

Better Memorization, Better Recall: A Lifelong Learning Framework for Remote Sensing Image Scene Classification

Dingqi Ye, Jian Peng, Haifeng Li, *Member, IEEE*, and Lorenzo Bruzzone, *Fellow, IEEE*

Abstract—To infer unknown remote sensing scenarios, most existing technologies use a supervised learning paradigm to train deep neural network (DNN) models on closed datasets. This paradigm faces challenges such as highly spatiotemporal variants and ever-changing scale-heterogeneous remote sensing scenarios. Additionally, DNN models cannot scale to new scenarios. Lifelong learning is an effective solution to these problems. Current lifelong learning approaches focus on overcoming the *catastrophic forgetting* issue (i.e., a successive increase in heterogeneous remote sensing scenes causes models to forget historical scenes) while ignoring the *knowledge recall* issue (i.e., how to facilitate the learning of new scenes by recalling historical experiences), which is a significant problem. This paper proposes a lifelong learning framework called asymmetric collaborative network (SCN) for lifelong remote sensing image classification. This framework consists of two structurally distinct networks: a preserving network (Pres-Net) and a transient network (Trans-Net), which imitates the long- and short-term memory processes in the brain, respectively. Moreover, this framework is based on two synergistic knowledge transfer mechanisms: triple distillation and prior feature fusion. The triple distillation mechanism enables knowledge persistence from Trans-Net to Pres-Net to achieve better memorization; the prior feature fusion mechanism enables knowledge transfer from Pres-Net to Trans-Net to achieve better recall. Experiments on three open datasets demonstrate the effectiveness of SCN for 3-, 6-, and 9-task-length learning. The idea of asymmetric separation networks and the synergistic strategy proposed in this paper are expected to provide new solutions to the translatability of the classification of remote sensing images in real world scenarios.

Index Terms—Remote sensing image scene classification, lifelong learning, catastrophic forgetting, knowledge recall, asymmetric collaborative network

I. INTRODUCTION

WITH the support of deep neural network (DNN) technology, remote sensing image scene classification has made remarkable progress in recent years [1–5]. In the most widely used paradigms, DNNs train on closed datasets by

supervised learning methods and the models are applied to future scenarios. Different issues may result from this process, including (1) *Data openness*, the continual and rapid increase and accumulation of remote sensing images (RSIs); (2) *Temporal openness*, the periodic or aperiodic changes of ground feature images in remote sensing images over time; (3) *Spatial openness*, the tendency of the same features to show completely different visual features in different places; and (4) *Spectral openness*, the variation in spectral measurements due to the phenological change in ground features. Traditional approaches address the above issue either by retraining models on a mixture of historical data and new data or fine-tuning models on new data. The former approach requires considerable computational cost and storage resources and cannot scale to edge computing devices [6]. The latter approach weakens the model's ability to recognize historical categories, which is known as the *catastrophic forgetting* issue [7–9], i.e., the model's performance on historical tasks sharply degrades when learning new tasks [10, 11].

Different from the traditional paradigm, lifelong learning technology incrementally learns new data while avoiding these problems. Existing approaches to lifelong learning can be organized into three categories: (1) Regularization-based approaches, which are based on design rules that constrain model parameters to reduce the impact of parameter coverage on historical tasks [12–15]; (2) Dynamic-architecture-based approaches, which are represented by progress neural networks (PNNs) [16], generate extra parameters for new tasks to reduce the impact of interference between parameters [16–19]; and (3) Memory-replay-based approaches, which construct a model memory system to alleviate catastrophic forgetting by simulating the interaction between the mammalian hippocampus and neocortex [20], are inspired by the complementary learning system (CLS) theory [21] in neuroscience.

Recently, remote sensing image classification based on lifelong learning has received growing attention and some promising results have been obtained [22–25]. However, the majority of current studies focus on overcoming the issue of catastrophic forgetting in neural networks, i.e. impaired performance on old tasks when learning new knowledge. However, these studies ignore the critical problem of *knowledge recall*. Indeed, it is still challenging to make full use of memorized historical knowledge to facilitate learning new knowledge due to two reasons: 1) *The stability-plasticity dilemma* [26], i.e., better memorization of historical knowledge facilitates learning new knowledge, where better learning of new knowledge

Portions of this work were presented at the IEEE International Geoscience and Remote Sensing Symposium IGARSS, in 2022, Asymmetric Collaborate Network: Transferable Lifelong Learning for Remote Sensing Images.

This work was supported by the National Natural Science Foundation of China under Grants 41871364 and 41871302 and a scholarship from the China Scholarship Council under Grant 201703170123. (Corresponding author: Jian Peng.)

Dingqi Ye, Jian Peng, and Haifeng Li are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China. (e-mail: 215011047@csu.edu.cn, pengj2017@csu.edu.cn, lihaifeng@csu.edu.cn).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

degrades memory. 2) *Knowledge entanglement*, i.e., the general knowledge shared among tasks and the specific knowledge related to individual task are entangled, which impairs the reuse of historical knowledge and specific task learning. In particular, the classification of RSIs in specific tasks is driven by the scale characteristic and channel properties. Therefore, it is important to focus more on the general characteristics for scaling to various tasks.

This research proposes a lifelong learning framework, asymmetric collaborative network (SCN), which is characterized by two core properties, independence and asymmetry. These properties aim to alleviate the stability-plasticity dilemma and the knowledge entanglement problem, respectively. Specifically, SCN exploits two separate networks and regularization strategies. For the network architecture, Pres-Net and Trans-Net encode historical knowledge and new knowledge, respectively, to avoid interference that can result in catastrophic forgetting between different types of knowledge. Moreover, the use of triple distillation loss for regularization allows us to capture the semantic similarity between historical and new knowledge in a feature space [27], ensuring better memorization of distilled knowledge (better memorization). By performing feature fusion prior to embedding old features and new features into the same low-dimensional feature space to learn their transferability across both networks, valuable knowledge from Pre-Net to complement the learning of Trans-Net is recalled (better recall). SCN also introduces asymmetric architectures. Pres-Net uses a simple network to facilitate the memorization of general features for the long-term retention of old tasks. Meanwhile, Trans-Net uses a more complex task-specific network to better learn the specific features for new tasks. In particular, the multiscale module Res2Block [28] and the multichannel attention module ECA [29] are added to Trans-Net to acquire more representative features.

In summary, the major contributions of our work are outlined as follows:

- 1) A new lifelong learning framework for remote sensing image classification is proposed to better overcome catastrophic forgetting while facilitating the learning of new unknown categories.
- 2) A new pair of asymmetric collaborative networks (SCNs) are proposed to simulate the abstraction mechanism in human long-term memory and the fine learning mechanism in short-term memory by asymmetric and independent network design. The synergy between the two networks has been achieved via a triple distillation mechanism and prior feature fusion mechanism.
- 3) Experiments on three open datasets show the effectiveness of the proposed method, which is promising for the classification of remote sensing images in real world scenarios.

II. RELATED WORK

In this section, we first introduce the current primary research in lifelong learning, which mainly addresses the problem of catastrophic forgetting. Then, we review research on lifelong learning oriented to remote sensing images.

A. Overcoming Catastrophic Forgetting in Lifelong Learning

Research on lifelong learning has been exploded in recent years. Related studies primarily address the problem of catastrophic forgetting [7]. Based on strategies for overcoming forgetting, these studies can be categorized into three types of approaches: regularization, dynamic structure, and memory replay.

Regularization. The core idea of regularization is to design rules for constrained parameters to reduce the impact of parameter coverage on historical tasks. A representative method includes the LwF model proposed by Li and Hoiem et al. [12], which distills knowledge from an old network into a new network by the similarity of outputs. Through the idea of knowledge distillation, this model is able to retain the features of historical tasks to a certain extent [30–32] while reducing its performance due to additional distillation loss [33]. Another idea is to constrain plasticity by calculating the importance of synapses to alleviate the weight coverage of old tasks caused by learning new tasks, which is representative of the elastic weight consolidation (EWC) [13] and intelligent synapse (IS) methods [14]. However, the computational complexity based on the Fisher matrix is too high. Pan et al. [34] proposed using a Gaussian process formula to obtain a higher efficiency of the regularization process. Wang et al. [29] proposed Adam-NSCL to update the candidate parameters to the null space of all previous tasks through two constraints on stability and plasticity to achieve continual learning. Another idea to mitigate the task-to-task interference by the model structure is in use. Adaptive RPS-Net [35] encourages the sharing of parameters between tasks through random path selection and achieves the purpose of unlocking the mutual influence between them. Although regularization-based methods are effective in mitigating catastrophic forgetting to a certain extent, they require a trade-off between old and new tasks due to the inclusion of an additional loss term [36]. In addition, these methods are still prone to forgetting in long-term-to-lifelong learning [37].

Dynamic architecture. This approach preserves task-specific parameters or modules to avoid knowledge interference between old and new tasks. One of the most representative methods is the progressive neural network (PNN) [16], in which each task is associated with a corresponding network and adapted by lateral links. PNN effectively alleviates the catastrophic forgetting issue since it completely avoids interference of task-specific parameters by separate subnetworks. However, its computational complexity grows exponentially with an increase in the number of tasks, making it difficult to deploy. Some other methods considered the problem of the dynamic construction of the structure. Xiao et al. [38] proposed a model that grows in a self-organized manner based on similarity, which enables it to learn faster while inheriting existing model features. However, this model can only grow at the top layer, which impairs training efficiency. Yoon et al. [18] proposed a dynamic expansion network (DEN), which expands the network with group sparse regularization. Similar studies in [39, 40] were also designed to break the upper bound of incremental learning heavy model capacity by a dynamic

structure.

