Scattering Prompt Tuning: A Fine-tuned Foundation Model for SAR Object Recognition

Weilong Guo[1,2], Shengyang Li[1,2,3,*], Jian Yang[1,2,3]
[1]Key Laboratory of Space Utilization, Chinese Academy of Sciences
[2]Technology and the Engineering Center for Space Utilization, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
{guoweilong19,shyli,yangjian20}@csu.ac.cn

Abstract

Synthetic Aperture Radar (SAR) serves as a vital tool in various earth observation applications, providing robust imaging under challenging weather conditions. While the fine-tuned foundation models excel in many downstream tasks, they struggle with SAR object recognition because of SAR's unique imaging and scattering characteristics. In this study, we propose a novel approach named Scattering Prompt Tuning (SPT) based vision foundation model. It uses SAR image scattering information as a prompt and integrates learnable parameters into the pre-trained model's input space to help learn SAR's unique information. We also employ a lightweight Residual AdapterMLP for fine-tuning, design a Sequential Feature Aggregation (SFA) to selectively fuse features from different transformer blocks effectively, and develop a Dynamic Distributional Contrast loss (DCLoss) to maintain the proper distance between different objects in feature space. Additionally, a four-stage training strategy, incorporating semi-supervised learning, is deployed to enhance SAR object recognition performance further. Our approach reaches a Top-1 accuracy of 37.9% *and an AUROC of 0.83 on the final dataset, winning the first place in the SAR Classification track of PBVS 2024 Multi-modal Aerial View Object Classification Challenge, which is better than the latest advanced fine-tuned foundation models.*

1. Introduction

Synthetic Aperture Radar (SAR) provides robust imaging unaffected by weather and time, addressing challenges faced by electro-optical (EO) systems in dealing with clouds, fog, and lighting changes [1, 2]. It plays an important role in observation applications like smart cities and ocean monitoring.

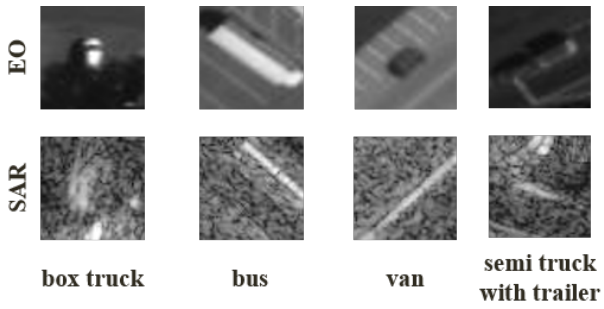


Figure 1. Comparison of different classes of EO images and SAR images

Despite its advantages, SAR image object recognition lags behind EO images for three main reasons [3]: (1) Limited availability of SAR data compared to the abundant EO images supporting large model training for learning generalized feature representation; (2) Inferior image quality in SAR, as shown in Fig. 1, characterized by lower resolution and higher noise, leading to blurred object details; (3) Different imaging mechanisms, rendering some EO object recognition findings inapplicable to SAR images. These challenges, coupled with the long-tail problem, contribute to the complexity of SAR object recognition.

Long-tail recognition aims to accurately represent objects in unbalanced data, where some classes (head classes) have abundant training samples, while others (tail classes) have limited samples. In addressing this challenge, various approaches can be categorized into three groups: 1) data processing [4, 5]. 2) representation learning [6, 7]. 3) model output tuning [8, 9]. Despite progress made in improving object recognition performance, a noticeable gap persists compared to models trained on balanced datasets [10].

Recently, large foundation models pre-trained on web-

*Corresponding Author.

This work was supported by Key Deployment Program of the Chinese Academy of Sciences: KGFZD-145-23-18.

scale datasets have revolutionized the field of computer vision, showing powerful zero-shot and few-shot generalizations [11]. These models have the potential to generalize tasks and data distributions beyond those seen during training. They have been successfully applied in various fields, including visual recognition [12–15], dense prediction [16–20], Reinforcement Learning (RL) [21–23], Robotics [24, 25], and etc.

Recent results from BALLAD [26], RAC [27], VL-LTR [28], and LPT [29] demonstrate that properly fine-tuning pre-trained foundation models can surprisingly improve the long-tail object recognition accuracy. For instance, LPT fine-tunes the vision transformer pre-trained on ImageNet, utilizing prompt tuning via two-stage training.

