remote sensing

MDPI

Article

Dynamic Expansion Mixture-of-Experts with Pre-Trained Vision Transformer for Few-Shot Class-Incremental Remote Sensing Scene Classification

Yunhao Wu [1,2], Xiang Li [1,2], Jianlin Zhang [1,*], Haorui Zuo [1], Hui Li [1,2] and Tong Tan [1,2]

[1] State Key Laboratory of Optical Field Manipulation Science and Technology, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China; wuyunhao24@mails.ucas.ac.cn (Y.W.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: jlin@ioe.ac.cn

Highlights

What are the main findings?

- We propose a Dynamic Expansion Mixture-of-Experts with Pre-trained Vision Transformer (DEM-ViT) framework to address the domain gap between natural images and remote sensing images.
- We design an Adapter-Based Mixture-of-Experts (ABMoE) module to capture the diverse visual patterns of remote sensing scenes, along with a Dynamic Expert Expansion (DEE) strategy to progressively expand the model capacity.

What are the implications of the main findings?

- The proposed method enables the pre-trained Vision Transformer to be effectively adapted for remote sensing scene classification, achieving superior recognition performance.
- The proposed method empowers the model to continually acquire new knowledge with limited samples while protecting previously acquired knowledge from being forgotten.

Abstract

Few-Shot Class-Incremental Learning (FSCIL) aims to sequentially learn new classes from very few labelled samples while preventing the forgetting of previously acquired knowledge, which has important practical value for remote sensing scene classification (RSSC). Recent studies have shown that applying a Vision Transformer (ViT) pre-trained on natural image datasets to FSCIL tasks can achieve significantly superior performance. Nevertheless, a substantial domain distribution gap exists between natural images and remote sensing images, which leads to severe performance degradation when such models are directly transferred to RSSC. To address the domain gap alongside FSCIL's inherent stability–plasticity dilemma and overfitting under data scarcity, we propose a Dynamic Expansion Mixture-of-Experts with Pre-trained Vision Transformer (DEM-ViT) framework. Specifically, to alleviate the domain discrepancy, DEM-ViT incorporates an Adapter-Based Mixture-of-Experts (ABMoE) module, which captures the diverse visual patterns of remote sensing scenes through feature reconstruction in the representation space and collaborative learning among multiple experts. Furthermore, to address the stability–plasticity dilemma in FSCIL, we propose a Dynamic Expert Expansion (DEE) strategy, which progressively expands the model capacity along the incremental sessions. DEE provides sufficient space for learning new knowledge while mitigating the forgetting of previous knowledge. In addition, we propose a Semantic-Guided Feature Alignment (SGFA) method to reduce the risk of overfitting under data-scarce conditions. SGFA leverages textual information to construct

Check for updates

Academic Editor: Zhenwei Shi

Received: 21 March 2026
Revised: 7 April 2026
Accepted: 9 April 2026
Published: 12 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

robust text prototypes and uses them to calibrate the visual feature space. Extensive experiments across three benchmarks indicate that our framework exhibits highly competitive performance compared with state-of-the-art methods.

Keywords: few-shot class-incremental learning; remote sensing scene classification; parameter-efficient fine-tuning; mixture-of-experts; pre-trained language model

1. Introduction

Remote Sensing Scene Classification (RSSC) aims to achieve semantic understanding and category discrimination of complex land-cover scenes within remote sensing images. It provides essential technical support for practical applications such as geological exploration [1], land-use analysis [2], and environment monitoring [3]. Owing to their robust modelling capabilities and superior performance, deep learning methodologies play a pivotal role in RSSC tasks [4]. However, conventional deep learning approaches typically rely on the "closed-world assumption," which means the training set is expected to cover all possible classes, and the model will not encounter any unseen classes during the testing phase [5]. This assumption presents a significant divergence from real-world remote sensing scenarios. As observation tasks proceed continuously, new land-cover classes inevitably emerge over time in the real world. Additionally, constrained by restricted access to observation platforms and the high costs of expert annotation, acquiring large-scale labelled samples for these emerging classes is often challenging [6]. The coexistence of dynamic class expansion and extreme sample scarcity poses a significant challenge to RSSC.

To cope with these problems, incorporating Few-Shot Class-Incremental Learning (FSCIL) into the RSSC has emerged as a promising and essential direction. FSCIL focuses on continuously acquiring knowledge of new classes with few labelled samples, while maintaining performance on previously learned classes [7]. In recent years, FSCIL has attracted widespread attention from researchers, and various advanced approaches have been developed. SoftNet [8] constructed soft sub-networks from the original architecture using non-binary masks, effectively addressing both overfitting and catastrophic forgetting. During the base session, both sub-network parameters and weight scores were learned. During the incremental sessions, only a subset of minor parameters within the sub-network was updated. Kukleva et al. [9] proposed a three-stage framework. The first two stages focused on learning the base classes and the new classes, respectively, employing model parameter constraints to prevent the forgetting of previous classes. In the third stage, a small set of stored exemplars was utilised for replay to improve the classifier's performance. Inspired by the theory of Neural Collapse, NC-FSCIL [10] aligned the set of prototypes formed in the course of neural collapse with the sample prototypes, thereby enhancing classifier performance. Furthermore, ABDS [11] assigned independent and learnable boundary parameters to each class. By dynamically shrinking these boundaries during training, it forced an increase in inter-class distances within the feature space, ensuring clear decision boundaries. In the field of FSCIL-RSSC, Zeng et al. [12] improved few-shot incremental aerial scene classification by combining relation-aware feature recalibration with continual prototype refinement to better balance old-class retention and new-class adaptation. HDCPAA [13] advances remote sensing few-shot class-incremental recognition by constructing hyperdimensional class prototypes and adaptively refining prototype matching, thereby promoting clearer inter-class separation during incremental updates.

However, many existing FSCIL-RSSC methods are built upon convolutional neural networks (CNNs). Constrained by the inherent local inductive bias and limited receptive

fields of convolutional operations, CNNs struggle to capture global contextual information [14]. Under data-scarce conditions, CNNs tend to focus on local textures or regions rather than the overall scene semantics, which limits their discriminative ability for complex classes during incremental sessions.

Recently, ViT-based architectures have attracted growing interest in FSCIL, benefiting from their superior ability to capture global semantics of images, compared with CNN-based approaches [15,16]. Due to the massive parameter count of ViT, training models from scratch is computationally expensive. Therefore, utilising ViT pre-trained on large-scale datasets as the backbone has become the standard paradigm in the current FSCIL landscape. Building upon this, Wang et al. [17] proposed an optimisation framework that minimised the upper bounds of transfer risk and consistency risk, and derived a set of principled guidelines for model design. These guidelines included preventing excessive feature concentration, reducing classification boundary discrepancy, and ensuring unbiased classification. Liu et al. [18] introduced the SEC-Prompt framework, which utilised D-Prompts to enhance the discriminability of key features while employing ND-Prompts to capture task-irrelevant, non-discriminative features to promote generalisation, thereby effectively balancing the learning of new concepts with the retention of prior knowledge under few-shot conditions. To fully exploit pre-trained knowledge, Chen et al. [19] proposed the BiMC method. By combining fine-grained text descriptions generated by the pre-trained language model with visual prototypes for Intra-modal Calibration, and fusing visual priors with linguistic knowledge for Inter-modal Calibration, they achieved robust incremental inference in a training-free manner.

Nevertheless, remote sensing images have unique features, such as overhead viewpoints, large-scale variations, and complex backgrounds, that differ significantly from those of natural images. Many existing methods built upon Vision Transformers adopt prompt-learning strategies, which mainly introduce learnable visual tokens into the input space. While this strategy is effective for processing natural images, prompt-based methods exhibit suboptimal performance on downstream tasks when adapting pre-trained ViTs [20]. Prompt-based methods are limited by the fixed weights of the ViT pre-trained on the natural image domain. As a result, the generic knowledge encoded in pre-trained ViT is often difficult to transfer effectively to highly specialized remote sensing tasks, leading to degraded discrimination performance on complex land-cover classes. For example, the state-of-the-art SEC-Prompt [18] method achieves an average accuracy of 89.58% on CIFAR100, but it only reaches 76.80% on the remote sensing dataset NWPU-RESISC45. Similarly, directly applying a ViT pre-trained on ImageNet-21K yields an average accuracy of 89.08% on miniImageNet, whereas it only achieves 82.62% on the UC Merced-LandUse dataset. These results indicate that existing ViT-based methods, especially prompt-based methods, still suffer a significant performance drop compared to their results on natural images. Therefore, bridging the domain gap and learning highly discriminative scene representations from limited samples remains a critical challenge for applying ViT-based FSCIL paradigms to RSSC tasks.

Furthermore, the main goal of FSCIL-RSSC is to help models continuously learn new knowledge with very few samples. To prevent the model from forgetting old knowledge, most FSCIL methods freeze the feature extractor to keep the feature space fixed. However, constantly adding new classes into this fixed space causes feature distribution collapse, which leads to category confusion [21]. Moreover, the extreme lack of data greatly increases the risk of overfitting on new classes. Remote sensing images usually have complex backgrounds and many distracting objects. When learning new land-cover classes from limited samples, the model is highly prone to overfitting local noise or unimportant details, inherently hindering the extraction of generalizable features. Ultimately, a successful

framework must solve the stability–plasticity dilemma, which is the fundamental challenge in FSCIL. On the one hand, the model needs high stability to remember old knowledge. On the other hand, it must have enough plasticity to learn new knowledge.

To address the challenges mentioned above in FSCIL-RSSC scenarios, we propose a Dynamic Expansion Mixture-of-Experts with Pre-trained Vision Transformer (DEM-ViT) framework for RSSC. DEM-ViT aims to learn discriminative remote sensing features while balancing the trade-off between plasticity and stability, thereby preventing catastrophic forgetting and overfitting under data-scarce incremental conditions. Specifically, to effectively mitigate domain discrepancy, we design an Adapter-Based Mixture-of-Experts (ABMoE) module. This module introduces trainable adapter experts that can reconstruct features in the representation space and adaptively aggregate multiple experts through a Mixture-of-Experts architecture. By increasing the capacity of trainable parameters, ABMoE enables comprehensive modelling of diverse visual patterns in complex remote sensing scenes, thereby effectively alleviating the severe domain shift encountered when transferring models pre-trained on natural images to remote sensing tasks. In addition, to address the well-known stability–plasticity dilemma in FSCIL, and to prevent the resulting feature distribution collapse and class confusion, we propose a Dynamic Expert Expansion (DEE) strategy, which progressively expands the number of experts and the model capacity along the incremental sessions. This strategy provides sufficient space for acquiring new concepts while effectively reducing forgetting of previously acquired knowledge. Furthermore, to mitigate the overfitting risk caused by extreme data scarcity in FSCIL-RSSC and to improve the overall stability of feature representations, we introduce a Semantic-Guided Feature Alignment (SGFA) method. SGFA fully exploits the rich and robust prior knowledge embedded in textual information generated by the pre-trained language model to construct stable class-level semantic constraints, and calibrates the visual feature space through a cross-modal alignment mechanism, thereby maintaining clear and robust decision boundaries.

