MiSSNet: Memory-Inspired Semantic Segmentation Augmentation Network for Class-Incremental Learning in Remote Sensing Images

Jiajun Xie, Bin Pan, *Member, IEEE*, Xia Xu, and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— With remote sensing images constantly being collected rapidly, the class-incremental semantic segmentation (CISS) task has attracted increasing attention. However, the semantic distribution shift problem of the background class in remote sensing images, which is a case of catastrophic forgetting, continues to limit available CISS algorithms. To address this challenge, we present a new memory-inspired semantic segmentation augmentation network (MiSSNet) for class-incremental learning in remote sensing images. The MiSSNet mainly includes two modules: the local semantic distillation (LSD) module and the class-specific regularization (CSR) module. LSD is a distillation structure that employs the local semantic features in retained memory to maintain correlation between pixels throughout the training process of incremental learning. It constructs a series of pixel-level correlation matrices and implicitly adjusts the semantic distribution shift problem of the background class. CSR is a classwise regularization term that utilizes the class-specific portion of the preserved memory to help the model keep repeating the learning of the old categories. It alleviates the background class shift problem by generating countless pixel-level instances of old classes. LSD and CSR work together to tackle the semantic distribution shift problem of background class from semantic information and class information aspects, respectively. In particular, MiSSNet only needs an additional single inference process for memory extraction and storage, and the whole algorithm does not add any new training parameters. Experimental results on three semantic segmentation datasets indicate the advantage of the proposed method.**

***Index Terms*— Incremental learning, remote sensing images, semantic segmentation.**

I. Introduction

REMOTE sensing images, acquired through satellite and aerial photography, provide an abundance of information on ground objects. Accurately identifying and segmenting various types of ground objects is crucial for diverse fields, such as urban planning [1], precision agriculture, and marine oil spill [2]. The primary objective of the semantic segmentation task for remote sensing images is to attribute category labels to each pixel corresponding to the ground objects of interest.

In recent years, researchers have proposed a lot of advanced works of pixel-level tasks on remote sensing images, such as spectral super-resolution [3], hyperspectral image classification [4], and change detection [5]. Zhou et al. [6] include multiscale fully CNN and multihop GCN to extract the multilevel information of hyperspectral images. Moreover, Guo et al. [7] combined a spatial subnetwork and a spectral subnetwork to extract the spatial–spectral features. And by combining a temporal feature encoder–decoder subnetwork, a bidirectional diff-changed feature representation module, and a multiscale attention fusion module, Luo et al. [8] improve the ability of feature representation for the hyperspectral image change detection task. Nonetheless, remote sensing images typically comprise dense and differently sized objects of interest [9], presenting a significant challenge for pixel-level tasks, especially the semantic segmentation task.

Yuan et al. [10] identified three strategies used to prevent the loss of spatial details from convolution and attain pixel-level accuracy in semantic segmentation task for remote sensing images [11], [12], [13], [14], [15]. These strategies are the multiscale strategy, fusion-based strategy, and postprocessing techniques. The multiscale strategy [16], [17], [18], [19] uses the dilated convolution module to produce superior features capable of addressing the loss of spatial details. Alternatively, this strategy utilizes an efficient upsampling technique or integrates edge maps into the segmentation process to achieve the spatial details. The fusion-based strategy [20], [21], [22], [23] aims to integrate geometry and spectral information to improve segmentation accuracy. The fusion always happens at the feature level or classification stage depending on the similarity of the input image structure or representation. Postprocessing techniques [24], [25], [26] involve utilizing methods, such as simple linear iterative clustering and conditional random fields that refine the segmentation results to improve classification accuracy and ensure smooth object boundaries.

However, when the traditional semantic segmentation models are required to constantly segment new categories and maintain the discriminative ability of the old categories, they

Manuscript received 17 December 2023; revised 17 January 2024; accepted 27 January 2024. Date of publication 31 January 2024; date of current version 9 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1003800 and Grant 2022ZD0160401; in part by the National Natural Science Foundation of China under Grant 62001251, Grant 62001252, and Grant 62125102; and in part by the Beijing–Tianjin–Hebei Basic Research Cooperation Project under the Grant F2021203109. *(Corresponding author: Bin Pan.)*

Jiajun Xie and Bin Pan are with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin 300071, China (e-mail: xiejiajuncqc@mail.nankai.edu.cn; panbin@nankai.edu.cn).

