remote sensing

MDPI

Article

A Memory-Efficient Class-Incremental Learning Framework for Remote Sensing Scene Classification via Feature Replay

Yunze Wei [1,2,3,4], Yuhan Liu [1,2,3,4], Ben Niu [1,2,3,*], Xiantai Xiang [1,2,3,4], Jingdun Lin [1,2], Yuxin Hu [1,2,3,4] and Yirong Wu [1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; weiyunze21@mails.ucas.ac.cn (Y.W.); liuyuhan@aircas.ac.cn (Y.L.)
[2] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China
[3] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: niuben@aircas.ac.cn

Highlights

What are the main findings?

- A novel memory-efficient feature-replay class-incremental learning framework is proposed for remote sensing scene classification by replaying compact feature embeddings instead of raw images, effectively alleviating catastrophic forgetting under strict memory and privacy constraints.
- A progressive multi-scale feature enhancement module, a specialized feature calibration network and a bias rectification strategy, are proposed to address representation ambiguity, feature space drift, and classifier bias, respectively. Together, they work synergistically to enhance the overall incremental learning performance.

What is the implication of the main finding?

- The proposed memory-efficient class-incremental learning framework substantially improves the feasibility of deployment on resource-constrained platforms, such as satellites and unmanned aerial vehicles (UAVs), where storage capacity is limited.
- Extensive validation on public benchmark datasets demonstrate the effectiveness and robustness of our method, confirming its substantial potential for practical implementation in real-world remote sensing scene classification scenarios.

Abstract

Most existing deep learning models for remote sensing scene classification (RSSC) adopt an offline learning paradigm, where all classes are jointly optimized on fixed-class datasets. In dynamic real-world scenarios with streaming data and emerging classes, such paradigms are inherently prone to catastrophic forgetting when models are incrementally trained on new data. Recently, a growing number of class-incremental learning (CIL) methods have been proposed to tackle these issues, some of which achieve promising performance by rehearsing training data from previous tasks. However, implementing such strategy in real-world scenarios is often challenging, as the requirement to store historical data frequently conflicts with strict memory constraints and data privacy protocols. To address these challenges, we propose a novel memory-efficient feature-replay CIL framework (FR-CIL) for RSSC that retains compact feature embeddings, rather than raw images, as exemplars for previously learned classes. Specifically, a progressive multi-scale feature enhancement (PMFE) module is proposed to alleviate representation ambiguity. It

Check for updates

Academic Editors: Pedro Melo-Pinto and Xiangtao Zheng

Received: 31 January 2026
Revised: 1 March 2026
Accepted: 12 March 2026
Published: 15 March 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.

adopts a progressive construction scheme to enable fine-grained and interactive feature enhancement, thereby improving the model's representation capability for remote sensing scenes. Then, a specialized feature calibration network (FCN) is trained in a transductive learning paradigm with manifold consistency regularization to adapt stored feature descriptors to the updated feature space, thereby effectively compensating for feature space drift and enabling a unified classifier. Following feature calibration, a bias rectification (BR) strategy is employed to mitigate prediction bias by exclusively optimizing the classifier on a balanced exemplar set. As a result, this memory-efficient CIL framework not only addresses data privacy concerns but also mitigates representation drift and classifier bias. Extensive experiments on public datasets demonstrate the effectiveness and robustness of the proposed method. Notably, FR-CIL outperforms the leading state-of-the-art CIL methods in mean accuracy by margins of 3.75%, 3.09%, and 2.82% on the six-task AID, seven-task RSI-CB256, and nine-task NWPU-45 datasets, respectively. At the same time, it reduces memory storage requirements by over 94.7%, highlighting its strong potential for real-world RSSC applications under strict memory constraints.

Keywords: class-incremental learning; catastrophic forgetting; remote sensing scene classification

1. Introduction

Unlike pixel-level and object-level interpretation techniques, remote sensing scene classification (RSSC) focuses on scene-level understanding by capturing the spatial distribution patterns of objects and their associated semantic attributes within remote sensing images (RSIs) [1]. Over the past decade, deep learning-based RSSC methods have achieved remarkable success in static, closed-set environments, extensively supporting applications such as land-use and land-cover mapping [2], environmental monitoring [3], urban planning [4,5], and related geospatial applications [6–10]. In real-world RSSC systems, new scene classes may emerge due to environmental changes, urban expansion, sensor upgrades, or the incorporation of new geographic regions into monitoring programs. In addition, the rapid advances in satellite sensors and earth observation technologies have led to a substantial increase in both the quantity and diversity of RSIs [11], thereby requiring RSSC systems to continuously adapt to newly emerging scene classes.

Most existing deep learning approaches for RSSC follow the offline learning paradigm, in which models are trained using all available data with fixed class sets. However, this paradigm becomes inadequate in dynamic open-world environments where training data are received as streams and new classes continually emerge. Retraining models from scratch using both historical and incoming data largely wastes the substantial computational resources previously invested in the old model. More critically, in many operational settings, historical data cannot be permanently stored or repeatedly accessed due to strict memory budgets or data privacy regulations.

These intrinsic limitations necessitate the adoption of the class-incremental learning (CIL) paradigm, which enables models to acquire new knowledge sequentially without requiring full retraining. Inspired by human cognition, CIL methods aim to retain prior knowledge while simultaneously integrating new information [12]. However, CIL methods are susceptible to catastrophic forgetting [13], where the incorporation of new classes leads to severe degradation in performance on previously learned classes. Therefore, effective CIL methods must balance model stability and plasticity [14,15], achieving a trade-off between preserving old knowledge and adapting to new classes.

Recently, incremental learning has attracted substantial interest in remote sensing applications. Notably, several studies have extended CIL to the task of RSSC [16–26]. To mitigate catastrophic forgetting, most existing methods achieve promising performance by employing replay-based strategy, where a subset of representative exemplars from past classes is stored in a limited memory buffer and jointly used with new data for mixed training. While effective in high-resource server environments, this strategy encounters significant bottlenecks in real-world deployment. Specifically, the requirement to store historical data frequently conflicts with strict memory constraints and data privacy protocols.

To address the inherent limitations of image-based replay in practical CIL scenarios, we propose a memory-efficient feature-replay framework that preserves compact feature embeddings instead of raw historical images. The overall workflow of our proposed framework is schematically illustrated in Figure 1. This feature-space rehearsal strategy provides a dual advantage. By retaining semantically rich feature embeddings, it effectively preserves the decision boundaries of previously learned classes while alleviating prediction bias toward newly introduced classes. Meanwhile, storing compact embeddings instead of raw data substantially enhances memory efficiency without sacrificing classification performance, thereby enabling practical continual adaptation in real-world remote sensing deployments.

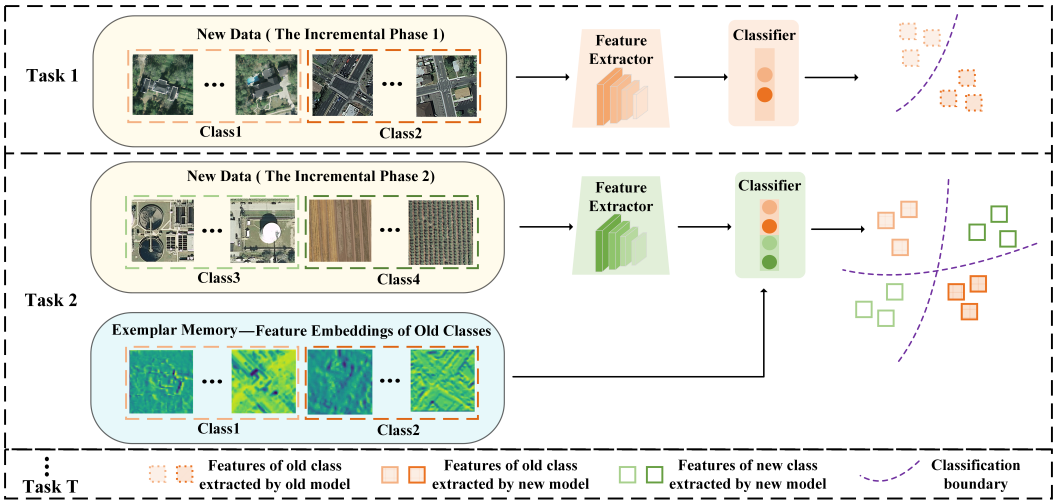


Figure 1. Overall workflow of the proposed framework. Compact feature embeddings, rather than raw images, are retained as exemplars for previously learned classes, highlighting how feature-level replay enables efficient and stable incremental learning.

Although retaining old feature embeddings is technically feasible, effectively incorporating them into mixed training with new data in the CIL setting remains challenging when applied to complex remote sensing scenarios. Specifically, the feature-replay strategy continues to face three key challenges: feature space drift, representation ambiguity and classifier bias. Firstly, despite its memory efficiency, our strategy of preserving feature descriptors confronts a critical challenge inherent to incremental learning: feature space drift. As the model is sequentially trained on new classes, the feature extractor is continually updated. This process inevitably alters the embedding space, rendering the stored descriptors from previous tasks obsolete. Consequently, a significant distributional gap emerges between the legacy feature space and the current one, making the two sets of representations incompatible. Secondly, the complex backgrounds, intra-class variability, and inter-class similarities inherent in RSIs necessitate the learning of well-structured and coherent feature representations. In addition, our feature replay strategy relies on replaying robust and discriminative features, which are essential for maintaining decision boundaries.

However, it becomes particularly challenging in CIL. As new classes emerge over time and the feature extractor is continuously updated across incremental stage, the feature representations of new and old classes become overlapped in the feature space, leading to ambiguity and making it difficult to distinguish between them [27]. Thirdly, new classes are typically trained with abundant data in CIL, whereas previous classes are represented by limited exemplars, leading to an imbalanced training distribution that skews predictions toward new classes. In addition, the commonly adopted knowledge distillation in replay-based methods relies on soft labels generated by earlier models, which become increasingly noisy due to error accumulation and feature distribution shifts. As a result, CIL methods are inherently prone to class imbalance and noisy distillation labels, both of which induce significant classifier bias.

To tackle the challenges inherent to feature-replay paradigm, we propose several specific technical components within our framework. First, we introduce a specialized feature calibration network (FCN) to compensate for feature space drift. The FCN efficiently adapt stored feature descriptors to the updated feature space, thereby bridging the distributional gap and facilitating a balanced training of the unified classifier. A key challenge in feature calibration is error accumulation: naive feature mapping strategies will accumulate errors over successive incremental steps, leading to performance degradation as the number of classes grows. To enhance the robustness of this mapping, we adopt a transductive learning strategy to train FCN. This strategy is notable as it exclusively leverages paired feature vectors from the current task, thereby obviating the need for original past-task images. Specifically, we model the feature space alignment as an orthogonal transformation, implemented via the Cayley transform, to enforce a manifold-preserving regularization constraint. This principled transformation preserves the intrinsic geometric structure and relative relationships among the feature vectors. To counteract representation ambiguity, we propose a progressive multi-scale feature enhancement (PMFE) module, which employs a progressive construction scheme to enable fine-grained and interactive feature enhancement, thereby yielding richer feature representations. To mitigate classifier bias, we employ a bias rectification (BR) strategy. Once the feature calibration process is complete, we fix the feature extractor and continue to optimize the classifier further only using the calibrated old class features and the new class features, effectively mitigating classification bias in the CIL task.

The primary contributions of our work are summarized as follows.

1. We propose a novel CIL framework for RSSC that retains compact feature embeddings, rather than raw images, as exemplars for previously learned classes. This memory-efficient feature replay method not only addresses data privacy concerns but also mitigates representation drift and classifier bias induced by data imbalance. Consequently, the framework maintains robust decision boundaries and significantly alleviates catastrophic forgetting.
2. A specialized FCN is trained in a transductive learning paradigm with manifold consistency regularization to adapt previously outdated feature descriptors to the current feature space. The FCN effectively compensates for feature space drift, thereby facilitating balanced and compatible unified classifier training. Following feature calibration process, we implement a BR strategy that mitigates the final prediction bias by exclusively optimizes the classifier on a balanced exemplar set.
3. To mitigate representation ambiguity, we propose a PMFE module. By adopting a progressive construction scheme, the PMFE module achieves fine-grained and interactive feature enhancement, yielding richer feature representations and a more comprehensive understanding of remote sensing scenes.

2. Related Work

In this section, we first provide a comprehensive overview of representative studies in incremental learning, primarily focusing on strategies to combat catastrophic forgetting. Subsequently, we present a overview of the existing CIL methods tailored for RSSC.

2.1. Incremental Learning Methods

Unlike statical offline training paradigms, incremental learning methods are designed to sequentially acquire new tasks while retaining knowledge learned from previous ones. Many incremental learning methods are conceptually inspired by cognitive processes in the human brain [12]. Existing strategies for alleviating catastrophic forgetting can be broadly categorized into three types: replay-based, regularization-based, and architecture-based methods. Each type exhibits distinct characteristics, reflecting inherent trade-offs in effectiveness, complexity, and resource requirements.