Extensive experiments are conducted on three standard RSSC benchmarks, namely WHU-RS19, UC Merced LandUse, and NWPU-RESISC45. Experimental results verify that our method is competitive with existing state-of-the-art approaches, validating its effectiveness and robustness in handling complex remote sensing scenes and dynamically evolving class-incremental settings. Our main contributions can be summarised as follows:

- To address the challenges of adopting pre-trained ViT in FSCIL-RSSC, we propose a Dynamic Expansion Mixture-of-Experts with Pre-trained Vision Transformer (DEM-ViT) framework with three synergistic components.
- We design an Adapter-Based Mixture-of-Experts module to better reconstruct features in the representation space, a Dynamic Expert Expansion strategy to progressively increase model capacity, and a Semantic-Guided Feature Alignment method to calibrate the visual space with textual information.
- Experiments on three benchmarks demonstrate that our method is competitive with existing state-of-the-art approaches.

2. Related Works

2.1. Few-Shot Class-Incremental Learning

Few-Shot Class-Incremental Learning (FSCIL) is dedicated to empowering models to sequentially learn new classes from few annotations while preserving previous knowledge. TOPIC [22] employed a neural gas network to preserve the topology of feature manifolds, thereby mitigating the distortion of old class representations. To further optimise the parameter space, WaRP [23] proposed transforming the original parameter space into a new coordinate system, effectively compacting prior knowledge into a few essential param-

eters, thereby preserving capacity for future incremental learning. SAVC [24] introduced a semantic-aware visual contrastive learning scheme, utilising virtual classes as placeholders to promote separation between new and base classes. TEEN [25] developed a task-adaptive feature extraction framework, enhancing the discriminability of new classes by fusing weighted base prototypes. To overcome the capacity limitations and the inherent risk of feature distribution crumbling in static architectures, DSN [21] introduced a compressive node expansion strategy for FSCIL. It tentatively expanded network nodes to accommodate new classes and subsequently compressed them via a self-activation mechanism, effectively supporting the feature space while mitigating overfitting. More recently, ViTs have been widely adopted owing to their robust feature representation capability. ASP [15] addressed the overfitting issue in few-shot prompt tuning by introducing a lightweight Prompt Encoder trained with an Information Bottleneck objective. This mechanism dynamically generates self-adaptive prompts directly from input images, effectively reducing the dependency on extensive training data. Meanwhile, PriViLege [16] incorporates a Prompt Modulation mechanism, designed to effectively capture domain-specific knowledge during the base session while preserving the general knowledge of the pre-trained model.

In the field of remote sensing, FSCIL for synthetic aperture radar (SAR) target recognition has been studied actively. CPL [26] constructed class prototypes using cosine similarity to capture azimuth-aware structural information of SAR targets, and introduced Pairwise Structure Separation and Prototype–Exemplar Distillation to balance model stability and plasticity in the cosine embedding space. FCPC [27] emphasised forward compatibility in SAR automatic target recognition by synthesising virtual classes via a Virtual-class Semantic Synthesiser and employing a Decoupled Margin Adaptation strategy to efficiently adapt to new classes while preserving low-level general representations.

In the broader context of RSSC, diverse frameworks have been proposed to address the challenges posed by complex spatial structures and significant scale variations. Zeng et al. [12] proposed a framework based on feature adaptation and prototype continuous optimisation, which dynamically adjusts feature weights via sample relation coefficients and employs a prototype correction mechanism to enhance discrimination and mitigate catastrophic forgetting during incremental stages. HDCPAA [13] introduced hyperdimensional computing to generate quasi-orthogonal prototypes and employs a prototype adaptive attention module to optimise decision boundaries, thereby enhancing feature discriminability. To address large distribution shifts and multiscale issues in scenes, G-MFCN [28] proposed a gradient-guided multiscale feature collaboration network, which adaptively fuses hierarchical features from parallel backbones. More recently, leveraging parameter-efficient fine-tuning, TerraSAP [29] developed a spatially aware prompt framework that injects spatial structural priors into the frozen backbone, strengthening the model's capability to capture complex topological structures in remote sensing scenes.

In this paper, our method explores FSCIL in the field of RSSC, utilising the pre-trained Vision Transformer as the backbone to effectively capture highly heterogeneous scene features, and employing the DEE strategy to prevent feature distribution crumbling and alleviate the stability–plasticity dilemma. Rather than adding entire layers, our DEE strategy progressively inserts lightweight Adapter experts into each Transformer block, achieving MoE-aligned capacity expansion. By initializing new experts and routers with the mean parameters of existing ones, the model safely absorbs novel concepts without disrupting accumulated knowledge. Benefiting from the bottleneck structure of Adapters, the parameter overhead remains minimal. Consequently, DEE serves as a highly efficient, MoE-based expansion strategy tailored for FSCIL-RSSC, distinguishing itself from traditional methods that suffer from massive parameter growth.

2.2. Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) aims to achieve effective adaptation of large-scale pre-trained models with a minimal number of learnable parameters, which originated in the field of Natural Language Processing [30,31]. Recently, with the emergence of ViTs, this paradigm has been widely applied to computer vision tasks.

Among representative approaches, LoRA [32] pioneered modelling weight updates as the product of two low-rank matrices, significantly reducing storage requirements. Taking this a step further, FacT [33] constructed the weight increments across all ViT layers into a unified 3D tensor and introduced tensor decomposition techniques to exploit inter-layer redundancy, thereby generating extremely lightweight update factors.

Beyond weight decomposition, another mainstream strategy involves embedding specific functional sub-modules within the network. AdaptFormer [20] proposed a parallel adapter architecture that inserted lightweight modules alongside MLP layers, achieving higher accuracy than full fine-tuning on standard action recognition datasets. To address the inherent lack of inductive bias in ViT, ConvPass [34] employed a bottleneck module with convolutional layers as a feature shaper, introducing locality priors to mitigate performance gaps in few-shot scenarios.

Beyond the modification of network weights, directly optimising inputs and intermediate feature representations constitutes another effective approach.VPT [35] proposed inserting learnable tokens at the input layer of the ViT, enabling them to participate in self-attention calculations alongside image patches to facilitate adaptation to downstream tasks. VQT [36] extended this concept to deeper network layers by introducing learnable 'query' tokens at each stage, utilising cross-attention mechanisms to aggregate intermediate features and enhance feature interaction capabilities. Furthermore, SSF [37] modulated feature distributions by inserting learnable scaling and shifting coefficients after each operation. These parameters were reparameterized into the backbone weights during the inference phase, achieving zero inference latency.

SSIAT [38] highlighted that the Adapter outperforms other PEFT techniques in incremental learning settings. Inspired by this, we employ the Adapter as our foundational paradigm for efficient fine-tuning. Furthermore, we incorporate prompt-tuning strategies to leverage semantic priors from the pre-trained language model, which serve as a robust regularizer to alleviate overfitting under data-scarce conditions.

To address the inherent instability and overfitting issues of pure visual feature reconstruction under extreme few-shot constraints, our SGFA method introduces a token-level PEFT scheme. Unlike traditional approaches that use text or semantic information to assist visual recognition, SGFA incorporates a semantic token parallel to the class token to specifically handle cross-modal alignment. Using KL divergence distillation, the semantic token is constrained to match the text prototypes generated by the pre-trained language model, providing essential text-driven regularization for new classes with limited data. During the inference phase, a dual-branch logits fusion integrates these semantic priors into the final decision. Therefore, SGFA goes beyond traditional cross-modal paradigms, offering a cross-modal semantic guidance approach designed for PEFT, specifically tackling the challenges of FSCIL-RSSC.

2.3. Mixture-of-Experts

Mixture-of-Experts (MoE) [39] has been widely adopted in vision tasks as a paradigm to enhance model capacity and efficiency. Early research primarily explored MoE construction based on CNN-based models. DeepME [40] addressed large-scale image classification by employing hierarchical clustering to assign semantically similar classes to the same expert and fusing predictions via a gating network.

With the widespread adoption of Vision Transformer, MoE architectures have experienced renewed interest and rapid advancement in the vision domain. V-MoE [41] introduced sparse gating mechanisms that scale model capacity to 15B parameters while reducing inference computation by approximately 50%. Berisha et al. [42] proposed extracting sub-networks as experts by clustering MLP layer activations, thereby converting dense ViTs into sparse MoE models without training from scratch.

Beyond basic architecture, the design of the Routing Mechanism is critical to performance. Liu et al. [43] systematically compared routing strategies in vision tasks, finding that Soft MoE often outperforms Sparse MoE under fixed compute budgets, and Expert Choice strategies generally surpass Token Choice. To further enhance the interpretability of the routing mechanism, Min et al. [44] introduced a spatially aware auxiliary loss to guide experts towards foreground regions, leveraging external semantic priors to improve routing quality.

Despite the immense potential of MoE, challenges remain regarding training stability and structural design. ViMoE [45] revealed that improper layer configurations lead to routing instability and suggested introducing shared experts to learn general information. Videau et al. [46] provided empirical design principles, identifying that medium-scale expert activation and placing MoE layers at the end of the network achieve the optimal performance-efficiency trade-off.

To mitigate the significant domain shift issues in remote sensing imagery, we propose the ABMoE module. Unlike traditional MoE methods, each expert in our proposed ABMoE is implemented as an Adapter rather than a full MLP branch. By inserting multiple parallel Adapter experts into each Transformer block while keeping the backbone parameters frozen, ABMoE achieves fine-grained feature reconstruction in a parameter-efficient manner, which is crucial for remote sensing deployment. Moreover, ABMoE is attached in parallel to the original MLP branch and integrated through a residual path, which allows the pre-trained ViT backbone to retain its general representation capability, while the experts focus on modeling remote-sensing-specific patterns. Therefore, ABMoE should be viewed as an MoE-Adapter design specifically addressing domain shift in FSCIL-RSSC, rather than a direct reuse of existing MoE architectures.

3. Methods

3.1. Preliminary

The objective of FSCIL is to train the model sequentially across T sessions. Formally, the entire labeled dataset is partitioned into a sequence of subsets $\{\mathcal{D}^{(0)}, \mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(T)}\}$, where each subset is denoted as $\mathcal{D}^{(t)} = \left\{\left(\mathbf{x}_j^{(t)}, y_j^{(t)}\right)\right\}_{j=1}^{|\mathcal{D}^{(t)}|}$. Specifically, the dataset $\mathcal{D}^{(0)}$ at $t = 0$ constitutes a large-scale dataset for the base session, whereas $\mathcal{D}^{(t)}$ for $t \geq 1$ are small-scale datasets containing several new classes.

Let $\mathcal{C}^{(t)}$ denote the label space of dataset $\mathcal{D}^{(t)}$. The label sets across different sessions are mutually disjoint, namely $\mathcal{C}^{(i)} \cap \mathcal{C}^{(j)} = \varnothing$ for all $i \neq j$. Furthermore, the dataset $\mathcal{D}^{(t)}$ in each session is partitioned into a training set $\mathcal{D}_{train}^{(t)}$ and a testing set $\mathcal{D}_{test}^{(t)}$.