Due to the large differences between general images and SAR images and the absence of labeled SAR image datasets, as far as we know, there is no fine-tuned vision foundation model available for fine-grained object recognition in SAR images. And existing fine-tuned models are designed from the perspective of the model structure [10, 30], ignoring the unique scattering information of SAR images.

Building upon these considerations, we propose a fine-tuned foundation model called Scattering Prompt Tuning (SPT) specifically tailored for object recognition in SAR images. The scattering information extracted from SAR images is converted into a textual description, serving as the prompt alongside the visual image description. These inputs are processed by the text encoder to extract semantic features, initializing the model head for better convergence. To guide the image encoder in learning scattering information, we introduce trainable parameters as a scattering characteristics prompt in the input space, facilitating the fine-tuning of the vision transformer model.

Acknowledging the disparities between web images and SAR images, we introduce a lightweight module named Residual AdapterMLP (RAMLP) within each transformer block. The pre-trained transformer is fine-tuned by updating RAMLP. Additionally, we implement a sequential feature aggregation module to selectively fuse feature outputs from different transformer blocks. This module adaptsively extracts rich hierarchical information from SAR images. To address the challenge of extreme inter-class similarity and intra-class differences, we develop the Dynamic Distributional Contrast Loss. This loss function ensures that features of objects from the same class and different classes maintain appropriate distances, enhancing class distinguishability while preserving intra-class differences to some extent.

Our fine-tuned foundation model, SPT, diverges from existing methods by effectively leveraging the scattering properties of SAR images. SPT mitigates domain shift issues between general web images and SAR images, proving to be effective and well-suited for fine-grained object recognition in SAR images.

In summary, our contributions are outlined as follows:

1. We propose a Scattering Prompt Tuning (SPT) based foundation model. As far as we know, it is the first fine-tuned foundation model for recognizing fine-grained objects in SAR images.
2. Different from existing methods, our method uniquely leverages SAR image scattering properties to markedly improve object recognition performance.
3. We combine and design novel modules like Residual AdapterMLP and Sequential Feature Aggregation with a Dynamic Distributional Contrast Loss, significantly enhancing semantic feature learning in SAR images.

2. Related Works

In our proposed SPT, we design modules for efficiently fine-tuning the current vision transformer model from the perspective of utilizing the scattering properties of SAR images and dealing with the long-tail recognition challenge. In this section, we perform a literature review of related works from two perspectives: the foundation models and long-tailed recognition.

2.1. Foundation Models

In recent years, there have been significant advancements in natural language processing (NLP) with the development of large language models (LLMs) [11] like GPT-3 [31], GPT-4 [32], and ChatGPT. These breakthroughs have sparked a revolution in the computer vision field, prompting researchers to explore vision foundation models [33–36]. These models leverage self-supervised, unsupervised, and image-text contrastive learning on vast web-scale datasets to pre-train vision transformers. They demonstrate robust generalization to downstream transfer learning tasks, even in few-shot or zero-shot scenarios [30]. However, training these state-of-the-art models from scratch or fully fine-tuning them for specific datasets, especially models like ViT-G/14 [33] with over 1.8 billion parameters, is impractical. To overcome this challenge, there is growing interest in parameter-efficient learning strategies (PETL). PETL aims to leverage pre-trained foundation models as a starting point and fine-tune only a subset of their parameters to achieve comparable or superior performance to fully tuned models [37–41]. Current PETL methods predominantly include adapter-based [39, 42] and prompting-based approaches [30]. However, these methods primarily focus on the model structure perspective. Different from existing methods, our proposed SPT uniquely leverages SAR image scattering properties to markedly improve object recognition performance.