2.1.1. Replay-Based Methods

Memory, the core mechanism in the human brain for retaining learned knowledge, provides the foundation for knowledge reuse and augmentation. This biological process offers conceptual inspiration for replay-based methods [12]. These methods be generally categorized into two types based on the knowledge storage mechanism: rehearsal learning and generative learning.

For rehearsal learning, methods mitigate catastrophic forgetting by storing and replaying a small subset of data from previously learned tasks, commonly termed exemplars. By periodically revisiting these exemplars during the training of new tasks, the model reinforces its understanding of past knowledge, thereby preserving performance on earlier tasks. Along this line, Rebuffi et al. [28] proposed iCaRL, a unified framework that simultaneously updates feature representations and classifiers. Notably, this work pioneered the integration of rehearsal mechanisms with knowledge distillation strategies. Rainbow memory (RM) [29] adopts an uncertainty-aware sample selection mechanism combined with data augmentation to enhance replay efficiency. Several methods incorporate bias correction strategies during the rehearsal process. For instance, BiC [30] mitigated classifier bias by learning adjustment coefficients from a balanced validation set, while WA [31] alleviated bias through weight normalization, avoiding the need for additional correction parameters. GD [32] addressed prediction bias by employing gradient scaling to mimic repeated data exposure during the class-balanced tuning phase. DRC [33] tackled the classifier bias issue through the implementation of a dynamic residual architecture.

In generative learning, representative samples are synthesized using generative models being explicitly stored. These models, such as generative adversarial networks (GANs) [34], are trained on data from previous tasks to learn and replicate their underlying statistical distributions. Liu et al. [35] proposed a method to preserve representations from previously learned tasks by learning a feature-level generative model, which synthesizes pseudo-features to support replay in continual learning. Shi et al. [36] constructed pseudo-features for rehearsal through random bidirectional interpolation between current class embeddings and stored historical prototypes.

2.1.2. Regularization-Based Methods

Neuroscience research suggests that synaptic plasticity is the key mechanism responsible for synchronizing previously acquired knowledge with new information. Drawing inspiration from this biological process, regularization-based methods have been proposed [12]. These methods operate by incorporating a regularization term into the loss function during the training of new tasks. This term penalizes significant changes to

model weights that are vital for prior tasks, thereby preventing catastrophic forgetting. These methods are broadly classified into two main categories: prior-focused methods and data-focused methods.

In prior-focused methods, previous model is utilized as prior information to constrain the learning process of the new task. These methods typically employ two main techniques: parameter importance estimation and subspace projection. Regarding the former, elastic weight consolidation (EWC) [37] was the pioneering parameter regularization method for alleviating catastrophic forgetting in CIL. By adopting Fisher information-based importance estimation, it penalized changes to parameters deemed essential for earlier tasks. Similarly, synaptic intelligence (SI) [38] estimated parameter importance by accumulating the impact of each parameter on reducing the training loss along the optimization trajectory. Furthermore, Pan et al. [39] introduced a functional regularization strategy that leverages a Gaussian Process formulation to select exemplars and construct a functional prior. For the latter, gradient episodic memory (GEM) [40] utilized episodic memory to constrain gradient updates, ensuring that the optimization of new tasks does not interfere with the performance on previously learned tasks.

Data-focused methods employ knowledge distillation to enforce consistency between the previous and current models at multiple levels, including preserving output probability distributions, maintaining feature representation integrity, and ensuring stable inter-sample relationships across incremental steps. iCaRL [28] combined rehearsal memory with knowledge distillation, utilizing output post-sigmoid probabilities to preserve old knowledge.

TwF [41] introduced a feature distillation strategy on intermediate layers, employing an attention map as a binary mask to selectively regulate the distillation process. Kang et al. [42] adopted a distillation strategy to constrain the drift of critical features, quantifying feature importance by estimating the upper bound of loss variation via Taylor approximation. Dong et al. [43] proposed a method to preserve old knowledge by encoding local exemplar relationships into an exemplar relation graph and enforcing consistency via a relational distillation loss. Co2L [44] utilized contrastive self-distillation on instance-wise relations, employing an asymmetric supervised contrastive loss to preserve relational structures across tasks.

2.1.3. Architecture-Based Methods

Architecture-based methods offer an effective mechanism for knowledge retention and reuse in dynamic environments. These methods draw inspiration from principles observed in biological learning systems, such as neurogenesis and structural modularity [12]. By integrating modularity with plasticity, architecture-based methods can effectively retain knowledge even in dynamic environments. Generally, they can be categorized into two streams: dynamic network methods and static network methods.

Dynamic network methods adapt to each new task by structurally expanding the model. This expansion typically involves adding new nodes [45], branches [46], and subnetworks [47]. Conversely, some methods incorporate pruning mechanisms to eliminate redundant nodes, layers, or parameters. Zhang et al. [48] introduced a multi-head architecture featuring a shared backbone coupled with task-specific prediction heads, where the final inference is derived through a weighted aggregation of their individual outputs. Yang et al. [49] proposed a dynamically expandable network that accommodates new features and employs a self-activation mechanism to mitigate network redundancy.

Static network methods employ a fixed architecture where a dedicated subnetwork is reserved for each task. During the training of a new task, the subnetworks associated with previous tasks are masked to prevent interference. Chen et al. [50] proposed a semantically guided convolution filter and normalization strategy to optimize the network's static

parameters. Liu et al. [51] facilitated the optimization of fixed parameters by defining residual propagation for sparse convolution within layers, complemented by a network-level uncertainty variable.

2.2. Class-Incremental Learning Methods for RSSC

Recently, incremental learning techniques have attracted increasing attention in remote sensing applications. In particular, several studies have extended CIL to the task of RSSC.

Ye et al. [18] introduced the asymmetric collaborative network (SCN), which employs asymmetric dual subnetworks to separately encode historical and current knowledge, facilitating effective memory interaction via triple distillation and feature fusion mechanisms. Lu et al. [16] proposed a lightweight incremental learning (LIL) framework to address the parameter redundancy issue, employing a task-common feature extractor and lightweight feature transfer modules to align data distributions across tasks. LIL mitigates catastrophic forgetting while introducing only minimal parameter growth. Liu et al. [22] proposed an open-set incremental learning framework utilizing prototype learning and uncertainty measurement. It maintains performance on old tasks via a controllable convex hull-based exemplar selection strategy. Ammour [20] proposed a CIL method based on data regeneration, which employs a variational autoencoder to capture the latent structure of old tasks. By replaying generated data during incremental updates, the method mitigates catastrophic forgetting. Ye et al. [24] developed an efficient CIL network tailored for RSSC. It enhances feature extraction and employs a dynamic structural expansion strategy to fit the residuals of new tasks, while maintaining stability on old classes through model compression to reduce redundancy. Wang et al. [26] proposed an efficient CIL framework to address the challenges of long-tailed distributions in RSSC. It leverages head-class contexts to enrich tail-class representations via scaling grafting and mitigates catastrophic forgetting by treating historical classes as pseudo-tail instances during the incremental update.

In summary, most of the aforementioned methods are effective in alleviating catastrophic forgetting. A prevalent and effective strategy among these methods involves retaining a subset of old class samples in an exemplar memory for joint training with incoming new data. However, implementing this strategy in real-world scenarios is often challenging. Specifically, the requirement to store historical data frequently conflicts with strict memory constraints and data privacy protocols. To the best of our knowledge, none of the existing research in RSSC has explored feature replay strategies to address these limitations, which restricts the practical applicability of incremental learning methods in realistic deployment settings.

3. Methodology

3.1. Problem Setting

Unlike the traditional learning paradigm, which utilizes all available data simultaneously, CIL involves training on a sequence of tasks with disjoint class sets.The ultimate goal is to learn a unified classifier capable of recognizing all classes encountered up to the current stage. Specifically, in task t, the training set comprises data $\mathcal{D}_t$ for new classes and an exemplar memory $\mathcal{M}$ for old classes. We define $\mathcal{D}_t = \{X_t, Y_t\} = \left\{\left(\boldsymbol{x}_t^{(i)}, y_t^{(i)}\right)\right\}_{i=1}^{n_t}$, where n_t represents the number of training samples for new classes in task t, while X_t and Y_t denote the input data and corresponding target labels, respectively. Notably, the class sets across different tasks are disjoint, i.e., $Y_i \cap Y_j = \emptyset$ for $i \neq j$. For task t, our classification model θ_t comprises a feature extractor $\boldsymbol{f}_t : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^d$ and a unified classifier $\boldsymbol{g}_t : \mathbb{R}^d \rightarrow \mathbb{R}^{C_t}$. Here, d denotes the feature dimension, and C_t indicates the cumulative

number of categories learned up to stage t. The model is formulated as the composition $\theta_t = g_t \circ f_t$. Consequently, the final prediction for a test sample x_{test} is obtained by:

$$\hat{y}_{\text{test}} = \arg\max_{y \in \mathcal{C}_t} z_{ty}(x_{\text{test}}, \theta_t) \tag{1}$$

where $z_y(\cdot)$ represents the logit corresponding to class y.

3.2. Method Overview

To address the challenge of catastrophic forgetting and effectively balance the stability-plasticity trade-off, we propose a novel CIL framework for RSSC, termed FR-CIL. It mainly consists of four key components: (1) the PMFE module, designed to construct fine-grained, interactive feature representations through parallel depthwise dilated convolutions with progressive connections; (2) the DSKR mechanism, which incorporates two synergistic distillation losses to preserve stability across both decision boundaries and the feature representation space; (3) the FCN, which trained in a transductive learning paradigm under manifold consistency regularization, efficiently adapting the previously stored feature descriptors to the updated feature space to reconcile the distributional gap for a unified classifier; (4) the BR strategy, which exclusively optimizes the classifier on a balanced exemplar set to mitigate prediction bias.

An overview of the proposed method (FRCIL) is illustrated in Figure 2, which is organized as a unified pipeline and proceeds through three sequential stages. (1) Backbone training stage. The backbone network is trained using a combination of new class data and retained feature descriptors. The optimization is governed by a modified cross-entropy loss augmented with cosine normalization and the DSKR mechanism. (2) Feature calibration stage. The FCN is employed to adapt historical feature descriptors from the previous latent space to the updated feature space, thereby ensuring distributional alignment. (3) Bias rectification stage. The feature extractor is fixed, and the classifier is further optimized exclusively using a balanced set of calibrated old features and new class features. This step effectively mitigates the classifier bias inherent in CIL tasks.

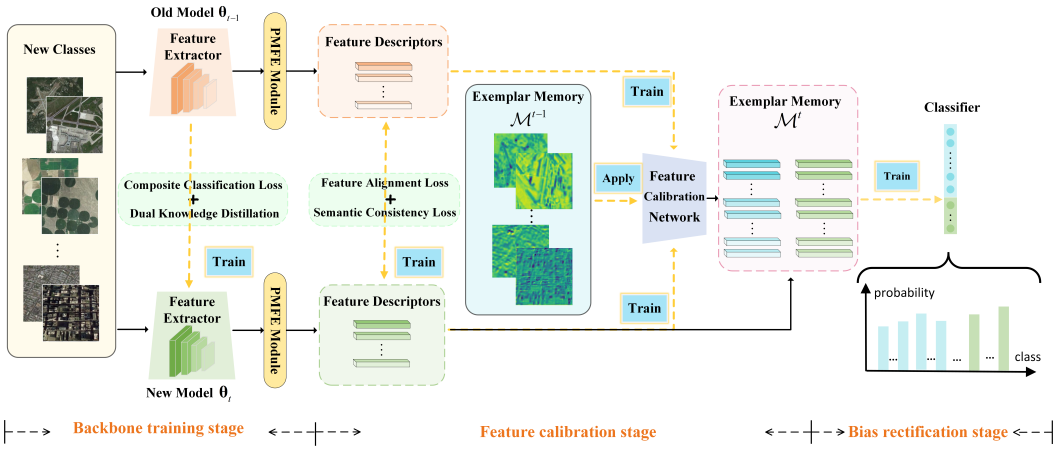


Figure 2. An overview of our method (FR-CIL). Given images of new classes, a new model is trained via distillation and classification losses. Feature descriptors extracted from the new images by both the old and new models are utilized to train a feature calibration network. This learned FCN is then applied to the preserved exemplars to adapt them to the current feature space. Finally, with features across all seen classes aligned in a unified space, a feature classifier is optimized to mitigate the classifier bias inherent in CIL tasks.

In summary, our proposed framework is designed to effectively mitigate both representation ambiguity and classifier bias in CIL for RSSC. In the subsequent sections, we provide a detailed elaboration of the key components and procedural steps.

3.3. Feature Learning

Given the complex semantic content inherent in remote sensing scenes, capturing features across a diverse range of spatial scales is essential for robust classification. However, conventional multi-scale feature extraction paradigms typically rely on parallel convolutional pathways with fixed receptive fields, often resulting in insufficient contextual correlation and limited feature interaction.