Memory replay. This approach considers that the memory of historical tasks is motivated by the primary difficulty in incremental learning, in which various factors limit access to historical data. The representative approach is DGR proposed by Shin et al. [41], which uses a generative network to generate pseudo-historical images for old tasks. Another solution is to store a subset of the representative samples from the previous task to reduce the data storage. Representative methods are the GEM algorithm proposed by Lopez-Paz et al. [42] and its variants [43, 44]. Although these methods avoid considerable storage of historical images, they fail to address hard convergence, long training sessions, and high hardware requirements issues. Moreover, in long-term learning, this type of replay network still faces the problem of forgetting, which will further lower the quality of replay and affect the final results. Thus, inspired by complementary learning system (CLS) theory, some works constructed a memory system to mitigate catastrophic forgetting by simulating the interaction between the mammalian hippocampus and neocortex. This approach was first validated in experiments by Hinton et al. [45], where a set of plasticity weights were assigned to memory and a set of rapid change weights were assigned to new knowledge. Furthermore, some recent research complemented this idea by generating data distribution [46], pseudo-feature [47], or learning of model parameters [48, 49].

B. Lifelong Learning in Remote Sensing

Some lifelong learning works for RSIs have been proposed in recent years. In terms of regularization-based strategies, Ammour [25] improved the effect on the old land cover classification task by maximizing the distance among tasks. Bhat et al. [24] utilized curriculum learning to design a learning order (i.e., from easy to complex) by computing the similarity between new and old classes to obtain better local optima and avoid forgetting. These methods can alleviate catastrophic forgetting to a certain extent. However, it is challenging to ensure the effectiveness of constraints in long sequence tasks, which would lead to considerable forgetting. In terms of dynamic structure, Lu et al. [23] suggested using fewer parameters to overcome forgetting, thus avoiding parameter explosion. With respect to memory replay-based methods, Ammour [25] proposed learning how to reproduce data from previous tasks by analyzing the potential data structure of new tasks to alleviate forgetting. Such a proposal still fails to avoid the problems of difficult training and slow convergence.

In summary, despite some progress achieved in remote sensing-oriented lifelong learning, most of the proposed methods only focus on overcoming the catastrophic forgetting issue. To the best of our knowledge, none of the previous research in RSIs considers the problem knowledge recall issue, limiting the performance of current lifelong learning algorithms for real world scenarios.

III. METHODOLOGY

A. Overview

To avoid interference between historical knowledge and new knowledge, we design a novel lifelong learning framework

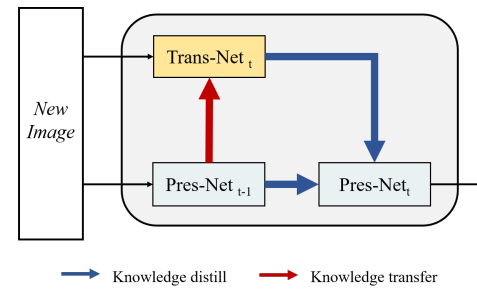


Fig. 1. Incremental learning procedure of the SCN framework. The orange rectangle indicates the Trans-Net learning of task t , and the first cyan rectangle indicates Pres-Net after learning task $t-1$. The learning is divided into two steps. The first step is shown by the red line: the new image is input to Trans-Net and Pres-Net. Trans-Net is learned, while Pres-Net is frozen, and the output features are migrated to Trans-Net to help its learning. The second step is shown by the blue lines, Trans-Net t and Pres-Net $t-1$ are integrated into a new Pres-Net t by knowledge distillation, after which Trans-Net will be released. At this point, the whole incremental learning process of the t -th task ends.

called asymmetric collaborative networks (SCNs), which mimics the human brain's segregate mechanism of long- and short-term memory. As shown in Fig. 1, SCN is an independent-asymmetric architecture with two subnetworks, the preserving network (Pres-Net) and the transient network (Trans-Net), which encode historical knowledge related to old tasks and current knowledge-related new tasks, respectively. Furthermore, Pres-Net acts as a task-general network ResNet [50] to preserve abstract knowledge. In contrast, Trans-Net, equipped with a multiscale feature extraction module and a multichannel attention feature module, acts as a task-specific network to learn fine-grained tasks. This asymmetric-independent architecture ensures the decoupling of accumulated general knowledge (long-term memory) and task-specific knowledge (short-term) while avoiding the fact that the learning of new tasks corrupts the encoding of historical knowledge.

Furthermore, SCN utilizes 2 collaborative strategies consisting of the triple distillation mechanism and the prior feature fusion mechanism on the two subnetworks, achieving effective knowledge recall (Pres-Net to Trans-Net) and fusion (Trans-Net to Pres-Net). The former mechanism uses a compound loss term to integrate Trans-Net into Pres-Net by maximizing the semantic similarity in the feature space, achieving knowledge persistence, and endowing the network with better memorization ability. The latter allows for integrating the prior features stored in Pres-Net with the features learned by Trans-Net to facilitate the learning of new tasks. To ensure the validity of the prior features and effective knowledge recall from Pres-Net to Trans-Net, we design a recall filter according to our previous work [51]. This filter embeds the prior and learned features into the same low-dimensional feature space to learn the transferability of features.

B. Asymmetric-independent Design of Dual Subnetworks

As mentioned before, SCN consists of two subnetworks with different structures, Pres-Net and Trans-Net, which are responsible for the memorization of old tasks and the learning of new tasks, respectively.

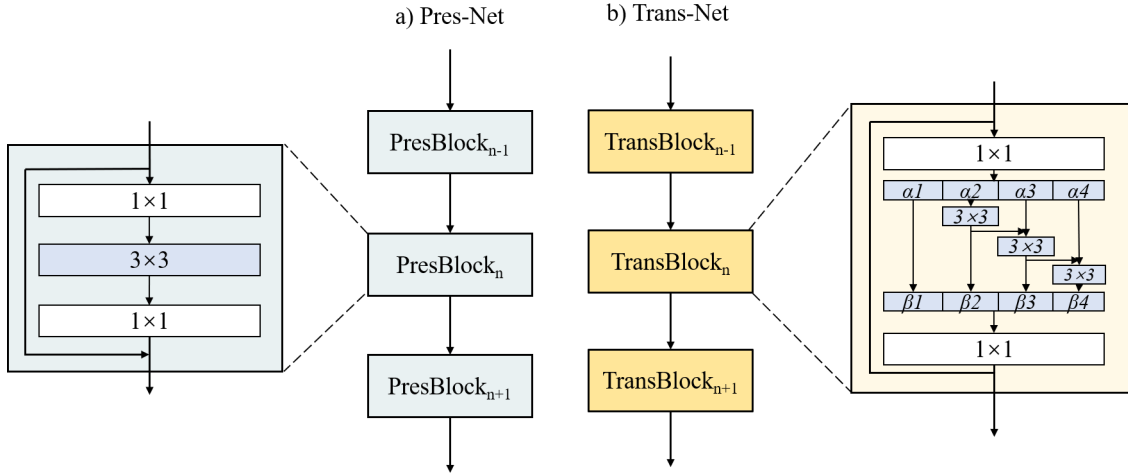


Fig. 2. Structure of Pres-Net and Trans-Net. The part on the left shows the basic structure of Pres-Net, which consists of ordinary residual blocks designed to maintain more historical memory in an abstract and simple structure. The right part shows the basic structure of Trans-Net, which replaces the 3×3 convolution in Pres-Net with a set of convolutions at different scales to obtain scale information from new images to help learning.

1) *Preserving Network*: The core function of Pres-Net is to learn abstract knowledge and thus implement long-term memory. Based on our assumption that more abstract knowledge can be acquired by forcing Pres-Net to learn using a simple network structure, we set Pres-Net as a simple network ResNet-50 composed of ordinary residual blocks [50] (ResBlock, shown in Fig. 2.left), denoted as:

$$F_n^P = \text{conv1}(\text{conv3}(\text{conv1}(F_{n-1}^P))) \quad (1)$$

where F_n^P denotes the output feature of the n -th block in Pres-Net, and conv1 and conv3 denote 1×1 and 3×3 convolution operations, respectively.

2) *Transient Network*: Trans-Net, which is dedicated to learning new tasks, is expected to extract rich semantic information from images. Considering the multiscale and multi-channel feature dependence of remote sensing images, we set up a multiscale feature extraction Res2Block (shown in Fig. 2 right) and a multichannel attention module ECA (shown in Fig. 3) to help Trans-Net better acquire scale and channel information in new tasks.

