Marine Navigation Radar Target Detection with Multi-scale Spatiotemporal Feature Fusion-based Deep Learning

1st Yangyang Liu
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, China
yyliu@shu.edu.cn

2rd Xiao Huang
China Ship Development and Design Center
Wuhan, China
huangxiao_88@outlook.com

3rd Hao Tao
China Ship Development and Design Center
Wuhan, China
chst722@163.com

4nd Zhiwei Xia
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, China
xzw_shu@shu.edu.cn

5nd Xinfeng Xu
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, China
xuxinfeng@shu.edu.cn

6nd Chun Liu
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, China
Chun_Liu@shu.edu.cn

2024 7th International Symposium on Autonomous Systems (ISAS) | 979-8-3503-6317-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS61044.2024.10552603

***Abstract*—This article endeavors to tackle the challenge of radar target detection in dynamic scenarios characterized by strong maneuvering, such as those involving rotating radar carriers or high-speed moving targets. A deep learning model, termed the multi-scale spatiotemporal fusion network (MSSF-Net), is proposed to effectively tackle the aforementioned scenarios. The proposed method encompasses several modules, with a notable inclusion being the improved Faster R-CNN module, which incorporates attention and spatial transformation mechanisms. This integration enables the extraction of spatial offset features from the current frame's echo image, facilitating the suppression of sea clutter and noise. Concurrently, the spatiotemporal feature fusion module within the network is designed to extract spatiotemporal information from historical frame echo images. It conducts multi-scale and multi-level feature fusion with the feature maps generated by other modules. This enables the capture of profound spatiotemporal features from a global perspective, thereby mitigating the decline in detection accuracy induced by the spatial shift of echo information resulting from strong maneuvering. Other modules within the method serve to manage the input and output of the model. Simulation results validate the enhanced accuracy of radar target detection achieved by the method under conditions of strong maneuvering.**

***Index Terms*—Radar target detection, deep learning, multi-scale spatiotemporal fusion network, spatiotemporal features, strong maneuvering.**

I. Introduction

Marine Navigation Radar Target Detection is of paramount importance in maritime operations. The ability to discern and locate objects amidst the vast expanses of the ocean underpins a multitude of activities, from civilian shipping [1] to military operations [2]. In [3], a constant false alarm rate (CFAR) detection algorithm, utilizing the correlation between linear measurements of the radar intermediate frequency signal and the sensing matrix, is proposed to realize target detection. Building upon CFAR, the strategy introduced in [4] incorporates a joint estimation approach for range-Doppler thresholds, effectively suppressing clutter. Adaptive doppler beam sharpening (DBS) technology, as employed in [5], addresses multi-target detection in marine environments through adaptive thresholds and density-based clustering techniques. However, the complex and dynamic nature of the marine environment introduces numerous disturbances and uncertainties. These factors often induce significant errors in target detection algorithms, thereby challenging the accuracy of target detection, including ranging and direction finding, in marine navigation radars.

In the wake of the swift advancements in the field of artificial intelligence, there has been a proliferation of object detection methodologies predicated on deep learning (DL). More often than not, these DL-based approaches demonstrate superior detection performance, outpacing traditional methods. This enhanced performance can be attributed to the ability of DL models to learn complex patterns and make sense of large

This work was supported by National Natural Science Foundation of China (62103250 and 62336005); Shanghai Sailing Program (21YF1414000); Project of Science and Technology Commission of Shanghai Municipality, China (22JC1401401).

volumes of data, thereby increasing the accuracy and reliability of object detection. By combining DL with CFAR to produce variable thresholds, [6] can precisely control the probability of false alarms, thereby achieving more accurate object detection results. In [7], a marine-faster R-CNN model is established, which extracts features from PPI images generated by radar echoes for target detection. In addressing the challenges of radar target detection in cluttered maritime environments, DL is employed in [8] to extract features from multiple data sources. This process of feature extraction and subsequent fusion significantly enhances the detection performance of ocean targets.