To address these limitations, we propose a progressive multi-scale feature enhancement (PMFE) module. This module adopts a progressive construction scheme to facilitate feature enhancement in a fine-grained and interactive manner, effectively remedying the deficiencies of static multi-branch architectures.

The structural design of the PMFE module is illustrated in Figure 3. PMFE employs a multi-branch architecture designed to systematically expand the receptive field. This is realized through a set of parallel 3×3 depthwise dilated convolutions configured with incrementally increasing dilation rates (e.g., $r \in \{1, 2, 4, 8\}$). By strategically introducing dilation within the kernel structure, depthwise dilated convolution effectively expands the receptive field without increasing the number of parameters. Furthermore, when synergized with pointwise convolution, this approach significantly curbs computational complexity. This structure enables the module to effectively capture a comprehensive range of multi-scale features, spanning from fine-grained local textures to broader global structural contexts. Formally, this operation is defined by the following equation:

$$\mathbf{F_i} = \begin{cases} \mathcal{P}(\sigma(\mathcal{D}^{3\times3}_{r=r_i}(\mathbf{F}))) & i = 1 \\ \mathcal{P}(\sigma(\mathcal{D}^{3\times3}_{r=r_i}(\mathcal{C}oncat[\mathbf{F}, \mathbf{F}_{i-1}]))) & i = 2, 3, 4 \end{cases} \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map, and Concat$[\cdot]$ represents channel-wise concatenation. The operator $\mathcal{D}^{3\times3}_{r=r_i}$ signifies a 3×3 depthwise dilated convolution with a dilation rate of $r_i = 2^{(i-1)}$, while $\mathcal{P}(\cdot)$ denotes a pointwise convolution employed to adjust the channel dimensionality to C. σ denotes ReLU activation function. Finally, $\mathbf{F}_i$ represents the enhanced feature output of the i-th branch.

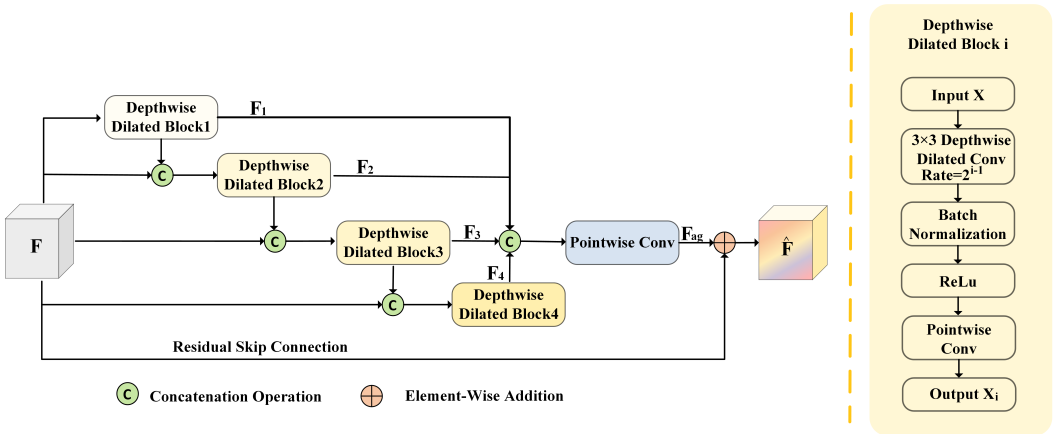


Figure 3. Structure of the PMFE module. Parallel 3×3 depthwise dilated convolutions with increasing dilation rates are employed to progressively expand the receptive field. By integrating pointwise convolutions, the structure minimizes computational cost while effectively facilitating fine-grained and interactive feature enhancement. The module features a streamlined architecture, ensuring high modularity and ease of integration.

The choice of exponentially increasing dilation rates is designed to rapidly expand the receptive field to capture the broad global context required for complex remote sensing scenes, without increasing computational overhead [52,53]. Furthermore, this progressive cascading structure naturally mitigates the gridding effect commonly associated with large dilation rates [53]. As each branch fuses the original input F with the densely extracted features F_{i-1} from the preceding smaller dilation rate, the spatial holes introduced by sparse sampling are continuously filled. This fine-grained and interactive enhancement preserves local continuity while capturing multi-scale structural dependencies.

Subsequently, the multi-scale feature maps $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3$, and $\mathbf{F}_4$ are concatenated along the channel dimension. The resulting tensor is then passed through a pointwise convolution $\mathcal{P}(\cdot)$ to fuse information across scales and adjust the channel dimensionality. This process yields the aggregated feature map $\mathbf{F}_{ag} \in \mathbb{R}^{C\times H\times W}$, formulated as follows:

$$\mathbf{F}_{ag} = \mathcal{P}(Concat[\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4]) \tag{3}$$

Finally, a residual skip connection is introduced to fuse the original input with the enhanced features. This design facilitates identity mapping, ensuring the preservation of original information while effectively mitigating the vanishing gradient problem to stabilize model training. Consequently, the final output $\hat{\mathbf{F}}$ can be calculated as follows:

$$\hat{\mathbf{F}} = \mathbf{F} + \mathbf{F_{ag}} \tag{4}$$

Conventional feature pyramid architectures, such as FPN [54] and PANet [55], are often characterized by intricate designs involving elaborate pathways. In contrast, the proposed PMFE module offers a streamlined architecture, ensuring high modularity and ease of integration. Notably, the construction process follows a coarse-to-fine cognitive paradigm, where global structural understanding is progressively enriched with local details. Consequently, the PMFE module is particularly adept at addressing the inherent challenges in RSIs, such as complex background interference, extreme scale variations, and high inter-class similarity.

3.4. Incremental Learning

CIL methods inherently suffer from catastrophic forgetting, characterized by a sharp degradation in performance on previously learned tasks when sequentially adapted to new data. To alleviate this, replay-based strategies mitigate forgetting by preserving a representative subset of old class samples within an exemplar memory. By interleaving these retained exemplars with incoming data during subsequent training phases, the model enables joint optimization over both past and current information, thereby maintaining old knowledge while while learning new ones.

Despite their effectiveness, replay-based strategies are hindered by inherent limitations. Firstly, storing raw samples imposes a substantial storage overhead that escalates linearly with the number of tasks. Secondly, the retention of historical RSIs may involve sensitive or proprietary information, raising critical privacy and security concerns. Finally, the data imbalance between the limited exemplar set and abundant new data leads to a severe prediction bias toward new classes.

To address these issues, we propose a novel replay-based framework for CIL that preserves low-dimensional feature descriptors, rather than raw images, as exemplars for previously learned classes. We employ ResNet-18 as the feature extractor, followed by a cosine classifier for final classification. The model is optimized using a composite cross-

entropy loss, which is applied jointly to both the new class training data $\mathcal{D}_t$ and the retained feature exemplars $\mathcal{M}^{t-1}$. The loss function is formulated as:

$$\mathcal{L}_{CE} = \sum_{(x,y)\in\mathcal{D}_t} -y\log(p_{1:t}(\mathbf{x},\theta)) + \sum_{(\mathbf{v},y)\in\mathcal{M}^{t-1}} -y\log(p_{1:t-1}(\mathbf{v},\boldsymbol{g}_t)) \tag{5}$$

where $p_{1:t}(\mathbf{x},\theta)$ denotes the probability distribution computed over all classes encountered up to stage t. The first term represents the standard classification loss for the new task data $\mathcal{D}_t$, and the second term imposes a supervision constraint on the retained feature exemplars $\mathbf{v}$, which bypass the feature extractor $\boldsymbol{f}_t$ and are evaluated by the unified classifier $\boldsymbol{g}_t$.

To further mitigate catastrophic forgetting and enhance feature discriminability, we augment this loss function with a cosine normalization strategy alongside a dual-space knowledge retention (DSKR) mechanism. The details of these components are elaborated below.

3.4.1. Cosine Normalization

During incremental learning, replay-based methods exhibit a pronounced prediction bias toward new classes. This phenomenon stems from the severe data disparity between the abundant new training data and the limited exemplar set, manifesting empirically as larger logit values for new classes compared to old ones.

This bias is structurally exacerbated by the standard fully connected layer. Specifically, the prediction logit for class i, denoted as $z_i = \boldsymbol{g}_i^\top \boldsymbol{f}_t(\boldsymbol{x}) + b_i$, is computed via the dot product between the feature vector $\boldsymbol{f}_t(\boldsymbol{x})$ and the classifier weight $\boldsymbol{g}_i$. As the dot product is sensitive to vector magnitudes, the larger weight norms of new classes directly inflate their output scores, regardless of the actual semantic similarity.

To mitigate this magnitude-driven bias, we reformulate the logit computation via cosine normalization. Specifically, the logit z_i' is computed via a scaled cosine similarity, effectively decoupling the prediction score from the vector magnitude:

$$z_i' = \langle \bar{\boldsymbol{g}}_i, \bar{\boldsymbol{f}}_t(x) \rangle \tag{6}$$

where $\bar{\boldsymbol{m}} = \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|_2}$ denotes the ℓ_2-normalized version of a vector $\boldsymbol{m}$, and the inner product $\langle \bar{m}_1, \bar{m}_2 \rangle = \bar{m}_1^{\mathrm{T}} \bar{m}_2$ quantifies the cosine similarity between a pair of normalized vectors. The final probabilities are then obtained by applying the softmax function to these rectified logits.

3.4.2. Dual-Space Knowledge Retention Mechanism

To effectively mitigate catastrophic forgetting, we employ a dual-space knowledge retention (DSKR) mechanism that integrates two complementary distillation strategies.

First, we introduce a prediction-space knowledge distillation loss, denoted as $\mathcal{L}_{KD}$. This objective serves as a regularization constraint, forcing the output probability distribution of the current model to align with that of the frozen old model. This process ensures the preservation of previously established decision boundaries. Formally, during the training of task t, $\mathcal{L}_{KD}$ is computed as:

$$\mathcal{L}_{KD} = D_{KL}(p_{1:t-1}^{\tau}(\mathbf{x},\theta_{t-1}) || p_{1:t-1}^{\tau}(\mathbf{x},\theta_t)) \tag{7}$$

where $D_{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. The term $p_{1:t-1}^{\tau}$ represents the temperature-scaled probability distribution computed over previously learned classes.

While prediction-space distillation preserves the final decision boundaries, feature representations may gradually drift. To address this, we introduce a feature-space distillation loss $\mathcal{L}_{FD}$ to impose a direct constraint on the feature representations, complementing

the output-level supervision. This loss encourages the current model to maintain discriminative feature representations consistent with the old model. Specifically, it minimizes the cosine distance between the normalized feature embeddings extracted by the model at the previous and current steps. The loss is formulated as:

$$\mathcal{L}_{FD} = 1 - \langle \bar{f}_t(x) \cdot \bar{f}_{t-1}(x) \rangle \tag{8}$$

As a result, the DSKR mechanism guarantees the stability of the decision boundaries and feature representation, providing a holistic defense against catastrophic forgetting.

3.4.3. The Total Loss

Combining these components, the total objective function $\mathcal{L}_{total}$ for optimizing the current model θ_t is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \mathcal{L}_{FD} \tag{9}$$

The hyperparameters λ_1 and λ_2 serve as weighting coefficients for the respective distillation losses.

3.5. Feature Calibration

A central challenge in feature replay is that feature descriptors stored from previous tasks become progressively misaligned with the evolving feature space as the model incrementally learns new classes. This distributional shift creates a discrepancy between legacy and current feature representations, rendering them incompatible. Consequently, the core objective is to adapt these historical feature descriptors into the current latent space, effectively bridging the distributional gap to facilitate the construction of a unified and robust classifier.

To resolve the feature space incompatibility, we propose a transductive feature calibration network that obviates the need for storing past raw images. This mechanism leverages the current task data $\mathcal{D}_t$ to generate a set of paired feature correspondences, which subsequently serve as proxy supervision for the feature calibration process. Specifically, for each sample $\boldsymbol{x}_k$ in the current dataset $\mathcal{D}_t$, we generate a pair of feature vectors: one obtained by passing the sample through the frozen prior extractor $\boldsymbol{f}_{t-1}$, and the other through the current extractor $\boldsymbol{f}_t$. As a result, two corresponding feature sets are generated as follows:

$$\mathcal{V}_{t-1} = \{\boldsymbol{f}_{t-1}(\boldsymbol{x}_k) \mid \boldsymbol{x}_k \in \mathcal{D}_t\}_{k=1}^{N_t} \tag{10}$$

$$\mathcal{V}_t = \{\boldsymbol{f}_t(\boldsymbol{x}_k) \mid \boldsymbol{x}_k \in \mathcal{D}_t\}_{k=1}^{N_t} \tag{11}$$

The discrepancy between $\mathcal{V}_{t-1}$ and $\mathcal{V}_t$ encapsulates the evolution of the feature space. The proposed feature calibration network (FCN) learns to bridge this gap by optimizing a transformation function that adapts the representation from the prior space $\mathcal{V}_{t-1}$ to the current space $\mathcal{V}_t$.