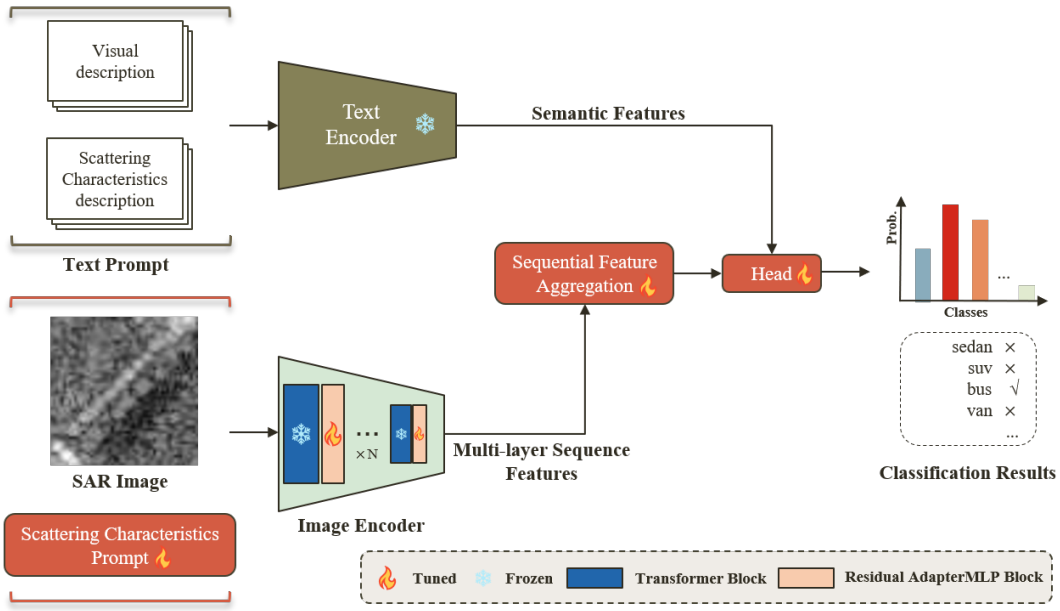


Figure 2. The framework of our proposed SPT.

2.2. Long-tail Recognition

Long-tail recognition aims to learn an accurate representation of objects within highly imbalanced datasets. Existing methods can be broadly classified into two categories: training from scratch (TFC) and fine-tuning (FT) pre-trained models. TFC methods involve training convolutional neural networks directly on long-tail datasets, incorporating strategies such as 1) data preprocessing [4, 5]. 2) representation learning [6, 7]. and 3) model output tuning, with notable contributions by [8, 9]. Conversely, recent FT approaches capitalize on the strong representational capabilities of pre-trained foundation models like CLIP [43] and ViT [12], fine-tuning them to improve performance in long-tail object recognition, with key studies by [26–28]. Despite these advancements, to our knowledge, there is currently no foundation model explicitly tailored for long-tail object recognition in SAR images, primarily due to the large difference between general images and SAR images and the lack of labeled SAR datasets.

3. Approach

3.1. Overall Framework

The framework of our SPT foundation model is shown in Fig. 2. (1) We convert scattering information extracted from SAR images into textual descriptions, used as prompts alongside visual descriptions. These inputs undergo processing by a text encoder, extracting semantic features to initialize the classifier for better convergence. (2) To guide the image encoder in learning scattering information, we introduce a small set of trainable parameters into the input space as scattering characteristics prompt. (3) A lightweight Residual AdapterMLP (RAMLP) module fine-tunes the vision transformer by updating its weights. (4) The sequential feature aggregation module selectively fuses outputs from different transformer blocks, capturing comprehensive information. (5) An optimized dynamic distributional contrast loss is designed to effectively address challenges posed by extreme intra-class differences and inter-class similarities in SAR image object recognition.