These existing methods leverage DL to extract features from various data sources, leading to improved detection performance for ocean targets. However, challenges persist, particularly in scenarios involving strong maneuvers and clutter. The spatial offset characteristics of radar echoes during intense movements pose accuracy limitations. Addressing these issues remains a significant research focus. The dynamic marine environment introduces additional complexities. Factors such as radar carrier rotation and high-speed target movement challenge the accuracy of target detection. Furthermore, clutter from ocean currents, sea winds, and other environmental variables exacerbates the situation. Researchers must grapple with these dynamic conditions to enhance detection accuracy and robustness in real-world scenarios .

The major contributions are summarized as follows:

1) Contrasted with traditional radar target detection algorithms designed for general detection scenarios [10], this article employs a deep learning network featuring multi-level and multi-scale feature fusion to address the problem associated with multi-target radar detection in high-maneuvering situations.
2) The model is capable of extracting spatiotemporal variation information from current and historical sequence echo images, and performing multi-scale feature fusion to capture global spatiotemporal variation features. This results in a feature vector with radar echo spatiotemporal information, providing a global perspective for radar target detection under strong maneuvers.
3) Attention mechanism is applied in the improved Faster R-CNN to extract pertinent feature information and captures spatial offset features caused by strong maneuvers from input echo images, simultaneously eliminating invalid features and clutter information.

The remainder of this paper is organized as follows. Radar echo image preprocessing is stated in Section II. In Section III, the multi-scale feature fusion-based radar target detection DL algorithm is elaborated in detail. Section IV presents the simulation results pertaining to the radar target detection algorithms. Concluding remarks and interpretations are subsequently presented in Section V.

II. Radar Echo Image Preprocessing

In this article, subsequent to receiving the raw echo signal, the radar undergoes minimalistic operations and filtering, followed by flattening the echo data received by the navigation radar into two-dimensional echo images. This approach eliminates the necessity for intricate data filtering and conversion processes. The intent is to meticulously preserve the characteristics of the original echo signal, mitigating the risk of intermediate feature loss induced by echo signal conversion. Fig. 1 delineates a comprehensive process for radar echo image processing, which is applicable for model input.

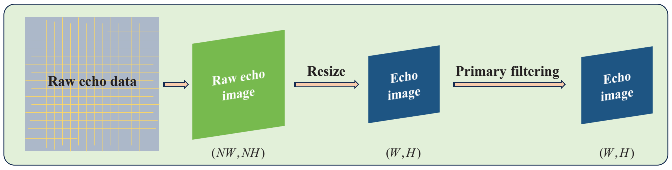

Fig. 1. The procedure for processing raw radar echo signal into radar echo image.

For the received radar echo signal, trim the weak echo signal at its periphery to mitigate interference from faint or extraneous information. This process is also beneficial for standardizing sizes, facilitating the subsequent flattening of echo images into a uniform format of (NW, NH). Following this, the echo image undergoes resizing to smaller dimensions of (W, H), where $NW > N, NH > H$, aimed at diminishing the input size of the model. Subsequently, a minimalistic filtering process is applied for noise reduction and to prevent overfitting. The filtering is conducted after adjusting the image size, enabling the application of filtering on smaller images to reduce computational overhead.

In this paper, object detection utilizes not only the echo image from the current frame but also incorporates the echo image from historical frames. This inclusion assists in extracting deeper and higher-scale global features, thereby enhancing the accuracy of object detection. Every preprocessed echo image is stored in a buffer with a capacity of n to serve as a historical echo image. For the initial echo image generated, its historical echo image input is itself.

III. Multi-scale Spatiotemporal Fusion-based Radar Target detection Algorithm

The proposed MSSF-Net consists of several modules, namely, modules data preprocessing module, improved faster R-CNN module, spatiotemporal feature fusion module, and target orientation output module. These modules are designed for distinct functions and are interconnected, enabling the completion of radar target detection tasks through multi-level and multi-scale feature fusion. The structure of the MSSF-Net, detailed in the subsequent subsections, is depicted in Fig. 3.

