, F. C. Qian and F. Xiao, "GS-RNN: A novel RNN optimization method based on vanishing gradient mitigation for HRRP sequence estimation and recognition," in *Proc. 2020 IEEE 3rd Int. Conf. on Electronics Technology (ICET)*, Chengdu, China, pp. 840–844, 2020.
[9] X. Peng, X. Z. Gao, Y. F. Zhang and X. Li. "An adaptive feature learning model for sequential radar high resolution range profile recognition," *Sensors*, vol. 17, no. 7, pp. 1675, 2017.
[10] W. A. Timothy and C. G. Steven, "Hidden markov models for classifying SAR target images," in *Proc. SPIE Algorithms for Synthetic Aperture Pader Imagery XI*, Orlando, Florida, USA, pp. 302–308, 2004.
[11] L. Du, H. W. Li, Z. Bao and J. Y. Zhang, "A Two-distribution compounded statistical model for radar HRRP target recognition," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2226–2238, 2006.
[12] P. Molchanov, K. Egiazarian, J. Astola, A. Totsky, S. Leshchenko *et al.,* "Classification of aircraft using micro-doppler bicoherence-based features," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 2, pp. 1455–1467, 2014.
[13] L. Lei, X. D. Wang, Y. Q. Xing and K. Bi, "Multi-polarized HRRP classification by SVM and DS evidence theory," *Control and Decision*, vol. 28, no. 6, pp. 861–866, 2013.
[14] Q. Xiang, X. D. Wang, Y. F. Song, L. Lei, R. Li *et al.,* "One-dimensional convolutional neural networks for high-resolution range profile recognition via adaptively feature recalibrating and automatically channel pruning," *International Journal of Intelligent Systems*, vol. 36, no. 1, pp. 332–361, 2021.
[15] C. Du, L. Tian, B. Chen, L. Zhang, W. Chen *et al.,* "Region-factorized recurrent attentional network with deep clustering for radar HRRP target recognition," *Signal Processing*, vol. 183, pp. 108010, 2021.
[16] M. Pan, A. L. Liu, Y. Z. Yu, P. H. Wang, J. J. Li *et al.,* "Radar HRRP target recognition model based on a stacked CNN–Bi-RNN with attention mechanism," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
[17] L. Zhang, C. Han, Y. H. Wang, Y. Li and T. Long, "Polarimetric HRRP recognition based on feature-guided transformer model," *Electronics Letters*, vol. 57, no. 18, pp. 705–707, 2021.
[18] Y. J. Diao, S. W. Liu, X. Z. Gao and A. F. Liu, "Position embedding-free transformer for radar HRRP target recognition," in *Proc. of 2022 IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, Kuala Lumpur, Malaysia, pp. 1896–1899, 2022.
[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang *et al.,* "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
[20] P. C. He, X. D. Liu, J. F. Gao and W. Z. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," arXiv Preprint arXiv:2006.03654, 2020.
[21] R. Müller, S. Kornblith and G. Hinton, "When does label smoothing help?" in *Proc. of 33rd Conf. on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.
[22] B. Erik, M. Uttam, Z. Edmund and V. Vincent, "Review of recent advances in AI/ML using the MSTAR data," in *Proc. of SPIE Algorithms for Synthetic Aperture Radar Imagery XXVII*, Orlando, Florida, USA, pp. 113930C, 2020.
[23] Y. F. Zhang, F. Xiao, F. C. Qian and X. Li, "VGM-RNN: HRRP sequence extrapolation and recognition based on a novel optimized RNN," *IEEE Access*, vol. 8, pp. 70071–70081, 2020.

[24] Y. F. Zhang, X. Z. Gao, X. Peng, J. Q. Ye and X. Li, "Attention-based recurrent temporal restricted boltzmann machine for radar high resolution range profile sequence recognition," *Sensors*, vol. 18, no. 5, pp. 1585, 2018.
[25] F. Karim, S. Majumdar, H. Darabi and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019.
[26] N. Elsayed, A. S. Maida and M. Bayoumi, "Deep gated recurrent and convolutional network hybrid model for univariate time series classification," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 654–664, 2019.
[27] S. J. Bai, J. Z. Kolter and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.
[28] H. X. Liu, Z. H. Dai, D. R. So and Q. V. Le, "Pay attention to mlps," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9204–9215, 2021.
[29] K. Fauvel, T. Lin, V. Masson, É. Fromont and A. Termier, "XCM: An explainable convolutional neural network for multivariate time series classification," *Mathematics*, vol. 9, no. 23, pp. 3137, 2021.
[30] M. H. Liu, S. Q. Ren, S. Y. Ma, J. H. Jiao, Y. Z. Chen *et al.,* "Gated transformer networks for multivariate time series classification," arXiv preprint arXiv:2103.14438, 2021.
[31] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. of the 34th Int. Conf. on Machine Learning (ICML)*, Sydney, NSW, Australia, pp. 1243–1252, 2017.
[32] Z. H. Dai, Z. L. Yang, Y. M. Yang, J. Carbonell, Q. V. Le *et al.,* "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2978–2988, 2019.
[33] J. L. Su, Y. Lu, S. F. Pan, A. Murtadha, B. Wen *et al.,* "RoFormer: Enhanced transformer with rotary position embedding," arXiv preprint arXiv:2104.09864, 2021.