Multiscale feature extraction module Res2Net. A multiscale feature extraction module Res2Net denoted as $G(a, S)$ is added to Trans-Net. This module can represent multiscale features with a finer degree of regularity by using a larger receptive field. As shown in Fig. 2, a larger receptive field is achieved by the convolution of a set of different scales, denoted as:

$$G(a, S) = K(\beta_1, \beta_2, \dots, \beta_s) \quad (2)$$

$$\beta_i = \begin{cases} \alpha_1 & i = 1 \\ \text{conv3}(\alpha_2) & i = 2 \\ \text{conv3}(K(\alpha_i + \beta_{i-1})) & 2 < i \leq s \end{cases} \quad (3)$$

where α denotes the feature after the 1×1 convolution layer, α_i denotes the i -th subfeature split by α , $K(\cdot)$ denotes the collocation of features, and S is the scale hyperparameter. The combined convolution first splits the input features into

s groups of subfeatures. A convolution operation with kernel size 3×3 is performed on all subfeatures $\alpha_i (i \neq 1)$, after which the result β_i will be combined with the next subfeature α_{i+1} for the next turn of the convolution operation.

In Trans-Net, the output F_n^T of each block is calculated as

$$F_n^T = \text{conv1}(G(\text{conv1}(X_{n-1}^T), S)) \quad (4)$$

where $G(\cdot)$ denotes the multiscale feature extraction module, S denotes the scale hyperparameter, conv1 denotes the 1×1 convolution, and X_{n-1}^T denotes the input feature of the block. It is worth noting that in Trans-Net, the output F_n^T of the n -th block is not equal to the input X_n^T of the $(n+1)$ -th block due to the migration mechanism that we have designed, as detailed in next Section.

Multichannel attention module ECA. Attention mechanisms improve the capability of networks to focus on critical information in complex contexts. As shown in Fig. 3, a lightweight channel attention module ECA [29] is built in the recall filter to model channel relationships by accessing local cross-channel interaction information among k neighbor channels. As shown in (2), the one-dimensional convolution is implemented as:

$$F_{n-1}^{P'} = \text{ECA}(F_{n-1}^P) = \text{AAP}(\text{conv1}(F_{n-1}^P)) \quad (5)$$

where F_{n-1}^P denotes the output features from the n -th block of Pres-Net, conv1 denotes the 1×1 convolution, and AAP denotes the adaptive average pooling.

C. Collaboration Between Dual Subnetworks via Transfer Recall Strategy ($P \rightarrow T$)

) and Triple Distill Strategy ($T \rightarrow P$) Two strategies are designed using SCN for collaboration between subnetworks to achieve knowledge recall without catastrophic forgetting: (1) *Transfer recall strategy*, which migrates prior knowledge from Pres-Net to Trans-Net by the recall filter and (2) *Triple distill strategy*, which integrates the newly learned knowledge from Trans-Net into Pres-Net through triple distill loss.

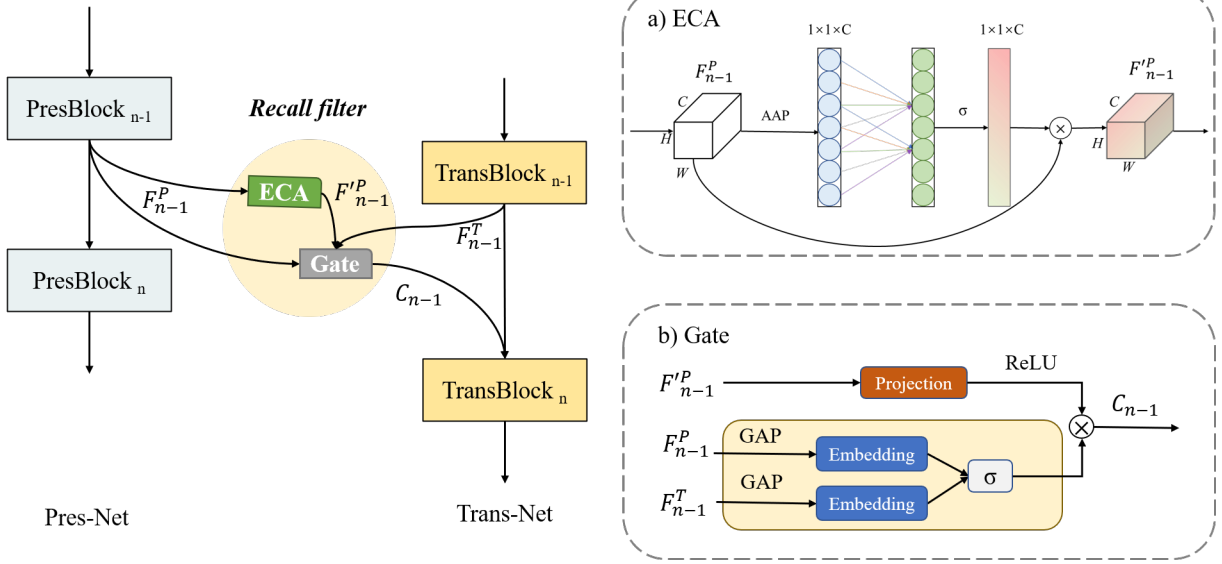


Fig. 3. Knowledge recall from Pres-Net to Trans-Net. Pres-Net and Trans-Net have the same number of layers and communicate between each layer by the recall filter. The recall filter consists of two main modules: a) A channel attention module ECA that captures local cross-channel interaction information by considering each channel and its k neighbors. b) A transfer gate that learns migrability by embedding Trans-Net (F_{n-1}^T) and Pres-Net (F_{n-1}^P) features into the same space while integrating the projection of features that have passed through the channel attention module ($F_{n-1}'^P$) and finally integrating them to achieve an effective recall of Trans-Net.

1) *Transfer Recall Strategy (P→T)*: The catastrophic forgetting and knowledge recall of historical knowledge are considered in SCN lifelong learning. Pres-Net utilizes groups of recall filters [51] to migrate features to Trans-Net while Trans-Net is learning new categories, i.e. before the feature fusion mechanism.

The core idea of the recall filter is to embed the features of both networks into the same low-dimensional feature space, learn the transferability between Pres-Net and Trans-Net features, and finally select the beneficial features for migration, ensuring the validity of migrated features. As shown in Fig. 3, this process contains two parts and is detailed as follows.

$$\begin{aligned} C_{n-1} &= f_{RF}(F_{n-1}'^P, F_{n-1}^P, F_{n-1}^T) \\ &= g_{n-1}(F_{n-1}^P, F_{n-1}^T) \otimes h_{n-1}(F_{n-1}'^P) \end{aligned} \quad (6)$$

where $f_{RF}(\cdot)$ denotes the recall filter function, F_{n-1}^P denotes the output of the n -th block of Pres-Net, $F_{n-1}'^P$ denotes the output of F_{n-1}^P after the channel attention module, F_{n-1}^T denotes the output of the n -th block of Trans-Net, and $g_{n-1}(\cdot)$ and $h_{n-1}(\cdot)$ represent the embedding and projecting processes in the recall filter, respectively (as shown in Fig. 4.2)

Embedding. The embedding process of the recall filter is represented as $g_{n-1}(\cdot)$. During this process, two linear layers embed features from Pres-Net and Trans-Net into the same space to learn the transferability between them. This process can be described as follows:

$$g_{n-1} = \sigma(E_{n-1}^P \text{GAP}(F_{n-1}^P) + E_{n-1}^T \text{GAP}(F_{n-1}^T) + b) \quad (7)$$

where F_{n-1}^P denotes the output features of the $(n-1)$ -th layer of Pres-Net, g_{n-1} denotes the recall filter, E_{n-1}^P and E_{n-1}^T denote the linear layer outputs of Pres-Net and Trans-

Net, respectively, b denotes bias, $\sigma(\cdot)$ is the sigmoid function, and GAP denotes global average pooling.

Projecting. The projecting process of the recall filter is represented as $h_{n-1}(\cdot)$. The main role of the projecting process is to project features from the channel attention module and achieve alignment. This process can be described as follows:

$$h_{n-1}(F_{n-1}'^P) = \text{ReLU}(P_{n-1} \cdot F_{n-1}'^P) \quad (8)$$

where $F_{n-1}'^P$ denotes the output of F_{n-1}^P after the channel attention module, $\text{ReLU}(\cdot)$ denotes the rectified linear unit function for activation, and P_{n-1} denotes the projection parameters.

In Trans-Net, the input X_n^T for the n -th block is:

$$X_n^T = C_{n-1} + F_{n-1}^T \quad (9)$$

where X_n^T denotes the features input of the n -th block when training Trans-Net, C_{n-1} denotes the knowledge migrated from Pres-Net, and F_{n-1}^T denotes the features output by the n -th block of Trans-Net.