3.2. Scattering information extraction

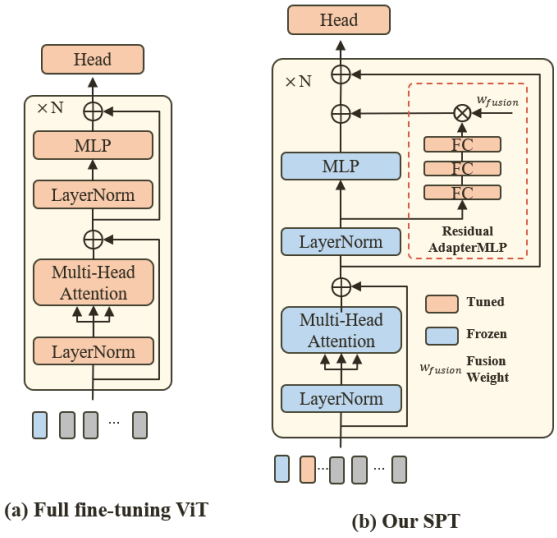
```python
def extract_scatter_points(sar_image_path,
                           threshold=15):
    # load sar images
    sar_image = cv2.imread(sar_image_path, cv2.IMREAD_GRAYSCALE)
    # Extracting scattering points as
    # scattering information for sar images
    _, binary_image = cv2.threshold(sar_image, threshold, 255,
                                    cv2.THRESH_BINARY)

    contours, _ = cv2.findContours(binary_image, cv2.RETR_EXTERNAL,
                                   cv2.CHAIN_APPROX_SIMPLE)
    scatter_points = []
    for contour in contours:
        M = cv2.moments(contour)
        if M["m00"] > 0:
            center_x = int(M["m10"] / M["m00"])
            center_y = int(M["m01"] / M["m00"])
            scatter_points.append((center_x, center_y))
    return scatter_points
```

Figure 3. Python code for scattering information extraction algorithm.

The scattering information of SAR images is the phenomenon of radar wave interaction with ground objects.

Different objects have different scattering characteristics. The scattering information describes the characteristics of the object surfaces, which is important for understanding their edges and structures in SAR images [44]. Embedding the scattering information into the prompt encoder may help to improve the object recognition performance of the model. We consider the scattering points extracted from the SAR image as its scattering information, and the python code for the extraction algorithm is illustrated in Fig. 3.

3.3. Scattering Characteristics Prompt



(a) Full fine-tuning ViT (b) Our SPT

Figure 4. Comparison of full fine-tuning ViT and our SPT. (a) Structure of full fine-tuning ViT. (b) Structure of our SPT.

For a plain vision transformer (ViT) [12] with N layers, an input image is divided into m fixed-sized patches. Each patch is then first embedded into d-dimensional latent space with positional encoding. Together with an extra learnable classification token ([CLS]), it is fed into the transformer for feature learning. In our SPT, as shown in Fig. 4, we add extra learnable vectors as the scattering characteristics prompt (SCP) in the input space to help the pre-trained model learn the scattering information of unseen SAR images. During the fine-tuning process, the parameters of scattering characteristics prompt is updated.

3.4. Residual AdapterMLP

Vision transformers are usually trained on large web-scale datasets and may lack exposure to SAR image information. To bridge this gap, as shown in Fig. 4, we introduce the Residual AdapterMLP (RAMLP) module in each transformer block, helping the model effectively capture accurate details from SAR images. With fully-connected layers, nonlinear activation functions, and residual feature fusion weights, RAMLP dynamically adjusts the vision transformer weights during fine-tuning.

Recent studies, like AdapterFormer, highlight the importance of the MLP in fine-tuned vision transformers for general image/video recognition. RAMLP not only prevents certain issues such as output degradation in vision transformers but also improves its performance. Our version of the RAMLP module in the fine-tuned vision transformer differs from AdapterFormer [42] by including more nonlinear layers and deeper embedding dimensions for better learning

3.5. Sequential Feature Aggregation

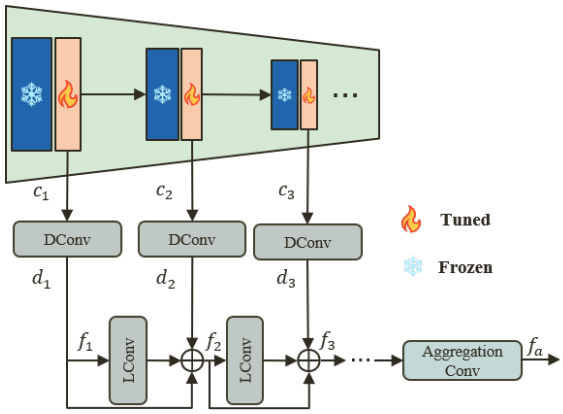


Figure 5. Details of our sequential feature aggregation module.

