Rethinking SAR ATR: A Target-Aware Frequency-Spatial Enhancement Framework with Noise-Resilient Knowledge Guidance

Yansong Lin[a], Zihan Cheng[a], Jielei Wang[a,b], Guoming Lu[a,b,*], Zongyong Cui[c]

[a]*the Institute of Intelligent Computing, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China*
[b]*Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province, Chengdu, 611731, China*
[c]*the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China*

Abstract

Synthetic aperture radar automatic target recognition (SAR ATR) is of considerable importance in marine navigation and disaster monitoring. However, the coherent speckle noise inherent in SAR imagery often obscures salient target features, leading to degraded recognition accuracy and limited model generalization. To address this issue, this paper proposes a target-aware frequency-spatial enhancement framework with noise-resilient knowledge guidance (FSCE) for SAR target recognition. The proposed framework incorporates a frequency–spatial shallow feature adaptive enhancement (DSAF) module, which processes shallow features through spatial multi-scale convolution and frequency-domain wavelet convolution. In addition, a teacher–student learning paradigm combined with an online knowledge distillation method (KD) is employed to guide the student network to focus more effectively on target regions, thereby enhancing its robustness to high-noise backgrounds. Through the collaborative optimization of attention transfer and noise-resilient representation learning, the proposed approach significantly improves the stability of target recognition under noisy conditions. Based on the FSCE framework, two network architectures with different performance emphases are developed: lightweight DSAFNet-M and high-

*Corresponding author

precision DSAFNet-L. Extensive experiments are conducted on the MSTAR, FUSARShip and OpenSARShip datasets. The results show that DSAFNet-L achieves competitive or superior performance compared with various methods on three datasets; DSAFNet-M significantly reduces the model complexity while maintaining comparable accuracy. These results indicate that the proposed FSCE framework exhibits strong cross-model generalization.

Keywords: Image Recognition, Synthetic Aperture Radar (SAR), Knowledge Distillation, Feature Enhancement

---

1. Introduction

Synthetic aperture radar (SAR) plays a critical role in geological and mineral exploration [1], ship monitoring and fishery management [2], disaster assessment [3], and agricultural monitoring [4] due to its all-weather and all-day imaging capabilities. Despite its broad use, SAR faces inherent challenges in image recognition due to its unique imaging mechanism. On the one hand, speckle noise is induced by coherent signal interference during image formation and blurs fine-grained textures while introducing spurious features, thereby misleading recognition models. On the other hand, the inherently low contrast between targets and background is further aggravated by variations in imaging angles and environmental conditions, which significantly complicates discriminative feature extraction.

Several studies have attempted to improve remote sensing image quality through super-resolution techniques to enhance recognition performance [5, 6, 7]. While convolutional neural network–based super-resolution methods have demonstrated effectiveness in natural image scenarios, they primarily focus on globally enhancing image resolution and do not explicitly emphasize target-specific feature enhancement. Consequently, super-resolution methods provide limited performance improvement for SAR image tasks.

In recent years, several studies have explored the use of attributed scattering center (ASC) [8, 9, 10] and physical properties [11, 12, 13] to model

target characteristics to further improve SAR ATR performance. ASC-based methods construct feature representations that reflect radar cross-section characteristics by quantifying the average backscattered energy of targets across different azimuth and elevation angles, while physical scattering models based on polarization decomposition characterize target materials and geometric structures by analyzing polarization-dependent scattering mechanisms. Although these methods are effective to a certain extent, they exhibit inherent limitations. ASC suppresses noise through averaging but inevitably discards fine-grained spatial structural information, making it difficult to distinguish targets with similar scattering characteristics but different geometric structures. Meanwhile, although polarization-based physical models are physically interpretable, they typically analyze targets in isolation and struggle to capture contextual interactions between the targets and the complex background, leading to limited robustness under strong interference conditions. In addition, these methods often ignore precise target localization and region-focused perception, and rely heavily on handcrafted prior knowledge and idealized imaging assumptions, which restricts their adaptability to dynamic real-world scenarios involving target pose variations and imaging parameter fluctuations. By contrast, frequency–spatial coupled modeling exhibits greater flexibility in feature representation [14]. Frequency-domain information emphasizes the response of the local texture and structural mutation areas of the target, whereas spatial-domain representations preserve geometric morphology and contextual relationships. Coupling these two domains enables complementary exploitation of fine-grained details and global target structure, yielding more discriminative feature representations in complex environments.

Therefore, to fully leverage the complementary advantages of frequency and spatial domain and address the limitations of weak target focusing and insufficient semantic information construction ability of existing methods, we propose a target-aware frequency-spatial enhancement framework with noise-resilient knowledge guidance (FSCE). The core component of FSCE is the frequency-spatial shallow feature adaptive enhancement module (DSAF), which employs

3

multi-scale spatial convolution to capture hierarchical texture and global structure, while using wavelet-based frequency decomposition to separate noise components from informative features. By integrating the convolutional block attention module (CBAM) [15], DSAF adaptively recalibrates channel and spatial features to focus on discriminative regions, thereby effectively smoothing background noise and enhancing target discrimination. Furthermore, the FSCE framework introduces online knowledge distillation method (KD) to dynamically guide the student network to focus its attention on the target region while transferring robust recognition knowledge from the teacher network. Unlike traditional offline KD, this real-time co-optimization strategy leverages adaptive feature enhancement to improve the stability and noise resistance of the student network, enabling more reliable recognition performance under high-noise conditions.

1. We propose a target-aware frequency–spatial enhancement framework with noise-resilient knowledge guidance (FSCE) that jointly enhances discriminative representations, suppresses noise, and generalizes well across different network architectures.

2. We propose a frequency-spatial shallow feature adaptive enhancement module (DSAF) to effectively capture and enhance target structure and texture and smooth coherent speckle noise in high-noise environments.

3. Based on the FSCE framework and the DSAF module, two SAR ATR models with different performance preferences are proposed:

   - **High-precision model DSAFNet-L:** Based on FSCE, it realizes cross-domain feature deep fusion and achieves competitive or superior performance on three datasets.

   - **Lightweight model DSAFNet-M:** By reducing parameters, it maintains high precision while balancing engineering practicality and performance.

4

2. Related Work

2.1. SAR ATR

In the early research of SAR ATR, traditional methods relied mainly on handcrafted feature extraction and classifiers [16, 17, 18, 19]. These methods were strongly dependent on image preprocessing and feature selection, with limited generalization capability and impaired recognition performance in complex environments or under large target pose variations.

Following the emergence of deep learning, convolutional neural networks (CNN) have demonstrated remarkable performance on benchmark datasets such as MSTAR, owing to their end-to-end feature learning and have significantly outperformed traditional methods [20, 21]. Building upon CNN architectures, researchers have further incorporated attention mechanisms, residual connections, and multi-task learning strategies to improve robustness and generalization under the high-noise and low signal-to-noise ratio conditions characteristic of SAR imagery. Chen *et al.* [22] proposed a fully convolutional network (A-ConvNets) consisting of only sparsely connected layers without using fully connected layers. It can achieve an average accuracy of 99% in the classification of MSTAR 10-class dataset and outperform traditional ConvNets in the classification of target configurations and version variants. Zhao *et al.* [23] proposed a Transformer-based instance-aware model and incremental learning framework to address the problems of small samples and continuous learning process, in order to cope with the data scarcity and incremental update demands of SAR ATR in practical applications. In addition, several studies have explored unsupervised and semi-supervised domain adaptation methods to mitigate the distribution gap between simulated and measured data, thereby enhancing model transferability in real-world deployment scenarios [24, 25, 26].

Beyond network architecture design, several studies have explored the integration of ASC characteristics [27] [28] and physical property knowledge into CNN. Gao *et al.* [29] proposed the ASC-RISE method guided by physical information, introducing ASC into the RISE method, and the heat map generated by

it effectively locates the decision features of the model and provides corresponding physical information. Qin *et al.* [13] adopted a two-stream architecture, including attribute-induced global semantic extraction and graph-based structure representation learning, to achieve physically embedded small-shot SAR target recognition. By constructing a physical scattering model based on geometric optics and scattering center theory, and matching scattering center or ASC features through multi-level region matching, their approach achieves interpretability and robustness. Although these methods showed high recognition accuracy under SOC and typical EOC conditions, their reliance on scattering models makes the performance significantly degrade when the background noise is complex or the model has errors.

Despite the substantial progress achieved by existing studies in feature extraction and recognition performance under SOC and typical EOC scenarios, their reliance on explicit scattering models or specific network architectures and data utilization strategies limits robustness in challenging real-world environments. Since Ranchin and Wald [30] introduced wavelet transform and its corresponding multi-resolution analysis into remote sensing image processing, they demonstrated that reversible wavelet decomposition can not only suppress speckle noise of SAR images, but also compress and reconstruct images without losing information. This laid a theoretical foundation for the application of frequency-domain features in subsequent deep learning frameworks.