The N-way K-shot FSCIL task can be formulated as follows: In the base session, the model is trained on the dataset $\mathcal{D}_{train}^{(0)}$ using all available samples, which allows the model to learn generalizable feature representations from classes with sufficient samples. In subsequent incremental sessions t ($t = 1, \ldots, T$) where $|\mathcal{C}^{(t)}| = N$, K examples are randomly sampled from each class in $\mathcal{D}_{train}^{(t)}$. The model is trained on this sampled N-way K-shot dataset, which means it only sees a few images of the new classes. For evaluation, the model is tested on the combined test sets from all previous sessions, denoted as

$\mathcal{D}_{test}^{t} = \mathcal{D}_{test}^{(0)} \cup \mathcal{D}_{test}^{(1)} \cup \cdots \cup \mathcal{D}_{test}^{(t)}$, and the classification accuracy $\mathcal{A}_t$ is then calculated. The ideal FSCIL model maintains strong performance across all seen classes.

3.2. Overview of the Proposed DEM-ViT Framework

The overall architecture of the proposed DEM-ViT framework is illustrated in Figure 1. We adopt the frozen pre-trained Vision Transformer as the backbone to extract discriminative representations and address the FSCIL task via parameter-efficient fine-tuning. Specifically, the DEM-ViT framework is built upon three key components: (1) an Adapter-Based Mixture-of-Experts fine-tuning module (Section 3.3), (2) a Dynamic Expert Expansion strategy (Section 3.4), and (3) a Semantic-Guided Feature Alignment method (Section 3.5).

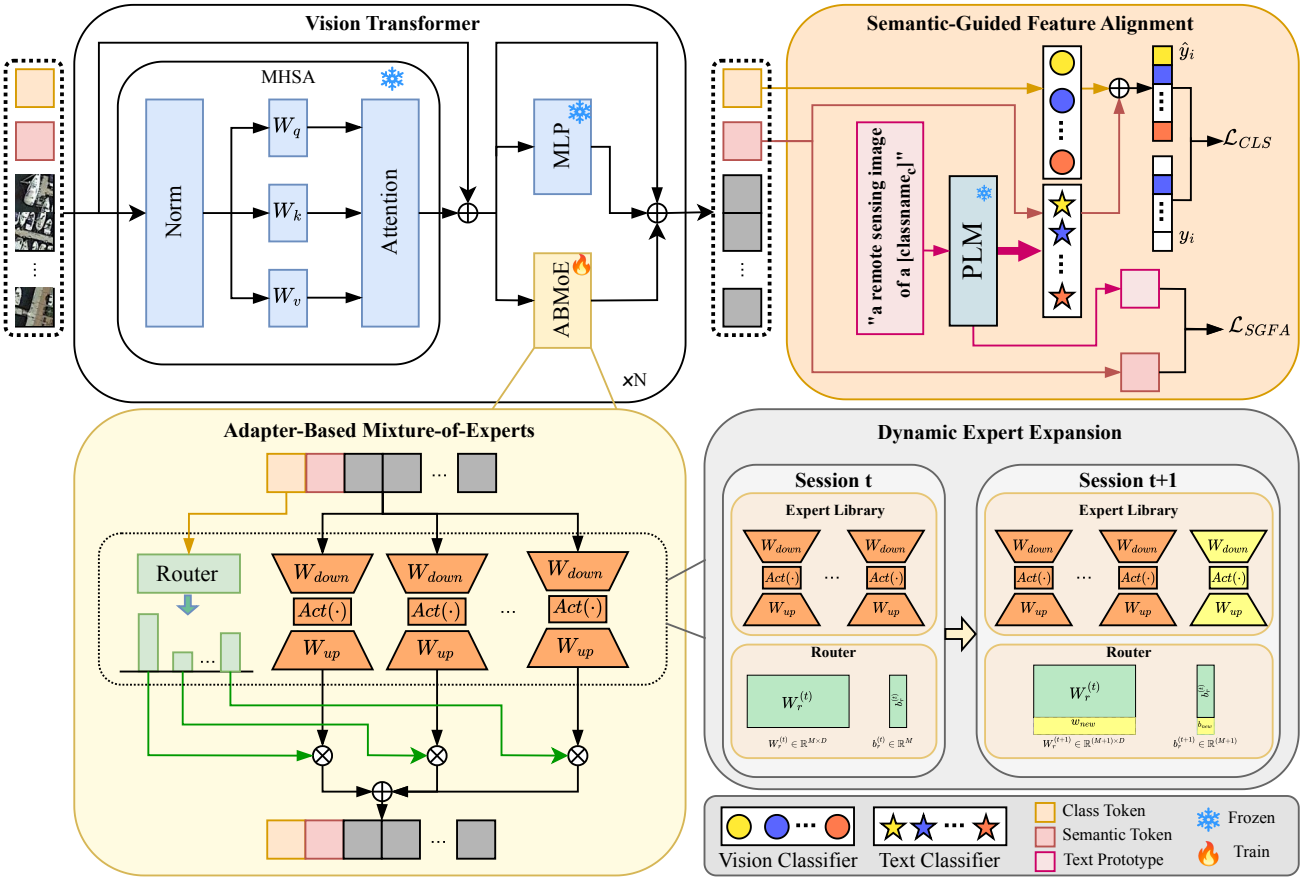


Figure 1. The overall framework of the DEM-ViT. The framework synergistically combines the Adapter-Based Mixture-of-Experts module, the Dynamic Expert Expansion strategy, and the Semantic-Guided Feature Alignment method.

Firstly, to achieve efficient task adaptation while overcoming the domain shift problem, we employ an Adapter-Based Mixture-of-Experts (ABMoE) fine-tuning module. Distinct from traditional full fine-tuning or single-adapter paradigms, we keep the pre-trained ViT weights entirely frozen and construct a Mixture-of-Experts module composed of multiple lightweight Adapters to fine-tune the backbone. In this design, each Adapter serves as an independent "expert" focusing on capturing specific feature patterns.

Secondly, to accommodate the incremental learning setting, we introduce a Dynamic Expert Expansion strategy. This strategy adds more experts every few sessions as the learning process continues, retaining previously acquired concepts while effectively learning

new knowledge, thereby alleviating catastrophic forgetting and ensuring sufficient capacity for learning new classes.

Finally, to address the potential risk of overfitting in few-shot settings, we introduce the Semantic-Guided Feature Alignment method. Specifically, this method leverages the semantic word embeddings corresponding to class labels as prior knowledge. By guiding the model to learn highly semantic features via cross-modal distribution alignment, it provides much-needed information for new classes and helps maintain strong feature discriminability from extreme data scarcity.

3.3. Adapter-Based Mixture-of-Experts

This section elucidates the architectural design of the Adapter-Based Mixture-of-Experts module, detailing the mechanism of individual Adapter experts and the routing strategy employed to aggregate their outputs dynamically.

The Adapter module adopts the classic bottleneck architecture. Given the input x to the Adapter module, it is first projected via a down-projection layer $W_{down} \in \mathbb{R}^{D \times d}$, where D represents the feature dimension and d represents the bottleneck dimension (where $d \ll D$). This is followed by a non-linear activation function $\sigma(\cdot)$. Finally, an up-projection layer $W_{up} \in \mathbb{R}^{d \times D}$ is employed to restore the features to their original dimension. Formally, this computational process is expressed as:

$$Adapter(x) = W_{up}^T \cdot \sigma(W_{down}^T \cdot x) \tag{1}$$

While a single Adapter provides an efficient fine-tuning method, it is limited to learning one fixed feature transformation. This is insufficient for remote sensing images, which exhibit high visual diversity and significant differences between scenes. Forcing a single Adapter to handle all classes makes it difficult to capture and decouple these diverse semantic features effectively. To enable the model to analyse and reassemble features from different perspectives simultaneously, we extend the single Adapter into a Mixture-of-Experts architecture. By integrating multiple experts specialised in different feature patterns, the model can construct more robust and comprehensive feature representations.

Specifically, within each layer of the ViT backbone, we construct an expert library denoted as $\mathcal{E} = \{E_1, E_2, \ldots, E_M\}$, comprising M experts. Each expert $E_i(\cdot)$ (for $i = 1, \ldots, M$) is an Adapter module, formulated as $E_i(x) = \text{Adapter}(x)$, and possesses independent weight parameters. This design empowers each expert to specialise in specific semantic patterns within the feature space, thereby achieving implicit feature disentanglement. The outputs from these experts are aggregated via a learnable routing module $\mathcal{R}(\cdot)$. This router computes importance scores for each expert utilising the input class token and subsequently combines the outputs from the experts in a weighted manner using the normalised scores.

This process can be formally described as follows: The Router maps the class token $x_{CLS} \in \mathbb{R}^D$ of the input sample into an expert score space via a linear layer and subsequently produces a normalised probability score using the Softmax function:

$$w(x) = Softmax(W_r \cdot x_{CLS} + b_r) \tag{2}$$

where $W_r \in \mathbb{R}^{M \times D}$ and $b_r \in \mathbb{R}^M$ denote the weight matrix and bias term of the Router, respectively, the vector $w(x) \in \mathbb{R}^M$ represents the importance weight of each expert, with $w_i(x)$ corresponding to the i-th element of $w(x)$. Consequently, the output of the ABMoE module is computed as the weighted sum of the activated experts' outputs:

$$MoE(x) = \sum_{i \in M} w_i(x) \cdot E_i(x) \tag{3}$$

By leveraging this expert weighting mechanism, the model dynamically combines the capabilities of distinct semantic experts based on sample characteristics, thereby effectively capturing the complex and diverse feature patterns inherent in remote sensing scenes.

We add the proposed ABMoE module to each block of the ViT backbone. Specifically, the ABMoE module is introduced as an additional branch parallel to the original MLP layer, operating on the same output features from the Multi-Head Self-Attention (MHSA) layer. Formally, for the l-th Transformer layer, let x_{l-1} denote the input features. First, the input undergoes standard MHSA processing followed by a residual connection:

$$x'_l = x_{l-1} + MHSA(x_{l-1}) \tag{4}$$

In the standard ViT architecture, the feature x'_l is processed by the MLP block and aggregated via a residual connection:

$$x_l^{ori} = x'_l + MLP(x'_l) \tag{5}$$

In contrast, upon incorporating the ABMoE module, x'_l is fed into the parallel MLP and ABMoE branches simultaneously. Consequently, the outputs from both modules are aggregated with the residual connection:

$$x_l = x'_l + MLP(x'_l) + ABMoE(x'_l) \tag{6}$$

Through this parallel design, the ABMoE acts as a fine-tuning module introduced via a residual path, enabling the model to flexibly refine and enhance feature representations without disrupting the weights of the backbone network.

3.4. Dynamic Expert Expansion

While the ABMoE module mentioned above enhances the model's capacity to represent complex features through multi-expert synergy, maintaining a fixed number of experts poses a significant challenge during the incremental session of FSCIL. Specifically, the model remains trapped in the stability–plasticity dilemma. On the one hand, forcing updates on existing expert parameters in subsequent incremental sessions inevitably disrupts the established structure, triggering catastrophic forgetting. On the other hand, a fixed number of experts constrains the model's plasticity, hindering the effective acquisition of novel concepts.