The pre-trained vision Transformer model produces diverse features in its different blocks, each containing distinct levels of semantic information from the SAR image. Considering these features as a sequential sequence, we design a sequential feature aggregation (SFA) module to selectively filter and merge the most relevant information for SAR object recognition. The details are shown in Fig. 5

For the output features $\{c_1, \cdots, c_N\}$ from the N transformer blocks, we initially apply a channel downsampling convolution (DConv) to get a uniform channel count for the feature sequence $\{d_1, \cdots, d_N\}$. Subsequently, short and long-term sequence feature screening is performed to derive features f_N, encompassing information from various positions within the sequence. Lastly, an aggregated convolution, formed by stacking multiple 1×1 convolutional layers, adapts to fuse the screened features effectively.

$$f_N = LConv(f_{N-1}) + f_{N-1} \tag{1}$$

3.6. Loss Function

The loss function $\mathcal{L}$ in our SPT model comprises two key components: L_A (Logit-Adjusted loss) [8] and L_{DC} (Designed Dynamic Distributional Contrast Loss, DCLoss).

L_A primarily addresses the challenges posed by the long-tailed distribution of data during training. On the other hand, L_{DC} is tailored to tackle the difficulties arising from the extreme inter-class similarity and intra-class differences in SAR image object recognition. For a detailed explanation of L_A, please refer to the paper [8]. Here, we focus on elucidating L_{DC}.

$$\mathcal{L} = L_A + L_{DC} \tag{2}$$

The primary objective of designing L_{DC} is to ensure that features from different object classes are widely separated, while features from the same object class are brought closer together, aligning with the actual data distribution. To address the local clustering issue within classes (intra-class differences), we incorporate a dynamic distance threshold screening strategy. This strategy allows for variations in distances between objects of the same class while ensuring they are kept distinct from features of other classes. For the features extracted from n SAR image objects, we measure their similarity using the Euclidean distance:

$$dist_{ij} = \|\|o_i - o_j\|\|_2^2 = \sum_{k=1}^{C}(o_{ik} - o_{jk})^2 \tag{3}$$

where i and j denote the index of the sar image, $\{1 \leqslant i, j \leqslant n\}$ and $\{i \neq j\}$. C is the channel number. To guide the learning of the similarity measure matrix $dist$, we build its ground truth, G_{sim}, based on the actual labels of each object. The value is set to 1 when the labels of two objects are the same and 0 otherwise. The calculation of L_{DC} can be expressed using the following equation:

$$L_{DC} = G_{sim} * (dist - \delta_p)^2 + (1 - G_{sim}) * max(\delta_q - dist, 0)^2 \tag{4}$$

δ_p and δ_q represent the dynamic thresholds for distances between objects of the same class and different classes, adapting to various data distributions. In our specific implementation, we use the average values of the distances between objects of the same class and different classes as δ_p and δ_q, respectively.

3.7. Training Strategy

To gradually adapt the pre-trained vision transformer model to SAR object recognition, our SPT model undergoes four stages of training.

In the first stage, we fine-tune our SPT initially on a mixed dataset of EO and SAR images, followed by another fine-tuning specifically on SAR images. This process yields an initial foundation model tailored to the SAR image domain.

Moving to the second stage, we create a balanced dataset by sampling images from various classes in the SAR image training dataset for further fine-tuning.

In the third stage, our focus is on understanding the differences between the training data and the val(development phase) / test(test phase) data. We capture features, prediction scores, and prediction categories from the model's last layer for all val/test data. By applying a threshold, we select high-scoring samples and employ Gaussian Mixture and K-means++ algorithms for feature clustering. The overlap of their results serves as the final clustering outcome. If a high-scoring sample is within a cluster, we assign the prediction categories of the high-scoring sample to all cluster members. This process generates reliable predictions, and we construct a balanced dataset by incorporating some training images for further fine-tuning.

In the fourth stage, we exclusively use high-scoring samples as reliable predictions to build a balanced dataset for additional fine-tuning. Since we find that the strategy in the third stage becomes less effective as the model's prediction accuracy improves. Consequently, we iterate through the fourth stage until the model's prediction accuracy reaches a plateau.

4. Experiments

4.1. Experimental Setup

Table 1. Details of the training dataset used in PBVS 2024 Multi-modal Aerial View Object Classification - C (SAR Classification) Challenge.