2) *Triple Distill Strategy (T→P)*: In the second stage of SCN, the new task knowledge preserved by Trans-Net is integrated with the historical knowledge preserved by Pres-Net through a knowledge distillation loss L_{distill} (see Fig. 4). This stage consists of three components:

$$L_{\text{distill}} = \gamma L_{\text{similar}} + L_{\text{dis}_P} + L_{\text{dis}_T} \quad (10)$$

L_{similar} : Semantic similarity generates similar activations in the feature space. For the output of the i -th layer of the independent network, its activation similarity is defined as:

$$G_i = \frac{\hat{X}_i \times \hat{X}_i^T}{\|\hat{X}_i \times \hat{X}_i^T\|_2} \quad (11)$$

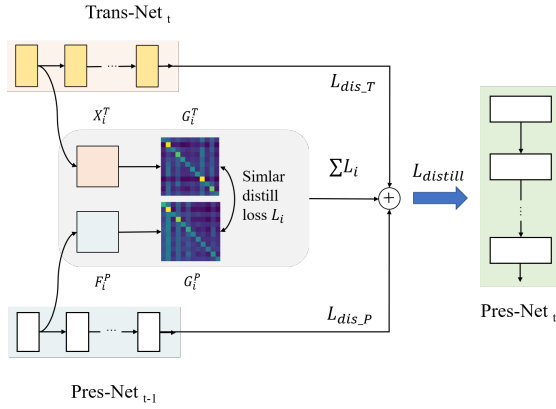


Fig. 4. Triple distillation loss. This loss is composed of three components. For each layer of the output of the two networks, we compute 2 similarity measures of the features and sum them to obtain $L_{similar}$. At the same time, we also consider the similarity of the output results of the two networks and calculate L_{dist_P} . In addition, to avoid the extreme case of poor performance of Trans-Net in the transfer recall phase, we calculate L_{dist_T} to make the data labels as close to the true distribution as possible.

where the output X_i of the i -th layer of the network has size $b \times c \times h \times w$, in which b denotes the batch size, c denotes the number of channels, h and w denote the spatial dimensions, X_i^T denotes the deformation of the output X_i of the i -th layer of this network, which is a two-dimensional vector of size $b \times [c \times h \times w]$, $\hat{X}_i \times \hat{X}_i^T$ is a matrix of size $b \times b$, and $\|\cdot\|$ denotes L2 regularization. In Pres-Net and Trans-Net, the similarity loss by G at each layer is defined as:

$$L_{similar} = \frac{1}{b^2} \sum \|G_i^T - G_i^P\|_F^2 \quad (12)$$

where G_i^T and G_i^P denote the activation similarity of the i -th layer of Trans-Net and Pres-Net, respectively, and $\|\cdot\|_F^2$ denotes the Frobenius norm.

The similarity L_{dis_P} is also considered in the output distribution of Pres-Net and Trans-Net. L_{dis_P} For Pres-Net, which is dedicated to preserving historical knowledge, the target loss is defined as:

$$L_{dis_P} = L_{CE}(\sum_{i=1} \sigma(\frac{X_i^{P'}}{\mathcal{T}}), \sum_{i=1} \sigma(\frac{X_i^P}{\mathcal{T}})) \quad (13)$$

where L_{CE} denotes the cross-entropy loss, P' denotes the updated Pres-Net, \mathcal{T} denotes the distillation hyperparameter temperature, and $\sigma(\cdot)$ denotes the sigmoid function.

L_{dis_T} : For Trans-Net, a hybrid loss L_{dis_T} is used that allows for both data distribution and output distribution to avoid poor learning of new images by Trans-Net. It is defined as follows:

$$L_{dis_T} = (1 - \beta)L_{CE}(\sigma(X^P), Y) + \beta t^2 L_{CE}(\sum_{i=1} \sigma(\frac{X_i^{P'}}{\mathcal{T}}), \sum_{i=1} \sigma(\frac{X_i^T}{\mathcal{T}})) \quad (14)$$

where β denotes the hyperparameter to control the effects of the data distribution and output distribution on the final loss and \mathcal{T} is the distillation hyperparameter temperature.

Algorithm 1 Pseudo Code for SCN Training and Inference on $N + 1$ Tasks

Training:

Input: Input image x and the corresponding label y of the task $N + 1$, the weights θ_N^P of Pres-Net that has been trained on previous N tasks, hyper-parameters $\mathcal{T}, \beta, scale$, the number of epochs.

Output: The weights $\theta_{(N+1)}^P$ of Pres-Net which can classify both new and old tasks.

initial Trans-Net weight θ^T

for $i = 1$ to the number of epochs do:

 Sample $B \in x, y$

$y = f_T(B, scale, \theta_N^P, \theta^T)$ #Pres-Net transfer to

Trans-Net

 Calculate $L = L_{binary_cross_entropy_with_logits}(y, \hat{y})$

 Update θ^T

end for

Extend classifier on Pres-Net (θ_N^P) to generate new model $\theta_{(N+1)}^P$

for $j = 1$ to the number of epochs do:

 Sample $B \in x, y$

$\tilde{y} = f_T(B, \theta^T); \bar{y} = f_P(B, \theta_N^P)$

 Calculate $L_{similarity}, L_{dis_T}, L_{dis_P}$ by y, \tilde{y}, \bar{y} ,

$L = \gamma L_{similarity} + L_{dis_P} + L_{dis_T}$

 Update $\theta_{(N+1)}^P$

End for

Discard Trans-Net

Inference:

Input: Input image x

Output: Output label y that corresponding to image x

$y = f_P(x)$

D. Pipeline of Training and Inference for Sequential Tasks

1) *Training phase:* As shown in Fig. 5, the training of SCN consists of two stages:

- **Transfer recall stage:** At the beginning of the stage, Trans-Net is initialized, and Pres-Net is frozen. While Trans-Net learns new tasks, Pres-Net transfers knowledge to Trans-Net through the recall filter (red line).
- **Triple Distill stage:** At this stage, the knowledge of Trans-Net is integrated into Pres-Net through knowledge distillation. A similarity knowledge distillation loss function, cross similar loss, is invoked to assist in the integration of knowledge (blue line).

2) *Inference phase:* As shown in Fig. 5, Trans-Net is removed, and only the Pres-Net remains for inference after completing the training of SCN, which avoids high computational cost and large parameter size.

The pseudocode for the training procedure is shown in Algorithm 1.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setting

We first compare the performance of SCN with some classical methods on class-incremental scenarios using three

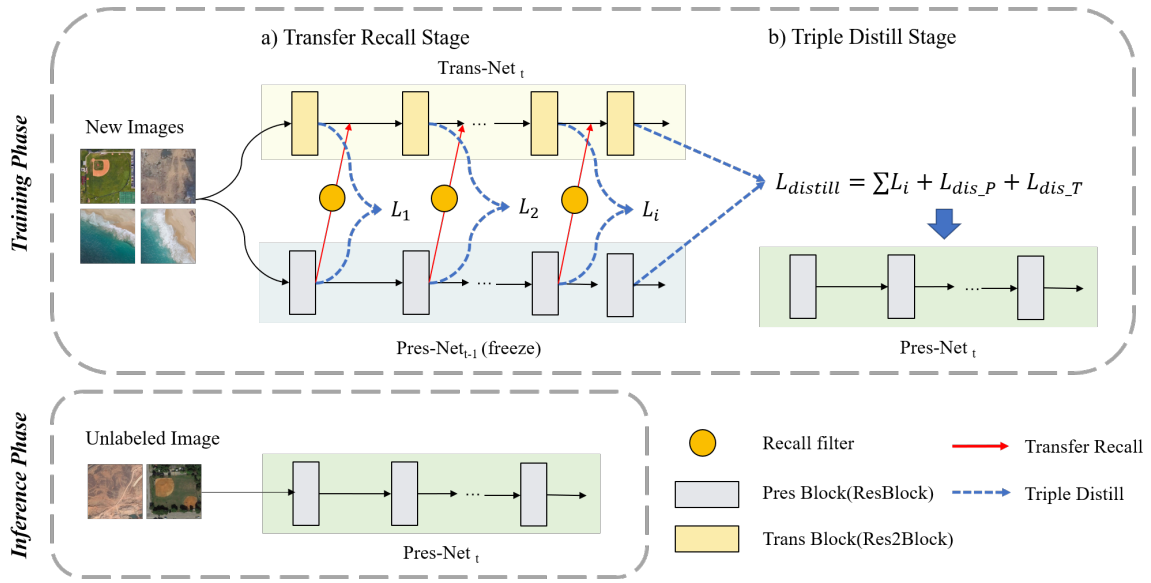


Fig. 5. SCN framework. This diagram shows the whole process of SCN in the training phase and the inference phase. The training phase contains two stages, namely, the transfer recall stage (red line) and the triple distill stage (blue line). The red line indicates the migration of historical experience from Pres-Net to Trans-Net by the recall filter (orange circles) at the transfer recall stage. The blue line indicates the loss of distillation at the triple distill stage, which distills to form Pres-Net_t that retains all current knowledge. In this phase, the parameters of Pres-Net_{t-1} are frozen. The lower part of the figure shows the inference phase of SCN. In this phase, only the Pres-Net_t network, which has saved the previous t tasks, is involved in inference.

TABLE I
DATASETS DESCRIPTION

Datasets	Properties	Characteristics	Experiments
UC Merced	21 categories, 100 images per category, spatial resolution 0.3	Representative dataset, small	IV-B1, IV-B3, IV-B4
AID	30 categories, 220 ~420 images per category, spatial resolution between 0.5 m and 8 m	Representative dataset, inhomogeneous dataset, difficult	IV-B1
NWPU	45 categories, 700 images per category, spatial resolution between 0.2 m and 30 m	Multiscale dataset	IV-B3, IV-B5
RSICB256	35 categories, 800 images per category, spatial resolution between 0.22 m and 3 m	Complex dataset, containing mislabeled noise	IV-B1, IV-B2

benchmark datasets for the task of remote sensing image scene classification. Then, we analyze the changes of model capabilities as the number of tasks growing, the impact of the individual modules of SCN and the robustness of our algorithm to hyperparameters. Finally, we visualize the differences and effectiveness of asymmetric network architectures on feature learning. Specifically, we set 3 different class-incremental learning scenarios for lifelong learning: 3 tasks with 7 classes for UC-Merced, 6 tasks with 5 classes for AID, and 9 tasks with 3 classes for RSICB-256.