In the field of computer vision, deep fusion of wavelet transform and CNN can preserve both spatial and frequency-domain information. Fujieda *et al.* [31] proposed Wavelet CNN, which regards convolution and pooling layers as special cases of frequency-domain analysis, and reconstructs multi-scale features with the help of discrete wavelet transform (DWT), so as to surpass conventional models with fewer parameters in texture classification tasks. Li *et al.* [32] introduced the WaveCNet framework, which not only significantly improved noise robustness on ImageNet and ImageNet-C by replacing the downsampling operation with pluggable DWT/IDWT layers, but also led to the performance improvement of object detection based on this structure on the COCO dataset.

6

The introduction of frequency-domain features provides a new idea for improving the robustness and generalization ability of the model. Remote sensing images usually contain complex backgrounds and rich texture information [33], which brings objective conditions for the introduction of frequency-domain in remote sensing image recognition. Recently, some researchers have also introduced frequency-domain into remote sensing image recognition. Zi *et al.* [34] proposed the WaveCNN-CR model, which replaced the traditional downsampling operation with discrete wavelet transform (DWT), achieved lossless multi-scale feature extraction and significantly improved thin cloud removal and classification accuracy. The FFDC-Net proposed by Song *et al.* [35] achieved robust recognition of remote sensing crop classification by directly converting feature maps to the spectral domain. However, these methods only process in a single frequency domain or adopt approaches such as static modeling in the frequency domain or direct removal of high-frequency noise. As a result, their performance is limited under complex noise spatial structures, highlighting the limitations of denoising processing. On the other hand, although some works combine frequency domain and spatial domain operations [36, 37, 38], they still lack a unified attention mechanism or spatial multi-scale fusion.

These studies demonstrate that introducing frequency domain features into image recognition models, especially in remote sensing image recognition tasks, can effectively improve the model's ability to capture high-frequency details and enhance its adaptability to complex backgrounds, thereby improving overall recognition performance.

2.2. Knowledge Distillation

With the rapid advancement of deep learning in remote sensing target recognition, model complexity and computational demands have increased accordingly. To reduce model complexity while maintaining high recognition accuracy, knowledge distillation technology was widely adopted for model compression and performance enhancement. Traditional knowledge distillation methods typically follow an offline paradigm, in which a high-capacity teacher model is

7

first trained and subsequently used to guide the student model through output supervision. However, offline distillation faces inherent limitations, including fixed teacher representations and inflexible training processes.

To this end, online knowledge distillation has been introduced, enabling models to learn collaboratively and improving training flexibility and efficiency. For instance, the Mutual Contrastive Learning (MCL) framework proposed by Yang *et al.* [39] achieves mutual enhancement of feature representations through contrastive learning between multiple networks, thereby improving visual recognition performance. In addition, Guo *et al.* [40] proposed a collaborative learning strategy that performs online knowledge distillation via joint training of multiple student models, leading to improved model generalization.

In the field of remote sensing target recognition, knowledge distillation also shows great potential. Song *et al.* [41] proposed the KDE-Net model, achieving 88% parameter reduction while maintaining high accuracy through logit-based KD technology, significantly improving the efficiency of remote sensing image classification. In addition, Lê and Pham *et al.* [42] applied KD to the target detection task of remote sensing images, evaluated the performance of various knowledge distillation methods on xView and VEDAI datasets, and verified the effectiveness of KD in remote sensing target detection.

3. The Proposed Method

In this paper, we propose a target-aware frequency–spatial enhancement framework with noise-resilient knowledge guidance (FSCE) for SAR ATR, which integrates adaptive feature enhancement with dynamic knowledge transfer. The overall framework is illustrated in Figure 1. At its core, the frequency-spatial shallow feature adaptive enhancement module (DSAF) processes shallow feature maps via multi-scale spatial convolutions and frequency-domain wavelet decomposition, suppressing noise while preserving target structures. Complementing this, an online knowledge distillation method (KD) facilitates real-time knowledge transfer between networks in a clean feature space, enhancing the
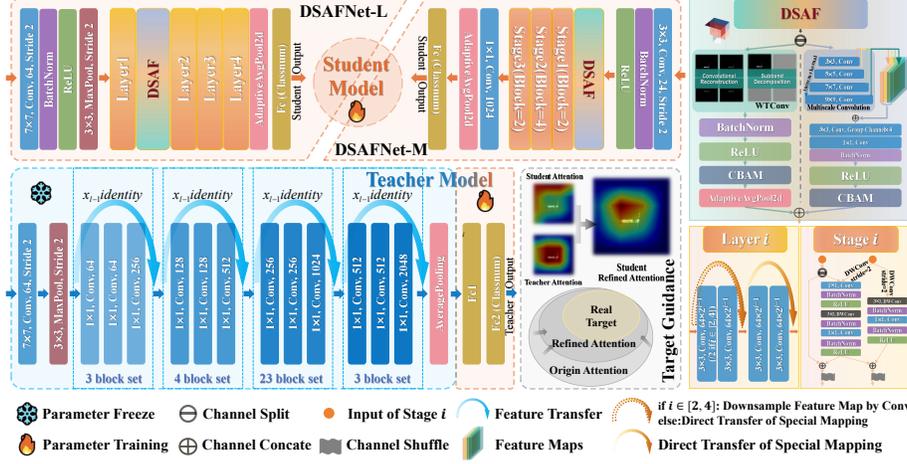
8

Figure 1: The overall architecture of our proposed method. DSAFNet-L and DSAFNet-M use ResNet18 and ShuffleNetV2 0.25x as the backbone network respectively, and use FSCE for focus area guidance to achieve spatial-frequency domain adaptive joint feature selection enhancement. The implementation path of each part is given.

robustness of recognition under high noise. Two student network architectures are developed under this framework. This section elaborates on the following four aspects: Data Preprocessing, Frequency-Spatial Collaborative Enhancement Framework; Student Model and Shallow Feature Enhancement; Teacher Model and KD Design.

3.1. Data Preprocessing

To enhance model robustness, we adopt a dual-view data augmentation strategy, wherein enhanced images are generated from both teacher and student perspectives during training. This method enables the model to learn multi-level and multi-angle features by applying different augmentation operations to the same image, thereby enhancing feature learning and generalization. Considering the characteristics of SAR imagery, we further introduce an augmentation strategy for interference factors such as adaptive speckle noise.

The teacher model augmentation emphasizes diversity of training data through geometric and visual transformations, including flipping, rotation, translation,

affine and perspective transformations, and random erasing, which collectively improve the model's ability to capture spatial structures and visual features. The student model augmentation focuses on frequency-domain and noise robustness, employing random cropping, Gaussian blurring, and random grayscale transformations to enhance resistance to interference such as noise and blur.

During the data loading, all images undergo uniform preprocessing, including resizing, center cropping, grayscale conversion, and conversion to tensor form, with an input size of 224×224.

3.2. Frequency-Spatial Collaborative Enhancement Framework

In the FSCE framework, a noise suppression pathway is established, where the teacher network provides semantic denoising guidance to dynamically assist the student network in learning noise-resistant features. This design addresses key challenges in SAR target recognition, such as the degradation of generalization ability caused by severe coherent speckle noise. As illustrated in Figure 1, FSCE innovatively coordinates spatial multi-scale convolution and frequency-domain wavelet decomposition, while introducing an online knowledge distillation mechanism that adaptively focuses on target regions and reinforces feature robustness. The core of FSCE is the frequency-spatial shallow feature adaptive enhancement module (DSAF), which effectively strengthens the student network's ability to handle noise, low contrast, and background interference in SAR images. Meanwhile, the semantic denoising knowledge transferred from the teacher network further amplifies the noise resilience and recognition stability of the student model.

After evenly dividing the input channels, **the spatial domain branch** applies four parallel convolution kernels of sizes 3×3, 5×5, 7×7, and 9×9, with padding set to 1, 2, 3, and 4, respectively, to maintain the spatial dimensions of the output feature maps consistent with the input. Multi-scale convolution kernels are employed to capture features at different spatial resolutions. Small-scale kernels effectively extract local details of SAR targets, such as edges, texture patterns, and scattering characteristics, which reflect the physical

10

structure and material properties of the targets. In contrast, large-scale kernels integrate broader contextual information, capturing target contours and their spatial relationships with the surrounding background, thereby enhancing discrimination under complex scattering conditions. Compared to using a single convolution kernel size, the multi-scale approach allows the network to simultaneously maintain sensitivity to high-frequency local variations and incorporate global structural information, which is particularly important for SAR imagery where targets often exhibit strong localized scattering while the background is dominated by low-frequency clutter. This design ensures adaptability to targets of varying sizes, shapes, and scattering characteristics. The output of each convolution is resized to a uniform spatial dimension via adaptive average pooling, with the 3×3 convolution serving as the reference to avoid excessive feature compression; thus, adaptive pooling is applied to the outputs of 5×5, 7×7, and 9×9 convolutions. The multi-scale feature maps are then fused using depthwise separable convolution followed by a 1×1 convolution and batch normalization. Nonlinearity is introduced via the ReLU activation, and feature representations of key regions are further refined using the CBAM attention mechanism.