To effectively handle this issue, we propose the Dynamic Expert Expansion strategy, which enables the model to evolve by progressively increasing the number of experts as incremental sessions advance, thereby effectively reconciling the stability–plasticity dilemma.

Suppose that at session $t-1$, the model comprises an expert library $\mathcal{E}_{t-1}$ consisting of M_{t-1} experts. Upon entering the new incremental session t, we append a novel expert module E_{new} to the expert library. The updated expert library becomes $\mathcal{E}_t = \mathcal{E}_{t-1} \cup \{E_{new}\}$, and the total count of experts increases to $M_t = M_{t-1} + 1$. We initialise this new expert utilising the mean parameters of the existing experts. Formally, the parameters θ_{new} for the new expert are computed as:

$$\theta_{new} = \frac{1}{M_{t-1}} \sum_{i=1}^{M_{t-1}} \theta_i \tag{7}$$

where θ_i denotes the parameter set of the i-th expert. Our initialisation strategy enables the new expert to inherit the prior knowledge of established experts, thereby significantly accelerating convergence under few-shot conditions and mitigating the risk of overfitting.

Along with the expansion of the expert library, the output dimension of the Router module requires corresponding adjustment. We expand the weight matrix $W_r^{(t-1)}$ and bias vector $b_r^{(t-1)}$ of the Router to align with the updated number of experts M_t. Following the

same strategy as the new expert module, we initialise the new Router parameters using the mean of the existing parameters. Formally, this is expressed as:

$$w_{new} = \frac{1}{M_{t-1}} \sum_{i=1}^{M_{t-1}} W_{r,i}^{(t-1)} \tag{8}$$

$$b_{new} = \frac{1}{M_{t-1}} \sum_{i=1}^{M_{t-1}} b_{r,i}^{(t-1)} \tag{9}$$

where $W_{r,i}^{(t-1)}$ and $b_{r,i}^{(t-1)}$ denote the weight vector and corresponding bias associated with the i-th expert from the previous session. $W_{r,i}^{(t-1)} \in \mathbb{R}^{M_{t-1} \times D}$, $b_{r,i}^{(t-1)} \in \mathbb{R}^{M_{t-1}}$, and $w_{new} \in \mathbb{R}^{D}$, $b_{new} \in \mathbb{R}$. The updated Router parameters are obtained via concatenation:

$$W_r^{(t)} = \text{Concat}(W_r^{(t-1)}, w_{new}^T) \tag{10}$$

$$b_r^{(t)} = \text{Concat}(b_r^{(t-1)}, b_{new}) \tag{11}$$

3.5. Semantic-Guided Feature Alignment

During the incremental sessions, the extreme scarcity of new class samples hinders the construction of robust feature representations. Conversely, the Pre-trained Language Model (PLM) contains broad semantic knowledge, which can provide robust and consistent prototypes for each class. To leverage this characteristic, we propose the Semantic-Guided Feature Alignment method.

To acquire accurate text prototypes, we design text prompts specific to the remote sensing domain. We employ the template $\mathcal{T}(c) =$ "a remote sensing image of a $[classname_c]$" and extract the corresponding text embeddings via the PLM to serve as text prototypes:

$$P_c^{text} = \Phi(\mathcal{T}(c)) \tag{12}$$

where Φ denotes the text encoder of the PLM.

To decouple the tasks of visual discrimination and semantic alignment within a single network, we introduce an additional learnable semantic token alongside the standard class token. Given an input image x_i, the input sequence X_i for the frozen ViT backbone is defined as:

$$X_i = [x_{cls,i}, x_{sem,i}, x_{p,i}^1, \ldots, x_{p,i}^N] + E_{pos} \tag{13}$$

where $x_{p,i}^j$ represents the patch embeddings of the image, E_{pos} denotes the position embeddings, and $x_{cls,i}$ is the class token. The term $x_{sem,i}$ denotes the semantic token, which is a randomly initialised learnable vector independent of the class token. In this design, the class token specialises in aggregating discriminative visual features, whereas the semantic token focuses on capturing high-level semantic features. While both tokens share contextual information within the encoder, they serve distinct roles during inference.

We employ knowledge distillation to constrain the output feature $z_{sem,i}$ of the semantic token, guiding it to align with the semantic prototype $P_{c_i}^{text}$ corresponding to the ground-truth c_i. The purpose is to minimise the Kullback–Leibler (KL) divergence loss:

$$\mathcal{L}_{SGFA} = \mathcal{L}_{KD}(z_{sem,i}, P_{c_i}^{text}) \tag{14}$$

To fully exploit the high-level semantic information captured by the semantic token, we design a dual-branch inference mechanism during the test phase. This mechanism aims to calibrate visual predictions using semantic cues. For a given test sample, the model computes logits from two complementary perspectives in parallel. On the one hand, the

class token generates visual logits $logits_{vis} = z_{cls,i} \cdot (P^{vis})^T$ via the vision classifier, where $P^{vis} = [p_1, p_2, \ldots, p_K]$ is the prototype classifier, and each prototype p_c is obtained by averaging the feature representations:

On the other hand, utilizing the text classifier $P^{text} = [P_1^{text}, P_2^{text}, \ldots, P_N^{text}]$, the semantic token calculates its similarity with the text prototypes of all classes, formulated as $logits_{sem} = z_{sem,i} \cdot (P^{text})^T$. The logits are obtained via a weighted fusion of both branches:

$$logits = logits_{vis} + \lambda_{sem} \cdot logits_{sem} \tag{15}$$

Finally, the network parameters are optimised using the Cross-Entropy loss. Let $\hat{y}_i = \text{Softmax}(logits)$ denote the predicted probability distribution. The loss objective is formulated as:

$$\mathcal{L}_{CLS} = L_{CE}(\hat{y}_i, y_i) \tag{16}$$

where y_i denotes the one-hot encoding of the ground truth label for the input image x_i.

Our inference mechanism ensures that the model retains access to rich semantic context even under data-scarce conditions, effectively mitigating the overfitting caused by limited sample availability.

The total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{CLS} + \lambda_{SGFA} \cdot \mathcal{L}_{SGFA} \tag{17}$$

4. Experiments

4.1. Experimental Settings

To comprehensively validate the performance and robustness of our proposed framework, we carried out a series of experiments on three widely utilised benchmark datasets in the field of RSSC: WHU-RS19 [47], UC Merced-LandUse [48], and NWPU-RESISC45 [49].

WHU-RS19, collected from Google Earth satellite imagery, is designed for remote sensing scene classification. It comprises 1005 images across 19 scene categories. Each image has a size of 600 × 600 pixels. The UC Merced-LandUse dataset is built from aerial imagery and includes 21 land-use classes, with 100 images per class at 256 × 256 pixels. The NWPU-RESISC45 dataset is constructed from Google Earth imagery and exhibits high scene diversity and rich image variations. It includes 45 scene categories with a total of 31,500 images, where each category contains 700 images at a fixed resolution of 256 × 256 pixels. Representative examples from the datasets are presented in Figure 2.

For all three datasets, we split the images into 80% for training and 20% for testing. For image preprocessing, images from the WHU-RS19 dataset are uniformly resized to 256 × 256 pixels. Subsequently, images from all datasets are cropped to 224 × 224 pixels before being fed into the model. Following the standard FSCIL settings, the experiments are carried out with a base session, followed by a series of incremental sessions, with detailed configurations summarised in Table 1. Specifically, for WHU-RS19, we utilise 9 classes for the base session, followed by 5 incremental sessions that each introduce 2 new classes. For UC Merced-LandUse, we designate 11 classes as base categories and 10 as new classes, similarly divided into 5 incremental sessions of 2 classes each. For NWPU-RESISC45, the first 20 classes are used for base training, while the remaining 25 classes are split into five incremental sessions, each introducing five new classes. During the incremental sessions across all datasets, the training set is constructed by randomly sampling 5 examples per class from the original training samples, while the testing set comprises the complete test samples of all seen classes.



Figure 2. Representative samples from the three benchmark datasets: (**a**) WHU-RS19, (**b**) UC Merced-LandUse, (**c**) NWPU-RESISC45.

Table 1. Configuration settings for FSCIL benchmarks on WHU-RS19, UC Merced-LandUse, and NWPU-RESISC45, where "Base + Inc." denotes the numbers of the base and incremental sessions.

Session	WHU-RS19	UC Merced-LandUse	NWPU-RESISC45
Base classes	9	11	20
Incremental	2-way 5-shot	2-way 5-shot	5-way 5-shot
Base + Inc.	1 + 5	1 + 5	1 + 5

4.2. Implementation Details

We employ a ViT-B/16 [50] pre-trained on the ImageNet-21K [51] dataset as the backbone feature extractor, with a feature dimension of 768. For semantic word embeddings, we utilise the BERT-base [52] model. In our proposed framework, the ABMoE module is incorporated into every layer of the ViT encoder. The initial number of experts is set to 2, with one new expert added every 2 sessions. Upon completion of training, we discard the original classification head and employ a prototype classifier. Input images are uniformly resized to 224×224. The network is optimised using the Adam optimiser [53]. Specifically, the initial learning rate is set to 1×10^{-2} for the base session and 1×10^{-4} for the incremental sessions. The batch size is set to 128. We train the model for 10 epochs during the base session and 3 epochs for each incremental session. The hyperparameters λ_{sem} and λ_{SGFA} are empirically set to 0.5 and 0.75, respectively. Experiments are carried out on a single NVIDIA RTX 4090 GPU under PyTorch 1.12.1 and CUDA 11.3.

4.3. Comparison Against Existing FSCIL Approaches

We compare our framework with various state-of-the-art FSCIL methods to evaluate its effectiveness. These approaches encompass representative works in the field, including CNN-based methods (i.e., ALICE [54], BiDist [55], FACT [56]) and ViT-based methods (e.g., PriViLege [16], Approximation [17]), particularly prompt-based methods (e.g., ASP [15], SEC-Prompt [18]). The detailed comparative results across the three benchmark datasets are reported in Tables 2, 3, and 4, respectively.

Table 2. Detailed Session-wise comparison of FSCIL methods on the WHU-RS19 dataset. The downward arrow indicates that lower values correspond to better performance.

Method	Accuracy in Each Session (%)						Avg. Acc	PD ↓
	0	1	2	3	4	5		
FACT [56]	94.95	77.08	60.50	56.63	57.52	48.11	65.80	46.84
CLOM [57]	89.90	81.51	72.14	61.73	55.91	52.43	68.94	37.47
ALICE [54]	88.89	83.19	77.14	67.28	61.83	60.68	73.17	28.21
TEEN [25]	92.93	67.32	59.75	56.16	56.93	54.56	64.61	38.37
BiDist [55]	92.93	84.03	70.00	61.11	58.60	53.88	70.09	39.05
Warp [23]	90.91	84.87	73.57	65.43	60.22	58.25	72.21	32.66
SAVC [24]	95.96	81.21	73.00	68.69	67.92	49.56	72.72	46.40
Comp [58]	96.97	69.75	60.00	56.79	57.53	54.37	65.90	42.60
PriViLege [16]	93.94	92.44	85.00	78.40	76.88	75.73	83.73	18.21
ASP [15]	99.99	97.48	92.86	87.04	84.95	83.98	91.05	16.01
Approximation [17]	97.98	93.24	89.25	83.48	81.72	83.11	88.13	16.26
SEC-Prompt [18]	99.99	83.19	82.86	82.10	77.96	77.18	83.88	22.82
Ours	99.99	99.16	96.43	94.44	93.01	91.75	95.80	8.25

Table 3. Detailed Session-wise comparison of FSCIL methods on the UC Merced-LandUse dataset. The downward arrow indicates that lower values correspond to better performance.