The entire calibration process adheres to the manifold consistency principle, which enforces that the intrinsic topology of the underlying data manifold remains invariant. This regularization ensures that the relative distances and structural relationships among the stored old descriptors are preserved, which guarantees a robust and structurally coherent calibration process.

To enforce this geometric constraint, we explicitly model the feature calibration process as an orthogonal operator, which is achieved by parameterizing the transformation weights

via the Cayley transform. Specifically, the orthogonal matrix $\boldsymbol{\phi}$ is generated from a learnable skew-symmetric matrix $\mathbf{S}$ using the formulation:

$$\boldsymbol{\phi} = (\mathbf{I} - \mathbf{S})(\mathbf{I} + \mathbf{S})^{-1} \tag{12}$$

where $\mathbf{I}$ denotes the identity matrix, and $\mathbf{S}$ satisfies the skew-symmetric property $\mathbf{S}^\top = -\mathbf{S}$.

During the calibration phase, with both the feature extractor and classifier frozen, we exclusively optimize the parameters of $\boldsymbol{\phi}$. The objective is to learn the optimal orthogonal transformation that aligns the calibrated old features with the current feature representations. To achieve this, a dual-component loss function is employed to enforce both feature alignment and semantic consistency. The formulation is given by:

$$\mathcal{L}_C = \sum_{v^{t-1} \in \mathcal{V}_{t-1}, v^t \in \mathcal{V}_t} (1 - \langle \boldsymbol{\phi} v^{t-1} \cdot v^t \rangle) + \alpha \sum_{(x_i, y_i) \in \mathcal{D}_t} -y_i \log g_t \circ (\boldsymbol{\phi}(f_{t-1}(x_{t,i}))) \tag{13}$$

where the hyperparameter α modulates the weight of the feature alignment loss. The first component is a feature alignment loss, which minimizes the spatial distance between the original and calibrated feature descriptors by leveraging cosine similarity. The second component is a semantic consistency loss that ensures the calibrated feature descriptors are still correctly classified by the current model. Consequently, this process ensures that the retained feature descriptors are adapted to the new latent space, without distorting its intrinsic topological structure or compromising its semantic consistency.

3.6. Bias Rectification

The commonly adopted knowledge distillation in replay-based methods relies on soft labels generated by earlier models, which become increasingly noisy due to error accumulation and feature distribution shifts. However, CIL methods are inherently susceptible to class imbalance and noisy distillation labels, both of which induce severe classifier bias.

To address these issues, we propose a bias rectification (BR) strategy. Upon completion of the feature calibration process, the memory buffer $\mathcal{M}_t$ is updated to adhere to specific memory constraints. We employ herding strategy [28] to select the most representative samples. This selection algorithm iteratively chooses a subset of feature descriptors such that their cumulative average best approximates the true class mean in the feature space.

In this stage, we freeze the feature extractor and exclusively optimize the classifier on the balanced exemplar set $\mathcal{M}_t$.

$$\mathcal{L}_R = -\sum_{(\mathbf{f_i}, y_i) \in \mathcal{M}^t} \sum_{c=1}^{C_t} \mathbb{I}(y_i = c) \log([g_t \circ (\mathbf{f}_i)]_c) \tag{14}$$

This optimization uses only true class labels, thereby avoiding the label noise inherent in distillation-based soft labels. As a result, the proposed BR strategy effectively mitigates prediction bias and yields consistent performance improvements across all previously learned classes.

The pseudo-code for the proposed method is presented in Algorithm 1. This procedure details the optimization steps required to address catastrophic forgetting.

Algorithm 1 Pseudocode for Our Method Training at Incremental Stage t

Input: New data $\mathcal{D}_t$; exemplar memory $\mathcal{M}^{t-1}$; old model $\boldsymbol{\theta}_{t-1}$; hyperparameters λ_1, λ_2 and α.

Initialize new module θ_t with θ_{t-1} and extend the classifier head for new classes.
for each incremental stage $t \in \{1, \ldots, T\}$ **do**
  if $t = 1$ **then**
    Train the first task:
    Train model $\boldsymbol{\theta}_1$ by minimizing $\mathcal{L}_{total}$, according to Equation (9).
    Construct exemplar memory $\mathcal{M}^1$ by herding strategy [28].
    Retain the model $\boldsymbol{\theta}_1$.
  else
    Train incremental tasks:
    Initialize new module $\boldsymbol{\theta}_t$ with $\boldsymbol{\theta}_{t-1}$.
    Train model $\boldsymbol{\theta}_t$ by minimizing $\mathcal{L}_{total}$, according to Equation (9).
    Train FCN $\boldsymbol{\phi}$ by minimizing $\mathcal{L}_C$, according to Equation (13).
    Update exemplar memory $\mathcal{M}^{t-1} = \boldsymbol{\phi}(\mathcal{M}^{t-1})$.
    Construct exemplar memory$\mathcal{M}^t$ by herding strategy [28].
    Jointly train classifier $\mathcal{F}^t$ by optimizing $\mathcal{L}_R$, according to Equation (14).
    Discard the old model $\boldsymbol{\theta}_{t-1}$.
    Retain the new model $\boldsymbol{\theta}_t$ for next stage.
  end if
end for
Output: New model $\boldsymbol{\theta}_t$; exemplar memory $\mathcal{M}^t$.

4. Experiments and Results

In this section, we present a comprehensive evaluation of the proposed FR-CIL across five datasets using diverse data split protocols. First, we outline the experimental setting, involving the datasets, selected baselines for comparison, evaluation metrics, and implementation details. Next, we compare FR-CIL against several state-of-the-art methods to demonstrate its superior performance. Following this, sequential learning experiments are conducted to analyze its stability and performance trends as tasks accumulate. Subsequently, we analyze resource efficiency by comparing memory footprint and evaluating sensitivity to the number of preserved data points. Finally, we perform ablation studies to examine the contribution of each proposed component to the overall performance and investigate model's robustness to changes in hyperparameters. Visual analysis further complements our findings, providing additional insights into the model's behavior.

4.1. Experimental Setting

4.1.1. Datasets

To comprehensively evaluate the effectiveness of the proposed method, experiments were conducted on five datasets: AID [56], RSICD-256 [57], NWPU-45 [11], PatternNet [58], and UC-Merced [59]. These datasets encompass a diverse range of spatial resolutions and scene complexities. The detailed characteristics of each dataset are summarized as follows.

The AID dataset contains 30 scene classes with spatial resolutions ranging from 0.5 to 8 m, characterized by high intra-class variability and high inter-class similarity. The RSICB256 dataset includes 35 classes featuring diverse angles, scales, and colors, providing rich sample diversity. The NWPU-45 dataset consists of 45 classes with spatial resolutions spanning 0.2–30 m, exhibiting significant variations in viewpoint, lighting, and occlusion. The PatternNet dataset comprises 38 classes with resolutions of 0.06–4.69 m, ensuring high object occupancy to minimize background noise while maintaining visual diversity. Specifically, the 'airplane' and 'baseball field' classes were excluded to ensure class balance across the incremental learning stages. Finally, the UC-Merced dataset is composed of

21 classes with a fixed spatial resolution of 0.3 m, covering a wide range of agricultural, urban, and natural textures under varying lighting and background contexts.

To provide a comprehensive and structured evaluation of the proposed method, five datasets are strategically distributed across different experimental analyses according to their characteristics and evaluation objectives. Specifically, AID, RSI-CB256, and NWPU-45 are used as the primary benchmarks for state-of-the-art comparisons and core ablation studies, as they represent large-scale and widely recognized RSSC datasets with varying incremental task lengths. PatternNet is utilized specifically for sequential learning analysis to validate the multiscale feature enhancement capabilities of the PMFE module given its broad spectrum of spatial resolutions and rich visual diversity. UC-Merced is utilized for qualitative visualization and hyperparameter sensitivity analysis, offering clear interpretability due to its moderate scale and well-structured class distribution.

4.1.2. Baselines

In the experiments, the proposed method was compared with eleven classical incremental learning methods. (1) **Joint (Retraining)**: This setting represents the ideal offline baseline, where all available data are utilized for simultaneous joint training. Consequently, it serves as the performance upper bound for evaluating CIL methods. (2) **Finetuning**: A naive incremental baseline that sequentially trains solely on new tasks. However, it typically suffers from severe catastrophic forgetting. (3) **GEM** [40]: A projection-based approach that computes gradients for both the memory buffer and the current task, projecting the current gradient to minimize its angle relative to the memory gradients. (4) **EWC** [37]: A pioneering regularization-based approach that utilizes the diagonal of the Fisher information matrix to approximate the posterior distribution, thereby penalizing changes to parameters deemed critical for previous tasks. (5) **LwF.MC** [60]: A representative distillation-based method that addresses catastrophic forgetting by incorporating a knowledge distillation term into the global loss function. (6) **iCaRL** [28]: A hybrid approach that combines a distillation loss with a reserved memory of exemplars to preserve old knowledge. (7) **WA** [31]: A post-processing method designed to correct prediction bias toward new classes. It introduces a weight aligning strategy to adjust the biased weights in the final layer, ensuring fairness between old and new classes. (8) **CwD** [61]: A regularization strategy applied at the initial stage to mitigate representation collapse. It compels the model to learn uniformly scattered features by decorrelating class-wise representations, thereby boosting generalization for subsequent incremental stages. In this work, we utilize the AANet-based version of CwD. (9) **DER** [62]: A structure-based approach that dynamically expands the network architecture. It freezes previous feature extractors to ensure stability while adding prunable new branches for plasticity, employing an auxiliary loss to enhance feature discrimination. (10) **MEMO** [63]: A parameter-efficient expansion framework that decomposes the backbone into shared shallow layers and expandable deep layers. It optimizes memory allocation between model parameters and exemplars, ensuring robust performance across varying memory budgets. (11) **EASE** [64]: A framework based on pre-trained models that expands the architecture using lightweight adapters. It constructs task-specific subspaces and utilizes semantic similarities to synthesize old class prototypes, enabling ensemble prediction without forgetting.

4.1.3. Evaluation Metrics

Following standard evaluation protocols in CIL [27], we assess performance based on two primary aspects: the overall discriminative capability across learned tasks and the stability of memory regarding previously acquired knowledge. To quantify these, we utilize accuracy, mean accuracy (mACC), and backward transfer (BWT) metrics.

(1) **Accuracy**: The classification accuracy for a specific task k serves as the fundamental performance metric.

$$Accuracy(k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{1}(\hat{y}_i = y_i) \tag{15}$$

where N_k represents the total number of test samples for task k, $\hat{y}_i$ denotes the predicted label for the i-th sample, y_i is the corresponding ground truth label, and $\mathbb{1}(\cdot)$ is an indicator function that outputs 1 if the condition is satisfied and 0 otherwise.

(2) **mACC**: Mean accuracy serves as a comprehensive metric for evaluating the overall performance of incremental learning models. Notably, it assesses the stability-plasticity trade-off by averaging the model's accuracy on all encountered tasks after the entire learning sequence is complete.

Let $a_{k,j} \in [0, 1]$ denote the accuracy on the test set of task j after the model has been trained on task k (where $j \leq k$). A higher mACC indicates superior prediction capability across all observed classes. mACC is calculated as:

$$mACC = \frac{1}{T} \sum_{j=1}^{T} a_{T,j} \tag{16}$$

(3) **BWT**: The backward transfer measures the model's resistance to forgetting and quantifies the impact of new tasks on previously learned tasks. BWT is calculated as:

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} (P_{i,i} - P_{T,i}) \tag{17}$$

where a positive BWT value indicates forgetting previous tasks, and a negative BWT value suggests that learning new tasks has improved performance on old tasks.

4.1.4. Implementation Details

To ensure a fair comparison, ResNet-18 is employed as the backbone for all compared methods, with the exception of EASE. For EASE, the ViT-B/16 architecture is utilized. Regarding data partitioning, all datasets are randomly divided into training and test sets with a ratio of 4:1. Unless otherwise specified, all methods adopted the default hyperparameter configurations reported in their respective original papers. Specifically, for EWC, the regularization weight was set to 1000, with $\lambda = 15$. For LwF, the temperature scaling factor T was set to 2. In WA, L_2 normalization was applied to the fully connected layers. For DER, 10 warm-up epochs were employed with $\tau = 5$. CwD was implemented using the AANet-based version. Finally, for EASE, the adapter projection dimension r was set to 16, with a trade-off parameter α of 0.1. All methods were optimized using stochastic gradient descent for 200 epochs, with a momentum of 0.9 and a weight decay of 2×10^{-4}. A batch size of 128 was utilized. Regarding the exemplar memory set, the herding strategy [28] was adopted to select representative samples after the completion of each training phase. The transductive learning strategy strictly and exclusively utilizes the training data from the current incremental stage. There is strictly no data leakage, no access to future tasks, and no peeking at the global test set at any point in the training process.