1) *Datasets*: We selected four remote sensing datasets for our experiments. The UC-Merced [52], AID [53], and

RSICB-256 [54] datasets are utilized for the comparison of algorithms. Furthermore, the NWPU [55] dataset are utilized for analyzing the effect of the Res2Block module in SCN. The characteristics of these datasets are listed in Table. I.

2) *Baseline*: coloredSix classical lifelong learning methods are chosen for comparison with SCN: **Joint**, **One**, **Fine-tuning**, **LwF**, **EWC**, and **NAmours(2020)**. Among them, **Joint** and **One** are the ideal-state methods. **Joint** learns all the data together, which avoids the catastrophic forgetting while is unable to facilitate new tasks learning. In contrast, **One** learns each task independently, which can avoid the catastrophic forgetting while is unable to learn new task. **Fine-tuning** is a common method that fine-tuning the model corresponding to the previous task to the new task. **LwF**[12] is the most typical function-based approach. **EWC**[13] is one of the most classic regularization-based approaches. The incremental learning framework proposed by NAmours (2020) is a typical regularization approach in remote sensing scene classification, which achieves incremental learning by regular constraints and distillation loss on parameters.

3) *Metrics*: Referring to [56], three metrics are utilized to evaluate the performance of algorithms: *accuracy*, *ACC*, and *FWT*.

- *accuracy*: accuracy represents the ratio of the number of correct model predictions to the total number of predictions, which evaluates the performance of algorithms on the new task, i.e.,

$$accuracy(T) = \frac{Num(\hat{y} = y)}{Num(y)} \quad (15)$$

where T denotes the task, \hat{y} denotes the model output label of the t -th task, and y denotes the ground truth label.

TABLE II
COMPARISON OF AVERAGE ACCURACY (ACC) ON CLASS-INCREMENTAL LEARNING

Number of sub-tasks	Settings	Joint	One	ACC(%)				
				Finetuning	LwF	EWC	NAmours	SCN
3 (UC-Merced)	256	97.14 ²	98.3 ¹	45.00	89.29	48.57	43.33	89.52 ³
	64	51.95	92.1 ¹	39.52	81.19 ³	35.48	52.86	82.86 ²
6 (AID)	256	96.80 ²	96.83 ¹	30.40	88.23	29.73	36.20	88.93 ³
	64	48.85	76.93 ¹	24.85	75.41 ²	26.56	43.90	66.55 ³
9 (RSICB-256)	256	99.56 ²	99.94 ¹	42.76	81.29	43.69	52.91	82.63 ³
	64	96.46 ²	99.02 ¹	41.9	76.39	44.72	50.82	82.44 ³

TABLE III
COMPARISON OF FWT ON CLASS-INCREMENTAL LEARNING

Number of sub-tasks	Settings	Joint	One	FWT(%)				
				Finetuning	LwF	EWC	NAmours	SCN
3 (UC-Merced)	256	\	\	1.43 ²	-0.71	-3.57	2.14 ¹	1.07
	64	\	\	-3.57 ¹	-9.29	-17.14	-4.64 ²	-4.64 ²
6 (AID)	256	\	\	1.42	0.72	-6.41	2.11 ¹	1.55 ²
	64	\	\	-4.20	-0.18 ²	-4.39	-4.18	0.23 ¹
9 (RSICB-256)	256	\	\	-0.35	0.07 ¹	0 ²	-0.14	0.07 ¹
	64	\	\	-1.07	-1.09	-3.39	0.05 ¹	0.02 ²

- *ACC* [42]: *ACC* represents the average accuracy of the model in sequential learning of historical tasks and reflects the model's ability to resist forgetting. It is calculated as follows:

$$ACC(i) = \frac{1}{T} \sum_{i=1}^T P_{T,i} \quad (16)$$

where T denotes the total number of tasks learned by the model, and $P_{T,i}$ denotes the accuracy of the model in learning the i -th task.

- *FWT* [42]: Forward Transfer (FWT) represents the average of the differences between the test accuracy of the model incrementally learning the task and the test accuracy of the model learning the task alone. This metric reflects the influence of the previously learned task on the later learned task and is calculated as follows:

$$FWT(i) = \frac{1}{T-1} \sum_{i=2}^T (P_{i,i} - m_i) \quad (17)$$

where T denotes the total number of tasks learned by the model, $P_{i,i}$ denotes the accuracy of the i -th task after learning the i task, and m_i denotes the accuracy of the i -th task through joint learning.

4) *Training details*: ResNet-50 was used as the backbone network in our experiments. All baselines were based on the same structure. The training and test sets were divided at a ratio of 4:1. The hyperparameters of the comparison methods were set as follows: λ is 15 in EWC, β is 0.08 in SCN, and \mathcal{T} is 10. The stochastic gradient descent strategy was used in all experiments, with an initialized learning rate in the range $\{0.1, 0.01, 0.0001\}$. The momentum was 0.9, and the weight decay was 1×10^{-5} . The number of training epochs was 100. All methods in the experiments were initialized

with *kaiming_uniform* [57]. The hyperparameter setting of baseline include: LwF $T = 2$, NAmours $\lambda_1 = 1, \lambda_2 = 1, T = 0.07$, and EWC $\lambda = 50$. The operating environments were Python 3.8 and PyTorch 1.7.1. We used 2 deep learning stations to run all experiments: DGX A100 Station with 4*80 G GPU and A100 station with 4*40 G GPU.

B. Results and Comparisons

1) *Performance comparison on class-incremental learning*: We compared the performance of SCN with the baselines. Tables II and III show the results in terms of *ACC* and *FWT* metrics, respectively.

To eliminate the interference of a single dataset, we evaluated the algorithm's lifelong learning ability on three datasets with various characteristics (Table. I). The 3-task test used the UC-Merced dataset, while the 6-task test used the AID dataset with an inhomogeneous data distribution, and the 9-task test used the RSICB-256 dataset with mislabeling noise.

In addition, to reduce the impact of image resolution on incremental learning results, we conducted experiments on all datasets in two learning methods. The first one is 256-size image learning, similar to some common testing processes in the field of remote sensing, where all images were resized to the size of 256x256 (denoted as 256-setting). And the second one is 64-size image learning, where all images were randomly cropped to the size of 64x64(denoted as 64-setting).

In the 256-setting, SCN got the first place in *ACC* among all incremental learning methods (except Joint and One). The incremental learning strategy LwF, which uses a similar distillation method as SCN, also achieved a better *ACC* score. This indicates that the distillation strategy had better characteristics in incremental learning. However, it performed poorly on *FWT* compared with SCN, which was due to