Method	Accuracy in Each Session (%)						Avg. Acc	PD ↓
	0	1	2	3	4	5		
FACT [56]	96.67	81.67	75.00	65.50	62.38	50.20	71.90	46.47
CLOM [57]	93.64	81.54	72.67	65.29	58.68	52.86	70.78	40.78
ALICE [54]	93.64	83.85	76.67	72.65	69.74	66.67	77.20	26.97
TEEN [25]	94.67	73.56	68.33	55.75	51.38	43.20	64.48	51.47
BiDist [55]	95.45	85.35	78.00	72.94	67.11	62.14	76.83	33.31
Warp [23]	88.64	78.85	69.67	64.41	58.68	52.14	68.73	36.49
SAVC [24]	98.67	84.22	80.00	73.38	70.75	63.80	78.47	34.87
Comp [58]	99.09	79.62	73.33	60.59	60.00	56.67	71.55	42.42
PriViLege [16]	95.91	90.39	80.00	78.82	77.37	76.19	83.11	19.72
ASP [15]	96.82	91.15	83.33	82.06	78.95	75.00	84.55	21.82
Approximation [17]	98.00	89.11	84.67	77.00	77.88	78.00	84.00	21.00
SEC-Prompt [18]	99.55	83.85	80.67	78.24	77.89	75.00	82.53	24.55
Ours	97.73	95.39	89.68	90.29	89.21	88.33	91.77	9.39

Across all three datasets, DEM-ViT demonstrates superior overall performance compared to other approaches. Specifically, on the WHU-RS19 dataset (Table 2), DEM-ViT achieves an average classification accuracy of 95.80%, outperforming Approximation (91.05%) by a significant margin. Furthermore, DEM-ViT limits the performance drop (PD) rate to 8.25%, which is considerably lower than that of other methods. This proves the effectiveness of our approach in mitigating catastrophic forgetting. Similarly, on the UC Merced-LandUse dataset, DEM-ViT yields an average accuracy of 91.77%, surpassing ASP by over 7%. Notably, even in the final incremental session, DEM-ViT maintains an accuracy of 88.33%, indicating that our method possesses a stronger ability to acquire knowledge of new classes compared to other approaches. On the NWPU-RESISC45 dataset, DEM-ViT exhibits robust performance, achieving an average accuracy of 86.78%.

Table 4. Detailed Session-wise comparison of FSCIL methods on the NWPU-RESISC45 dataset. The downward arrow indicates that lower values correspond to better performance.

Method	Accuracy in Each Session (%)						Avg. Acc	PD ↓
	0	1	2	3	4	5		
FACT [56]	96.46	84.37	74.48	66.51	62.38	56.06	73.38	40.40
CLOM [57]	94.93	87.57	77.79	69.88	66.98	62.03	76.53	32.90
ALICE [54]	96.11	87.37	79.07	73.31	70.52	66.08	78.74	30.03
TEEN [25]	91.61	80.03	70.26	62.67	58.70	53.56	69.47	38.05
BiDist [55]	97.00	88.23	77.95	70.73	67.04	61.14	77.02	35.86
Warp [23]	91.21	83.00	72.60	63.82	60.30	54.97	70.98	36.25
SAVC [24]	98.61	90.91	82.98	77.49	74.88	69.33	82.37	29.27
Comp [58]	95.75	82.97	70.93	59.69	56.82	51.25	69.57	44.50
PriViLege [16]	96.29	89.17	83.64	80.63	79.23	76.52	84.25	19.76
ASP [15]	95.82	88.46	81.29	78.37	76.84	74.60	82.56	21.22
Approximation [17]	98.86	87.66	79.71	73.71	70.27	65.57	79.30	33.29
SEC-Prompt [18]	97.54	78.14	75.12	72.18	69.21	68.59	76.80	28.95
Ours	98.07	93.31	87.76	82.06	81.14	78.30	86.78	19.77

Overall, our method achieved the highest average accuracy across all three datasets, demonstrating significant improvements over other models. These consistent results validate that our DEM-ViT framework effectively mitigates the domain gap between natural and remote sensing images, enabling the pre-trained model to better adapt to complex remote sensing scene classification tasks.

Figure 3 illustrates the few-shot class-incremental learning performance of the proposed framework across three benchmark datasets. Overall, compared to other state-of-the-art methods, DEM-ViT demonstrates more stable performance throughout the incremental learning process. While delivering competitive results in the base session, DEM-ViT outperforms other approaches across subsequent incremental sessions, characterised by a more gradual decline in accuracy. Excessive emphasis on fitting base categories during the base session tends to lead to model rigidity, thereby potentially impairing its plasticity for acquiring new knowledge. In contrast, DEM-ViT effectively mitigates catastrophic forgetting during the incremental sessions, resulting in a more favourable trade-off between model stability and plasticity.

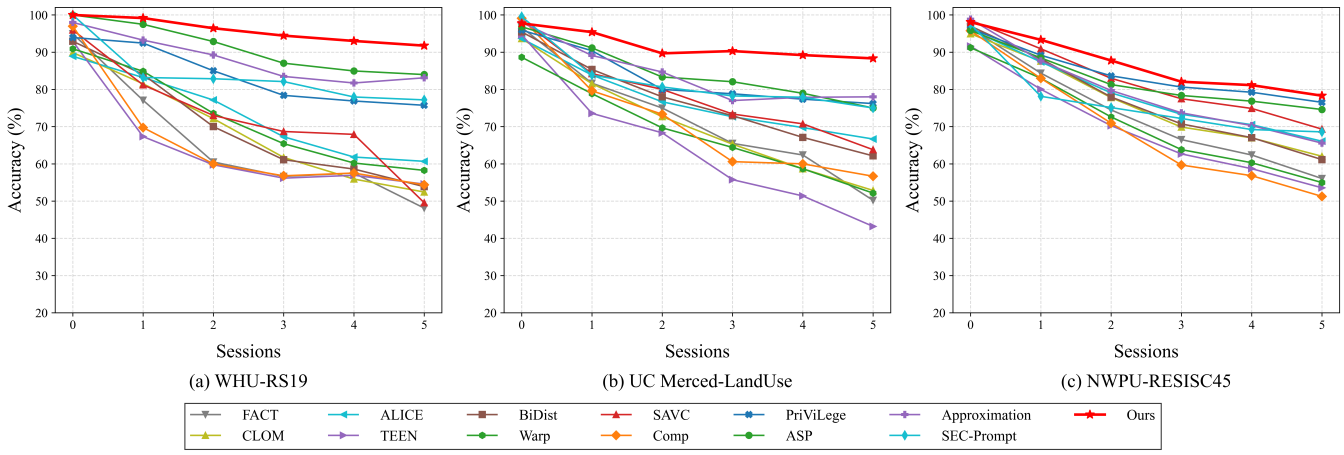


Figure 3. Classification accuracy of each few-shot incremental session on the three benchmarks.

4.4. Ablation Study

We conducted ablation studies on the UC Merced-LandUse dataset to investigate the independent contributions of each component within the DEM-ViT framework. The

results highlight: (1) the effectiveness of the Adapter-Based Mixture-of-Experts module, the Dynamic Expert Expansion strategy, and the Semantic-Guided Feature Alignment method; (2) the synergistic contribution of these components in tackling overfitting and catastrophic forgetting in FSCIL.

Our proposed DEM-ViT framework comprises the Adapter-Based Mixture-of-Experts module, the Dynamic Expert Expansion strategy, and the Semantic-Guided Feature Alignment method. The ABMoE module is designed to efficiently adapt to downstream tasks by freezing the ViT backbone and inserting multiple lightweight Adapters in parallel. Specifically, this module utilises a Router to dynamically assign weights, enabling distinct Adapter experts to specialise in capturing domain-specific semantic knowledge. The Dynamic Expert Expansion mechanism aims to address the stability–plasticity dilemma by incrementally increasing the number of experts to allocate storage space for new knowledge. However, in the extremely data-scarce FSCIL setting, simply increasing the number of experts may cause them to overfit the limited samples. To mitigate this problem, we introduce the Semantic-Guided Feature Alignment method. By leveraging text prototypes generated by the pre-trained language model as prior knowledge, this method guides the model to learn highly generalizable features through cross-modal alignment, thereby serving as a potent regularisation strategy to alleviate overfitting.

To more comprehensively evaluate the trade-off between the model's plasticity and stability, we introduce the average Harmonic Mean (aHM) metric. In FSCIL, the average accuracy is often dominated by the well-trained base classes, thereby obscuring the model's performance on new classes. The Harmonic Mean effectively addresses this issue by calculating the harmonic mean of the base class accuracy and the new class accuracy during the incremental sessions, thereby providing a more reasonable assessment of the trade-off between plasticity and stability. The Harmonic Mean of the i-th session can be expressed as:

$$HM_i = \frac{2 \times A_i^{base} \times A_i^{new}}{A_i^{base} + A_i^{new}} \tag{18}$$

The aHM is then obtained by averaging HM_i across all incremental sessions.

As shown in Table 5, we conducted comprehensive ablation studies on the three proposed components (ABMoE, DEE, and SGFA), where $\Delta\mathcal{A}_{avg.}$ denotes the change in average accuracy relative to the baseline configuration with none of these components enabled. Specifically, incorporating the ABMoE module significantly improves average classification accuracy by 6.33% over the baseline, validating the superiority of the ABMoE architecture for feature extraction.

Table 5. Ablation study of the three proposed components. The downward arrow indicates that lower values correspond to better performance, and – indicates no change.