A. Improved faster R-CNN

Faster R-CNN demonstrates relatively high accuracy in radar target detection tasks owing to its second-order precision in the two-stage detection process. Each input echo image in this module typically contains channel information, with the data shape represented as (C, H, W), where C signifies the channel. However, for radar echo images, channel information is often inconsequential. Hence, the channel is consistently set to $C = 1$ to reduce computational complexity and enhance model efficiency.

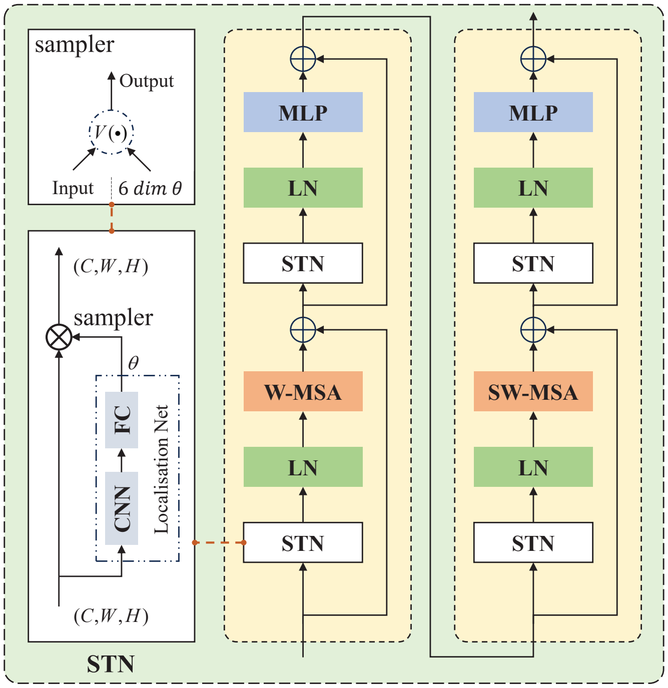

Fig. 2. Two consecutive Swin-Spatial-Transformer blocks.

In the model design, the convolutional layers (backbones) in the Faster R-CNN front-end module are prone to causing subsequent modules to lose the spatial offset features induced by rotational maneuvering in the initial input echo image. Consequently, the Swin-Spatial-Transformer is introduced to replace these layers for feature extraction. This transformation aims to extract crucial feature information through an attention mechanism, filter out irrelevant sea clutter information, and extract spatial features from echo images. This approach addresses, to a certain extent, the spatial offset problem in input echo images caused by rotation maneuvers.

The conventional Swin-Transformer [11] comprises image block segmentation (patch partition), linear embedding, and Swin-Transformer blocks. The core component is the Swin-Transformer Block, which, while capable of extracting certain spatial features, predominantly focuses on capturing dependency information between global features, lacking explicit representation of spatial features. To address this limitation, this study introduces the Swin-Spatial-Transformer (SST) block, a modification of the Swin-Transformer. This block is designed to explicitly extract spatial features by processing spatial transformations, forming the SST module. Fig. 2 illustrates two consecutive SST blocks, each consisting of multilayer perceptron (MLP), layer normalization (LN), convolutional neural networks (CNN), fully connected (FC), window multi-head self attention (W-MSA), and shifted window multi-head self attention (SW-MSA). While preserving the attention mechanism, the spatial transformer network (STN) [12] is embedded in the pre-feature extraction and convergence segments of the Swin-Transformer to enhance spatial feature representation during feature extraction. In radar target detection, to obtain the spatially transformed feature map, affine transformation is applied to the input feature map. Consequently, the θ in STN can employ six-dimensional parameters.

The input and output of STN are both feature maps. Assuming the input shape is (C, H, W) and the output shape is (C, H', W'), each feature V_i^c in channel C of the output feature map V can be represented as a function of the input feature map U with respect to the neural network parameters:

$$\begin{gathered} V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y), \\ \forall i \in [1 \dots H'W'], \forall c \in [1 \dots C] \end{gathered} \tag{1}$$

where U_{nm}^c represents the feature of input feature map U located at position (m, n) in channel C, the term $k(\cdot)$ represents a linear interpolation method, and Φ_x, Φ_y represent the parameters of the linear interpolation method. To ensure the network can perform backpropagation, $k(\cdot)$ can utilize bilinear interpolation [12].