Given an input feature map $F \in \mathbb{R}^{B \times C \times H \times W}$, where B is the batch size, C is the number of channels, and H, W are the spatial dimensions. The Convolutional Block Attention Module (CBAM) sequentially applies channel and spatial attention mechanisms to refine the representation.

First, the channel attention map $M_c(F)$ is generated as:

$$M_c(F) = \sigma^{(c)}\big(\mathrm{MLP}\big(\mathrm{AvgPool}_{H \times W}(F) + \mathrm{MaxPool}_{H \times W}(F)\big)\big) \qquad (1)$$

where $\sigma^{(c)}(\cdot)$ denotes a sigmoid activation, and MLP is a shared two-layer fully connected network with a reduction ratio.

The intermediate feature map refined by channel attention is:

$$F' = M_c(F) \odot F, \qquad (2)$$

where $\odot$ represents element-wise multiplication with channel-wise broadcasting.

11

Next, spatial attention $M_s(F')$ is computed by applying a 7×7 convolution over pooled descriptors:

$$M_s(F') = \sigma^{(s)}\Big(\text{Conv}_{7\times7}\big(\text{AvgPool}_{\text{chan}}(F') + \text{MaxPool}_{\text{chan}}(F')\big)\Big), \quad (3)$$

and the final CBAM-refined output is:

$$F_{\text{CBAM}} = M_s(F') \odot F'. \tag{4}$$

where M_c and M_s are channel and spatial attention maps, and $\odot$ is element-wise multiplication.

The CBAM attention mechanism comprises channel and spatial attention modules, which emphasize discriminative feature channels and salient spatial regions, enabling the model to focus on both critical feature types and target-relevant areas in SAR images.

Relying solely on spatial-domain representations is insufficient to fully exploit the distinctive characteristics of SAR imagery. SAR backscattering is best modeled as a piecewise stationary process dominated by localized singularities rather than smooth textures. The piecewise constant basis functions of the Haar wavelet effectively capture these localized singularities, and its four directional sub-bands (LL, LH, HL, HH) correspond well to the horizontal, vertical, and diagonal high-frequency backscattering components in SAR images. Compared with higher-order wavelets such as Daubechies or Coiflet, which favor smooth approximations, Haar wavelets preserve sharp transitions and local structural details inherent to SAR targets, while maintaining low computational complexity suitable for online adaptive enhancement.

Thus, for **the frequency-domain branch**, the WTConv2d module based on the Haar wavelet transform aligns with the intrinsic scattering patterns of SAR targets and provides an efficient mechanism for multi-resolution analysis of target features.

Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, where B, C, H and W denote the batch size, channel number, and spatial dimensions, respectively. For each

12

channel X_c, a two-dimensional discrete wavelet transform (DWT) using the Haar wavelet is applied. The corresponding low-pass filter $g[n]$ and high-pass filter $h[n]$ generate four directional filters:

$$
\begin{aligned}
\phi_{LL} &= g \otimes g, \\
\phi_{LH} &= g \otimes h, \\
\phi_{HL} &= h \otimes g, \\
\phi_{HH} &= h \otimes h
\end{aligned}
\tag{5}
$$

These four sub-bands respectively capture the low-frequency background clutter (LL) and high-frequency directional scattering components (LH, HL, HH), which are closely related to SAR target edges, structural boundaries, and anisotropic scattering behaviors. Each channel is convolved and downsampled as:

$$
WT(X_c) = \left[X_c * \phi_{LL}, \ X_c * \phi_{LH}, X_c * \phi_{HL}, \ X_c * \phi_{HH} \right] \downarrow 2
\tag{6}
$$

where $\downarrow 2$ denotes spatial downsampling by a factor of 2. This operation achieves multi-resolution analysis, suppressing redundant background information while preserving discriminative high-frequency responses. The resulting wavelet coefficients are aggregated into $X^{(1)} \in \mathbb{R}^{B \times C \times 4 \times \frac{H}{2} \times \frac{W}{2}}$ and reshaped as $\tilde{X}^{(1)} \in \mathbb{R}^{B \times 4C \times \frac{H}{2} \times \frac{W}{2}}$.

A depthwise convolution is then applied independently to each sub-band to adaptively reweight different frequency components, followed by a learnable channel-wise scaling factor γ:

$$
\hat{X}^{(1)} = \gamma \cdot \text{Conv}(\tilde{X}^{(1)})
\tag{7}
$$

The processed coefficients are reshaped back and reconstructed via the inverse wavelet transform:

$$
X' = IWT(\hat{X}^{(1)}) \in \mathbb{R}^{B \times C \times H \times W}
\tag{8}
$$

A residual fusion strategy is adopted to integrate frequency-domain information with spatial-domain convolutional features:

$$Y = \mathrm{BaseConv}(X) + X'$$

(9)

When stride > 1, average pooling is applied to maintain consistent resolution:

$$Y_{\mathrm{out}} = \mathrm{AvgPool}(Y)$$

(10)

The fused features are further normalized and activated, and enhanced by the CBAM attention mechanism. Finally, frequency-domain features are then concatenated with spatial-domain features along the channel dimension after dimensional alignment, forming a unified representation that jointly exploits spatial semantics and frequency-aware scattering cues.

By coupling spatial and frequency domain information, our DSAF module achieves a complementary balance between geometric integrity preservation and incoherent noise suppression, yielding more discriminative and noise-robust target representations.

For knowledge guidance, ResNet101 is employed as the teacher network, whose deep hierarchical structure naturally suppresses speckle noise through successive nonlinear transformations [43]. By continuously distilling dynamically updated noise-resistant semantic knowledge from the teacher, the student network progressively forms a hierarchical response pattern characterized by noise suppression and target enhancement across both shallow and deep layers. This process is illustrated in Figure 3, while detailed designs and validations are presented in the subsequent sections and ablation studies.

3.3. Student Model and Shallow Feature Enhancement

In SAR ATR, we aim to balance accuracy and complexity, ensuring suitability for real-time or resource-constrained deployment. ResNet18 and ShuffleNetV2 0.25x are selected as the primary student architectures due to their

14

complementary characteristics: ResNet18 provides strong feature extraction capability and high representational power, particularly for shallow spatial and frequency-enhanced features, while ShuffleNetV2 0.25x offers a highly lightweight design with minimal FLOPs, enabling faster inference and lower memory usage. This analysis demonstrates that the chosen student networks maintain a favorable balance between precision and efficiency: ResNet18 is suitable for scenarios prioritizing recognition performance, while ShuffleNetV2 0.25x provides a lightweight alternative for latency-sensitive applications. Coupling these architectures with the proposed DSAF module and online knowledge distillation ensures that both networks leverage spatial-frequency enhancements and teacher-guided semantic denoising, maximizing accuracy without incurring prohibitive computational cost.

Furthermore, selecting the Layer1 feature map of ResNet18 or the Conv layer feature map of ShuffleNetV2 as input has important significance. Owing to their high spatial resolution and fine-grained representation capacity, shallow features effectively capture local edges, textures, and structural details that are critical for target discrimination. By preserving such fine-grained spatial information, shallow feature maps provide a more discriminative basis for recognizing targets in complex and cluttered SAR scenes, thereby improving recognition accuracy and robustness.
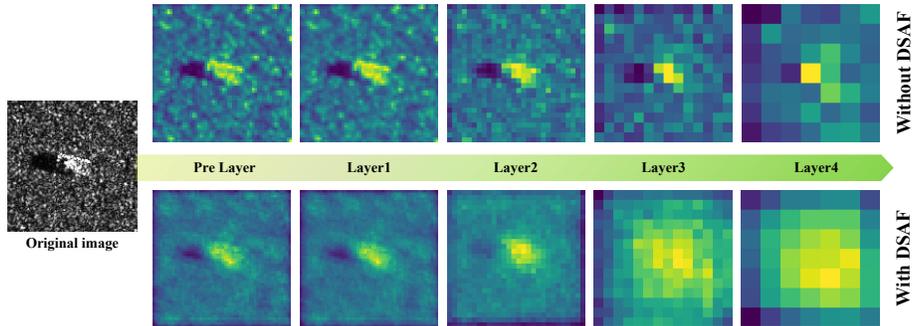
Figure 2: ResNet18 is used to display feature maps of different depths and their feature maps after passing through the DSAF module. The gray image on the left is the image in MSTAR, and heatmaps on the right. heatmaps from left to right are Pre layer, Layer1, layer2, Layer3, Layer4.