To ensure a fair and reproducible comparison, we adopt a fixed total memory budget for all methods. Specifically, image-based replay is allocated 300 MB, while feature-based replay (FR-CIL) is allocated 16 MB. For image rehearsal, RGB images are stored in their raw format. For feature-based replay, we store feature embeddings using 32-bit floating-point precision. When new classes are introduced, the memory is evenly reallocated across all observed classes to maintain a fixed total memory constraint. The class order was randomized using a fixed numPy random seed, thereby enforcing an identical task

sequence across all methods. Each experiment was repeated three times, and the mean results are reported. No data augmentation strategies were utilized. All experiments were implemented in Python 3.8 using the PyTorch 1.9 framework and executed on a single NVIDIA RTX 3090 GPU.

4.2. Results and Analysis

To provide a comprehensive assessment of FR-CIL, we compare it with several state-of-the-art baselines. Experiments are conducted across three datasets to ensure a fair and robust evaluation, thereby reducing potential bias from any single dataset. The mACC and BWT results are summarized in Table 1.

Table 1. Performance comparison of different CIL methods on three datasets, where the upward arrow indicates that a larger value corresponds to better performance, and the downward arrow indicates the opposite.

Method \ Dataset	AID 6 Tasks with 5 Classes		RSI-CB256 7 Tasks with 5 Classes		NWPU-45 9 Tasks with 5 Classes	
	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓
Joint (Retraining)	95.31	-	96.42	-	87.69	-
Finetuning	21.65	80.58	19.92	84.65	10.61	92.57
GEM	27.92	75.51	21.57	81.93	9.32	93.82
EWC	27.31	73.84	23.58	81.67	13.83	88.65
LwF.MC	51.62	48.33	47.07	57.08	26.95	75.67
iCaRL	61.17	42.81	64.02	30.74	43.17	50.02
WA	66.82	35.07	69.29	26.42	53.98	47.01
CWD	74.06	20.97	80.92	16.82	68.79	29.93
DER	78.92	20.92	87.93	12.82	77.81	28.63
MEMO	81.37	20.29	84.63	12.97	74.14	28.85
EASE	85.21	12.67	87.98	14.27	76.96	14.81
Ours	88.96	11.08	91.07	7.96	80.63	16.07

As expected, Joint (Retraining) achieves the highest mACC scores. Since the model is trained using all accumulated data simultaneously, it does not suffer from catastrophic forgetting and therefore has no corresponding BWT value. Consequently, it serves as an upper bound for CIL performance. In contrast, finetuning focuses solely on the current task without any mechanism to mitigate catastrophic forgetting, resulting in the lowest mACC values and the most severe forgetting among all compared methods. Prior-focused methods, such as GEM and EWC, demonstrate limited effectiveness and offer negligible gains over finetuning. Notably, GEM exhibits even poorer BWT performance on NWPU-45, as the feasible gradient space for subspace projection becomes increasingly constrained with increasing tasks. EWC, which estimates parameter importance via the Fisher information matrix to restrict updates on critical parameters, similarly yields only a marginal improvement in mACC relative to finetuning. In contrast, LwF employs knowledge distillation during incremental updates, yielding a substantial improvement in mACC over finetuning and demonstrating the effectiveness of this strategy. iCaRL further enhances performance by maintaining an exemplar memory, resulting in higher mACC scores and highlighting the efficacy of exemplar replay. WA builds upon iCaRL by addressing class imbalance through bias correction, thereby achieving additional performance gains. CWD constrains the model to generate uniformly distributed features by explicitly decorrelating class-wise representations. This mechanism enhances generalization across incremental stages, leading to improved overall performance. Architecture-based methods (i.e., DER, MEMO, and EASE) employ expandable architectures and consequently deliver competitive

results. These findings highlight that greater representational capacity plays a crucial role in alleviating forgetting and improving continual learning robustness. Among them, EASE distinguishes itself by integrating pre-trained models, resulting in superior performance across all datasets. Our method surpasses existing state-of-the-art methods, achieving mACC values closer to those of the Joint upper bound and attaining the lowest BWT scores, thereby validating the effectiveness of the proposed framework.

4.3. Sequential Learning Analysis

CIL methods proceed in a sequential manner, where each incremental stage introduces new classes. To evaluate the dynamic learning behavior, we employ sequential learning analysis, which serves as a dynamic visualization tool for characterizing how well the model retains knowledge of previously learned classes while adapting to new ones. After completing the t-th incremental stage, we evaluate the model using mACC over all classes learned up to that point. While the mACC metric provides a global summary, it fails to elucidate the specific timing and magnitude of catastrophic forgetting. Consequently, we analyze the sequential learning trajectory on the ten-task AID dataset to provide a more detailed view of performance evolution as the number of learned classes increases.

As illustrated in Figure 4, the mACC curves generally exhibit a downward trend as the number of learned classes increases. Notably, our method maintains consistently high and stable mACC values throughout the incremental tasks, indicating its robust capability to adapt to new tasks without compromising old knowledge. Conversely, naive finetuning lacks any strategies to prevent catastrophic forgetting and therefore exhibits the poorest performance among all evaluated methods. Under long-sequence CIL settings, GEM behaves similarly to finetuning, failing to preserve knowledge of earlier tasks. EWC, which estimates parameter importance, achieves only a marginal improvement over finetuning. The above two methods suggest that such simple regularization is insufficient for the challenging CIL setting. The incremental improvements across LwF, iCaRL, and WA highlight the effectiveness of their strategies. These methods exhibit a progressive relationship: LwF employs knowledge distillation strategy; iCaRL improves upon this by incorporating exemplar replay; and WA further enhances performance via postprocessing for bias correction. Additionally, CwD mitigates representation collapse during the initial phase, outperforming methods that rely exclusively on replay or distillation. Architecture-based methods, including DER, MEMO, and EASE, rank among the top methods and achieve competitive results. By maintaining additional network branches or adapters, they sustain high mACC scores even as the total number of learned classes increases.

4.4. Impact of Memory Footprint

We propose a novel replay-based framework for CIL that preserves low-dimensional feature embeddings, rather than raw images, as exemplars for previously learned classes. The primary goal is to substantially reduce the memory consumption while maintaining performance. Given that different replay-based methods vary in their storage mechanisms, we standardize the memory budget to ensure a fair comparison. Specifically, considering that the NWPU-45 dataset imposes the most stringent storage demands due to its extensive class number and extended incremental sequence, we compare our approach with three representative methods employing distinct storage strategies on the NWPU-45 dataset in terms of the exemplar memory footprint.

As illustrated in Figure 5, our method demonstrates competitive mACC scores while maintaining a significantly lower memory footprint compared to conventional image-based replay methods. It is important to note that the reported memory footprint encompasses all preserved data (i.e., raw images or feature embeddings) across all classes. Specifically,

storing a single $256 \times 256 \times 3$ image in uint8 format requires 192 KB. However, for GPU-based training, these images are typically normalized and converted to float32 tensors, escalating the storage requirement to 768 KB per sample. In contrast, a floating-point feature embedding with dimension $d = 512$ occupies merely 2 KB, which represents less than 1% of the storage needed for the original image. Consequently, as demonstrated in Figure 5, our method reduces memory requirements by at least an order of magnitude while improving the mACC in most cases.

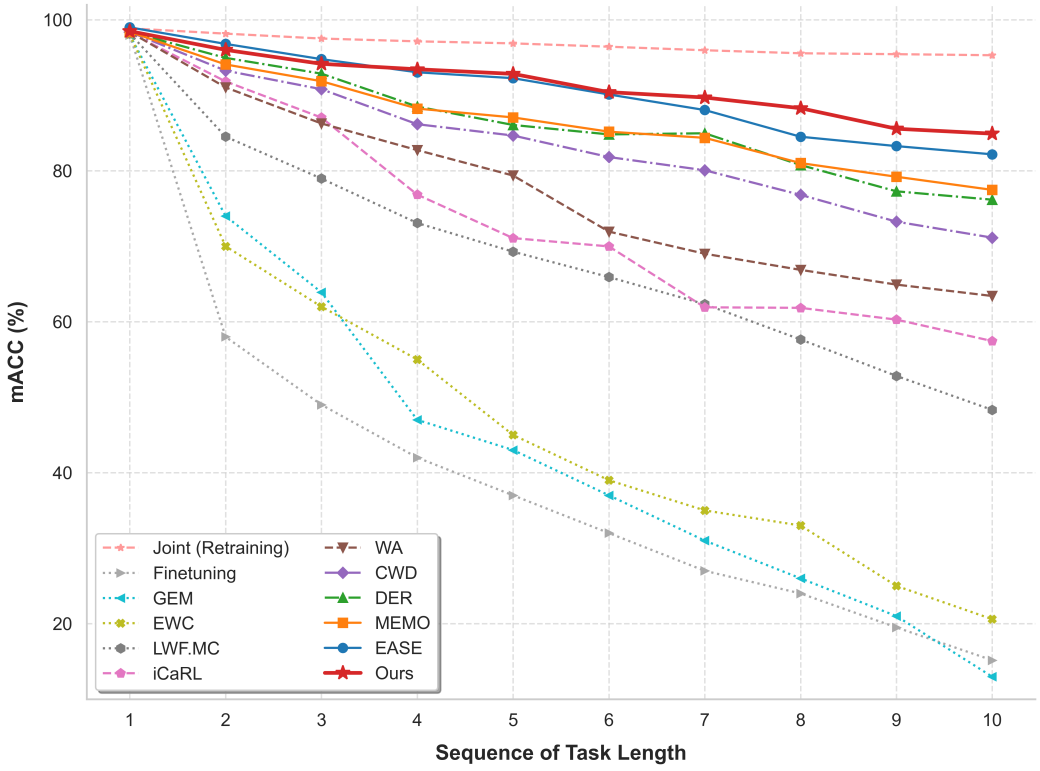


Figure 4. Performance comparison of different CIL methods on the ten-task AID dataset. The horizontal axis represents the task sequence length (incremental steps), while the vertical axis indicates the mean average accuracy (mACC).

The experimental results presented in Figure 5 illustrate the trade-off between model performance and storage overhead. As the quantity of preserved exemplars increases, both mACC and memory consumption rise. Notably, the proposed method achieves performance saturation at a memory usage of approximately 16 MB. In contrast, conventional image-based replay methods necessitate nearly 300 MB to achieve comparable stability. Beyond these thresholds, the marginal gains in mACC diminish significantly while memory cost continues to grow unnecessarily.

Given the substantial discrepancy in storage density between raw images and compact feature embeddings, strictly enforcing an identical memory budget for all comparisons proves impractical. Specifically, a budget sufficient for images implies an excessive number of features, whereas a budget optimized for features is insufficient for images. Consequently, throughout all experiments on the NWPU-45 dataset, we allocate a memory budget of 16 MB for feature-based replay and 300 MB for image-based replay, respectively. To maintain experimental consistency, the memory constraints for all other datasets are calibrated based on these benchmarks, with the number of preserved exemplars adjusted accordingly.

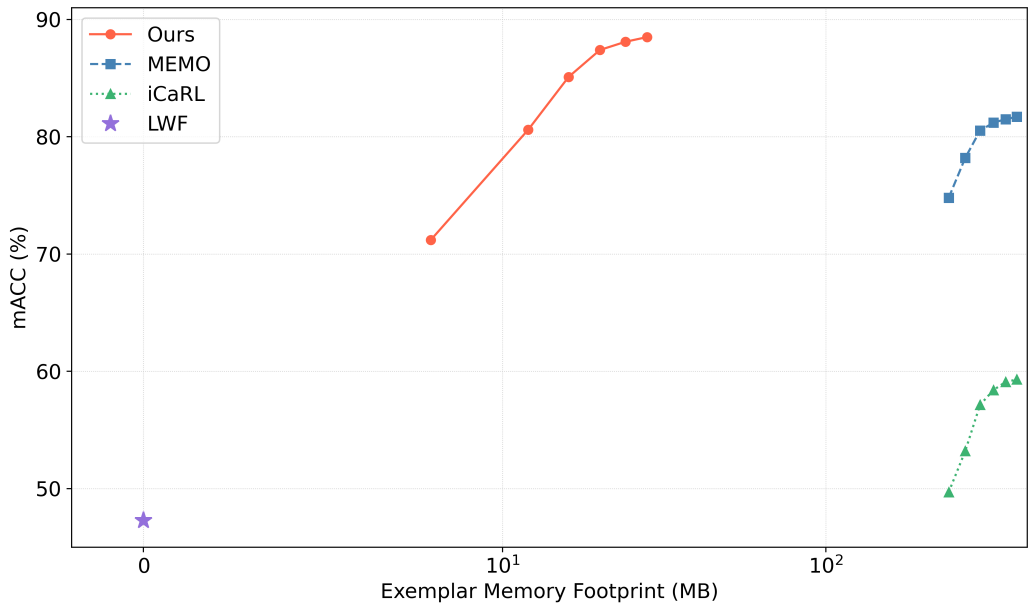


Figure 5. Exemplar memory footprint versus mean average accuracy on the nine-task NWPU-45 dataset for different replay-based methods with distinct storage mechanisms. The memory budget is varied by adjusting the number of preserved feature descriptors for our method and the number of stored images for the compared methods.The horizontal axis is plotted on a logarithmic scale.