The SST block incorporates attention mechanisms for feature extraction. STN operates both before and after the attention mechanism in SST, extracting multi-level, more abstract, and essential spatial features. The feature map output by Swin Spatial Transformer, denoted as $F1$, has a shape of (C_f, H_f, W_f) and is subsequently passed to both the spatiotemporal feature fusion module and the subsequent improved faster R-CNN module.

For the feature maps entering the subsequent modules of the improved faster R-CNN, deeper feature extraction is performed on the current frame echo image. The output of the post ROI pooling layer, denoted as $F3$, is a one-dimensional vector flattened from a two-dimensional feature map with a size of $(1, h)$. The feature $F3$ represents a comprehensive set of features extracted from the input echo image using the improved Faster R-CNN. In contrast to traditional Faster R-CNN, where separate proposal box features and classification features are provided, our approach fuses $F3$ with the global feature $F4$ generated by the spatiotemporal feature fusion module. This fusion strategy offers several advantages. It results in more comprehensive features, enhancing target detection accuracy by considering both target location and contextual information. Additionally, the complementary nature of $F3$ and $F4$ allows us to effectively capture target diversity and complexity. By reducing redundancy, it also improve overall efficiency in target detection.

B. Spatiotemporal feature fusion

The spatiotemporal feature fusion module, depicted in the proposed MSSF-Net shown in Fig. 3, is designed to integrate the feature maps of the current frame echo image and the historical frame echo image. Consequently, it represents a global feature map containing both temporal and spatial transformation features. This integration enhances the accuracy

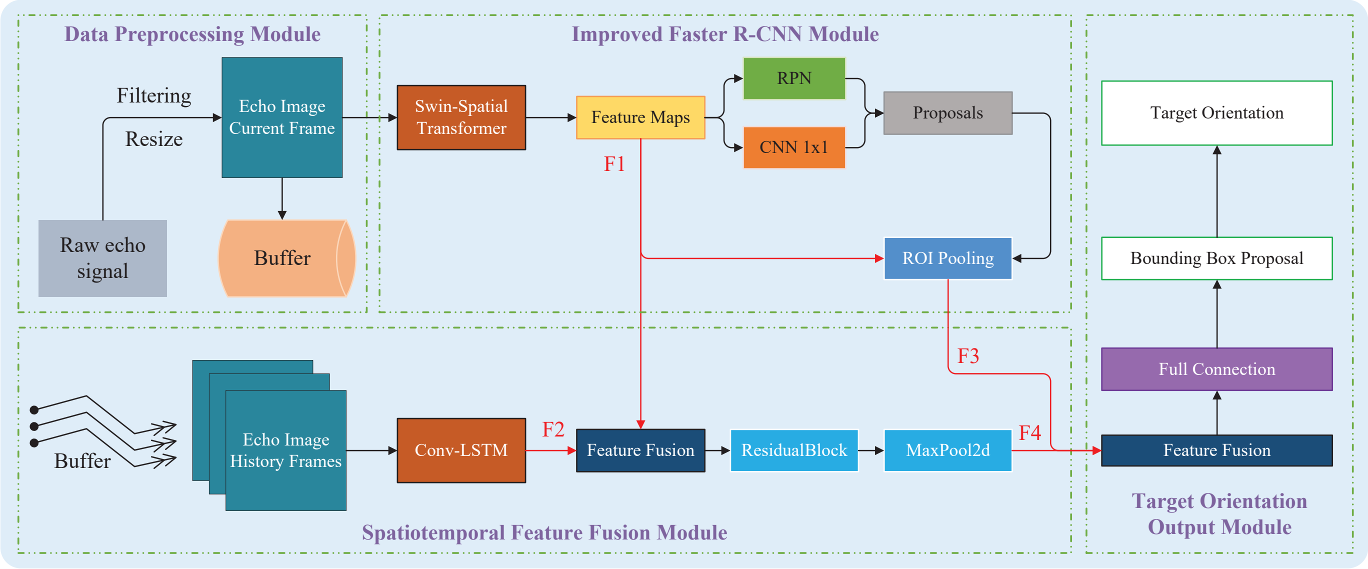

Fig. 3. The structure of the MSSF-Net.

of radar target detection, especially in the final suggestion box regression and azimuth output, thereby mitigating the challenge of reduced accuracy in target detection caused by strong motion.