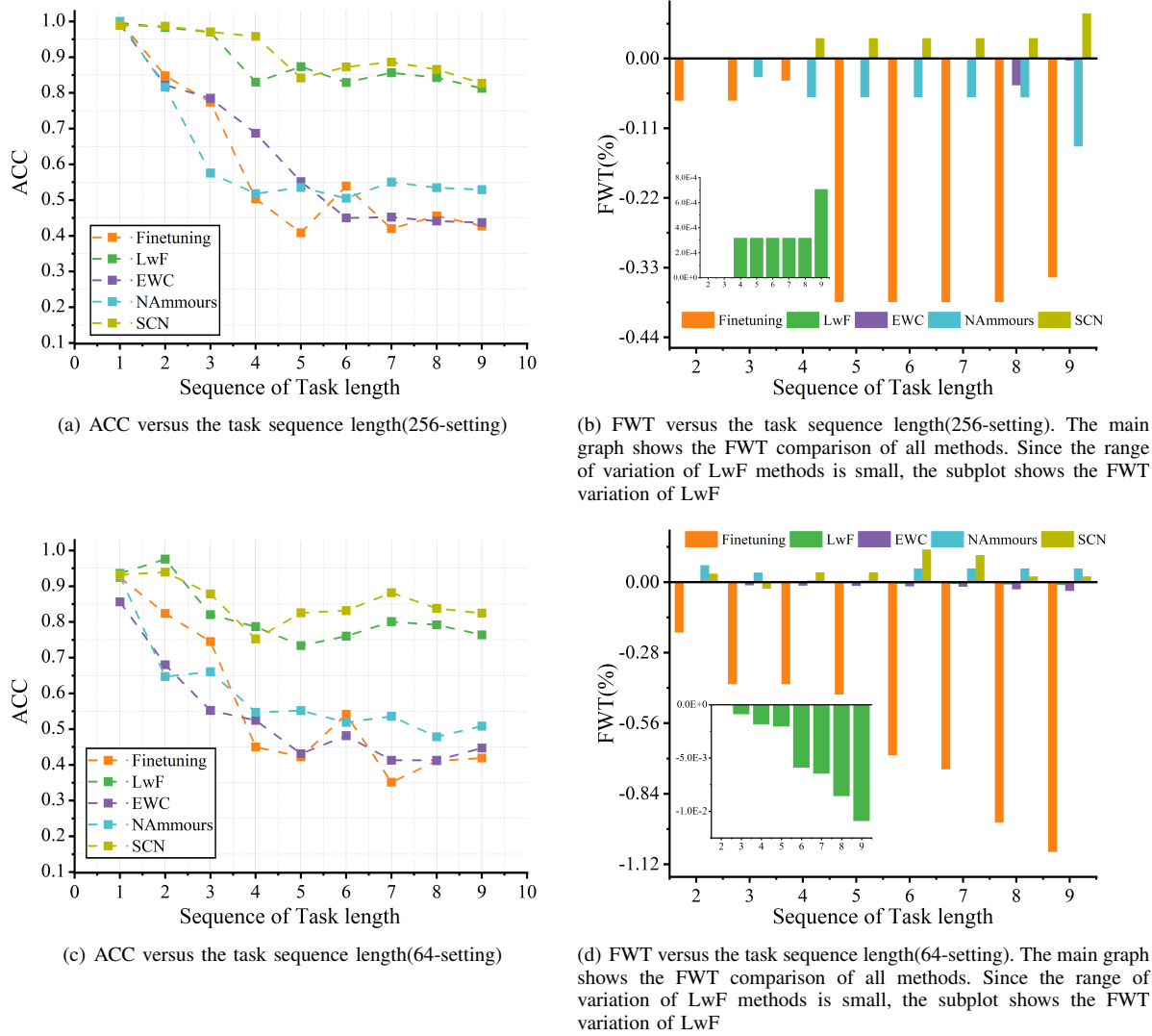


Fig. 6. Performance with sequence length

SCN's transfer strategy that benefitted the learning of new tasks. Similarly, the remote sensing method NAMmours (2020) exhibited a higher score on *FWT*, while the *ACC* results were 22.65% ~ 31.61% lower than SCN. It indicated that its distance constraints at the independent task level and parameter constraints at the shared level are more flexible, and have less impact on subsequent tasks.

SCN also showed similar results under 64-setting. SCN preserved the high average precision (*ACC*) after sequential tasks and stayed to the top two among similar incremental learning algorithms (except One, Joint), which indicated that SCN had a high ability to overcome catastrophic forgetting. Likewise, the advantage of SCN was more obvious on *FWT*. Its values reached positive 0.23 and 0.02 in 6 and 9 task sequence learning, and 3 task was also 0% ~12.5% higher compared to other methods. This indicated that SCN was not only resistant to interference from historical tasks in this sequential task learning, but also improves from historical tasks. This was attributed to our knowledge transfer using Trans-Net to Pres-Net in the new task learning twinning,

which somewhat generalized and utilized the experience of the historical task.

We noticed that NAMmours(2020) has a gap in overcoming forgetting and obtain positive migration compared to SCN. In this result we focused on two points: Firstly, we focused on the fact that 3-UCMerced did not have the high accuracy as in the original NAMmours (2020), which was due to the fact that we are using UC-Merced for 3-task learning training and the individual task difficulty was higher compared with the 7-task training set by NAMmours (2020). Secondly, we noticed that the *ACC* of NAMmours(2020) 256-setting was lower than that of 64-setting in the 3-UC-Merced and 6-AID dataset tests. TableIV shows the accuracy of NAMmours(2020) and SCN on each historical task after learning 6 tasks in sequence. The 256-setting underwent strong forgetting, while the new task was well learned; 64-setting restrained memory, while the new task learning is greatly impaired. This indicated that NAMmours (2020) fails to reach a balance between over-constraint and over-relaxation of the old task, while SCN showed more stable performance. This was also reflected in

TABLE IV
NAMMOURS AND SCN INCREMENTAL LEARNING AFTER 6 TASKS(AID),
THE ACCURACY OF EACH TASK

	Settings	1	2	3	4	5	6
NAmmours	256	21.82	23.08	24.68	25.94	23.17	98.53
	64	21.82	30.77	46.84	51.87	51.52	60.59
SCN	256	78.79	72.12	90.51	97.59	96.34	98.24
	64	51.82	69.23	68.99	74.60	65.55	69.12

TABLE V
TASK LEARNING CONTENT OF UC-MERGED

Num	Task content	Number of shadow-prone categories
1	agricultural, airplane , baseball diamond, beach, building , chaparral, dense residential	3
2	forest, freeway, golf course, harbor, intersection, medium residential , mobile home park	1
3	overpass , parking lot, river, runway, sparse residential , storagetanks , tennis court	3

the metrics of the *FWT*.

It is worth mentioning that One showed the best score on *ACC* in all three test settings, because One learns each task independently, without sharing of parameters, which means it did not suffer from the catastrophic forgetting while is also unable to learn to tasks. We only used it as a reference in an ideal state.

2) *Analysis of sequential learning on various tasks*: We further analyzed how the model's learning ability changes as the tasks increases on RSICB-256. Fig. 6 shows the changes on *ACC* and *FWT* for each model as the number of tasks increases.

In terms of *ACC* metrics, SCN maintains *ACC* at a relatively stable and high level with increasing task volume, both in 256-setting and in 64-setting. In contrast, other methods such as EWC showed a task-volume-related decrease in *ACC*. It indicates that simple regular methods' limitations in overcoming catastrophic forgetting as tasks increases. Therefore, these models perform limited ability to learn a large number of tasks, which may have fatal consequences for the learning of continuously generated remote sensing images.

In terms of *FWT* metric, SCN showed a small improvement in *FWT* with increasing tasks and obtains positive *FWT* scores in most cases, which indicated that SCN is more capable of maintaining the learning ability of new tasks in incremental learning compared to the ordinary regular incremental learning methods. It is worth noting that in the case of 256-setting, the *FWT* scores of methods such as EWC is almost close to 0 in the first few tasks. This is due to the fact that the individual tasks in sequential learning are simpler, the model can easily reach an accuracy of 1.0, resulting in the inability to measure the negative migration impact. The regular variation of *FWT* with task length is more obvious under 64-setting. As shown in Fig. 6, the *ACC* and *FWT* score of finetuning, EWC and LwF decrease as the number of tasks

increases. This indicated that the constraint that incremental learning tends to be complex as the number of tasks increases somewhat hinders the learning of new tasks.

3) *Ablation study for facilitating better learning*: In this paper, we claim that SCN can learn new tasks better without forgetting historical knowledge, primarily due to three modules: (1) ECA, a multichannel feature transfer learning module in the Pres-Net to Trans-Net migration phase (2) Res2Net, a multiscale feature learning module in Trans-Net; and (3) Distillation regularization, an objective function for knowledge distillation in Trans-Net integration into the Pres-Net stage. This section analyzed the impact of different modules for learning a new task separately.

Effect of the multichannel attention module ECA. Fig. 7 top shows the performances of SCN with and without the channel attention module, ECA, with the 3-task length setting on the UC-Merced dataset. Fig. 7(a) shows the behavior of *ACC* during the learning process, and Fig. 7(b) shows the accuracy changes on the coming task. The results show that ECA increases *ACC* by 5.71% and 0.97% in the first and third tasks, respectively, whereas there is almost no increase in *ACC* in the second task.

We argue that these results depends on the difference between these 3 tasks (see Table. V). In the first task, there are more scenes, such as airplanes, houses, and dense residential areas, that are prone to shadows. In contrast, relatively few scenes are prone to shadows in the second and third tasks. Note that the addition of the channel attention module ECA improves the anti-interference ability of SCN to cope with problems such as image shadows.

Effect of the multiscale feature extraction module Res2Net. Fig. 7 bottom shows the performance comparison of SCN with and without the multiscale module Res2Net, in the setting of 5-task length with the NWPU dataset. NWPU is a remote sensing dataset containing 0.2 m ~ 30 m spatial resolution images. Thus, it is appropriate for multiscale feature learning. Fig. 7(c) shows the *ACC* during the learning process, and Fig. 7(d) shows the accuracy of the model on each task after sequential learning. The results suggest that the accuracy with Res2Net is overall 2.68% higher than the accuracy with ResNet. This indicates that Res2Net to SCN in remote sensing classification task datasets with multiscale features.

Effect of similarity-based knowledge distillation loss. As mentioned in Section III, the triple distillation loss in SCN achieves knowledge integration between Trans-Net and Pres-Net. To test its knowledge preservation ability, we compare it with the normal cross-entropy distillation loss and sigmoid loss. Fig. 8 shows the *ACC* of these three distillation losses in 3 tasks with 7 classes for the UC-Merced dataset. Empirical results show that the *ACC* associated with triple distillation loss is 0.42% and 1.01% higher than those associated with cross and sigmoid losses, respectively. This suggests that with the triple distillation loss, SCN can better retain and integrate the knowledge of Trans-Net and Pres-Net to achieve better memorization.