As shown in Figure 2, processing the shallow feature map of ResNet18 through the DSAF module results in a smoother background distribution while prominently highlighting the main texture features, leveraging the high-resolution nature of shallow representations. These characteristics originate from the low-level expression of the input signal by the shallow layers, which retain pixel-level details that form a rich foundation for subsequent feature enhancement. Moreover, shallow feature maps exhibit high stability during information transmission. As the network depth increases, the abstract level of feature expression gradually increases. Information faces the risks of gradient vanishing and feature fragmentation during cross-layer transmission. By contrast, the Layer1 output of ResNet18 and the initial convolutional layers of ShuffleNetV2 are early in the network hierarchy, with minimal nonlinear modulation and short transmission paths, effectively mitigating information attenuation. Such low-level representations not only support efficient gradient backpropagation but also provide purer input information, enhancing feature learning and classification performance. Hence, architectures based on shallow feature maps achieve an optimal balance between detail preservation and discriminative feature representation, providing a reliable technical path for high-precision interpretation of complex remote sensing images.

16

3.4. Teacher Model and KD Design

ResNet101 exhibits strong feature extraction capabilities and superior performance across multiple domains. However, its high computational cost limits its deployment on resource-constrained or real-time edge devices. Therefore, we adopt ResNet101 as the teacher model, using the same training parameters as the student network. Its deep architecture and precise target representation compensate for the student network's limited capacity.
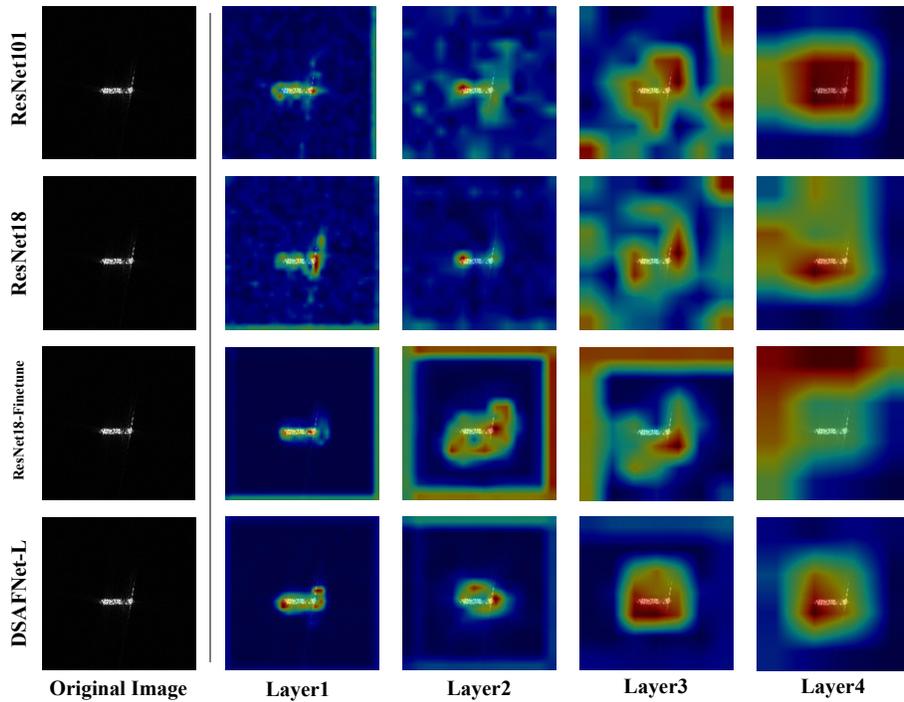

Figure 3: Grad-CAM is used to compare the heat maps of the interest areas of different layers of each model. From top to bottom, the models are ResNet101 pre-trained model, ResNet18 pre-trained model, ResNet18 model without pretraining but with the same training configuration as DSAFNet-L, and DSAFNet-L.

In contrast to conventional offline distillation, online distillation updates the teacher and student networks synchronously within a single training loop, dynamically modulating knowledge transfer to improve training stability. From the outset, the student receives real-time semantic guidance from the teacher,

17

accelerating convergence and mitigating overfitting to noise. This co-evolution allows the teacher to progressively provide cleaner semantic cues while the student incrementally refines its noise-resistant feature representation.

To illustrate the effect of model complexity and depth on noise robustness and focus area acquisition ability, we use Grad-CAM to visualize focus-area heatmaps for pre-trained ResNet18 and ResNet101, ResNet18-Finetune that was fine-tuned under the same conditions as DSAFNet-L without pre-training, and DSAFNet-L at different layers. As shown in Figure 3, deeper and more complex models more accurately capture target regions. In particular, ResNet101 demonstrates superior focus on the main target at Layer4 compared with ResNet18. Comparing ResNet18-Finetune and DSAFNet-L, the key difference is the presence of the DSAF module. Without DSAF, ResNet18-Finetune is more susceptible to coherent speckle noise, reducing its focus on the main target. These observations validate our approach: using deep ResNet101 as the teacher network, transferring soft-label knowledge, and leveraging category probability distributions smooths early noise interference and guides the student network to more accurately capture target features.

In this study, we adopt a Kullback-Leibler (KL) divergence-based knowledge distillation to transfer semantic denoising knowledge. During training, the soft labels output by the teacher network guide student feature learning, and the distillation loss is defined as:

$$\mathcal{L}_{\mathrm{KD}} = T^2 \cdot \mathrm{KL}\left(\mathrm{Softmax}\left(\frac{t}{T}\right) \parallel \mathrm{LogSoftmax}\left(\frac{s}{T}\right)\right) \qquad (11)$$

Among them, $\mathbf{s}$ and $\mathbf{t}$ denote the outputs of the student and teacher networks, respectively, α and $\mathbf{T}$ are temperature parameters. Increasing $\mathbf{T}$ smooths the soft labels distribution, providing richer relative information between categories. The total loss is a weighted combination of the classification loss and the distillation loss of the student model:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \alpha \cdot \mathcal{L}_{\mathrm{KD}} \qquad (12)$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss, and α controls the influence of the distillation objective during training.

To fully leverage the feature extraction and noise resistance capabilities of the student network and DSAF module, as well as to obtain focus and collaborative noise resistance capabilities from the teacher model, we fix the weight of $\mathcal{L}_{\mathrm{CE}}$ at 1, and scale $\mathcal{L}_{\mathrm{KD}}$ by α. This is because KD introduces a semantic denoising mechanism through soft labels transfer. Deeper teacher networks implicitly filter speckle noise and generate smoother probability distributions that reflect clear semantic associations. Student networks learn from these soft labels, suppressing responses to noise-dominated regions and focusing attention on the true target regions. This process acts as implicit regularization, mitigating overfitting to noisy features and promoting stable recognition even under high-noise conditions.

4. Experiments

In this section, numerous experiments are conducted and experimental settings are provided to clearly demonstrate the effectiveness of our method. First, three SAR ATR datasets used for experiments are introduced, as well as their similarities and differences. Then, hyperparameter and ablation experiments are conducted to evaluate the effectiveness of the proposed dual-domain attention feature enhancement module (DSAF) and online knowledge distillation method (KD). Finally, multiple comparative experiments with different types of methods were conducted to verify the superiority of our method.

4.1. Dataset Details

To evaluate the effectiveness of the FSCE framework, three publicly available SAR datasets were employed: MSTAR, OpenSARShip, and FUSARShip.

The MSTAR dataset [44] is a set of SAR images provided by the Defense Advanced Research Project Agency (DARPA) and the Air Force Research Laboratory (AFRL). This dataset collected images of ten different categories of

ground military vehicles (armored personnel carriers: BMP2, BRDM2, BTR60, and BTR70; tanks: T62, T72; air defense units: ZSU23_4; bulldozers: D7; rocket launchers: 2S1; and trucks: ZIL131) based on the synthetic aperture radar sensor platform of Sandia National Laboratories, as shown in Figure 4.
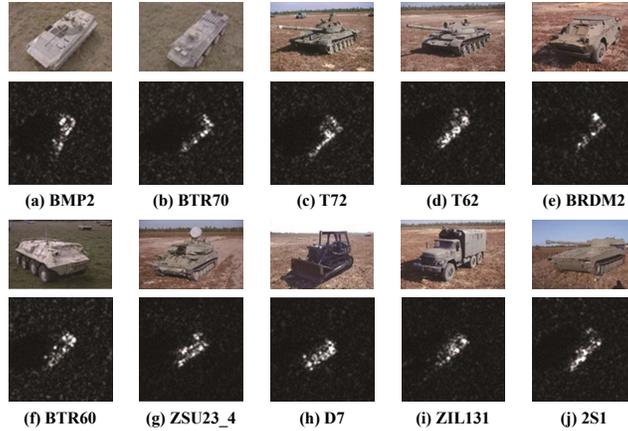

Figure 4: MSTAR 10-classification dataset.