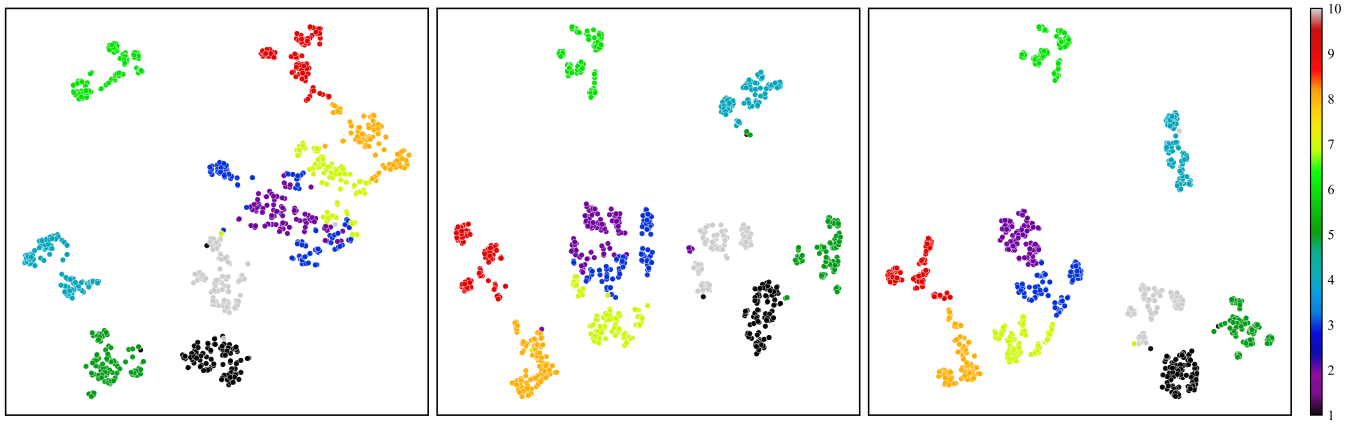
ABMoE	DEE	SGFA	PD ↓	aHM	$\mathcal{A}_{avg.}$	$\Delta\mathcal{A}_{avg.}$
			19.61	78.67	84.44	–
✓			13.14	87.20	90.77	+6.33
	✓		20.56	78.41	84.28	−0.16
		✓	13.01	84.53	87.84	+3.40
✓	✓		13.14	87.18	90.65	+6.21
✓		✓	10.11	89.00	91.31	+6.87
	✓	✓	11.36	85.67	88.78	+4.34
✓	✓	✓	9.39	89.60	91.77	+7.33

Notably, to alleviate the inherent stability–plasticity dilemma in incremental learning, we introduce the DEE strategy, which dynamically expands the number of experts as

incremental sessions progress to accommodate new knowledge while preventing the catastrophic forgetting of previous knowledge. However, under the extreme data-scarce setting of FSCIL, merely increasing the number of experts may cause them to overfit to the limited samples, thereby leading to a slight degradation in accuracy. To this end, we introduce the SGFA method to overcome this bottleneck. Incorporating semantic priors generated by the pre-trained language model provides explicit semantic guidance and cross-modal calibration, thereby effectively preventing the model from overfitting. Therefore, DEE endows the model with the necessary capacity to absorb new knowledge, while SGFA provides it with the much-needed semantic knowledge. The synergy between the two enables the model to fully enhance its continuous learning capability while mitigating the overfitting problem. The combination of DEE and SGFA yields a significantly higher average accuracy of 88.78%, effectively overcoming the slight performance drop caused by using DEE in isolation. This indicates that the semantic priors provided by SGFA successfully calibrate the visual features. Consequently, our DEE strategy is able to fully leverage its plasticity for acquiring new knowledge while ensuring stability.

The complete framework incorporating all three components achieves the best performance, with an average accuracy of 91.77%, and a 7.33% improvement over the baseline. This demonstrates that our method successfully mitigates the severe domain gap between natural and remote sensing images. The average Harmonic Mean (aHM) reaches a peak of 89.60%, representing a 10.93% improvement over the baseline. Consequently, alongside improving the overall performance, our model achieves a better trade-off between the plasticity of acquiring new knowledge and the stability of retaining previous knowledge.

To provide an intuitive understanding of the contributions of each proposed component, we present the t-SNE visualisation results of feature embeddings on the UC Merced-LandUse dataset. Classes 1–6 represent base classes, while classes 7–10 denote new classes. Figure 4a illustrates the visualisation results of the Baseline. The feature distribution of the baseline model exhibits a relatively loose structure, with a substantial number of outliers scattered away from their respective cluster centres. This indicates that in the few-shot incremental setting, relying solely on visual information makes it difficult for the model to establish clear decision boundaries between visually similar classes. A striking example is observed between Class 2 and Class 3, where the feature clusters exhibit severe overlap and confusion, implying that the baseline model lacks sufficient discriminative power to distinguish between these ambiguous classes. Figure 4b shows the effects of introducing the SGFA method. Evidently, the overall structure of the feature space becomes significantly more organised. The most notable improvement lies in the separation of the Class 2 and Class 3, where scattered sample points are effectively realigned towards their respective cluster centres. This is attributed to the mechanism of SGFA, where the text embeddings generated by the PLM serve as stable semantic prototypes. These prototypes provide robust high-level semantic guidance, constraining visual features to align with consistent semantic definitions rather than drifting with visual noise. Figure 4c presents the feature embedding visualisation of our complete model. The clusters achieve minimal intra-class variance and maximal inter-class distance. In particular, the separation between Class 2 and Class 3 becomes distinct and unambiguous, characterised by a significant margin. Furthermore, some outliers have been realigned into their corresponding class clusters. This enhancement stems from the ABMoE architecture, which enables distinct experts to specialise in capturing diverse feature patterns. The combination of semantic guidance from SGFA and the structural adaptability of the DEM-ViT framework ultimately yields a highly discriminative feature space.



(a) Baseline (b) w/ SGFA (c) w/ ABMoE and DEE

Figure 4. The t-SNE visualisation on the UC Merced-LandUse dataset. (**a**) t-SNE results of the baseline. (**b**) t-SNE results with the SGFA method. (**c**) t-SNE results with the ABMoE module and DEE strategy.

We conducted further experiments to determine the optimal insertion layers for the ABMoE module within the ViT backbone. As shown in Table 6, inserting the ABMoE module across all layers yields optimal performance, achieving the highest average accuracy and average Harmonic Mean, while maintaining the lowest PD rate. This indicates that full-level feature adaptation effectively captures domain-specific knowledge relevant to remote sensing. Notably, fine-tuning only the deep layers resulted in the poorest performance. This suggests that adjusting only high-level semantics is insufficient to bridge the domain gap between remote sensing images and natural images. Furthermore, although fine-tuning shallow or middle layers outperformed deep-layer fine-tuning, the results remained inferior to the full-layer configuration. By inserting ABMoE into all layers, the model can refine both shallow features and deep semantics at the same time, ensuring it achieves the highest accuracy as new tasks are added.

Table 6. Ablation study on the insertion position of the ABMoE module. The layers are indexed from 0 to 11. The downward arrow indicates that lower values correspond to better performance, and – indicates no change.

Stage	Layers	PD ↓	aHM	$\mathcal{A}_{avg.}$	$\Delta\mathcal{A}_{avg.}$
Shallow	0–2	12.08	84.14	87.48	−4.29
Shallow-Mid	0–5	10.84	86.75	89.84	−1.93
Mid-Deep	6–11	12.51	86.51	89.70	−2.07
Deep	9–11	15.93	82.21	84.35	−7.42
Full	0–11	9.39	89.60	91.77	–

We also investigate the impact of hyperparameters λ_{sem} and λ_{SGFA}. As indicated in Equations (15) and (17), the hyperparameter λ_{sem} determines the contribution of semantic logits to the final prediction, while λ_{SGFA} governs the intensity of the semantic alignment loss during training. We evaluated the performance of DEM-ViT under varying values for these two hyperparameters. As shown in Figure 5a, with the increase in λ_{sem}, the accuracy exhibits a trend of initially increasing and then decreasing. This suggests that appropriate semantic guidance facilitates visual discrimination, whereas excessive semantic weight may introduce unnecessary noise. DEM-ViT achieves optimal results when $\lambda_{sem} = 0.5$. Similarly, for the regularisation weight, as illustrated in Figure 5b, the performance follows a similar pattern of increasing followed by a decline. The method attains peak performance at $\lambda_{SGFA} = 0.75$. This indicates that utilising semantic alignment as a regularisation

strategy effectively supplements knowledge for new class samples and mitigates overfitting. Meanwhile, the regularisation weight constrains the alignment intensity, preventing it from hindering feature learning.

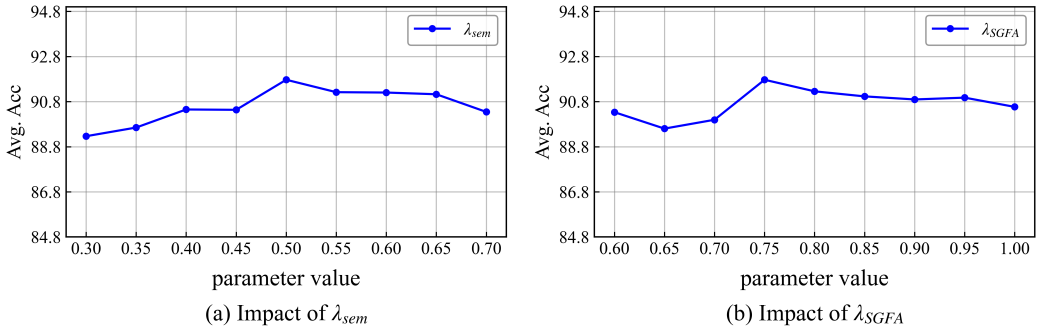


Figure 5. Impact of the hyperparameters λ_{sem} and λ_{SGFA} on DEM-ViT performance.

To investigate the impact of the initial number of experts within the ABMoE module on model performance, we conducted ablation studies on the initial number of experts M in the base session. We varied $M \in \{1, 2, 3, 4\}$ and reported the average accuracy and performance drop rate across all incremental sessions. As shown in Figure 6, when $M = 1$, the model degenerates into a standard single-Adapter structure. In this case, both the average accuracy and the aHM are relatively low, accompanied by a high performance drop (PD) rate. As discussed in Section 3.3, a single Adapter struggles to simultaneously capture the complex textures and diverse object scales inherent in remote sensing scenes. This results in insufficient feature decoupling capability, thereby limiting the upper bound of the model's generalisation. Model performance peaks at an expert count of two. This indicates that introducing the MoE architecture indeed enhances the robustness of feature representation. Multiple experts can collaborate and specialise, facilitating the capture of diverse semantic patterns and thus constructing a more comprehensive feature space. However, as the number of experts increases to 3 and 4, performance exhibits a downward trend, with a significant increase in PD. This is because excessive experts introduce redundant parameters, which may induce a tendency to overfit the base classes, thereby impairing the model's generalisation performance. Consequently, we adopt a starting configuration of 2 experts for the base session.

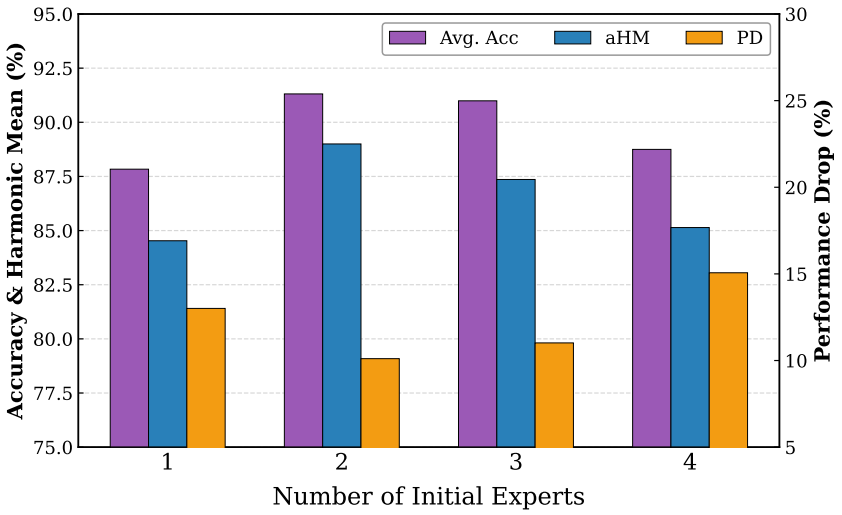


Figure 6. Performance of the DEM-ViT framework with varying numbers of initial experts.