Class ID	Class Name	Number	Percent(%)
0	sedan	364291	79.95
1	suv	43401	9.53
2	pickup truck	24158	5.30
3	van	16890	3.71
4	box truck	2896	6.36
5	motorcycle	1441	0.32
6	flatbed truck	898	0.20
7	bus	612	0.13
8	pickup truck with trailer	695	0.15
9	semi truck with trailer	353	0.08

Dataset. The dataset employed for the PBVS 2024 Multi-modal Aerial View Object Classification - C (SAR Classification) Challenge encompasses aerial view SAR images of 10 fine-grained vehicles. Tab. 1 outlines the specifics of the training dataset, which includes 455,635 images exhibiting a pronounced long-tail distribution. Notably, Class 0 constitutes approximately 80% of the dataset, while Class 9 comprises less than 0.08%. The image sizes in the dataset exhibit some variability, averaging around 56×56 pixels.

Evaluation Metrics. To quantitatively evaluate our proposed method, four evaluation metrics are established for this challenge: (1) Top-1 Accuracy. (2) Area Under the Re-

ceiver Operating Characteristic curve (AUROC). (3) True negative rate (TNR) at 95% true positive rate (tpr95). and (4) Total Score. Total Score is a combination of Top-1 accuracy and AUROC. Higher values of these four metrics indicate better performance of the model. Note that the total score determines the final ranking in the PBVS 2024 SAR classification challenge.

Implementation details. All experiments are conducted using a GeForce RTX3090 GPU. Our baseline model, PEL, utilize CLIP-ViT-Base as its backbone. We employ the SGD optimizer with a batch size of 128 and an initial learning rate of 0.01. The training process involved multiple stages. The first stage run for 40 epochs. The second and third stages are each set to 10 epochs. The fourth stage comprise 7 epochs. Gaussian Mixture is configured with 10 components, and K-means++ utilize 10 clusters with 100 iterations. The threshold for identifying high-score prediction samples ranged from 0.85 to 0.98.

4.2. Ablation Study

To assess the effectiveness of key components, hyperparameters, and modules in our proposed method for SAR object recognition, we conduct a series of experiments. Unless specified otherwise, all evaluations are performed on the val dataset (development phase) and limited to the first training stage.

Table 2. Classification performance of different backbones and image scales in development phase (val dataset).

Backbone	Image Scale	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score
ResNet101	56 × 56	11.83	0.413	0.02	0.19
ResNet101	112 × 112	25.25	0.382	0.003	0.28
ResNet101	224 × 224	26.4	0.496	0.03	0.32
CLIP(Vit-B)	224 × 224	28.4	0.613	0.07	0.37

Backbone Selection. In the development phase, we conduct experiments to assess the performance of different backbone networks. Initially, we adopt ResNet101 [45], following the approach of the first-place solution in the PBVS 2023 SAR classification challenge [46]. Additionally, we test the performance of CLIP (ViT-B) [43], a widely used backbone in foundation models.

As depicted in the Tab. 2, CLIP outperforms ResNet101 across all four metrics—Top-1 Accuracy, AUROC, TNR at tpr95, and Total Score. This clear superiority highlights the effectiveness of CLIP, leading us to choose it as our backbone network.

Image Scale Exploration. Commonly utilized backbone networks like ResNet, CLIP, etc., are typically trained on images with a scale of 224 × 224 pixels. However, the PBVS 2024 SAR classification dataset contains images of approximately 56 × 56 pixels.

In Tab. 2, we investigate the impact of different image scales on SAR object recognition, employing ResNet101 as the backbone network on the val dataset. The results reveal a gradual increase in object recognition accuracy as the image scale enlarges. Consequently, for all subsequent experiments, we standardize the image scale to 224 × 224 pixels.

Table 3. Classification performance of different loss functions in development phase (val dataset).

Loss	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score
CBLoss [47]	25.83	0.57	0.02	0.34
LDAM [48]	26.70	0.56	0.01	0.34
GRW [49]	27.56	0.59	0.04	0.35
LADE [50]	5.48	0.59	0.16	0.19
LA [8]	28.4	0.613	0.07	0.37

Exploring Loss Functions for Long-Tail SAR Object Recognition. In our exploration of the impact of different loss functions on object recognition accuracy for SAR images, we utilize PEL as the benchmark method and CLIP (Vit-B) as the backbone network. The evaluated loss functions include CBLoss [47], LDAM [48], GRW [49], LADE [50], and LA [8].