For each category, SAR images were collected with all-around azimuth coverage. The spatial resolution distributions of the training and test sets were summarized in Table 1.

Table 1: MSTAR Dataset Distribution

Serial No.	Training		Test	
	Elevation	Number	Elevation	Number
ZSU_23_4	17°	299	15°	274
BRDM_2	17°	298	15°	274
BTR60	17°	256	15°	195
BTR70	17°	233	15°	196
BMP2	17°	233	15°	195
D7	17°	299	15°	274
ZIL131	17°	299	15°	274
T62	17°	299	15°	273
T72	17°	232	15°	196
Total		2747		2425

The FUSARShip high-resolution SAR ship dataset [45] was acquired by the Gaofen-3 satellite under dual-polarization modes (DH and DV). It contained 15 major ship categories, 98 subcategories, as well as a variety of non-ship marine targets. The spatial resolution for all images ranged from 1.7-1.754m×1.124m (range×azimuth). Bulk Carrier, Container Ship, General Cargo, and Tanker were selected. The image size was resized to 256×256 in the data processing section, and then the center was cropped to 224×224 pixels for better feature learning and target recognition. Representative samples from the datasets are shown in Figure 5.

(a) Bulk Carrier   (b) Container Ship   (c) General Cargo   (d) Tanker
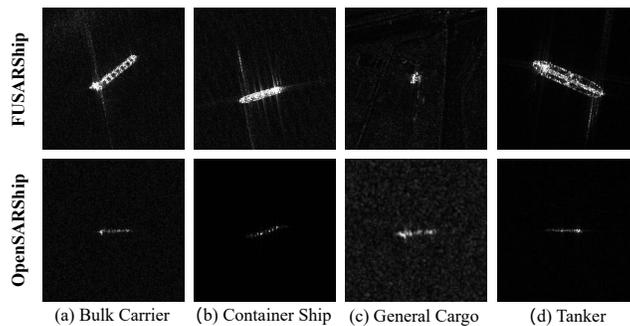
Figure 5: OpenSARShip and FUSARShip 4-classification datasets

The OpenSARShip dataset [46] was acquired by the Sentinel-1 satellite. The dataset contains 17 types of targets in ground range detection (GRD) and single-look complex (SLC) modes. The spatial resolution is 20m×22m (range×azimuth) in GRD mode and 2.7-3.5m×22m (range×azimuth) in SLC mode. The dataset exhibits a highly imbalanced category distribution (e.g., 8,470 cargo ships versus only 214 tugboats). To mitigate the impact of severe class imbalance, we selected Bulk Carrier, Container Ship, General Cargo, and Tanker. Following the same preprocessing pipeline as FUSARShip, the images were resized to 256×256 and then the center-cropped to 224×224 pixels. Representative samples from the dataset are shown in Figure 5.

The overall situation of the FUSARShip and the OpenSARShip dataset used in this paper is shown in Table 2.

21

Table 2: Distribution of Samples in OpenSARShip and FUSARShip Datasets

Dataset	Bulk Carrier	Container Ship	General Cargo	Tanker	Total
OpenSARShip	910	326	227	647	2110
FUSARShip	262	56	464	240	1022

To ensure a fair comparison with existing methods, the training and test splits of the MSTAR dataset follow the standard protocol summarized in Table 1. For the FUSARShip and OpenSARShip datasets, the samples are randomly divided into training and test sets, with 80% of the data used for training and the remaining 20% reserved for testing.

4.2. Experimental Setup

During training, the number of iterations was set to 300 with a batch size of 64. The learning rate was initialized to 2.5e-4 for the teacher network and 2.5e-3 for the student network. The teacher network employed a Cosine Annealing learning rate scheduler to gradually decay the learning rate and enhance generalization, while the student network adopted the OneCycle learning rate scheduler, which rapidly increases the learning rate in the early training stage and then gradually decreases it. The AdamW optimizer was used.

All experiments were conducted using the PyTorch deep learning framework on an NVIDIA GeForce RTX 4060 GPU. The experimental environment was configured with PyTorch 1.13, Python 3.9, CUDA 11.7, and Windows 11.

4.3. Comparison with Representative Methods

With the rapid development of deep learning, recognition algorithms based on deep neural networks continue to emerge. To thoroughly evaluate the effectiveness of the proposed model, we conducted comparative experiments against representative mainstream methods on three SAR ATR datasets: MSTAR, OpenSARShip, and FUSARShip.

For maritime target recognition, we selected representative SAR ATR models, including A-ConvNet [22], ESENet [47], FENNet [48], PAD-SE [49], as well

as the classic optical network DenseNet [50], for comparison on the OpenSAR-Ship and FUSARShip datasets. Since PAD-SE adopted a cross-domain dataset fusion strategy, the single-domain accuracy reported in its original paper is used for comparison. The experimental results are summarized in Table 3, where the best and second-best results are highlighted in red and blue, respectively.

Table 3: Accuracy (%) and parameter comparison of different models on OpenSARShip and FUSARShip datasets. The best and second-best results are highlighted in red and blue, respectively. ↑ indicates that a higher value denotes better performance, and ↓ indicates that a lower value denotes better performance.

Model	OpenSARShip		FUSARShip	
	Accuracy (%) ↑	Params (M) ↓	Accuracy (%) ↑	Params (M) ↓
A-ConvNet [22]	71.05	0.304	83.93	0.304
DenseNet [50]	73.45	0.80	83.62	0.80
ESENet [47]	73.06	0.54	82.74	0.54
FENNet [48]	71.69	0.66	84.02	0.66
PAE-SD [49]	75.61	–	85.05	–
Ours(DSAFNet-L)	76.42	11.4	89.32	11.4
Ours(DSAFNet-M)	74.76	0.17	86.41	0.17

Table 4: FLOPs comparison of different models. The best and second-best results are highlighted in red and blue, respectively. ↓ indicates that a lower value denotes better performance.

Metric	A-ConvNet	DenseNet	ESENet	FENNet	Ours(DSAFNet-L)	Ours(DSAFNet-M)
FLOPs (M) ↓	455.19	2895.99	295.41	584.03	2374.05	343.87

To further assess the suitability of the proposed models for edge deployment and eliminate hardware-induced variability, their floating-point operations (FLOPs) were measured, as shown in Table 4. All models were evaluated under a unified input specification of three channels with 224×224 resolution to ensure fair comparison.

The results indicate that the proposed models achieve a favorable trade-off between accuracy and complexity. Specifically, DSAFNet-L achieves the highest accuracy on both datasets, reaching 76.42% on OpenSARShip, 0.81% higher than the second-best result of 75.61%. And attaining 89.32% on FUSAR-

Ship, outperforming the second-best result among external methods (85.05%) by 4.27%. Meanwhile, DSAFNet-M maintains an extremely compact parameter size of only 0.17M while delivering competitive performance on both datasets. Notably, on FUSARShip, DSAFNet-M attains an accuracy of 86.41% with only one-fifth of the parameters of DenseNet, demonstrating its strong lightweight advantage and effective feature extraction capability.

Table 5: Accuracy Comparison of Different Models for SAR ATR in MSTAR Dataset. The best and second-best results are highlighted in red and blue, respectively. ↑ indicates that a higher value denotes better performance.

Method-Based	Model	Accuracy (%) ↑
CNN-Based	ConvNeXt-v2 [51]	96.05
	VGG19 [52]	97.50
	Inception-v3 [53]	94.56
	ResNet34 [54]	97.61
GCN-Based	GraphSAGE [55]	73.25
Transformer-Based	mViT [56]	85.35
	Swin Transformer [57]	81.88
Proposed for SAR ATR	A-ConvNet [22]	99.13
	ResNet-DTL [58]	99.46
	ResNet18+IFTS [59]	98.90
	CA-MCNN [8]	97.81
	GSP-IF [60]	98.56
	Ours(DSAFNet-L)	99.59

The experimental results on the MSTAR dataset are shown in Table 5. We compare our method with various approaches, including: 1) CNN-Based models such as ConvNeXt-v2 [51], VGG19 [52], Inception-v3 [53], and ResNet34 [54]; 2) GCN-Based models such as GraphSAGE [55]; 3) Transformer-Based models such as mViT [56] and Swin Transformer [57], and 4) SAR-specific recognition models such as A-ConvNet [22], ResNet-DTL [58], and GSP-IF [60]. Among them, ResNet-DTL and ResNet18-IFTS share a similar backbone with this work but differ in the feature types introduced.