We also show the impact of the expert expansion interval within the DEE strategy on model performance by conducting ablation studies on the expansion interval I during the incremental sessions. We varied the interval $I \in \{2, 3, 5\}$, where one new expert is added every I sessions, and compared the results with a configuration without expansion ("w/o New Experts"). For visual clarity, only the results from Sessions 2–5 are shown, as the accuracy of Session 1 differs substantially from that of the other sessions. The dotted line indicates the connection between the results of Sessions 1 and 2. As shown in Figure 7, when no new experts are added ("w/o New Experts"), the model exhibits the lowest accuracy across all incremental sessions. Without structural expansion, the fixed number of experts struggles to handle continuous new concepts, leading to a model performance drop as new knowledge interferes with previously learned representations. Model performance peaks when the interval is set to 2. This indicates that a timely expansion of the expert library effectively supports the learning of new remote sensing classes. Adding a new expert every two sessions provides just enough model capacity to capture novel semantic patterns while preserving prior knowledge. Consequently, we adopt an expansion interval of $I = 2$ as the default configuration for the incremental sessions.

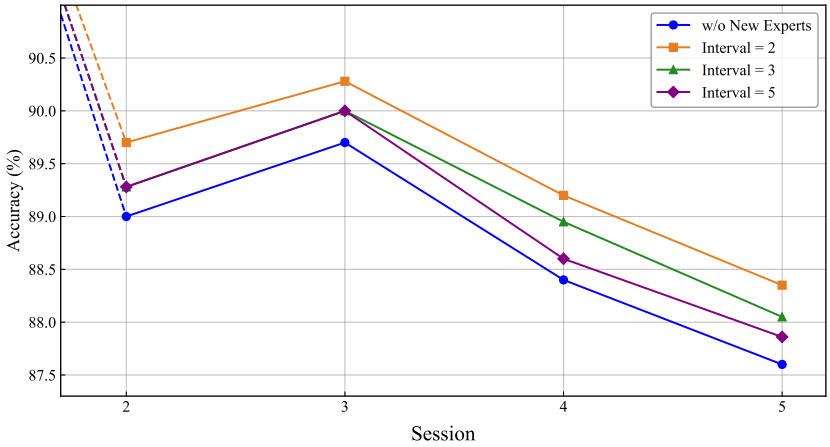


Figure 7. Performance of the DEM-ViT framework with varying expert expansion intervals.

To verify the advantage of initializing new experts with the mean parameters of existing ones, we compared this method with random initialization. As shown in Figure 8, the "Mean" initialization strategy outperforms "Random" initialization across all metrics, achieving higher average accuracy and aHM, alongside a lower PD. This is primarily because, in few-shot learning scenarios where training data for new classes is extremely scarce, a randomly initialized new expert lacks prior knowledge and is difficult to train effectively, which easily leads to performance fluctuations during the incremental sessions. In contrast, initializing with the mean parameters of existing experts effectively transfers the knowledge already accumulated by the model, providing a stable starting point for the new expert. This approach enables the DEM-ViT framework to learn new classes while effectively minimizing interference with the previously established knowledge structure.

To explicitly demonstrate the effectiveness of our SGFA method in mitigating overfitting, and to evaluate how the quality of text prototypes affects the classification results, we conducted an ablation study on the text prototypes within the SGFA method. As shown in Table 7, we compared different text templates. The results demonstrate that compared to a generic template ("a photo of a ...") or a template that might introduce incorrect semantic biases ("a satellite photo of a ...", as the dataset primarily contains aerial imagery rather than typical high-altitude satellite photos), providing accurate contextual information ("a remote sensing image of a ...") helps BERT generate higher-quality text prototypes, which

significantly boosts the average accuracy of new classes ($A_{avg.new}$) to 87.38%. The significant performance improvement for new classes clearly demonstrates that high-quality semantic priors can effectively alleviate the overfitting issue. Additionally, to directly simulate suboptimal text prototype quality, we injected random Gaussian noise with intensities of 0.01 and 0.05 into the standard text features. As the noise increases, the model's performance declines significantly; under noise intensities of 0.01 and 0.05, $A_{avg.new}$ drops to 85.43% and 82.91%, respectively. This significant performance degradation strongly proves that our method genuinely utilizes the high-level semantic information provided by the pre-trained language model. These semantic priors provide crucial guidance for recognizing rare new classes under few-shot conditions, thereby substantially improving their classification accuracy.

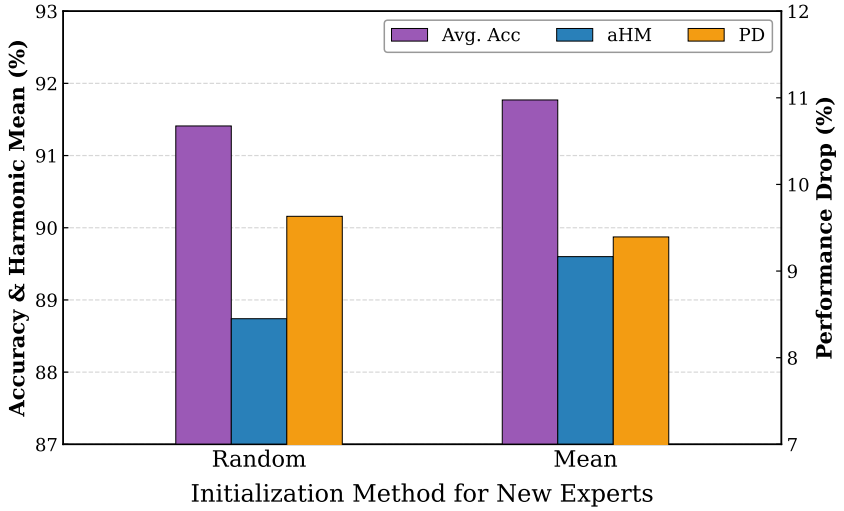


Figure 8. Performance comparison of different initialization strategies for new experts.

Table 7. Impact of text prototype quality on model performance within the SGFA method. The downward arrow indicates that lower values correspond to better performance.

Template	PD ↓	aHM	$\mathcal{A}_{avg.}$	$\mathcal{A}_{avg.new}$
"a photo of a [classname]"	12.23	85.43	89.42	80.82
"a satellite photo of a [classname]"	11.32	85.55	89.11	81.46
"an aerial view of a [classname]"	12.42	86.17	90.68	83.00
"a remote sensing image of a [classname]" (with 0.05 formatting noise)	11.06	86.94	90.26	82.91
"a remote sensing image of a [classname]" (with 0.01 noise)	8.94	88.35	90.96	85.43
"a remote sensing image of a [classname]" (without noise)	9.39	89.60	91.77	87.38

We compared the complexity of our model at the final incremental session with recent state-of-the-art ViT-based methods, PriViLege [16] and ASP [15]. Table 8 details the Total Parameters, Trainable Parameters, Inference Time, and Storage footprint for each configuration. Although the dynamic expert expansion strategy introduces new structural units, the total parameter count of DEM-ViT is 91.81 M, which is substantially lower than ASP's 173.61 M and comparable to PriViLege's 85.84 M. The overall storage requirement is also kept manageable at 350.23 MB. In terms of inference latency, DEM-ViT takes 1.63 ms, slightly slower than PriViLege (1.23 ms) but faster than ASP (2.44 ms). These results confirm that DEM-ViT achieves superior recognition performance without introducing impractical memory or latency overheads, thereby ensuring its reliable scalability in applications.

Table 8. Complexity analysis of the DEM-ViT framework compared to recent ViT-based methods.

Configuration	Total Params (M)	Trainable Params (M)	Time (ms)	Storage (MB)
PriViLege [16]	85.84	14.20	1.23	327.45
ASP [15]	173.61	2.02	2.44	662.28
DEM-ViT	91.81	6.01	1.63	350.23

5. Discussion

Our proposed DEM-ViT framework achieves promising performance in the FSCIL-RSSC task. Here, we discuss the underlying mechanism of its success, its potential limitations, and directions for future research.

The main goal of this study is to leverage a pre-trained large model (ViT) to address remote sensing scene classification under an incremental learning setting with extremely limited data. We face two main challenges. On the one hand, there is a large domain gap between remote sensing images and the natural images used to pre-train the ViT. On the other hand, the model must effectively learn new knowledge during incremental sessions with very few samples, while preserving previously learned knowledge, in order to alleviate the stability–plasticity dilemma.

To address the challenges mentioned above, we propose targeted solutions. First, to mitigate the significant domain gap, ABMoE aggregates lightweight adapters through a soft routing mechanism, enabling implicit decoupling and reconstruction of complex remote sensing features. However, applying experts with fixed capacity to continuous incremental tasks inevitably causes previously learned remote sensing patterns to be overwritten by new knowledge, leading to feature distribution collapse. To address this issue, we adopt the DEE strategy, which dynamically increases model capacity to accommodate new knowledge and thus alleviates the stability–plasticity dilemma. A key challenge then arises: expanding the number of experts under extremely limited data makes the new experts highly prone to overfitting. Moreover, we believe that purely visual feature reconstruction is inherently unstable in few-shot settings. Therefore, we introduce the cross-modal SGFA method, which leverages reliable semantic prior knowledge from a pre-trained language model to calibrate the visual feature space across modalities, effectively reducing the severe overfitting of new experts. As a result, these three components are not simply combined engineering modules, but a collaborative framework driven by clear scientific motivations.

While our method achieves promising results, we acknowledge the potential limitations. A crucial factor in our cross-modal alignment is the quality of semantic priors. Since remote sensing terms can sometimes be ambiguous, directly applying a general pre-trained language model without constraints may, in theory, introduce semantic noise. Although our domain-specific templates successfully guide the model to extract robust knowledge, to completely eliminate semantic ambiguity in highly specialized scenarios, integrating a visual-language model pre-trained on large-scale remote sensing corpora remains the best approach for enhancing model performance in the future.

Additionally, both our structural expansion approach via DEE and the routing strategy within ABMoE are specifically customized for the current state of FSCIL-RSSC tasks. At present, the most comprehensive remote sensing datasets (such as the NWPU-RESISC45 [49] with up to 45 classes) involve relatively short incremental sequences. Under these conditions, our method maintains high efficiency while controlling parameter overhead. However, if we transition to ultra-long, open-world scenarios with hundreds of continuous incremental tasks, the complexity of the structure will naturally increase. In such extreme FSCIL settings, researching advanced MoE dynamic mechanisms, such as routing

optimization to ensure expert balance and post-learning expert management strategies like pruning or merging, will be essential for maintaining a highly compact model architecture.