As illustrated in the Tab. 3, CBLoss, LDAM, and GRW achieve comparable object recognition accuracy, while LADE exhibits poor performance in SAR object recognition. Remarkably, the LA loss function stands out as the optimal choice across all evaluation metrics. Consequently, we adopt LA as our base loss function and employ it throughout all training stages. Our baseline method combines PEL with the LA loss function.

Effect of Different Modules. Tab. 4 illustrates the impact of different modules on the SAR object recognition results of our proposed SPT on the val dataset. The model's overall object recognition accuracy experiences significant enhancement when individual modules, namely SCP, RAMLP, SFA, and DCLoss, are incorporated.

Remarkably, the Top-1 accuracy improves from 28.4% to 32.75%, and the overall score rises from 0.37 to 0.40 when all modules are combined, compared to the baseline method. This improvement underscores the effectiveness of our proposed method.

Results of Different Training Stages. We conduct tests on the model at different training stages, evaluating its performance on both the val and test datasets. The results, presented in the Tab. 5, reveal a gradual increase in the model's overall accuracy as training progresses.

Table 4. Classification performance of different modules in development phase (val dataset).

SCP	RAMLP	SFA	DCLoss	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score
				28.4	0.61	0.07	0.37
✓				28.72	0.64	0.08	0.37
	✓			29.29	0.64	0.08	0.38
		✓		29.44	0.64	0.05	0.38
			✓	29.87	0.63	0.06	0.38
✓	✓	✓	✓	32.75	0.61	0.06	0.40

Table 5. Classification performance of different training stages in development phase and test phase.

Stage				development phase (val dataset)				test phase (test dataset)			
1	2	3	4	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score
✓				32.75	0.61	0.06	0.40	21.22	0.67	0.03	0.33
	✓			34.92	0.60	0.04	0.41	23.35	0.68	0.04	0.34
		✓		37.04	0.55	0.03	0.42	31.04	0.55	0.03	0.37
			✓	37.52	0.62	0.07	0.44	33.1	0.61	0.02	0.40

Notably, in the third and fourth stages, where we implement distinct strategies to incorporate reliable samples from the val/test data into the training set, there is a substantial and comprehensive enhancement in the model's performance. After the fourth stage of training, the model achieves a top-1 accuracy of 37.52% on the val dataset, accompanies by an total score of 0.44. On the test dataset, the model's top-1 accuracy reaches 33.1%, yielding an total score of 0.40.

4.3. Comparison with methods for Fine-Tuning Foundation Model

The object recognition accuracy of different fine-tuning methods on both the val and test datasets is presented in the table, with all method codes derived from . Notably, SSF-LN, SSF-MLP, and SSF-Attention represent LayerNorm, MLP, and self-attention in the multi-head self-attention layer (MHSA) within the fine-tuned transformer block. Team A, Team B, and Team C represent the results of top methods in the test phase of this challenge.

Observing the results in Tab. 6, SSF-LN, BitFit, VPT, and PEL, based on the pre-trained base model, achieve the highest total object recognition score on the val dataset, all reaching 0.37. However, PEL stands out for its quicker training, requiring fewer than 20 epochs to achieve comparable results.

Our SPT method, after four stages of training, surpasses all others, achieving optimal SAR object recognition results on the val dataset with a total score of 0.44 and on the test dataset with a total score of 0.40. This represents a significant improvement over our benchmark method, PEL. Additionally, SPT* denotes the result of 20 iterations of our fourth-stage training strategy, boasting an impressive total score of 0.49. These results affirm the validity and superiority of our proposed method.

5. Conclusion

The fine-tuned foundation models excel in many downstream tasks. However, they struggle with SAR object recognition because of SAR's unique imaging and scattering characteristics. In this work, we introduce a novel approach named Scattering Prompt Tuning (SPT) based vision foundation model. It utilizes SAR image scattering information as a prompt, integrating learnable parameters into the pre-trained model's input space to help learn SAR's unique information. We also introduce lightweight modules to fine-tune the pre-trained foundation model. Additionally, a four-stage training strategy, incorporating semi-supervised learning, is deployed to enhance SAR object recognition performance further. The experimental results demonstrate the outstanding performance of our approach. Future work needs to investigate more effective foundation model fine-tuning methods for SAR object recognition.

https://github.com/shijxcs/PEL

Table 6. Classification performance of different fine-tuning methods in development phase and test phase.