Xia Xu is with the College of Computer Science, Nankai University, Tianjin 300071, China (e-mail: xuxia@nankai.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3360701

may suffer catastrophic forgetting [27]; that is, their ability to discriminate the old categories will significantly decline after the training of learning new categories. Due to the ability to enable the model to continuously learn new categories and maintain discriminative ability against old ones, class-incremental learning, which is successful in solving the problem of catastrophic forgetting, has acquired significant attention. In the specific case of class incremental learning, the difficulty is that in the subsequent training of the neural network, new categories will appear, and at this time, the data used in the previous training steps often cannot be used again or only the part saved in advance can be used. Meanwhile, the network needs to avoid catastrophic forgetting and maintain the ability to discriminate all known categories. According to the literature [28], class-incremental learning methods can be categorized into three groups: regularization-based methods [29], [30], [31], bias-correction methods [32], [33], [34], and rehearsal-based methods [35], [36]. Rehearsal-based methods focus on effectively utilizing generated or saved data and our proposed method is based on a memory-inspired strategy, which is a part of the rehearsal-based methods. Currently, class-incremental learning has a wide range of applications in the field of remote sensing, including scene classification [37], [38], [39], [40], target recognition [41], [42], [43], and object detection [44]. Class-incremental learning offers a promising approach to consecutively perform semantic segmentation on new categories.

The class-incremental semantic segmentation (CISS) task is an area of increasing concern because of the widespread use in real-world applications. This is particularly true in the field of remote sensing, where different geographical environments exhibit distinct ground object categories, making it challenging to collect all desired categories at once. The CISS task allows for the continuous collection and learning of new categories of interest without discarding old models and retraining [45], [46], [47], [48]. Recent developments in the CISS task for remote sensing images have led to numerous studies. Shan et al. [49] designed an algorithm based on pseudolabeling that used historical information from the old model to obtain pseudolabels for new data and combined them with ground truth as supervised information. Rong et al. [50] combined historical information with a label reconstruction strategy to retrieve label pixels belonging to old classes. Furthermore, Li et al. [51] introduced an auxiliary binary classification task and a diversity distillation loss to the model. In addition, Yang and Tang [52] proposed a dynamic expansion strategy for the network structure to accommodate new classes.

However, recent CISS algorithms tend to ignore the semantic distribution shift problem of the background class [53], which is a specific aspect of catastrophic forgetting, especially they do not value the use of memory saved from old data. Zhu et al. [32] and Yang et al. [54] adopted memory-inspired strategies and successfully realized exceptional outcomes in semantic segmentation knowledge distillation or incremental learning. Nonetheless, while they were able to preserve and utilize old data information, they could not combat the problem of catastrophic forgetting while still completing the semantic segmentation task satisfactorily. We attempt to utilize the efficiently stored memory of old data that requires minimal storage to alleviate the semantic distribution shift problem in the background class of CISS tasks.

In this article, we present a memory-inspired semantic segmentation augmentation network (MiSSNet) which constructs local semantic distillation (LSD) and class-specific regularization (CSR) for class-incremental learning in remote sensing images. Our model aims to tackle the semantic distribution shift problem of the background class on both the semantic and category levels with two modules: LSD module and CSR module. The LSD module overcomes this challenge by employing the retained local semantic features to create correlation matrices between the output features of old and new models at the pixel level. The generated correlation matrices are then integrated into relevant distillation losses. To further overcome the semantic distribution shift problem of the background class, the CSR module utilizes the previously extracted local semantic features to acquire prototype class-specific features for each known category. Then, CSR generates countless pixels of the old classes while building a regularization term to add constraints, forcing the model to repeat the learning of the old classes which are considered the background class during the new training phase. Finally, the LSD and CSR are integrated naturally into the process of memory preservation. With the development of commercial drones and space exploration, the CISS task studied in this article is facing greater challenges. To verify the validity of our algorithm, in addition to the traditional remote sensing image dataset, we also conducted related experiments on a low-altitude aerial image dataset and a Mars dataset.