Although most networks achieve high accuracy on MSTAR due to its rel-

atively uniform class distribution, our proposed DSAFNet-L achieves 99.59%
recognition accuracy, the best among 13 methods. CNN-Based, GCN-Based,
and Transformer-Based methods, primarily designed for optical images, are not
optimized for SAR, resulting in lower overall performance. Compared with other
SAR ATR methods, DSAFNet-L excels because purely spatial-domain models
like A-ConvNet rely on convolution depth or sparse connections for noise sup-
pression and do not leverage frequency-domain information, making it difficult
to separate high-frequency noise from target structures. Models using ASC or
physical priors often suffer from reduced robustness under strong noise due to
over-reliance on artificial designs. In contrast, our method adaptively models
both frequency and spatial domains, extracting discriminative features without
excessive reliance on prior knowledge, demonstrating strong modeling capability
and high robustness in noisy and complex environments.

For marine target recognition, both the FUSARShip and OpenSARShip
datasets present more challenging scenarios due to sea clutter, noise, and target
shape variations. As shown in Table 3 and Table 5, overall accuracy is lower than
MSTAR. Among the compared methods, PAE-SD method achieves the highest
accuracy, with accuracies of 85.05% and 75.61%, respectively. Compared with
ResNet-DTL, our method shows slight improvement of 0.13% on MSTAR, a
moderate gain of 0.81% on OpenSARShip, and a substantial improvement of
4.27% on FUSARShip compared with the second-best method.

The differences in improvement magnitude arise from dataset characteris-
tics: MSTAR has uniform image sizes and nearly balanced class divisions, leav-
ing little room for improvement; OpenSARShip has multiple polarization modes
(VH and VV) and varying spatial resolutions, complicating feature alignment;
FUSARShip has consistent spatial resolution with larger pixel coverage, allow-
ing the DSAF module to extract richer features, yielding larger accuracy gains.
While DenseNet has fewer parameters and performs well on optical tasks, it
tends to overfit in SAR recognition due to its design focus. PAE-SD leverages
incoherent phase alignment and visual attention to learn comprehensive fea-
tures, but struggles with local blur and occlusion. In contrast, the DSAFNet

series is specifically designed for SAR imagery, effectively capturing multi-scale structures and spatial-frequency information, achieving stable and superior performance across multiple benchmark datasets.
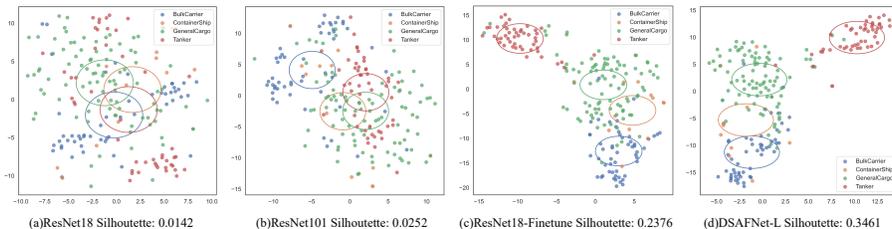

Figure 6: T-SNE visualization and silhouette score. The higher the silhouette score, the better the clustering effect. (a) Pre-trained ResNet18 (b) Pre-trained ResNet101 (c) Fine-tuned ResNet18 without pre-training (d) DSAFNet-L.

We employed T-SNE visualization and the Silhouette Score to evaluate feature clustering quality on the FUSARShip. The Silhouette Score quantifies clustering effectiveness, with higher values indicating better inter-class separability and intra-class compactness. As shown in Figure 6, DSAFNet-L exhibits the most discriminative feature distribution among the four compared models, achieving the highest Silhouette Score. Specifically, its score improves by 0.3319 compared with the pre-trained ResNet18, demonstrating the superior clustering capability of the proposed method.

4.4. FSCE Framework Transferability Experiment

To verify the synergistic effect of the DSAF module and online knowledge distillation method (KD) on model generalization, module migration experiments were conducted on two remote sensing ship datasets OpenSARShip and FUSARShip with substantial domain discrepancies.

As shown in Table 6, a comparison of different model variants with and without the DSAF module and KD demonstrates that DSAF consistently yields notable performance gains across both datasets by jointly integrating spatial multi-scale feature fusion and wavelet-based domain feature extraction. Taking

26

Table 6: Accuracy comparison with and without KD on OpenSARShip and FUSARShip. The best results of each model are highlighted in red. ↑ indicates that a higher value denotes better performance.

Dataset	Model	With KD (%) ↑	Without KD (%) ↑
OpenSARShip	A-ConvNet	65.80	62.50
	A-ConvNet + DSAF	73.11	72.17
	ResNet34	75.24	75.00
	ResNet34 + DSAF	76.18	75.47
	ResNet18	75.47	72.17
	Ours (ResNet18 + DSAF)	76.42	74.76
FUSARShip	A-ConvNet	81.07	77.18
	A-ConvNet + DSAF	83.50	82.52
	ResNet34	83.98	83.50
	ResNet34 + DSAF	84.95	84.47
	ResNet18	83.98	83.01
	Ours (ResNet18 + DSAF)	89.32	84.47

ResNet18 as an example, incorporating DSAF increases the accuracy on Open-SARShip from 72.17% to 74.76% (+2.59%) without KD, and from 83.01% to 84.47% (+1.46%) on FUSARShip. These results indicate that the local domain dual-branch structure of DSAF effectively enhances robustness to scale variation and domain shift in SAR imagery.

Furthermore, the introduction of KD further amplifies the advantages of DSAF through focal-region guidance. Under the ResNet18+DSAF configuration, KD yields an accuracy gain of 1.66% on OpenSARShip (74.76% → 76.42%), which is substantially higher than the 0.71% improvement observed for ResNet34+OnlineDSAF (75.47% → 76.18%). This verifies that KD can effectively improve the ability of lightweight models to grasp the subject in high-noise environments through the focus area guidance and implicit regularization effect of dynamic teacher-student interaction.

Comparing the A-ConvNet experiment, it is found that the improvement of DSAF on the traditional CNN model (OpenSARShip: +5.31%) is significantly higher than that of the ResNet series (+1.42-2.13%), indicating that its multi-

scale fusion mechanism can effectively compensate for the insufficient feature expression ability of shallow networks, offering a practical and efficient solution for deploying lightweight SAR ATR models.

4.5. Hierarchical Sensitivity Experiment

In order to investigate how the DSAF and KD enhance feature representations at different network depths, DSAF was inserted at multiple stages of the ResNet18+DSAF architecture, including the preprocessing layer (Pre) and the 1st, 2nd, and 5th residual blocks (Layer1, Layer2, and Layer5). The corresponding experimental results are reported in Table 7.

Table 7: results on different stages of DSAF. The best results are highlighted in red.

Dataset	Pre (%)	Layer1 (%)	Layer2 (%)	Layer4 (%)
OpenSARShip	74.29	76.42	74.76	75.94
FUSARShip	86.41	89.32	83.98	82.52

Through the layered performance comparison, Layer1 achieved the best performance on both datasets, which was 76.42% on OpenSARShip and 89.32% on FUSARShip, verifying that the shallow embedding strategy of DSAF is universal. This phenomenon is related to the physical characteristics of the WTConv wavelet transform: shallow features contain richer frequency-domain information, and the coefficient decomposition of WTConv can more accurately suppress the coherent speckle noise unique to SAR images. On OpenSARShip, inserting DSAF at Layer1 improves accuracy by 2.13% over the preprocessing layer, whereas embedding it at Layer5 results in a slight performance degradation to 75.94% (-0.48%). A similar trend is observed on FUSARShip, where Layer1 achieves 89.32%, exceeding the preprocessing layer by 2.91% (86.41%), while Layer5 suffers a substantial decline to 82.52% (-6.80%).

These results indicate that the synergy between DSAF and KD is most significant in shallower feature space, where the model directly operates on low-level, cross-domain-invariant cues such as edges and textures. As network depth

increases from the layer2 to layer5, performance on both datasets gradually declines. This is because it is difficult for the local multi-scale convolution of DSAF to effectively parse high-level semantic information due to the high abstraction of deep features, and the soft label supervision of KD may cause gradient diffusion in deep layers, resulting in reduced learning efficiency of domain-invariant features.

4.6. Ablation Experiment

The ablation results on the FUSARShip and OpenSARShip datasets are presented in Table 8 and Table 9. We conduct a comparative analysis from the perspective of feature-mode decoupling and supervision-signal coupling to elucidate the interactions between the proposed components.

Table 8: The Ablation Experiments on the FUSARShip Dataset. The best results are highlighted in red.

ID	Frequency	Spacial	KD	Accuracy (%)
1	×	×	×	83.01
2	✓	×	×	83.98
3	×	✓	×	83.50
4	×	×	✓	83.98
5	✓	✓	×	84.47
6	✓	×	✓	85.92
7	×	✓	✓	86.41
8	✓	✓	✓	89.32

The incorporation of wavelet-based frequency-domain feature enhancement demonstrates robust performance gains on both datasets. When KD is disabled, introducing frequency-domain features alone yields a 0.97% improvement on FUSARShip (id=2: 83.01%→83.98%) over the baseline, while a more pronounced gain is observed on OpenSARShip (+2.36%). This indicates that wavelet decomposition effectively captures frequency-domain invariant characteristics in SAR imagery.