4.5. Ablation Experiments

To validate the effectiveness of the proposed FR-CIL, we conduct a comprehensive ablation study to systematically evaluate the specific contribution of each key component: the PMFE module, the FCN, the DSKR mechanism and the BR strategy.

4.5.1. Effect of PMFE

We conducted ablation experiments on the PMFE module across three datasets, with the results reported in Table 2. Compared to the variant without PMFE (w/o PMFE), our method achieves a consistent increase in mACC alongside a reduction in BWT values. These findings indicate that PMFE effectively improves incremental learning performance by promoting fine-grained and interactive feature enhancement.

Table 2. Ablation results of the PMFE module on three datasets, where the upward arrow indicates that a larger value corresponds to better performance, and the downward arrow indicates the opposite.

Method \ Dataset	AID 6 Tasks with 5 Classes		RSI-CB256 7 Tasks with 5 Classes		NWPU-45 9 Tasks with 5 Classes	
	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓
w/o PMFE	86.74	13.31	88.08	10.47	76.78	23.19
FR-CIL	88.96	11.08	91.07	7.96	80.63	16.07

To further validate the effectiveness of PMFE, we perform a sequential learning analysis on the six-task PatternNet dataset. PatternNet encompasses a broad spectrum of spatial resolutions (0.06–4.69 m) and is characterized by high object occupancy and rich visual diversity. These attributes make it particularly suitable for evaluating multiscale feature learning in remote sensing scenarios.

As illustrated in Figure 6a, the model incorporating PMFE maintains consistently higher mACC values throughout the incremental learning process. Furthermore, Figure 6b

details the final accuracy achieved for each task after completing all incremental learning phases. The proposed method demonstrates significant performance gains across all six tasks, yielding improvements of 4.04%, 2.83%, 1.66%, 4.02%, 1.99%, and 1.14%, respectively. These performance enhancements are attributed to the multiscale and fine-grained feature construction process, which enhances the model's capacity to learn discriminative representations for RSIs. The performance gains on all tasks provide strong empirical evidence for the effectiveness of PMFE.

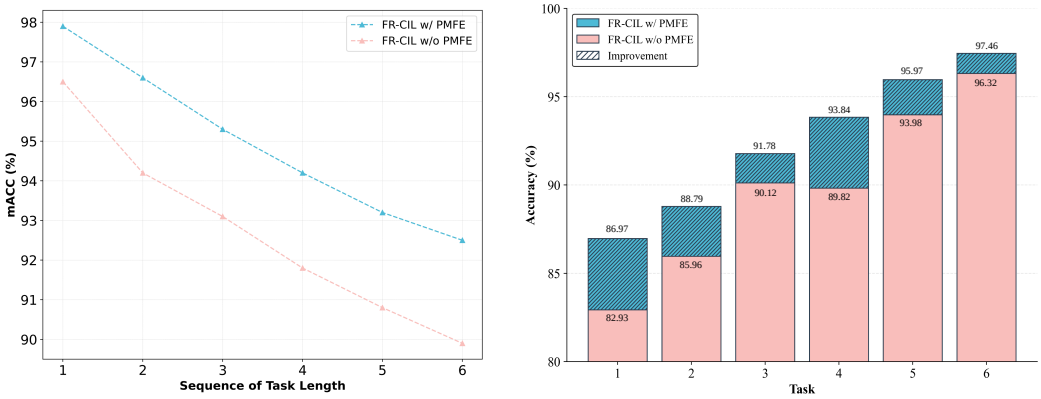


Figure 6. Ablation study of the PMFE module on the six-task PatternNet dataset. Sequential learning analysis is conducted. (**a**,**b**) report the mACC and accuracy comparisons, respectively, with and without the PMFE module.

4.5.2. Effect of FCN

To evaluate the effect of FCN, we assess the quality of feature calibration by computing the average similarity between adapted feature vectors and their corresponding ground-truth representations. Specifically, for a given image x, the ground-truth vector is the feature representation extracted by the current model using the original image. This vector is compared against the adapted feature vector of x, which evolves across incremental stages. The quality of feature calibration is quantified using cosine similarity, serving as a metric to determine how accurately the model approximates the true feature distribution.

Figure 7 illustrates the cosine similarity of the five classes from the initial task, evaluated after completing all learning phases on the AID, RSI-CB256, and NWPU-45 datasets. The results demonstrate that the proposed method consistently maintains higher average cosine similarity for old classes relative to the variant without FCN. This improvement is particularly pronounced on the NWPU-45 dataset, where our method attains average cosine similarities ranging from 80.39% to 85.04%. In contrast, the variant without FCN yields significantly lower values, falling within the range of 51.29% to 60.35%.

These results indicate that the proposed FCN efficiently adapts the previously stored feature descriptors to the updated feature space to reconcile the distributional gap. Consequently, the adapted features retained in the memory bank maintain high semantic fidelity to the original class distributions, thereby facilitating robust incremental learning.

To further validate if its contribution is justified, we employ t-distributed stochastic neighbor embedding (t-SNE) for feature visualization. This qualitative analysis complements our quantitative findings by providing intuitive insights into the model's internal representation learning. Specifically, we utilize t-SNE to project the high-dimensional feature embeddings from the final layer of FR-CIL into a two-dimensional latent space.

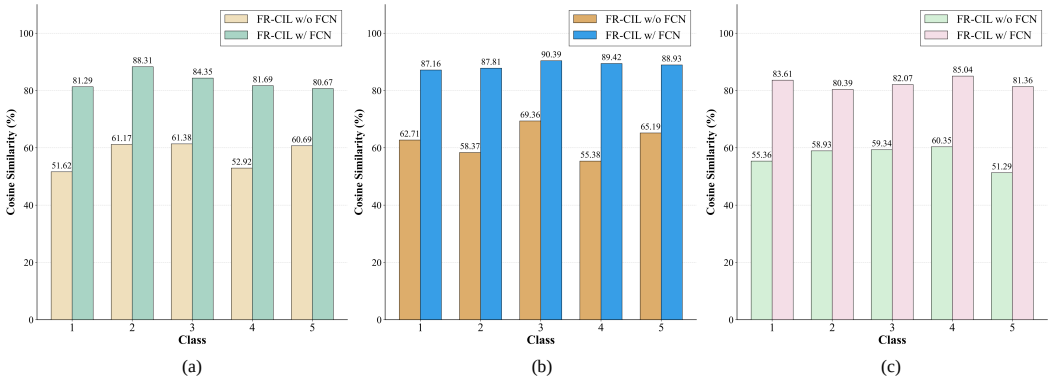


Figure 7. Cosine similarity of the five classes from the initial task evaluated after completing all incremental learning phases on the six-task AID dataset (**a**), seven-task RSI-CB256 dataset (**b**), and nine-task NWPU-45 dataset (**c**). Comparisons are conducted between adapted feature vectors and their corresponding ground-truth representations.

Figure 8 presents the t-SNE visualization of feature embeddings on the UC-Merced dataset. The feature distribution in Figure 8a appears less organized and diffuse, characterized by ambiguous class boundaries. In contrast, Figure 8b demonstrates that the proposed FR-CIL generates a highly structured and regularized embedding space. The visualization reveals significantly enhanced inter-class separability alongside a high degree of intra-class compactness. These geometric properties ensure that the learned features will not lose discrimination as feature space evolves. Consequently, the proposed method effectively enhances feature stability, allowing the model to maintain robust classification boundaries even as the data distribution evolves.

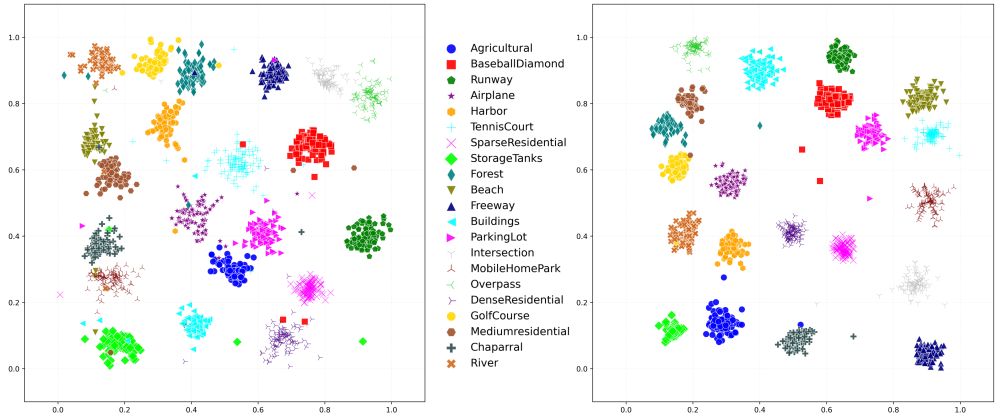


Figure 8. Ablation analysis of the FCN on the UC-Merced dataset. t-SNE is used to visualize feature embeddings of 21 classes, where each point corresponds to a sample and different colors and marker shapes indicate distinct classes. The legend, consisting of class-specific markers, is shown at the center. (**a**) Our method without the FCN. (**b**) Our method with the FCN.

4.5.3. Effect of DSKR

During the incremental learning stage, the proposed method incorporates two complementary knowledge distillation losses within the DSKR mechanism to ensure robust knowledge transfer. Specifically, $\mathcal{L}_{KD}$ facilitates distillation within the output probability space, whereas $\mathcal{L}_{FD}$ operates directly in the feature representation space. To verify the effectiveness of this dual-constraint mechanism, we performed ablation experiments across three datasets. The detailed results are presented in Table 3.

Table 3. Ablation results of the DSKR mechanism on three datasets, ✓ indicating that the corresponding loss is included, where the upward arrow indicates that a larger value corresponds to better performance, and the downward arrow indicates the opposite.

$\mathcal{L}_{total}$			AID (6 Tasks)		RSI-CB256 (7 Tasks)		NWPU-45 (9 Tasks)	
$\mathcal{L}_{CE}$	$\mathcal{L}_{KD}$	$\mathcal{L}_{FD}$	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓
✓			86.29	15.06	87.95	12.24	79.22	22.93
✓	✓		87.83	12.98	89.91	8.92	79.81	18.62
✓		✓	87.57	13.15	89.49	9.37	79.38	19.04
✓	✓	✓	88.96	11.08	91.07	7.96	80.63	16.07

Compared to models optimized solely with classification loss, the incorporation of either distillation loss yields notable improvements in mACC while effectively reducing BWT. These findings underscore the effectiveness of distillation mechanisms in preserving knowledge via both output semantic probabilities and latent feature representations. In particular, $\mathcal{L}_{KD}$ consistently achieves slightly superior performance compared to $\mathcal{L}_{FD}$, suggesting that constraints imposed at the semantic output space play a more critical role in retaining class discrimination. Notably, when both distillation losses are jointly employed, mACC improves by 2.67%, 3.12%, and 1.41%, while BWT decreases by 3.98%, 4.28%, and 6.86% across the three datasets, respectively. These gains surpass those obtained using either distillation strategy alone, demonstrating the complementary and synergistic benefits of the proposed dual-space distillation mechanism.

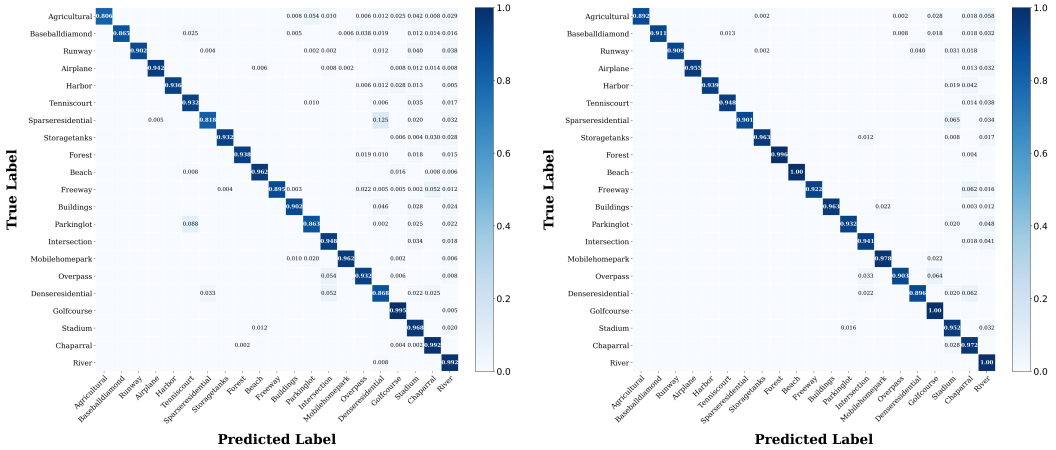
4.5.4. Effect of BR

We assess the effectiveness of the proposed BR strategy on the performance of classifier across three datasets. As detailed in Table 4, the incorporation of the BR strategy yields additional improvement in ACC and BWT. Notably, the impact of the BR strategy is particularly pronounced on the AID dataset compared to the other datasets. This distinction can be attributed to the class imbalance inherent in AID, characterized by a highly uneven distribution of samples across different scene classes. Specifically, such imbalance typically causes model parameters to bias towards classes with a higher quantity of samples during incremental learning. The BR strategy effectively mitigates this bias by optimizing the classifier on the balanced exemplar set, thereby enhancing overall accuracy and alleviating catastrophic forgetting for old classes.