In summary, the major contributions of MiSSNet are as follows.

1) We propose a memory-inspired semantic distillation and CSR network for CISS in remote sensing images, which can extract the local semantic features and class-specific features efficiently from old data stored as memories.
2) We propose an LSD module to construct correlation matrices at the pixel level and alleviate the semantic distribution shift problem of the background class in CISS.
3) We propose a CSR module to reinforce the learning of the old classes and further tackle the problem of semantic distribution shift of the background class.

II. Methodology

A. Problem Definition

CISS involves learning a model in $t = 1, \ldots, T$ steps, where the model can progressively segment more categories as the number of steps increases. For each step t, let $\mathcal{D}_t$ be the input dataset which consists of the input images $\mathcal{I}^t$ and the labels $\mathcal{Y}^t$. The size of the input images $\mathcal{I}^t$ and the corresponding ground truth segmentation mask $\mathcal{Y}^t$ is $H \times W$, where H and W denote the image height and width, respectively. In the context of incremental learning for semantic segmentation, the set $\mathcal{Y}^t$ only contains the labels of the current classes $\mathcal{C}^t$. Therefore, $\mathcal{Y}^t$ does not include any previous classes $(\mathcal{C}^1, \ldots, \mathcal{C}^{t-1})$, or classes that have not yet

appeared ($\mathcal{C}^{t+1}, \ldots, \mathcal{C}^{T}$). Moreover, all of the classes that are not in $\mathcal{C}^t$ make up the background class $\mathcal{C}^{B^t}$. We use N^t to denote the number of images in $\mathcal{I}^t$. Especially, the training set $\mathcal{D}_t$ with N^t images and labels are available only during the training of step t. So within a supervised framework, the problem at each step can be defined as given a training dataset $\mathcal{D}_t = \{\mathcal{I}_n^t, \mathcal{Y}_n^t\}_{n=1}^{N^t}$, we need to learn a network to associate seen segmentation labels ($\mathcal{C}^{1:t}, \mathcal{C}^{B^t}$) of each pixel in any test images.

Generally, a deep network at step t can be decomposed of a feature extractor and a classifier. Specifically, the feature extractor f, parameterized by θ, maps the input $I \in \mathcal{I}^t$ into a dense feature map $Z = f_\theta(I) \in \mathbb{R}^{h\times w\times m}$ in the deep feature space $\mathcal{Z}$, where h, w, and m denote the height, width, and number of channels, respectively; the classifier g, parameterized by ϕ, and transforms I into a probability logit map $G = g_\phi(Z) \in \mathbb{R}^{h\times w\times(c+1)}$, where c denote the number of all seen classes without the background class. Denote the overall parameters by $\Theta = (\theta, \phi)$. We use a Deeplab-V3 [55] architecture with a ResNet-101 backbone. Consequently, feature map Z_l can be extracted at any layer l of the feature extractor $f_l^t(I)$, $l \in \{1, \ldots, L\}$, especially we are going to abbreviate $Z_L = f_L^t(I)$ as Z. Using the softmax function and upsampling, the final output segmentation mask is $Y = upsampling(softmax(G))$.

B. Overall Framework

Unfortunately, one of the major challenges of the CISS model is its susceptibility to semantic distribution shift of the background class, which is a unique aspect of catastrophic forgetting. Catastrophic forgetting refers to the inability of the network to maintain good discrimination ability for both old ($\mathcal{C}^{1:t-1}$) and new ($\mathcal{C}^t$) classes during new training steps. This is further complicated by the fact that pixels from old classes are set as the background class when new data arrives, causing the semantic distribution of the background class to shift. Consequently, the background classes for each step ($\mathcal{C}^{B^1}, \ldots \mathcal{C}^{B^T}$) differ from one another.

Our architecture, which is based on memory, is shown in Fig. 1. After the training of each step is completed, using an additional inference process, we utilize the new model to extract local semantic features from each instance by classwise pooling. Notably, the local semantic features only include the foreground categories of the current step, and the local semantic features of the background class are extracted solely during the initial phase. We then compute the mean value and covariance matrix for the local semantic features of each class to acquire the class-specific features, and store the semantic and class features in our saved memory. Accordingly, our memory retrieval and preservation process requires little computational cost and does not require any new training parameters.