29

Table 9: The Ablation Experiments on The OpenSARShip Dataset. The best results are highlighted in red.

ID	Frequency	Spacial	KD	Accuracy (%)
1	×	×	×	72.17
2	✓	×	×	74.53
3	×	✓	×	74.29
4	×	×	✓	75.47
5	✓	✓	×	74.76
6	✓	×	✓	73.58
7	×	✓	✓	74.29
8	✓	✓	✓	76.42

Moreover, a clear synergistic interaction exists between spatial multi-scale convolution and frequency-domain components. On FUSARShip, their combination (id=5) surpasses single-modality configurations by 0.97–1.49%, whereas on OpenSARShip—where background clutter is more severe—the standalone spatial branch (id=3) slightly outperforms the joint configuration (74.53% vs. 74.29%). This suggests that large-kernel convolutions (e.g., 9×9) enhance long-range contextual modeling for sparsely distributed ship targets (FUSARShip), but may introduce additional noise in dense small-target scenarios (OpenSAR-Ship), leading to marginal accuracy degradation.

The proposed KD exhibits nonlinear enhancement behavior in terms of focus-area guidance and multimodal fusion of lightweight models. On FUSARShip, KD combined solely with frequency-domain (id=6) achieves suboptimal performance (85.92%), whereas tri-modal fusion with KD (id=8) reaches the peak accuracy of 89.32%, outperforming the spatial+KD configuration (id=7) by 0.97%. This confirms that KD can adaptively balance frequency–spatial feature contributions and guide discriminative attention via soft-target entropy regularization, thereby mitigating overfitting caused by single-modality dominance. Conversely, the performance of the frequency-domain + KD combination (id=6) on OpenSARShip has decreased (73.58% vs single KD 75.47%), indicating that excessive frequency filtering may suppress discriminative details when domain

discrepancies are limited. The tri-modal configuration (id=8) compensates for detail information through spatial convolution and finally reaches the optimal 76.42%, revealing the necessity of component coupling for complex domain adaptation scenarios.

Table 10: The Experiments of different convolution kernel sizes. The best results are highlighted in red, respectively. ↑ indicates that a higher value denotes better performance.

Size of convolution kernels	Accuracy (%) ↑	
	OpenSARShip	FUSARShip
3×3 **and** 5×5 **and** 7×7 **and** 9×9	76.42	89.32
3×3 and 5×5 and 7×7 and 11×11	75.24	84.95
3×3 and 5×5 and 9×9 and 11×11	73.82	84.47
3×3 and 7×7 and 9×9 and 11×11	73.35	83.98
5×5 and 7×7 and 9×9 and 11×11	75.47	85.92
3×3 and 5×5 and 7×7	74.76	84.95
5×5 and 7×7 and 9×9	73.82	84.47
3×3 and 7×7 and 11×11	75.24	86.41

Furthermore, we investigated different combinations of convolutional kernel sizes to validate the rationality of the selected configuration. Table 10 shows that the adopted kernel-size combination achieves the highest accuracy on Open-SARhip and FUSARShip. Together, these kernels enable effective multi-scale feature representation from detailed textures to global structures, allowing the model to adapt to targets with varying sizes and shapes. In contrast, introducing an 11×11 kernel with an excessively large receptive field leads to performance degradation, likely due to the over-inclusion of background clutter and a mismatch with the intrinsic target scale. Similarly, reducing the kernel-size diversity results in inferior accuracy, as it limits the model's capacity to capture scale-dependent features. These results further confirm the effectiveness and rationality of the selected kernel-size combination.

4.7. KD optimal hyperparameter experiment

This section investigates the optimal settings for the temperature parameter T and the loss weight α in KD. The specific results are shown in Table 11.

Table 11: Hyperparameter T and α Tuning Results. The best results for different hyperparameter settings are highlighted in red.

FUSARShip					
T (fixed $\alpha = 0.5$)	1	3	5	7	9
Accuracy (%)	85.92	89.32	84.47	86.41	84.95
α (fixed $T = 3$)	0.1	0.3	0.5	0.7	0.9
Accuracy (%)	85.44	84.47	89.32	85.44	85.44
OpenSARShip					
T (fixed $\alpha = 0.5$)	1	3	5	7	9
Accuracy (%)	73.82	76.42	73.35	75.47	74.76
α (fixed $T = 3$)	0.1	0.3	0.5	0.7	0.9
Accuracy (%)	73.35	73.35	76.42	76.18	75.94

In the temperature parameter experiment, both datasets achieve peak performance at T=3 (FUSARShip: 89.32%, OpenSARShip: 76.42%), indicating that moderate temperature values ($T \in [3, 5]$) are most suitable for SAR ATR. The same hyperparameter setting was also applied to MSTAR and yielded the best performance among all compared methods. This is because T=3 provides moderate probability smoothing, which suppresses overfitting while preserving the discriminative relationships between target and background categories, playing a role analogous to a low-pass filter in the frequency-domain.

For the loss weight, both datasets achieve optimal performance at α=0.5, corresponding to a balanced mutual-information state between soft labels and hard labels. Under this setting, the student network not only benefits from the intrinsic feature enhancement of the DSAF module but also effectively absorbs the semantic knowledge conveyed by the teacher through soft labels, while maintaining discriminative supervision from ground-truth labels. This balance enables collaborative noise suppression and avoids semantic drift during domain transfer. When α approaches 0 or 1, the training becomes biased toward either the student or the teacher, leading to performance degradation. Specifically, on OpenSARShip, the accuracy drops to 75.94% at α=0.9 and 73.35% at α=0.1,

corresponding to decreases of 0.48% and 3.07%, respectively; on FUSARShip, the accuracy decreases by 3.88% in both cases. These results demonstrate that fixing the classification loss weight to 1 and setting the distillation loss weight to $\alpha=0.5$ maximizes the focus guidance and collaborative noise-resistance capability of the proposed FSCE framework.

5. Conclusion

We propose a target-aware frequency-spatial enhancement framework with noise-resilient knowledge guidance (FSCE) for SAR automatic target recognition, with the frequency-spatial shallow feature adaptive enhancement module (DSAF) as its core. Leveraging FSCE and online knowledge distillation, two model variants, DSAFNet-L and DSAFNet-M, were developed. Extensive experiments on MSTAR, FUSARShip, and OpenSARShip demonstrate that DSAFNet-L achieves competitive or superior recognition accuracy, while DSAFNet-M provides a favorable trade-off between performance and model compactness. Results confirm that joint frequency- and spatial-domain feature modeling effectively suppresses speckle noise, and KD further enhances target discrimination through focus-area guidance. Limitations remain under significant scale variations, likely due to semantic ambiguity introduced during resizing. Future work will explore scale-aware representations and adaptive alignment strategies to improve robustness.

6. Acknowledgement

References

[1] L. Wei, F. Wang, C. Tolomei, S. Liu, C. Bignami, B. Li, D. Lv, E. Trasatti, Y. Cui, G. Ventura, et al., Displacements of fushun west opencast coal mine revealed by multi-temporal insar technology, Geo-Spatial Information Science 27 (3) (2024) 585–601.

[2] A. Elyouncha, G. Broström, H. Johnsen, Synergistic utilization of spaceborne sar observations for monitoring the baltic sea flow through the danish straits, Earth and Space Science 11 (10) (2024) e2024EA003794.

[3] B. Huang, P. Li, H. Lu, J. Yin, Z. Li, H. Wang, Waterdetectionnet: a new deep learning method for flood mapping with sar image convolutional neural network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[4] O. Sunantha, Z. Shao, P. Pattama, A. Potchara, X. Huang, A. Zeeshan, Machine learning-based estimation of soil organic carbon in thailand's cash crops using multispectral and sar data fusion combined with environmental variables, Geo-spatial Information Science (2025) 1–23.

[5] J. Wang, X. Chen, X. Huang, R. Zhang, Y. Wang, T. Lu, Rethinking the role of panchromatic images in pan-sharpening, IEEE Transactions on Multimedia.

[6] Z. Liu, S. Wang, Y. Li, Y. Gu, Q. Yu, Dsrkd: Joint despecking and superresolution of sar images via knowledge distillation, IEEE Transactions on Geoscience and Remote Sensing.

[7] G. Dong, Y. Wang, H. Liu, S. Liu, Complex-valued sar image superresolution via sub-aperture learning and fusion, IEEE Transactions on Geoscience and Remote Sensing.

[8] Y. Li, L. Du, D. Wei, Multiscale cnn based on component analysis for sar atr, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–12.