Regarding future research directions, we plan to address the challenges mentioned above from two perspectives. Given the reliance on text embeddings, we intend to prioritise integrating Vision-Language Models, specifically those pre-trained on remote sensing corpora, to provide more precise cross-modal supervision. Additionally, improving the efficiency and scalability of expert expansion remains an important direction for future work. We will deeply investigate valuable issues such as the optimization of the soft routing mechanism, as well as strategies for pruning, consolidating, or managing experts. By implementing these expert management mechanisms after continuous learning, we can maintain a highly compact model size without compromising the structural plasticity required for open-world scenarios.

6. Conclusions

In this paper, we propose a novel Dynamic Expansion Mixture-of-Experts with Pre-trained Vision Transformer framework, which aims to address the domain distribution discrepancy between natural images and remote sensing images in FSCIL-RSSC tasks. Our framework primarily comprises the ABMoE module, the DEE strategy, and the SGFA method. The ABMoE module adapts the model to remote sensing scenes by reconstructing features and leveraging collaborative experts, effectively bridging the domain gap. The DEE strategy progressively increases model capacity along incremental sessions, providing sufficient representational ability for learning new classes while alleviating catastrophic forgetting. Meanwhile, the SGFA scheme uses text information from the pre-trained language model as text prototypes. These prototypes regularise the visual feature space, preventing the model from overfitting. On the WHU-RS19, UC Merced-LandUse, and NWPU-RESISC45 datasets, the proposed DEM-ViT achieves average accuracies of 95.80%, 91.77%, and 86.78% under the 2-way 5-shot, 2-way 5-shot, and 5-way 5-shot settings, respectively. We also perform extensive ablation studies, verifying the effectiveness of the proposed framework. In the future, we plan to incorporate vision–language models pre-trained on remote sensing corpora to further improve performance on FSCIL-RSSC tasks and to explore strategies to maintain a controllable model size over incremental sessions.

Author Contributions: Conceptualisation, Y.W.; methodology, Y.W.; software, Y.W.; validation, Y.W.; formal analysis, Y.W.; investigation, Y.W.; resources, Y.W.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W., X.L., J.Z., H.Z., H.L. and T.T.; visualisation, Y.W.; supervision, Y.W.; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Frontier Research Fund of Institute of Optics and Electronics, China Academy of Sciences (Grant No. C24K003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to express our gratitude to the editor and reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, Y.; Wang, Y.; Zhang, F.; Dong, Y.; Song, Z.; Liu, G. Remote sensing for lithology mapping in vegetation-covered regions: Methods, challenges, and opportunities. *Minerals* **2023**, *13*, 1153. [CrossRef]
2. Manandhar, R.; Odeh, I.O.; Ancev, T. Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement. *Remote Sens.* **2009**, *1*, 330–344. [CrossRef]
3. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A review of remote sensing for environmental monitoring in China. *Remote Sens.* **2020**, *12*, 1130. [CrossRef]
4. Nogueira, K.; Penatti, O.A.; Dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **2017**, *61*, 539–556. [CrossRef]
5. Shu, Y.; Shi, Y.; Wang, Y.; Huang, T.; Tian, Y. P-odn: Prototype-based open deep network for open set recognition. *Sci. Rep.* **2020**, *10*, 7146. [CrossRef] [PubMed]
6. Zhu, Z.; Wang, P.; Diao, W.; Yang, J.; Wang, H.; Sun, X. Few-shot incremental learning with continual prototype calibration for remote sensing image fine-grained classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 210–227. [CrossRef]
7. Hu, Y.; Yang, G.; Tan, Z.; Huang, X.; Huang, K.; Wang, Q.F. Covariance-based Space Regularization for Few-shot Class Incremental Learning. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2025; pp. 9566–9576.
8. Kang, H.; Yoon, J.; Madjid, S.R.H.; Hwang, S.J.; Yoo, C.D. On the soft-subnetwork for few-shot class incremental learning. *arXiv* **2022**, arXiv:2209.07529.
9. Kukleva, A.; Kuehne, H.; Schiele, B. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual Event, 11–17 October 2021; pp. 9020–9029.
10. Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; Tao, D. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv* **2023**, arXiv:2302.03004.
11. Li, L.; Tan, Y.; Yang, S.; Cheng, H.; Dong, Y.; Yang, L. Adaptive Decision Boundary for Few-Shot Class-Incremental Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 28 February–5 March 2025; Volume 39, pp. 18359–18367.
12. Zeng, Y.; Xue, B.; Han, W. Few-shot incremental learning for aerial image scene classification based on feature adaptation and prototype continuous optimization. In *Proceedings of the 2022 IEEE International Conference on Unmanned Systems (ICUS)*; IEEE: New York, NY, USA, 2022; pp. 510–515.
13. Li, P.; Feng, C.; Hu, X.; Feng, W. HDCPAA: A few-shot class-incremental learning model for remote sensing image recognition. *Neurocomputing* **2025**, *636*, 130043. [CrossRef]
14. Rosy, N.A.; Balasubadra, K.; Deepa, K. Are vision transformers replacing convolutional neural networks in scene interpretation?: A review. *Discov. Appl. Sci.* **2025**, *7*, 932. [CrossRef]
15. Liu, C.; Wang, Z.; Xiong, T.; Chen, R.; Wu, Y.; Guo, J.; Huang, H. Few-shot class incremental learning with attention-aware self-adaptive prompt. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–18.
16. Park, K.H.; Song, K.; Park, G.M. Pre-trained vision and language transformers are few-shot incremental learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 23881–23890.
17. Wang, X.; Ji, Z.; Liu, X.; Pang, Y.; Han, J. On the approximation risk of few-shot class-incremental learning. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 162–178.
18. Liu, Y.; Yang, M. SEC-Prompt: SEmantic Complementary Prompting for Few-Shot Class-Incremental Learning. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 25643–25656.
19. Chen, Y.; Ding, T.; Wang, L.; Huo, J.; Gao, Y.; Li, W. Enhancing Few-Shot Class-Incremental Learning via Training-Free Bi-Level Modality Calibration. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 9881–9890.
20. Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. Adaptformer: Adapting vision transformers for scalable visual recognition. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16664–16678.
21. Yang, B.; Lin, M.; Zhang, Y.; Liu, B.; Liang, X.; Ji, R.; Ye, Q. Dynamic support network for few-shot class incremental learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2945–2951. [CrossRef] [PubMed]
22. Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; Gong, Y. Few-shot class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12183–12192.
23. Kim, D.Y.; Han, D.J.; Seo, J.; Moon, J. Warping the space: Weight space rotation for class-incremental few-shot learning. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.

24. Song, Z.; Zhao, Y.; Shi, Y.; Peng, P.; Yuan, L.; Tian, Y. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 24183–24192.
25. Wang, Q.W.; Zhou, D.W.; Zhang, Y.K.; Zhan, D.C.; Ye, H.J. Few-shot class-incremental learning via training-free prototype calibration. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 15060–15076.
26. Zhao, Y.; Zhao, L.; Ding, D.; Hu, D.; Kuang, G.; Liu, L. Few-shot class-incremental SAR target recognition via cosine prototype learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5212718. [CrossRef]
27. Guan, D.; Feng, R.; Xie, Y.; Zheng, X.; Li, B.; Xiang, D. Few-Shot Class-Incremental SAR Target Recognition with a Forward-Compatible Prototype Classifier. *Remote Sens.* **2025**, *17*, 3518. [CrossRef]
28. Wang, W.; Zhang, L.; Fu, S.; Ren, P.; Ren, G.; Peng, Q.; Liu, B. Gradient guided multiscale feature collaboration networks for few-shot class-incremental remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611912. [CrossRef]
29. Zeng, J.; Tan, Y.; Yang, L.; Zhang, S.; Liang, L. TerraSAP: Spatially Aware Prompt-Based Framework for Few-Shot Class-Incremental Learning in Remote Sensing Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *19*, 3143–3156. [CrossRef]
30. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2019; pp. 2790–2799.
31. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691. [CrossRef]
32. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual Event, 25–29 April 2022; Volume 1, p. 3.
33. Jie, S.; Deng, Z.H. Fact: Factor-tuning for lightweight adaptation on vision transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1060–1068.
34. Jie, S.; Deng, Z.H. Convolutional bypasses are better vision transformer adapters. *arXiv* **2022**, arXiv:2207.07039. [CrossRef]
35. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 709–727.
36. Tu, C.H.; Mai, Z.; Chao, W.L. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7725–7735.
37. Lian, D.; Zhou, D.; Feng, J.; Wang, X. Scaling & shifting your features: A new baseline for efficient model tuning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 109–123.
38. Tan, Y.; Zhou, Q.; Xiang, X.; Wang, K.; Wu, Y.; Li, Y. Semantically-shifted incremental adapter-tuning is a continual vitransformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 23252–23262.
39. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
40. He, M.; Lv, G.; He, W.; Fan, J.; Zeng, G. DeepME: Deep Mixture Experts for Large-scale Image Classification. In Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021), Montreal, QC, Canada, 19–27 August 2021; pp. 722–728.
41. Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; Houlsby, N. Scaling vision with sparse mixture of experts. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8583–8595.
42. Berisha, U.; Mehnert, J.; Condurache, A.P. Efficient Data Driven Mixture-of-Expert Extraction from Trained Networks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 20082–20091.
43. Liu, T.; Blondel, M.; Riquelme, C.; Puigcerver, J. Routers in vision mixture of experts: An empirical study. *arXiv* **2024**, arXiv:2401.15969. [CrossRef]
44. Min, C.; Wang, W.; Liu, Y.; Ye, W.; Sangineto, E.; Wang, Q.; Zhao, Y. Guiding the Experts: Semantic Priors for Efficient and Focused MoE Routing. *arXiv* **2025**, arXiv:2505.18586. [CrossRef]
45. Han, X.; Wei, L.; Dou, Z.; Sun, Y.; Han, Z.; Tian, Q. Vimoe: An empirical study of designing vision mixture-of-experts. *IEEE Trans. Image Process.* **2025**, *34*, 7209–7221. [CrossRef] [PubMed]
46. Videau, M.; Leite, A.; Schoenauer, M.; Teytaud, O. Mixture of Experts in Image Classification: What's the Sweet Spot? *arXiv* **2024**, arXiv:2411.18322.
47. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [CrossRef]

48. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
49. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
50. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
51. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
52. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MA, USA, 2–7 June 2019; Volume 1 (Long and Short Papers), pp. 4171–4186.
53. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
54. Peng, C.; Zhao, K.; Wang, T.; Li, M.; Lovell, B.C. Few-shot class-incremental learning from an open-set perspective. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 382–397.
55. Zhao, L.; Lu, J.; Xu, Y.; Cheng, Z.; Guo, D.; Niu, Y.; Fang, X. Few-shot class-incremental learning via class-aware bilateral distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11838–11847.
56. Zhou, D.W.; Wang, F.Y.; Ye, H.J.; Ma, L.; Pu, S.; Zhan, D.C. Forward compatible few-shot class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9046–9056.
57. Zou, Y.; Zhang, S.; Li, Y.; Li, R. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27267–27279.
58. Zou, Y.; Zhang, S.; Zhou, H.; Li, Y.; Li, R. Compositional few-shot class-incremental learning. *arXiv* **2024**, arXiv:2405.17022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.