This article proposes LSD, which is based on local semantic features, to overcome the semantic distribution shift problem of the background class common in CISS. LSD first constructs a correlation matrix at the pixel level and then distills the correlation matrices of the old and new models at each training stage. To further tackle the problem of semantic distribution shift of the background class and enhance the stability and plasticity of the network, this article introduces CSR, which utilizes preserved class-specific features for regularization. Our proposed method is based on the existing incremental learning semantic segmentation method Pseudo-label and Local POD (PLOP) [45], and we will summarize the specific form of the final loss function at the end of this section. The following details explain the method.

C. Local Semantic Distillation Module

In the task of incremental learning for semantic segmentation, some studies have explored constructing correlation matrices between pixels to distill information. However, these methods focus primarily on modeling pixel relationships within a single instance while neglecting relationships between pixels of different images. To break this limitation, we propose a new LSD module, which models relationships between pixels of multiple images. This method utilizes the local semantic memory saved at each step, allowing us to retrieve the embeddings of pixels from before images in the current step's online memory bank.

In practice, due to the limited batch size on each GPU (such as 1, 2, or 4), we save all local semantic memories after each training step to build cross-image correlation. This approach overcomes the limitation of insufficient pixel richness caused by the limited sample size, resulting in more complete information between pixels. At step t, the local semantic memory we save is

$$\mathcal{R}_t = \bigcup_{C\in\mathcal{C}^t} \bigcup_{I\in\mathcal{I}^t} \xi_C(f_\theta^t(I)) \tag{1}$$

where ξ means classwise average pooling, in other words, it averagely pools the pixel embeddings belonging to each class in a single image.

It is worth emphasizing that the classwise local semantic features in $\mathcal{R}_t$ only contain the foreground classes at the current step, whereas we only extract background classes $\mathcal{R}_0$ at the first step. Also, we do not need to do any further feature extractions at the last step. So at step t ($t > 1$), the online local semantic features used in training are

$$R_t = \bigcup_{i=0}^{t-1} \mathcal{R}_i. \tag{2}$$

For convenience, we can stack the pixels embeddings in set R_t as a sequence, that is, rewrite it as $R_t \in \mathbb{R}^{(\sum_C N_C)\times m}$, where $C \in \mathcal{C}^{1:t-1} \cup \mathcal{C}^{B^1}$, $\mathcal{C}^{1:t-1}$ means all seen classes from step 1 to step $t-1$, N_C is the number of pixel embeddings of class C and m is the embedding size. And, we denote the number of all known classes as M_C.

During each iteration, we extract a specific number of local semantic features classwise from R_t to form the matrix of prototypes. To be specific, given the number K_R of samples for all M_C categories, the sample number of each class is $r = K_R|M_C$. For class C, we will sample r local semantic embeddings from R_t