The module initially extracts spatiotemporal features from historical frame echo images with a shape of (T, C, W, H) through Conv-LSTM, where $C = 1$ and $T = n$ represents the input frame number of the historical echo image. Subsequently, the output feature map $F2$ of this module takes the shape $(1, C_f, H_f, W_f)$. $F2$, the output feature map of Conv-LSTM, is fused with the received output feature map $F1$ from SST. Given the necessity to use this fused feature for multi-scale feature extraction in the future, the fusion method here adopts the concatenation method, preserving information from each feature. ResidualBlock within the spatiotemporal fusion module enhance the model's expressive power, alleviate representation bottlenecks, and address gradient vanishing problems. Concurrently, the number of channels for fused features is reduced to 1 for subsequent processing. MaxPool2d is applied to reduce the dimensionality of the data while retaining essential feature information, thereby reducing computational complexity and preventing overfitting. Fig. 4 illustrates the changes in shape of feature fusion and data flow within the spatiotemporal feature fusion module.

C. Target orientation output

The target orientation output module in the MSSF-Net integrates features from all pre-existing modules to obtain multi-scale and multi-level deep features. These features are then used to extract suggestion box information and orientation details of the target. The feature fusion incorporates the enhancement of the output feature $F3$ from Faster R-CNN, which is an k-dimensional vector, along with the output feature $F4$ from the spatiotemporal feature extraction fusion module. The shape of the fused deep feature is $(1, H_f \times W_f/4 + k)$.

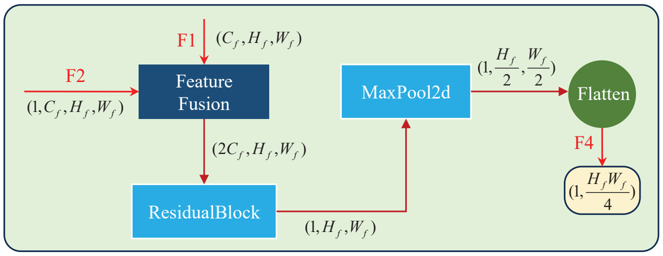

Fig. 4. The feature fusion and data flow in the spatiotemporal feature fusion module.

The multi-scale and multi-layer fusion features output the spatial position of the target's suggestion box through the final fully connected layer. Each suggestion box can be represented as a quaternion (X_i, Y_i, h_i, w_i), where (X_i, Y_i) denotes the ith coordinates of the center point of the suggestion box, and h_i, w_i represent the length and width of the corresponding suggestion box, respectively. Based on the target's coordinates (X_i, Y_i), the distance D_i from the target to the radar carrier and its direction φ_i can be calculated by:

$$\begin{aligned} D_i &= \sqrt{X_i^2 + Y_i^2} \\ \varphi_i &= \operatorname{atan2}(Y_i, X_i) \end{aligned} \tag{2}$$

The proposed MSSF Net extracts extensive global feature information through multi-level and multi-scale spatiotemporal feature fusion, enhancing object detection accuracy. While the model doesn't directly provide the orientation of the target, it deduces it through a suggestion box, which also aids in identifying the size and shape of the target. This capability is crucial for various marine tasks, particularly those requiring collision avoidance and path planning.

IV. Simulation results

Due to the high cost of acquiring an extensive radar echo dataset for simulation development, we augmented our existing dataset with open-source datasets from [13–15], primarily sourced from X-band microwave radar. Additionally, simulated radar echoes generated by the Matlab toolbox have also been utilized. The input radar echo signal undergoes the initial data preprocessing module in the deep network, where the radar data is processed and filtered, as illustrated in Fig. 5 - 6. The filter employed here is a basic high-pass filter, as more intricate and precise noise and clutter suppression is accomplished in subsequent networks.