Table 4. Ablation results of the BR strategy on three datasets, where the upward arrow indicates that a larger value corresponds to better performance, and the downward arrow indicates the opposite.

Method \ Dataset	AID 6 Tasks with 5 Classes		RSI-CB256 7 Tasks with 5 Classes		NWPU-45 9 Tasks with 5 Classes	
	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓
w/o BR	86.51	14.03	89.45	10.98	78.68	19.87
FR-CIL	88.96	11.08	91.07	7.96	80.63	16.07

The effect of the BR strategy is further illustrated through confusion matrix analysis on the UC-Merced dataset. Compared to Figure 9a,b exhibits significantly fewer off-diagonal elements and a more uniform diagonal distribution. Notably, the diagonal elements corresponding to old classes exhibit higher values in Figure 9b. This observation confirms that incorporating the BR strategy effectively mitigates classifier bias, ensuring that the model retains high accuracy for previously learned classes while simultaneously learning new ones.



(a) (b)

Figure 9. Comparative analysis of confusion matrices on the UC-Merced dataset. (**a**) Our method without the BR strategy. (**b**) Our method with the BR strategy.

4.5.5. Synergistic Effect Among PMFE, FCN, and BR

To further investigate the interaction among the three proposed components, we conduct combination ablation experiments by incrementally integrating PMFE module, FCN, and BR strategy.

As detailed in Table 5, the results reveal strong synergistic interactions among the three proposed components beyond their individual contributions. Notably, there is a pronounced synergy between PMFE and FCN. The PMFE module functions as a stabilizing foundation by enhancing discriminative multi-scale features. When the FCN subsequently performs transductive mapping to calibrate historical embeddings into the current feature space, it directly benefits from this enhanced representation. As PMFE inherently increases intra-class compactness and inter-class separability, the calibration process within FCN becomes more stable and significantly less susceptible to distortion. Furthermore, the BR strategy specifically targets the classifier to rectify biased decision boundaries. Its effectiveness becomes more pronounced when the feature space is both discriminative and stabilized. When FCN and PMFE are jointly applied, the classifier processes embeddings with reduced drift and improved separability, thereby enabling the BR strategy to more effectively mitigate classification bias.

Table 5. Ablation results of PMFE, FCN, and BR on three datasets, ✓ indicating that the corresponding loss is included, where the upward arrow indicates that a larger value corresponds to better performance, and the downward arrow indicates the opposite.

Baseline	PMFE	FCN	BR	AID (6 Tasks)		RSI-CB256 (7 Tasks)		NWPU-45 (9 Tasks)	
				mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓	mACC (%) ↑	BWT (%) ↓
✓				62.17	40.82	60.38	33.57	58.92	43.84
✓	✓			66.91	34.57	70.25	30.81	65.04	36.87
✓		✓		73.54	25.91	77.17	23.08	75.61	25.87
✓			✓	67.15	31.85	69.67	29.04	66.41	34.17
✓	✓	✓		86.51	14.03	89.45	10.98	78.68	19.87
✓	✓		✓	84.21	17.38	86.79	15.07	73.54	27.95
✓		✓	✓	86.74	13.31	88.08	10.47	76.78	23.19
✓	✓	✓	✓	88.96	11.08	91.07	7.96	80.63	16.07

4.5.6. Effect of λ_1 and λ_2

As formulated in Equation (9), the overall loss function incorporates two hyperparameters, λ_1 and λ_2. These terms serve as weighting coefficients for the respective distillation

losses. Specifically, λ_1 governs the contribution of knowledge distillation within the output probability space, while λ_2 modulates the impact of distillation in the feature representation space.

These hyperparameters are critical for navigating the stability–plasticity dilemma. When λ_1 and λ_2 are set to small values, insufficient regularization leads to pronounced catastrophic forgetting. Conversely, excessively large values enforce overly strong constraints on knowledge retention, leading to stricter loss function, more rigid constraints, and lower level of plasticity.

To evaluate the sensitivity of the proposed framework to hyperparameters λ_1 and λ_2, we conducted ablation experiments on the three-task UC-Merced dataset. Preliminary empirical observations indicated that the model achieves a favorable stability-plasticity trade-off within the ranges of $\lambda_1 \in [1.4, 2.0]$ and $\lambda_2 \in [0.5, 1.1]$. Consequently, we performed a fine-grained grid search within these ranges, incrementing the values by 0.1 at each step to precisely identify the optimal configuration.

As illustrated in Figure 10, our method achieves optimal performance on the UC-Merced dataset when λ_1 is set to 1.8 and λ_2 to 0.8. Therefore, these hyperparameter settings are employed as the default hyperparameters throughout the experiments.

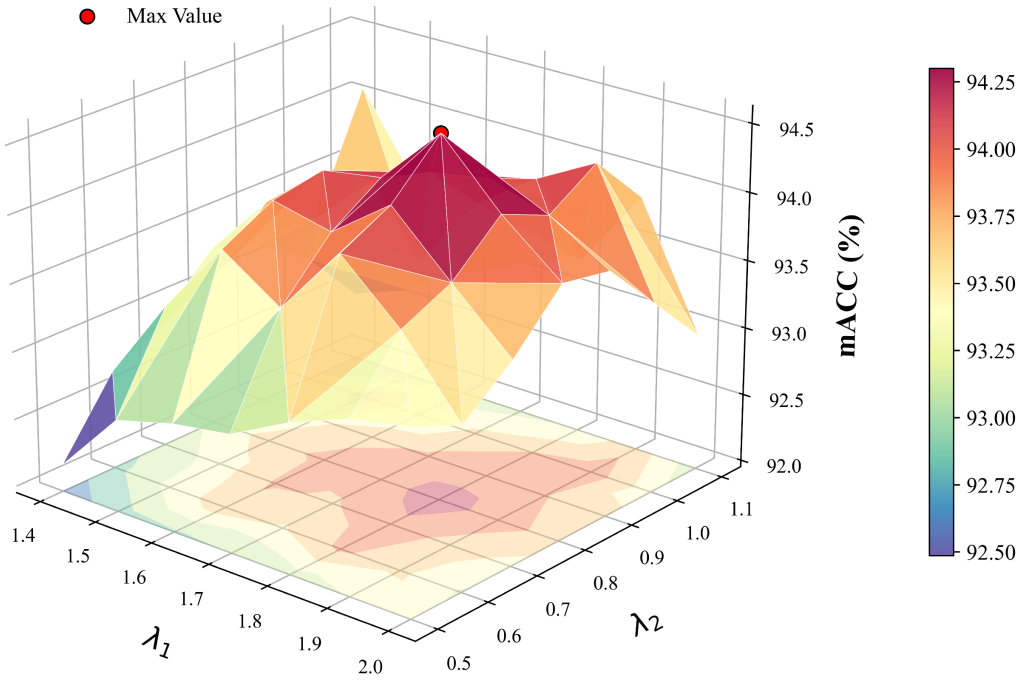


Figure 10. Sensitivity analysis of the mean average accuracy with respect to hyperparameters λ_1 and λ_2 on the three-task UC-Merced dataset.

4.5.7. Effect of α

As formulated in Equation (13), the hyperparameter α modulates the weight of the feature alignment loss, thereby regulating the degree of emphasis placed on feature similarity during the training of the feature calibration network. To evaluate the model's sensitivity to α, we conducted ablation studies on the UC-Merced dataset.

As presented in Table 6, the mACC demonstrates robustness across a relatively broad range of $\alpha \in [1, 5]$. Notably, the optimal performance is achieved when $\alpha = 3$. Consequently, we adopt this value for all experiments reported in our work.

Table 6. mACC versus the hyperparameter values of α on the three-task UC-Merced dataset

Hyperparameter α	0	0.5	1	2	3	4	5	7	10
mACC (%)	78.61	89.31	93.45	94.57	94.63	94.52	93.84	91.12	88.93

5. Discussion

5.1. Dataset-Specific Sensitivity Analysis

Through comprehensive comparisons with several state-of-the-art methods, extensive experiments conducted on multiple public datasets demonstrate both the effectiveness and robustness of the proposed method, as well as the contribution of each key component, indicating its strong potential for real-world RSSC applications. To supplement the general performance overview across these datasets, we will provide an additional discussion on the dataset-specific behavior of different methods. In particular, we identify which methods maintain stable performance across diverse datasets and which exhibit notable performance fluctuations under certain data distributions. Such performance variations are worthy of careful investigation, as they often reflect the intrinsic compatibility between a given incremental learning strategy and the underlying characteristics of the data.

As illustrated in Table 1, the performance variations observed across the AID, RSI-CB256, and NWPU-45 datasets highlight the impact of inherent dataset characteristics, including intra-class variability, inter-class similarity, class distribution balance, and the granularity of task partitioning.

For the AID dataset, the primary challenges arise from significant fluctuations in spatial resolution (ranging from 0.5 m to 8 m) and class imbalance (220–420 samples per class). These factors adversely affect prior-focused regularization methods, such as GEM and EWC. Under conditions of high intra-class variability, the gradient-based constraints employed by these methods become unreliable, leading to substantial performance degradation. In contrast, other methods demonstrate robust resistance to catastrophic forgetting under these conditions. The RSI-CB256 dataset features a balanced class distribution that mitigates the adverse effects of class imbalance, thereby enhancing the efficacy of data-replay strategies. Consequently, iCaRL achieves a slightly higher mACC on RSI-CB256 than on the AID dataset (64.02% vs. 61.17%), and the performance gap between iCaRL and WA narrows significantly. However, as the sequence length and number of classes increase, the limitations of pure regularization methods (GEM, EWC, and LwF) become pronounced, leading to sustained performance degradation. The NWPU-45 dataset poses the most rigorous challenge due to its high scene complexity and extended incremental sequence. The larger number of incremental stages leads to cumulative task interference, thereby exacerbating the stability–plasticity dilemma. Notably, methods such as iCaRL and WA, which perform well on the AID and RSI-CB256 datasets, exhibit a significant performance drop on NWPU-45 dataset, suggesting that strategies heavily reliant on exemplar replay lose effectiveness as task interference intensifies. Architecture-based methods, including DER, MEMO, and EASE, consistently rank among the top-performing methods and achieve competitive results, indicating that maintaining additional network branches or adapters can be highly effective, despite additional storage overhead. Our method maintains consistently high and stable mACC values throughout incremental tasks without relying on additional storage, indicating that the proposed feature replay strategy effectively preserves discriminative representations while mitigating catastrophic forgetting.

This discussion facilitates a deeper understanding of the underlying causes of the observed performance fluctuations and offers guidance for the development of more robust and generalizable RSSC-oriented CIL methods.

5.2. Real-World Application Scenarios Analysis

To further clarify the practical value of the proposed method, we highlight its applicability in real-world remote sensing deployment scenarios, particularly in resource-constrained platforms, such as satellites and unmanned aerial vehicles (UAVs). In such platforms, storage capacity and computational resources are strictly limited, and continuous data transmission to ground stations for full retraining is impractical due to bandwidth constraints and latency considerations. As a result, models deployed in such environments must support incremental updates while operating under strict memory constraints.

Long-term earth observation missions necessitate adaptive models update to accommodate newly emerging scene classes, such as evolving urban structures, disaster-affected regions, and newly monitored geographic zones. However, retaining large volumes of historical raw imagery for rehearsal is frequently infeasible due to limited on-board storage and strict regulatory restrictions on data retention. Under these constraints, memory efficiency becomes a decisive factor for operational deployment rather than a purely theoretical advantage.

The proposed FR-CIL framework is explicitly designed to accommodate such deployment constraints. By storing compact feature embeddings instead of raw images, the framework achieves substantial memory reduction while preserving knowledge of previously learned classes. This enables incremental adaptation without periodic retraining or permanently storing historical imagery. Consequently, FR-CIL offers a robust and practical solution for resource-constrained remote sensing platforms, empowering them with long-term, adaptive scene understanding capabilities.

5.3. Computational Efficiency Analysis

The proposed memory-efficient design is not achieved at the expense of substantial computational overhead. Both the FCN and the PMFE module incur only a little computational cost during training and negligible impact during inference.

The FCN operates as an orthogonal transformation applied to 512-dimensional feature embeddings rather than high-resolution images. As a result, its computational cost is minimal compared with the backbone network, which dominates overall training complexity. The calibration process therefore introduces only marginal overhead in training, and at inference time the FCN performs a simple forward transformation, leading to negligible additional latency. The PMFE module enhances multiscale representation within the feature extraction stage using depthwise and dilated convolutions. These operations are computationally efficient and widely adopted in lightweight CNN architectures. In particular, depthwise convolutions significantly reduce parameter count and FLOPs compared with standard convolutions. Consequently, the overhead introduced by PMFE remains small relative to the total computational cost of the backbone. Importantly, conventional image-replay methods require repeated forward and backward propagation of stored raw images through the entire backbone during each incremental stage. In contrast, the proposed feature-replay strategy directly reuses precomputed feature embeddings, thereby eliminating redundant backbone processing of historical data. Consequently, this computational saving effectively offsets the additional training time introduced in each incremental task, leading to comparable overall training time in practice.