$$\left\{z_k^C \in \mathbb{R}^m\right\}_{k=1}^{r}. \tag{3}$$

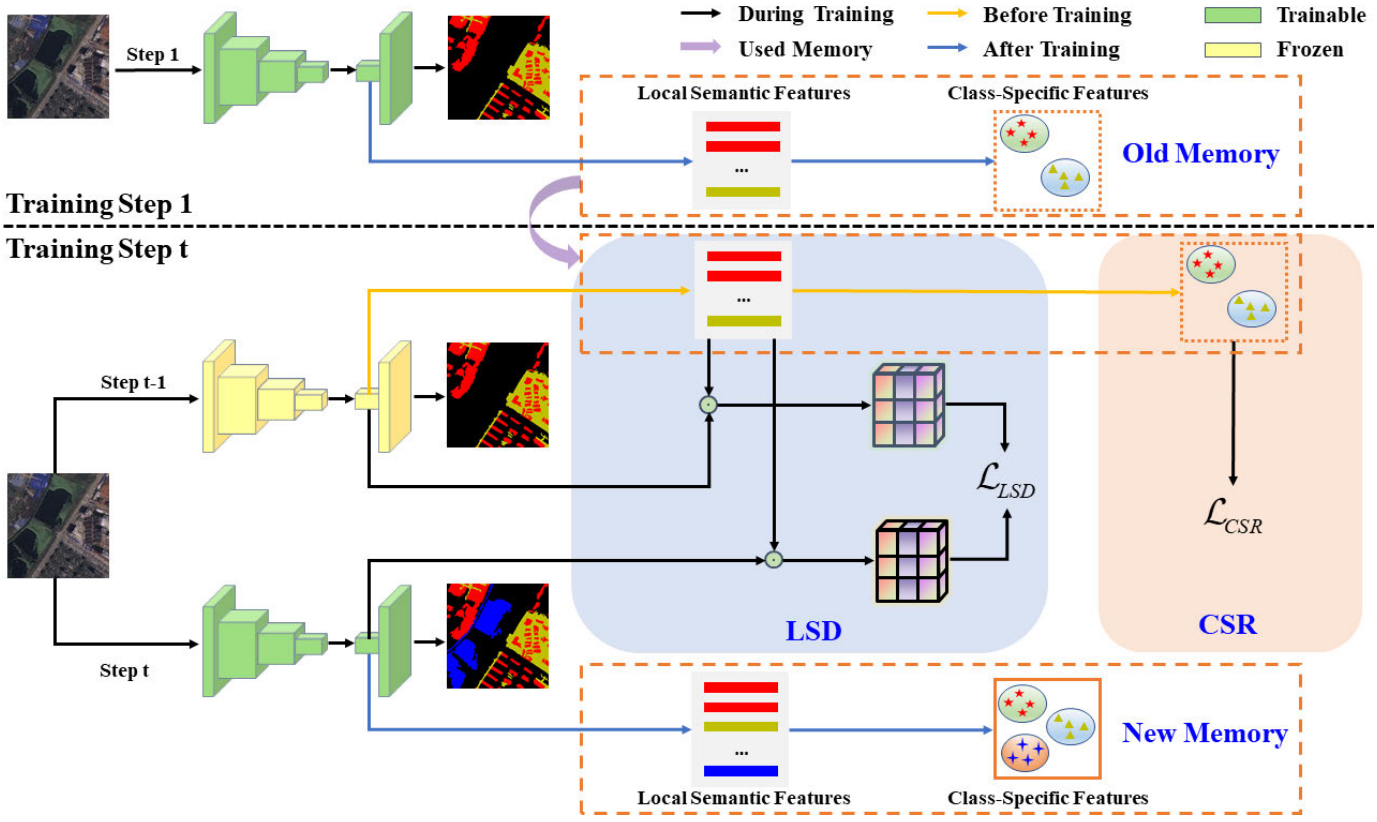

Fig. 1. Overview of our memory-inspired approach. The top half is the first step of the training, and the bottom half is the second and subsequent training. When the training of step t is carried out, the old model parameters obtained at step $t-1$ will be fixed. After the training of each step, we will save the memory, which includes two parts: local semantic features and class-specific features. With the local semantic features and class-specific features, we design LSD for knowledge distillation and CSR for regularization, respectively. One or more new classes are added at each new step, as shown in blue in the semantic graph.

With a class-balanced manner, the sample set of all categories is

$$\bigcup_{C}\left\{z_k^C\right\}_{k=1}^{r}=\left\{z_1, z_2, \ldots, z_{K_R}\right\} \tag{4}$$

and we can rewrite the extracted embeddings set as a matrix

$$\mathcal{S}_t=\left[z_1, z_2, \ldots, z_{K_R}\right] \in \mathbb{R}^{K_R \times m}. \tag{5}$$

In practice, we ensure that we extract the same number of embeddings for each class during the incremental learning process so that the value of K_R gradually increases.