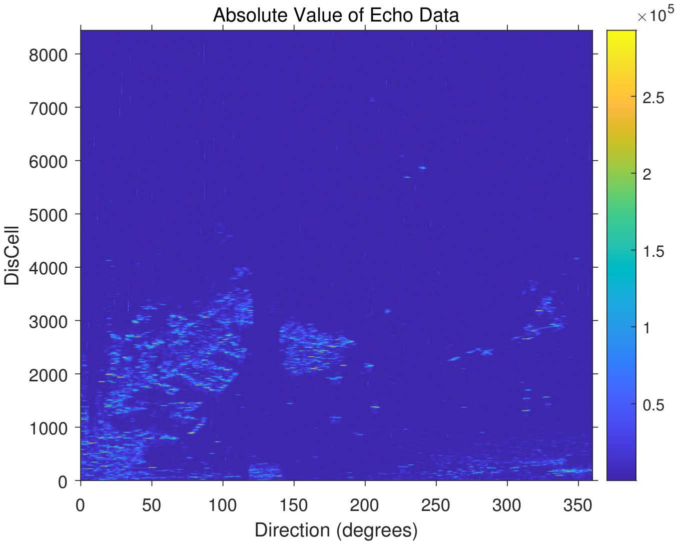

Fig. 5. 2D echo image from echo data

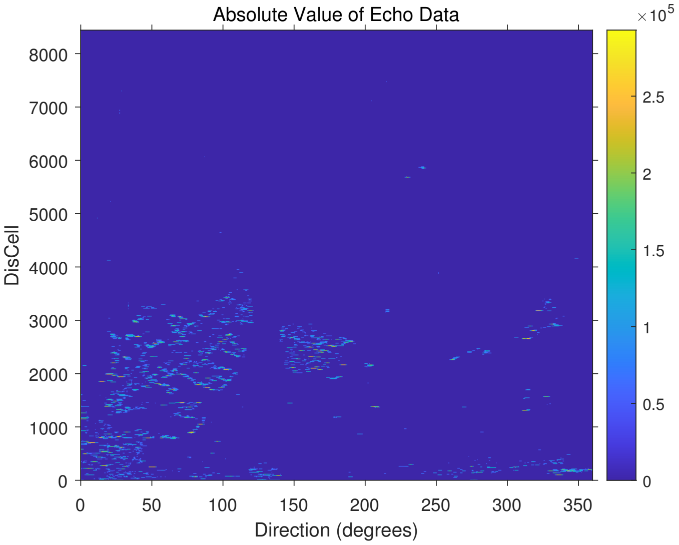

Fig. 6. Input echo image after simple filtering

As the detection distance increases, the accuracy of target detection gradually diminishes due to the reflective nature of radar echoes. Notably, the change in detection accuracy with increasing distance may not be consistent when comparing far and near distances. Consequently, the evaluation of target detection error in this paper is stratified into two detection ranges: namely, the near range $(0 - 2\text{km})$ and the far range $(> 2\text{km})$. TABLE I displays the distance measurement error and direction measurement error of the radar toward the target in the case of strong maneuvering, involving high-speed maneuvers of the radar carrier or the target. In contrast, TABLE II presents the same errors in scenarios without strong maneuvering. The distance measurement error and direction measurement error decrease under both non-strong maneuvers and strong maneuvers conditions, with a more significant reduction observed, especially under strong maneuvers. In strong maneuvering situations, the distance measurement error decreases by over 20% in both near and far ranges, while the direction measurement error decreases by 25%. The proposed MSSF Net demonstrates superior performance in both key indicators of radar target detection, demonstrating its effectiveness.