4) *Effect of hyperparameters on the model*: SCN contains three hyperparameters, namely, β , \mathcal{T} , and *scale*, where β denotes the weight of the data distribution and S-Net output

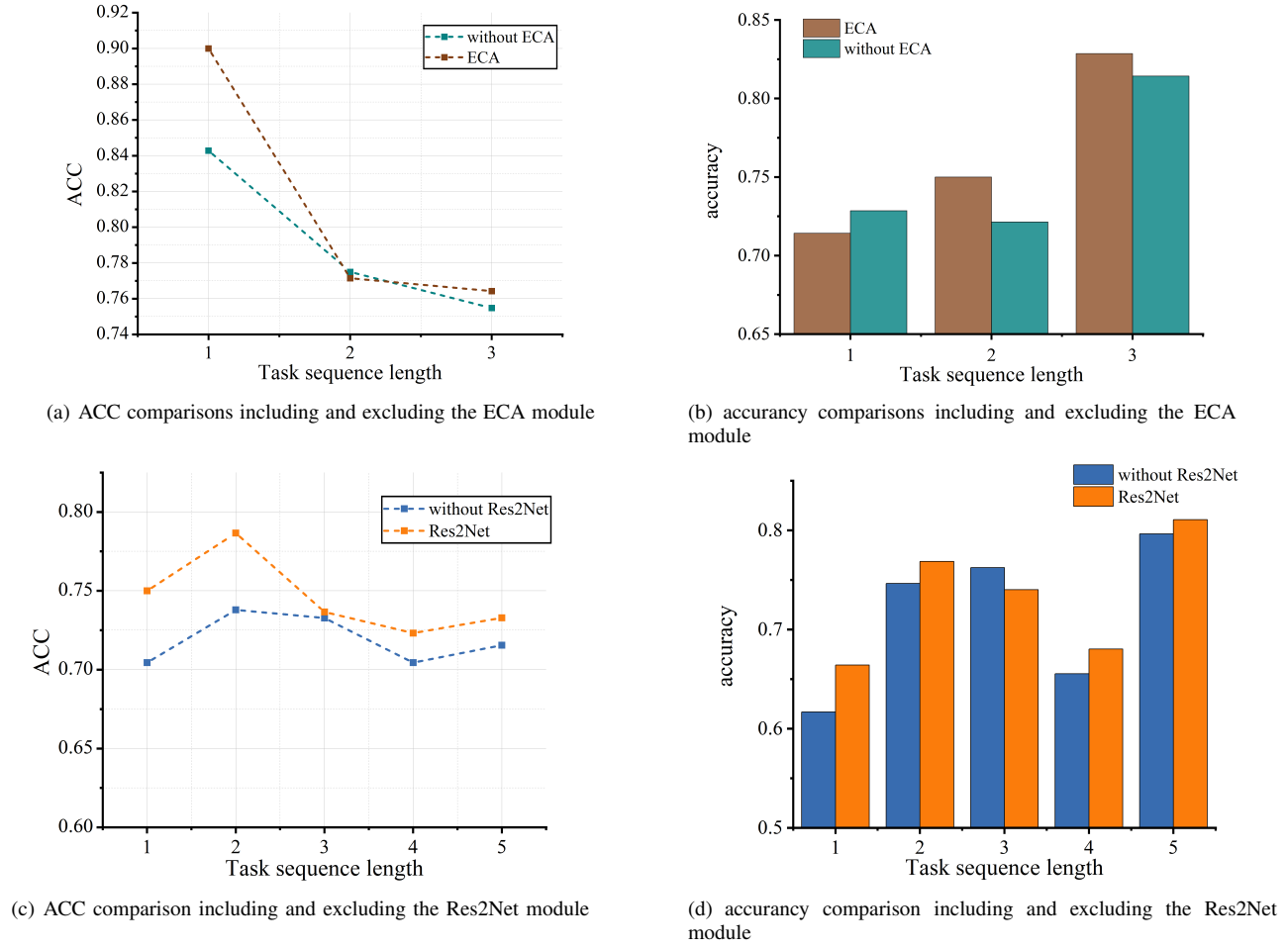


Fig. 7. Feature Extraction Ablation Experiments. The ECA channel ablation experiments were conducted under a 3-task UC-Merced experimental setup. The Res2Net module ablation experiments were conducted under a 5-task NWPU dataset with multi-scale remote sensing images experimental setup

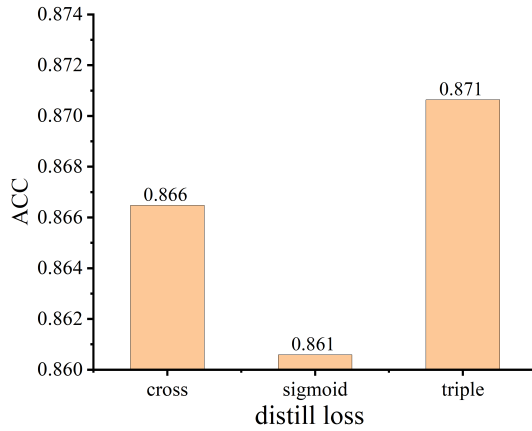


Fig. 8. Knowledge Recall from Pres-Net to Trans-Net test under UC-Merced dataset

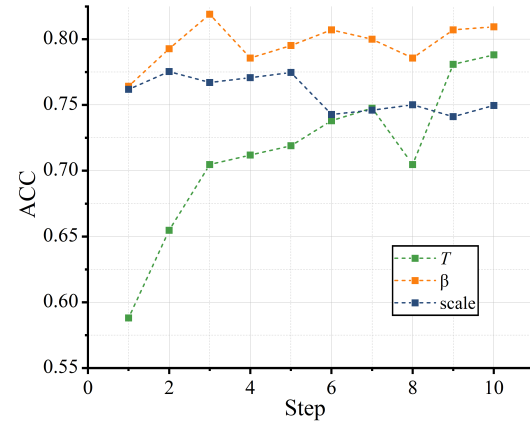


Fig. 9. Behavior of ACC versus the hyperparameter value, where T step = 1.0, β step = 0.1, and $scale$ step = 1. This test is under UC-Merced dataset

distribution, T denotes the temperature in the knowledge distillation stage, and $scale$ denotes the scale in Res2Net.

We analyzed the effect of these hyperparameters on model performance by controlling for single variable changes. Fig. 9 shows the curves of ACC versus the β , T , and $scale$ values:

- (1) The effect of β on ACC ranges from 0.76 to 0.82, and the highest value is obtained when β is 0.3;
- (2) The effect of T on ACC is more significant and ranges from 0.57 to 0.8, reaching the maximum values when T is 10, and
- (3) The effect of $scale$ on ACC on the NWPU dataset is very

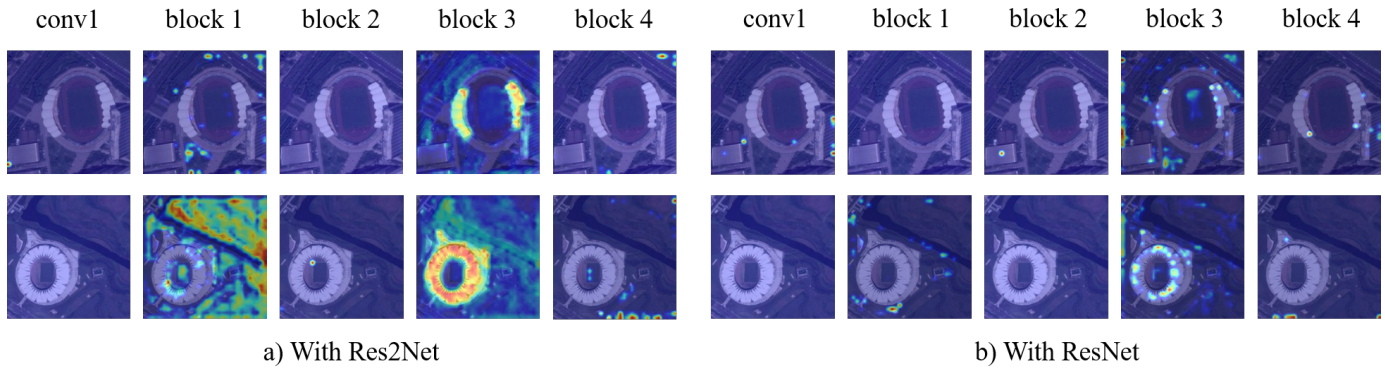


Fig. 10. Grad CAM Visualization Results Using Res2Net and ResNet SCN for Stadium Scenes(NWPU)

limited, ranging from 0.74 to 0.77. In summary, SCN is more robust to hyperparameters β and $scale$ compared with the hyperparameter \mathcal{T} . This also indicates that the value of \mathcal{T} is crucial for the performance of SCN.

5) *Feature visualization analysis of the effectiveness of asymmetric structures*: To verify the effectiveness of the asymmetric design, we visualized SCN with symmetric and asymmetric structures by Grad CAM [58], which is one of the most classic visualization tools. Trans-Net is set up as Res2Net and ResNet to test on the multiscale dataset NWPU. Fig. 10 shows the heatmap of two different scales on the stadium scenes on the final Pres-Net. It shows that the model using Res2Net can locate the information of the overall stadium at different scales more precisely than the model with ResNet. The latter performs relatively worse, focusing on the fragmented parts of the stadium rather than on the stadium itself. It also shows that with the use of Res2Net, the asymmetric design of SCN can be better adapted to the multiscale characteristics of remote sensing images, hence performing the classification work of remote sensing scenes better.

V. CONCLUSION

This paper proposes SCN, a lifelong learning framework for remote sensing image scene classification, aiming to achieve higher classification accuracy while overcoming catastrophic forgetting. SCN consists of a couple of asymmetric collaborative dual subnetwork functions for learning the new task and memorizing old tasks, called Trans-Net and Pres-Net, respectively. Trans-Net enables better learning of the specific coming task by introducing multiscale and multichannel attention mechanisms into a complex network. At the same time, Pres-Net preserves long-term and general knowledge related to historical tasks via a simpler network. The dual subnetworks cooperate via the triple distillation mechanism, which maximizes the similarity in the feature space and the output distribution to integrate new knowledge of Trans-Net into Pres-Net, achieving effective knowledge integration. Meanwhile, it utilizes the recall filter to learn the transferability between the historical knowledge and new knowledge in networks to

integrate prior knowledge from Pres-Net and assists in Trans-Net learning, achieving effective knowledge recall.