Overall, the proposed method maintains comparable training time and inference latency while significantly reducing memory consumption, ensuring that its storage efficiency does not come at the cost of substantial computational burden.

6. Conclusions

In this article, we introduce a memory-efficient CIL framework for RSSC that stores compact feature embeddings, rather than raw images, as exemplars for previously learned classes. This strategy mitigates privacy concerns, representation drift, and classifier bias, thereby preserving decision boundaries and alleviating catastrophic forgetting. The proposed framework comprises four key components: the PMFE module, the DSKR mechanism, the FCN, and the BR strategy. The PMFE module enables fine-grained and interactive feature enhancement through a progressive construction scheme, yielding richer representations and a more comprehensive understanding of remote sensing scenes. The DSKR mechanism integrates two complementary distillation losses to jointly preserve decision boundaries and feature representation stability. A specialized FCN is trained in a transductive learning paradigm with manifold consistency regularization to adapt stored feature descriptors to the updated feature space, thereby bridging the distributional gap for a unified classifier. Finally, the BR strategy mitigates prediction bias by exclusively optimizing the classifier on a balanced exemplar set. The framework proceeds in three sequential stages. It first learns discriminative representations via joint optimization with the DSKR mechanism, then calibrates old features into the current feature space via FCN, and finally mitigates classifier bias by optimizing the classifier on a balanced exemplar set.

Extensive experiments on five public datasets demonstrate the superior performance and robustness of the proposed framework. Comprehensive analyses further validate its effectiveness across diverse remote sensing scenarios. In future work, we plan to extend this framework to few-shot class-incremental learning, where only limited samples of new classes are available while preserving previously acquired knowledge. This setting more closely reflects real-world remote sensing applications.

Author Contributions: Conceptualization, Y.W. (Yunze Wei); Methodology, Y.W. (Yunze Wei) and X.X.; Software, Y.W. (Yunze Wei); Validation, Y.W. (Yunze Wei), B.N., X.X. and Y.W. (Yirong Wu); Formal analysis, Y.W. (Yunze Wei); Investigation, Y.W. (Yunze Wei); Resources, Y.W. (Yunze Wei); Data curation, Y.W. (Yunze Wei), Y.L., B.N. and J.L.; Writing—original draft, Y.W. (Yunze Wei); Writing—review & editing, Y.W. (Yunze Wei), Y.L., B.N. and J.L.; Visualization, Y.W. (Yunze Wei), Y.L., B.N. and X.X.; Supervision, Y.W. (Yunze Wei), B.N., J.L., Y.H. and Y.W. (Yirong Wu); Project administration, Y.W. (Yunze Wei), B.N., Y.H. and Y.W. (Yirong Wu); Funding acquisition, Y.W. (Yunze Wei), Y.L., B.N. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available from their respective sources via the following DOIs: https://doi.org/10.1109/TGRS.2017.2685945; https://doi.org/10.3390/s20061594; https://doi.org/10.1109/JPROC.2017.2675998; https://doi.org/10.1145/1869790.1869829; and https://doi.org/10.1016/j.isprsjprs.2018.01.004, which is accessed on 31 January 2026.

Conflicts of Interest: The authors declare no conflict 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. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
2. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
3. Lv, Z.Y.; Shi, W.; Zhang, X.; Benediktsson, J.A. Landslide Inventory Mapping From Bitemporal High-Resolution Remote Sensing Images Using Change Detection and Multiscale Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1520–1532. [CrossRef]

4. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very High Resolution Multiangle Urban Classification Analysis. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1155–1170. [CrossRef]
5. Pham, H.M.; Yamaguchi, Y.; Bui, T.Q. A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landsc. Urban Plan.* **2011**, *100*, 223–230. [CrossRef]
6. Liu, E.; Zheng, Y.; Pan, B.; Xu, X.; Shi, Z. DCL-Net: Augmenting the Capability of Classification and Localization for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7933–7944. [CrossRef]
7. Yuan, J.; Wang, S. HCFPN: Hierarchical Contextual Feature-Preserved Network for Remote Sensing Scene Classification. *Remote Sens.* **2023**, *15*, 810. [CrossRef]
8. Chen, W.; Ouyang, S.; Tong, W.; Li, X.; Zheng, X.; Wang, L. GCSANet: A Global Context Spatial Attention Deep Learning Network for Remote Sensing Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1150–1162. [CrossRef]
9. Wang, Q.; Liu, Y.; Xiong, Z.; Yuan, Y. Hybrid Feature Aligned Network for Salient Object Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624915. [CrossRef]
10. Xu, C.; Shu, J.; Wang, Z.; Wang, J. A Scene Classification Model Based on Global-Local Features and Attention in Lie Group Space. *Remote Sens.* **2024**, *16*, 2323. [CrossRef]
11. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
12. Aslam, M.A.; Hamza, M.; Zhu, S.; Hu, H.; Xu, W.; Irfan, M.; Zheng, J.; Aslam, S. Continual Learning Inspired by Brain Functionality: A Comprehensive Survey. *Int. J. Intell. Syst.* **2025**, *2025*, 3145236. [CrossRef]
13. Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; Sanner, S. Online continual learning in image classification: An empirical survey. *Neurocomputing* **2022**, *469*, 28–51. [CrossRef]
14. Abraham, W.C.; Robins, A. Memory retention—The synaptic stability versus plasticity dilemma. *Trends Neurosci.* **2005**, *28*, 73–78. [CrossRef]
15. Mermillod, M.; Bugaiska, A.; Bonin, P. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Front. Psychol.* **2013**, *4*, 504. [CrossRef] [PubMed]
16. Lu, X.; Sun, X.; Diao, W.; Feng, Y.; Wang, P.; Fu, K. LIL: Lightweight Incremental Learning Approach Through Feature Transfer for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5611320. [CrossRef]
17. Ammour, N.; Bazi, Y.; Alhichri, H.; Alajlan, N. Continual Learning Approach for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8000905. [CrossRef]
18. Ye, D.; Peng, J.; Li, H.; Bruzzone, L. Better Memorization, Better Recall: A Lifelong Learning Framework for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5626814. [CrossRef]
19. Tang, J.; Xiang, D.; Zhang, F.; Ma, F.; Zhou, Y.; Li, H. Incremental SAR Automatic Target Recognition with Error Correction and High Plasticity. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1327–1339. [CrossRef]
20. Ammour, N. Continual Learning Using Data Regeneration for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8012805. [CrossRef]
21. Pan, Q.; Liao, K.; He, X.; Bu, Z.; Huang, J. A Class-Incremental Learning Method for SAR Images Based on Self-Sustainment Guidance Representation. *Remote Sens.* **2023**, *15*, 2631. [CrossRef]
22. Liu, W.; Nie, X.; Zhang, B.; Sun, X. Incremental Learning with Open-Set Recognition for Remote Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622916. [CrossRef]
23. Fu, Y.; Liu, Z.; Wu, C.; Wu, F.; Liu, M. Class-Incremental Recognition of Objects in Remote Sensing Images with Dynamic Hybrid Exemplar Selection. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, *60*, 3468–3481. [CrossRef]
24. Ye, Z.; Zhang, Y.; Zhang, J.; Li, W.; Bai, L. A Multiscale Incremental Learning Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5606015. [CrossRef]
25. Wei, Y.; Pan, Z.; Wu, Y. Class Bias Correction Matters: A Class-Incremental Learning Framework for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5616518. [CrossRef]
26. Wang, W.; Song, Y.; Wang, J.; Fu, S.; Ren, P.; Qin, H.; Li, W.; Ou, W. Continually Evolved Feature and Classifiers Learning for Long-Tailed Class-Incremental Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5636213. [CrossRef]
27. Wang, L.; Zhang, X.; Su, H.; Zhu, J. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5362–5383. [CrossRef]
28. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5533–5542. [CrossRef]
29. Bang, J.; Kim, H.; Yoo, Y.; Ha, J.W.; Choi, J. Rainbow Memory: Continual Learning with a Memory of Diverse Samples. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Montreal, QC, Canada, 19–25 June 2021; pp. 8214–8223. [CrossRef]

30. Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; Fu, Y. Large Scale Incremental Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 374–382. [CrossRef]
31. Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; Xia, S.T. Maintaining Discrimination and Fairness in Class Incremental Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 13205–13214. [CrossRef]
32. Lee, K.; Lee, K.; Shin, J.; Lee, H. Overcoming Catastrophic Forgetting with Unlabeled Data in the Wild. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 312–321. [CrossRef]
33. Chen, X.; Chang, X. Dynamic Residual Classifier for Class Incremental Learning. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 18697–18706. [CrossRef]
34. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 214–223.
35. Liu, X.; Wu, C.; Menta, M.; Herranz, L.; Raducanu, B.; Bagdanov, A.D.; Jui, S.; van de Weijer, J. Generative Feature Replay For Class-Incremental Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 915–924. [CrossRef]
36. Shi, W.; Ye, M. Prototype Reminiscence and Augmented Asymmetric Knowledge Aggregation for Non-Exemplar Class-Incremental Learning. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 1772–1781. [CrossRef]
37. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
38. Zenke, F.; Poole, B.; Ganguli, S. Continual learning through synaptic intelligence. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 3987–3995.
39. Pan, P.; Swaroop, S.; Immer, A.; Eschenhagen, R.; Turner, R.; Khan, M.E. Continual deep learning by functional regularisation of memorable past. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4453–4464.
40. Lopez-Paz, D.; Ranzato, M. Gradient episodic memory for continual learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6470–6479.
41. Boschini, M.; Bonicelli, L.; Porrello, A.; Bellitto, G.; Pennisi, M.; Palazzo, S.; Spampinato, C.; Calderara, S. Transfer Without Forgetting. In *Proceedings of the Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 692–709.
42. Kang, M.; Park, J.; Han, B. Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 16050–16059. [CrossRef]
43. Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; Gong, Y. Few-Shot Class-Incremental Learning via Relation Knowledge Distillation. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1255–1263. [CrossRef]
44. Cha, H.; Lee, J.; Shin, J. Co2L: Contrastive Continual Learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9496–9505. [CrossRef]
45. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsell, R. Progressive neural networks. *arXiv* **2016**, arXiv:1606.04671. [CrossRef]
46. Yoon, J.; Yang, E.; Lee, J.; Hwang, S.J. Lifelong learning with dynamically expandable networks. *arXiv* **2017**, arXiv:1708.01547.
47. Fini, E.; Da Costa, V.G.T.; Alameda-Pineda, X.; Ricci, E.; Alahari, K.; Mairal, J. Self-Supervised Models are Continual Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 9611–9620. [CrossRef]
48. Zhang, W.; Li, D.; Ma, C.; Zhai, G.; Yang, X.; Ma, K. Continual Learning for Blind Image Quality Assessment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2864–2878. [CrossRef]
49. 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.* **2023**, *45*, 2945–2951. [CrossRef]
50. Chen, P.; Zhang, Y.; Li, Z.; Sun, L. Few-Shot Incremental Learning for Label-to-Image Translation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 3687–3697. [CrossRef]
51. Liu, L.; Zheng, T.; Lin, Y.; Ni, K.; Fang, L. INS-Conv: Incremental Sparse Convolution for Online 3D Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 18953–18962. [CrossRef]
52. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the ICLR, San Juan, PR, USA, 2–4 May 2016.

53. Yu, F.; Koltun, V.; Funkhouser, T. Dilated Residual Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 636–644. [CrossRef]
54. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
55. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]
56. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
57. Li, H.; Dou, X.; Tao, C.; Wu, Z.; Chen, J.; Peng, J.; Deng, M.; Zhao, L. RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data. *Sensors* **2020**, *20*, 1594. [CrossRef]
58. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
59. 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. [CrossRef]
60. Li, Z.; Hoiem, D. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2935–2947. [CrossRef]
61. Shi, Y.; Zhou, K.; Liang, J.; Jiang, Z.; Feng, J.; Torr, P.; Bai, S.; Tan, V.Y. Mimicking the Oracle: An Initial Phase Decorrelation Approach for Class Incremental Learning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16701–16710. [CrossRef]
62. Yan, S.; Xie, J.; He, X. DER: Dynamically Expandable Representation for Class Incremental Learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 3013–3022. [CrossRef]
63. Zhou, D.W.; Wang, Q.W.; Ye, H.J.; Zhan, D.C. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv* **2022**, arXiv:2205.13218.
64. Zhou, D.W.; Sun, H.L.; Ye, H.J.; Zhan, D.C. Expandable Subspace Ensemble for Pre-Trained Model-Based Class-Incremental Learning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 23554–23564. [CrossRef]

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.