TABLE I
Detection errors under strong maneuvering conditions

Detection conditions	**Near range**		**Far range**	
Detection erro	Distance	Direction	Distance	Direction
Faster R-CNN	12.47Discell	4.23°	1.46%Range	2.55°
MSSF-NET	9.79Discell	3.18°	1.13%Range	1.71°
Improved	21.46%	25.14%	22.60%	32.94%

TABLE II
Detection errors under non strong maneuvering conditions

Detection conditions	**Near range**		**Far range**	
Detection erro	Distance	Direction	Distance	Direction
Faster R-CNN	9.16Discell	2.18°	1.17%Range	1.12°
MSSF-NET	8.48Discell	1.89°	1.05%Range	1.03°
Improved	7.39%	13.30%	10.13%	8.24%

V. Conclusion

This article proposes an MSSF Net to address the challenge of radar target detection, specifically aiming to reduce errors in distance and direction detection during strong maneuvering situations. In the proposed radar target detection method, there are two data input channels and multiple functional modules involved. The radar echo signal undergoes a unified preprocessing and a simple filtering process to generate echo images for both the current and historical frames, which serve as inputs for the MSSF-Net. For the echo image of the current frame, an improved Faster R-CNN network is designed, incorporating an SST block for clutter suppression and spatial feature extraction. In contrast, historical frame echo images leverage a combination of Conv-LSTM and residual blocks to extract spatiotemporal variation features. These features are then fused at multiple levels and scales within the network, ultimately feeding into the target orientation output module to generate target orientation information. The simulation results demonstrate an improvement in object detection accuracy compared to the traditional Faster R-CNN method under identical detection conditions. Given the limitations of the dataset in this article, our future work will prioritize training and validation on more diverse and general radar datasets.

References

[1] Q. Zhang, Y. Li, C. Guo, S. Yin, L. Ma, and Y. Zhu, "Marine radar monitoring iot system and case study of target detection based on ppi images," *Expert Systems*, 2023.

[2] J. Wang and S. Li, "Maritime radar target detection in sea clutter based on cnn with dual-perspective attention," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[3] Z. Cao, J. Li, C. Song, Z. Xu, and X. Wang, "Compressed sensing-based multitarget cfar detection algorithm for fmcw radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9160–9172, 2021.

[4] C. Kuang, C. Wang, B. Wen, Y. Hou, and Y. Lai, "An improved ca-cfar method for ship target detection in strong clutter using uhf radar," *IEEE Signal Processing Letters*, vol. 27, pp. 1445–1449, 2020.

[5] A. Pirkani, D. Kumar, L. Daniel, E. Hoare, M. Cherniakov, and M. Gashinova, "Dynamic multi-target detection and focus in maritime conditions," in *2023 20th European Radar Conference (EuRAD)*, pp. 510–513, 2023.

[6] Y. Zhu, Y. Li, and Q. Zhang, "False-alarm-controllable radar target detection by differentiable neyman pearson criterion for neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[7] X. Chen, X. Mu, J. Guan, N. Liu, and W. Zhou, "Marine target detection based on marine-faster r-cnn for navigation radar plane position indicator images," *Frontiers of Information Technology & Electronic Engineering*, vol. 23, no. 4, SI, pp. 630–643, 2022.

[8] X. Wang, Y. Wang, X. Chen, C. Zang, and G. Cui, "Deep learning-based marine target detection method with multiple feature fusion," 2023.

[9] Z. Esmaeilbeig, A. Eamaz, K. V. Mishra, and M. Soltanalian, "Moving target detection via multi-irs-aided ofdm radar," in *2023 IEEE Radar Conference (RadarConf23)*, pp. 1–6, 2023.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2016.

[13] J. Guan, N. B. Liu, G. Q. Wang, and et al. "Sea-detecting X-band radar and data acquisition program," *Journal of Radars*, vol. 8, no. 5, pp. 656-667, 2019.

[14] J. Guan, H. Ding, Y. Huang, and et al. "Annual Progress of Sea-detecting X-band Radar and Data Acquisition Program," *Journal of Radars*, vol. 10, no. 1, 2021.

[15] J. Guan, N. B. Liu, G. Q. Wang, and et al. "Sea-detecting radar experiment and target feature data acquisition for dual polarization multistate scattering dataset of marine targets," *Journal of Radars*, vol. 12, no. 2, pp. 456-469, 2023.