The independent network structure design enables SCN to effectively reduce parameter interference during sequential task learning, i.e., it alleviates catastrophic forgetting. Furthermore, the knowledge recall mechanism allows SCN to use the historical experience to facilitate learning new tasks. In addition, the asymmetric separation of long-term memory networks from short-term memory networks and the synergistic strategy provide a new perspective for remote sensing image scene classification in the real world.

REFERENCES

- [1] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [2] L. Chen, H. Li, G. Zhu, Q. Li, J. Zhu, H. Huang, J. Peng, and L. Zhao, "Attack selectivity of adversarial examples in remote sensing image scene classification," *IEEE Access*, vol. 8, pp. 137 477–137 489, 2020.
- [3] J. Peng, X. Mei, W. Li, L. Hong, B. Sun, and H. Li, "Scene complexity: A new perspective on understanding the scene semantics of remote sensing and designing image-adaptive convolutional neural networks," *Remote Sensing*, vol. 13, no. 4, p. 742, 2021.
- [4] J. Chen, H. Huang, J. Peng, J. Zhu, L. Chen, C. Tao, and H. Li, "Contextual information-preserved architecture learning for remote-sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [5] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification," *arXiv preprint arXiv:2009.13364*, 2020.
- [6] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [7] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

- [8] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.
- [9] W. C. Abraham and A. Robins, "Memory retention—the synaptic stability versus plasticity dilemma," *Trends in neurosciences*, vol. 28, no. 2, pp. 73–78, 2005.
- [10] J. Peng, B. Tang, H. Jiang, Z. Li, Y. Lei, T. Lin, and H. Li, "Overcoming long-term catastrophic forgetting through adversarial neural pruning and synaptic consolidation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [11] H. Li, H. Jiang, X. Gu, J. Peng, W. Li, L. Hong, and C. Tao, "Clrs: Continual learning benchmark for remote sensing image scene classification," *Sensors*, vol. 20, no. 4, p. 1226, 2020.
- [12] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [13] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [14] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.
- [15] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," *Neural Networks*, vol. 116, pp. 56–73, 2019.
- [16] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.
- [17] G. Zhou, K. Sohn, and H. Lee, "Online incremental feature learning with denoising autoencoders," in *Artificial intelligence and statistics*. PMLR, 2012, pp. 1453–1461.
- [18] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," *arXiv preprint arXiv:1708.01547*, 2017.
- [19] T. J. Draelos, N. E. Miner, C. C. Lamb, J. A. Cox, C. M. Vineyard, K. D. Carlson, W. M. Severa, C. D. James, and J. B. Aimone, "Neurogenesis deep learning: Extending deep networks to accommodate new classes," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 526–533.
- [20] Y. Chang, Y. Wang, J. Peng, Z. Dong, H. Li, and W. Li, "Mfs: A brain-inspired memory formation system for gan," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2021.
- [21] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory," *Psychological review*, vol. 102, no. 3, p. 419, 1995.
- [22] N. Ammour, Y. Bazi, H. Alhichri, and N. Alajlan, "Continual learning approach for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [23] X. Lu, X. Sun, W. Diao, Y. Feng, P. Wang, and K. Fu, "Lil: Lightweight incremental learning approach through feature transfer for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [24] S. D. Bhat, B. Banerjee, S. Chaudhuri, and A. Bhattacharya, "Cilea-net: A curriculum-driven incremental learning network for remote sensing image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.
- [25] N. Ammour, "Continual learning using data regeneration for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, 2021.
- [26] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in psychology*, vol. 4, p. 504, 2013.
- [27] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.
- [28] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," 2020.
- [30] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5138–5146.
- [31] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong gan: Continual learning for conditional image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2759–2768.
- [32] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, and T. Moon, "Ss-il: Separated softmax for incremental learning," *arXiv e-prints*, pp. arXiv–2003, 2020.
- [33] E. Belouadah and A. Popescu, "Il2m: Class incremental learning with dual memory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 583–592.
- [34] P. Pan, S. Swaroop, A. Immer, R. Eschenhagen, R. E. Turner, and M. E. Khan, "Continual deep learning by functional regularisation of memorable past," *arXiv preprint arXiv:2004.14070*, 2020.
- [35] J. Rajasegaran, M. Hayat, S. Khan, F. S. Khan, and L. Shao, "Random path selection for incremental learning," *Advances in Neural Information Processing Systems*, 2019.
- [36] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–

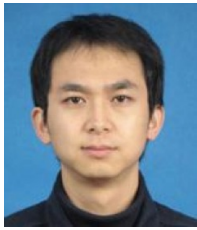
- 71, 2019.
- [37] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.
- [38] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 177–186.
- [39] Y. Yang, D.-W. Zhou, D.-C. Zhan, H. Xiong, and Y. Jiang, "Adaptive deep models for incremental learning: Considering capacity scalability and sustainability," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 74–82.
- [40] O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi, "Learning to remember: A synaptic plasticity driven framework for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 321–11 329.
- [41] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *arXiv preprint arXiv:1705.08690*, 2017.
- [42] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, pp. 6467–6476, 2017.
- [43] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. El-hoseiny, "Efficient lifelong learning with a-gem," *arXiv preprint arXiv:1812.00420*, 2018.
- [44] C. D. Kim, J. Jeong, and G. Kim, "Imbalanced continual learning with partitioning reservoir sampling," in *European Conference on Computer Vision*. Springer, 2020, pp. 411–428.
- [45] G. E. Hinton and D. C. Plaut, "Using fast weights to deblur old memories," in *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, 1987, pp. 177–186.
- [46] L. Wang, K. Yang, C. Li, L. Hong, Z. Li, and J. Zhu, "Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5383–5392.
- [47] A. Cheraghian, S. Rahman, S. Ramasinghe, P. Fang, C. Simon, L. Petersson, and M. Harandi, "Synthesized feature based few-shot class-incremental learning on a mixture of subspaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8661–8670.
- [48] W. Hu, Z. Lin, B. Liu, C. Tao, Z. Tao, J. Ma, D. Zhao, and R. Yan, "Overcoming catastrophic forgetting for continual learning via model adaptation," in *International Conference on Learning Representations*, 2018.
- [49] J. von Oswald, C. Henning, J. Sacramento, and B. F. Grewe, "Continual learning with hypernetworks," *arXiv preprint arXiv:1906.00695*, 2019.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [51] J. Peng, D. Ye, B. Tang, Y. Lei, Y. Liu, and H. Li, "Overcome anterograde forgetting with cycled memory networks," *arXiv preprint arXiv:2112.02342*, 2021.
- [52] Y. Yang and S. Newsam, "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*, 2010, pp. 270–279.
- [53] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [54] H. Li, X. Dou, C. Tao, Z. Hou, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large scale remote sensing image classification benchmark via crowdsourcing data," *arXiv preprint arXiv:1705.10450*, 2017.
- [55] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [56] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.



Dingqi Ye received a B.Sc. from Central South University, Changsha, China, in 2021 and is currently pursuing an M.Sc. degree. Her research interests include computer vision, continual learning, and remote sensing image processing.



Jian Peng received a B.S. in geographic information science from YunNan University, Kunming, China, in 2015. He is currently working towards a Ph.D. in surveying and mapping from Central South University, Changsha, China. His research topic is intelligent image understanding inspired by brain memory mechanisms. His research interests include neural computation, continual learning and remote sensing image processing.



Haifeng Li received a master's degree in transportation engineering from the South China University of Technology, Guangzhou, China, in 2005, and a Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009. He is currently a professor at the School of Geosciences and Info-Physics, Central South University, Changsha, China. He was a research associate with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, in 2011, and a visiting scholar with the University

of Illinois at Urbana-Champaign, Urbana, IL, USA, from 2013 to 2014. He has authored over 30 journal papers. His current research interests include geo/remote sensing big data, machine/deep learning, and artificial/brain-inspired intelligence. He is a reviewer for many journals.



Bruzzone Lorenzo received a Laurea (M.S.) degree in electronic engineering (summa cum laude) and his Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively. He is currently a full professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications.

Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory (<https://rslab.disi.unitn.it/>) in the Department of

Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the principal investigator of many research projects. Among the others, he is currently the principal investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the Jupiter ICy moons Explorer (JUICE) mission of the European Space Agency (ESA) and the science lead for the High Resolution Land Cover project in the framework of the Climate Change Initiative of ESA.

He is the author (or coauthor) of 294 scientific publications in referred international journals (221 in IEEE journals), more than 340 papers in conference proceedings, and 22 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. His papers are highly cited, as proven from the total number of citations (more than 40000) and the value of the h-index (92) (source: Google Scholar). He was invited as keynote speaker in more than 40 international conferences and workshops. Since 2009 he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS), where since 2019 he is Vice-President for Professional Activities. Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since then, he has been a recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards and the 2019 WHISPER Outstanding Paper Award. Dr. Bruzzone was a guest co-editor of many special issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He was the founder of the IEEE Geoscience and Remote Sensing Magazine for which he was Editor-in-Chief between 2013 and 2017. Currently, he is an associate editor for the IEEE Transactions on Geoscience and Remote Sensing. He was a distinguished speaker of the IEEE Geoscience and Remote Sensing Society between 2012 and 2016. He is a fellow of IEEE.