[9] H. Zhongling, Y. Xiwen, H. Junwei, Progress and perspective on physically explainable deep learning for synthetic aperture radar image interpretation, 雷达学报 11 (1) (2021) 107–125.

[10] S. Feng, K. Ji, F. Wang, L. Zhang, X. Ma, G. Kuang, Electromagnetic scattering feature (esf) module embedded network based on asc model for robust and interpretable sar atr, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15.

[11] R. Theagarajan, B. Bhanu, T. Erpek, Y.-K. Hue, R. Schwieterman, K. Davaslioglu, Y. Shi, Y. E. Sagduyu, Integrating deep learning-based data driven and model-based approaches for inverse synthetic aperture radar target recognition, Optical Engineering 59 (5) (2020) 051407–051407.

[12] Z. Huang, C. O. Dumitru, J. Ren, Physics-aware feature learning of sar images with deep neural networks: A case study, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 1264–1267.

[13] J. Qin, B. Zou, Y. Chen, H. Li, L. Zhang, Scattering attribute embedded network for few-shot sar atr, IEEE Transactions on Aerospace and Electronic Systems.

[14] Z. Miao, M. Zhao, Time–space–frequency feature fusion for 3-channel motor imagery classification, Biomedical Signal Processing and Control 90 (2024) 105867.

[15] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[16] J. Cheng, L. Li, H. Li, F. Wang, Sar target recognition based on improved joint sparse representation, EURASIP Journal on Advances in Signal Processing 2014 (2014) 1–12.

[17] C. Lin, B. Wang, X. Zhao, M. Pang, Optimizing kernel pca using sparse representation-based classifier for mstar sar image target recognition, Mathematical Problems in Engineering 2013 (1) (2013) 847062.

[18] Y. Deng, Y. Deng, A method of sar image automatic target recognition based on convolution auto-encode and support vector machine, Remote Sensing 14 (21) (2022) 5559.

[19] X. Cheng, Z. Li, Modeling information flow from multispectral remote sensing images to land use and land cover maps for understanding classification mechanism, Geo-spatial Information Science 27 (5) (2024) 1568–1584.

[20] S. Shi, B. Chen, S. Bi, J. Li, W. Gong, J. Sun, B. Chen, L. Du, J. Yang, Q. Xu, et al., A spatial–spectral classification framework for multispectral lidar, Geo-Spatial Information Science 27 (5) (2024) 1460–1474.

[21] M. Yang, M. Cao, L. Cheng, H. Jiang, T. Ai, X. Yan, Classification of urban interchange patterns using a model combining shape context descriptor and graph convolutional neural network, Geo-Spatial Information Science 27 (5) (2024) 1622–1637.

[22] S. Chen, H. Wang, F. Xu, Y.-Q. Jin, Target classification using the deep convolutional networks for sar images, IEEE transactions on geoscience and remote sensing 54 (8) (2016) 4806–4817.

[23] X. Zhao, X. Lv, J. Cai, J. Guo, Y. Zhang, X. Qiu, Y. Wu, Few-shot sar-atr based on instance-aware transformer, Remote Sensing 14 (8) (2022) 1884.

[24] L. Zhao, Q. He, D. Ding, S. Zhang, G. Kuang, L. Liu, Selecting pseudo supervision for unsupervised domain adaptive sar target classification, EURASIP Journal on Advances in Signal Processing 2022 (1) (2022) 84.

[25] M. Kim, S. Kim, Synthetic sar data domain randomization for unseen sar atr, in: Algorithms for Synthetic Aperture Radar Imagery XXXI, Vol. 13032, SPIE, 2024, pp. 180–184.

[26] X. Zhang, Y. Luo, G. Li, Energy score-based pseudo-label filtering and adaptive loss for imbalanced semi-supervised sar target recognition, arXiv preprint arXiv:2411.03959.

[27] C. Li, L. Du, Y. Li, A novel method combining global visual features and local structural features for sar atr, IEEE Geoscience and Remote Sensing Letters 20 (2023) 1–5.

[28] J. Li, Z. Yu, L. Yu, P. Cheng, J. Chen, C. Chi, A comprehensive survey on sar atr in deep-learning era, Remote Sensing 15 (5) (2023) 1454.

[29] Y. Gao, W. Guo, D. Li, W. Yu, Asc-rise: Physical information guided explanation of sar atr models, in: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2024, pp. 2159–2162.

[30] T. Ranchin, L. Wald, The wavelet transform for the analysis of remotely sensed images, International Journal of Remote Sensing 14 (3) (1993) 615–619.

[31] S. Fujieda, K. Takayama, T. Hachisuka, Wavelet convolutional neural networks for texture classification, arXiv preprint arXiv:1707.07394.

[32] Q. Li, L. Shen, S. Guo, Z. Lai, Wavelet integrated cnns for noise-robust image classification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7245–7254.

[33] J. Wang, Y. Lin, C. Chen, X. Huang, R. Zhang, Y. Wang, T. Lu, From forgotten to pan-sharpening, Pattern Recognition (2025) 112653.

[34] Y. Zi, H. Ding, F. Xie, Z. Jiang, X. Song, Wavelet integrated convolutional neural network for thin cloud removal in remote sensing images, Remote Sensing 15 (3) (2023) 781.

[35] B. Song, S. Min, H. Yang, Y. Wu, B. Wang, A fourier frequency domain convolutional neural network for remote sensing crop classification considering global consistency and edge specificity, Remote Sensing 15 (19) (2023) 4788.

[36] C. He, Z. Shi, T. Qu, D. Wang, M. Liao, Lifting scheme-based deep neural network for remote sensing scene classification, Remote Sensing 11 (22) (2019) 2648.

[37] Y. Dang, X. Zhang, H. Zhao, B. Liu, Dctransformer: A channel attention combined discrete cosine transform to extract spatial–spectral feature for hyperspectral image classification, Applied Sciences 14 (5) (2024) 1701.

[38] G. Li, M. Ye, 3d wavelet convolutions with extended receptive fields for hyperspectral image classification, arXiv preprint arXiv:2504.10795.

[39] C. Yang, Z. An, H. Zhou, F. Zhuang, Y. Xu, Q. Zhang, Online knowledge distillation via mutual contrastive learning for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (8) (2023) 10212–10227.

[40] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, P. Luo, Online knowledge distillation via collaborative learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11020–11029.

[41] H. Song, C. Wei, Z. Yong, Efficient knowledge distillation for remote sensing image classification: a cnn-based approach, International Journal of Web Information Systems 20 (2) (2023) 129–158.

[42] H.-Â. Lê, M.-T. Pham, Knowledge distillation for object detection: from generic to remote sensing datasets, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2023, pp. 6194–6197.

[43] P. Wang, H. Zhang, V. M. Patel, Sar image despeckling using a convolutional neural network, IEEE Signal Processing Letters 24 (12) (2017) 1763–1767.

[44] D. Nistér, Preemptive ransac for live structure and motion estimation, Machine Vision and Applications 16 (5) (2005) 321–329.

[45] X. Hou, W. Ao, Q. Song, J. Lai, H. Wang, F. Xu, Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition, Science China Information Sciences 63 (2020) 1–19.

[46] B. Li, B. Liu, L. Huang, W. Guo, Z. Zhang, W. Yu, Opensarship 2.0: A large-volume dataset for deeper interpretation of ship targets in sentinel-1 imagery, in: 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), IEEE, 2017, pp. 1–5.

[47] L. Wang, X. Bai, F. Zhou, Sar atr of ground vehicles based on esenet, Remote Sensing 11 (11) (2019) 1316.

[48] Z. Zeng, J. Sun, C. Xu, H. Wang, Unknown sar target identification method based on feature extraction network and kld–rpa joint discrimination, Remote Sensing 13 (15) (2021) 2901.

[49] L. Wang, Y. Yang, Z. Liu, Multiresolution sar target recognition based on physical attention enhancement and scale distillation, IEEE Transactions on Aerospace and Electronic Systems 60 (3) (2024) 3081–3094.

[50] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[51] Y. Zhu, K. Yuan, W. Zhong, L. Xu, Spatial–spectral convnext for hyperspectral image classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2023) 5453–5463.

[52] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[55] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30.

[56] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6824–6835.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[58] Z. Huang, C. O. Dumitru, Z. Pan, B. Lei, M. Datcu, Classification of large-scale high-resolution sar images with deep transfer learning, IEEE Geoscience and Remote Sensing Letters 18 (1) (2020) 107–111.

[59] J.-H. Choi, M.-J. Lee, N.-H. Jeong, G. Lee, K.-T. Kim, Fusion of target and shadow regions for improved sar atr, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–17.

[60] C. Jingyi, Z. Yang, Y. Yanan, W. Yamin, Y. Feng, R. Weijia, L. Jun, Target recognition method based on graph structure perception of invariant features for sar images, 雷达学报 13 (2025) 1–23.