Method	development phase (val dataset)				test phase (test dataset)			
	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score	Top-1 Accrucy(%)	AUROC	TNR at tpr95	Total Score
SSF-LN [41]	29.29	0.61	0.08	0.37	-	-	-	-
SSF-MLP [41]	28.43	0.59	0.05	0.36	-	-	-	-
SSF-Attention [41]	27.71	0.53	0.04	0.34	-	-	-	-
Lora [40]	23.81	0.54	0.07	0.31	-	-	-	-
LN Tuning	28.86	0.53	0.03	0.35	-	-	-	-
BitFit [51]	28.43	0.64	0.06	0.37	-	-	-	-
Adapter [39]	28.57	0.59	0.05	0.36	-	-	-	-
VPT [30]	29.44	0.61	0.06	0.37	-	-	-	-
PEL [10]	28.4	0.61	0.07	0.37	21.22	0.49	0.01	0.28
Team A	-	-	-	-	38.80	0.24	0.01	0.35
Team B	-	-	-	-	35.10	0.49	0.04	0.39
Team C	-	-	-	-	38.85	0.69	0.24	0.46
SPT(ours)	37.52	0.62	0.07	0.44	33.1	0.61	0.02	0.40
SPT*(ours)	-	-	-	-	37.9	0.83	0.22	0.49

References

[1] Linbin Zhang, Xiangguang Leng, Sijia Feng, Xiaojie Ma, Kefeng Ji, Gangyao Kuang, and Li Liu. Domain knowledge powered two-stream deep network for few-shot sar vehicle recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021. 1

[2] Shengyang Li, Zhuang Zhou, Manqi Zhao, Jian Yang, Weilong Guo, Yixuan Lv, Longxuan Kou, Han Wang, and Yanfeng Gu. A multi-task benchmark dataset for satellite video: Object detection, tracking, and segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1

[3] Zhongling Huang, Chong Wu, Xiwen Yao, Zhicheng Zhao, Xiankai Huang, and Junwei Han. Physics inspired hybrid attention for sar target recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 207:164–174, 2024. 1

[4] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 1, 3

[5] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1, 3

[6] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 1, 3

[7] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 1, 3

[8] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1, 3, 4, 5, 6

[9] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 1, 3

[10] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*, 2023. 1, 2, 8

[11] Zhiyuan Yan, Junxi Li, Xuexue Li, Ruixue Zhou, Wenkai Zhang, Yingchao Feng, Wenhui Diao, Kun Fu, and Xian Sun. Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[14] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[16] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[17] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[18] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[20] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[21] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 2

[22] Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022.

[23] Ruihan Yang, Minghao Zhang, Nicklas Hansen, Huazhe Xu, and Xiaolong Wang. Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers. *arXiv preprint arXiv:2107.03996*, 2021. 2

[24] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022. 2

[25] Sudeep Dasari and Abhinav Gupta. Transformers for one-shot visual imitation. In *Conference on Robot Learning*, pages 2071–2084. PMLR, 2021. 2

[26] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021. 2, 3

[27] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022. 2

[28] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 2, 3

[29] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. *arXiv preprint arXiv:2210.01033*, 2022. 2

[30] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 8

[31] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[32] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[33] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2

[34] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[35] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[36] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 2

[37] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022. 2

[38] Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, et al. Hyperprompt: Prompt-based task-conditioning of transformers. In *International Conference on Machine Learning*, pages 8678–8690. PMLR, 2022.

[39] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 8

[40] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023. 8

[41] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline

for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 2, 8

[42] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2, 4

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[44] Han Wang, Silei Liu, Yixuan Lv, and Shengyang Li. Scattering information fusion network for oriented ship detection in sar images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 4

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[46] Feng Cai, Keyu Wu, Haipeng Wang, and Feng Wang. A three-stage framework with reliable sample pool for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 479–486, 2023. 6

[47] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 6

[48] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 6

[49] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 6

[50] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 6

[51] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 8