At step t, given an input image $I \in \mathcal{I}^t$, the feature extractors f^t and f^{t-1} maps I to $Z^t \in \mathbb{R}^{h\times w\times m}$ and $Z^{t-1} \in \mathbb{R}^{h\times w\times m}$. Moreover, we reshape the spatial dimension of $Z \in \mathbb{R}^{h\times w\times m}$ to $Z \in \mathbb{R}^{s\times m}$, where $s=h\times w$, and preprocess each pixel embeddings in Z^t and Z^{t-1} by $l2$-normalization. We can now construct the correlation matrix from Z^t, Z^{t-1}, and $\mathcal{S}_t$ to $\mathbf{M}^t$ and $\mathbf{M}^{t-1}$

$$\mathbf{M}^t=Z^t\mathcal{S}_t^{\top} \in \mathbb{R}^{s\times K_R}, \mathbf{M}^{t-1}=Z^{t-1}\mathcal{S}_t^{\top} \in \mathbb{R}^{s\times K_R}. \tag{6}$$

To combat the problem of forgetting learned knowledge, we adopt the distillation strategy in combination with the constructed correlation matrices. Specifically, we aim for the model to generate a correlation matrix after training for the current step that is highly similar to the one generated before training. Furthermore, we apply softmax normalization on each row distribution of $\mathbf{M}^t$ and $\mathbf{M}^{t-1}$ and conduct distillation by using the KL-divergence loss. In this way, we indirectly maintain in the model that has been updated the capability of distinguishing the classes that belong to the previous steps and alleviate the semantic distribution shift problem of the background class. It is formulated as follows:

$$\mathcal{L}_{\mathrm{LSD}}=\frac{1}{s}\sum_{i=1}^{s} KL\left(\sigma\left(\frac{\mathbf{M}_{i,:}^t}{\tau}\right)\|\sigma\left(\frac{\mathbf{M}_{i,:}^{t-1}}{\tau}\right)\right) \tag{7}$$

where τ is the temperature and σ is the softmax normalization.

As we extract memories after each step of training, preserving the memory involves an additional inference of the entire training set. In addition, due to the use of average pooling for local semantic feature extraction, we require only a small memory capacity for our stored memory. In practice, we will retrieve and save the memory after the last epoch of each step. In addition, since a large proportion of the training images contain only some categories at the same time, it is advantageous to use the average pool for local semantic feature extraction, and we only need a small memory capacity to store memory. Therefore, LSD fully utilizes the information of old data in constructing the correlation matrix in the distillation process, which effectively mitigates the problem of semantic distribution shift of the background class while keeping the computation and memory capacity requirements at a reasonable level.

D. Class-Specific Regularization Module

To alleviate the semantic distribution shift of the background class, we design a regularization module based on the class-specific features. In particular, the classifier cannot properly recognize all known classes due to the foreground class containing only the new classes and the old classes being set as the background class. To solve this problem, we introduce the CSR module that balances foreground and background classes using the saved class-specific features memory. The class-specific features memory includes the class mean and covariance, which is derived from the local semantic features memory. Formally, at step $t > 1$, for each old class $C \in \mathcal{C}^t$, we can compute the class mean (μ_C) and covariance (Σ_C) of it with the local semantic features memory $\mathcal{R}_t$

$$\begin{aligned} \mu_C &= \text{mean}(\mathcal{R}_t^C) \\ \Sigma_C &= \text{cov}(\mathcal{R}_t^C) \end{aligned} \tag{8}$$

where $\mathcal{R}_t^C$ is the set of all the pixel embeddings in set $\mathcal{R}_t$ belonging to class C, mean is the operation of averaging vectors, and cov represents the computation of covariance matrices with some vectors. We can write the mean and the covariance at step t as $\{\mu, \Sigma\}_t$.

For class $C \in \mathcal{C}^{1:t-1}$, it is an old class in step t and we can generate K_C pixel embeddings instances from its class mean and covariance

$$\boldsymbol{p}_C \sim \mathcal{N}(\mu_C, \gamma \Sigma_C) \in \mathbb{R}^m \tag{9}$$

where γ is a nonnegative coefficient. To facilitate input into the semantic segmentation classifier, the generated pixel embedding is copied s times and arranged as a feature map. Formally,

$$\hat{Z}_C = \text{stack}(\text{copy}(\boldsymbol{p}_C, s)) \in \mathbb{R}^{h \times w \times m} \tag{10}$$

where $s = h \times w$. Then, we generate K_C instance in the deep feature space $\mathcal{Z}$.

Notably, the classifier in the network structure is implemented as a convolutional layer instead of a fully connected layer. During the forward propagation, this 1×1 convolutional layer can be regarded as a fully connected layer. Formally, the parameter of classifier g is ϕ which can be formally decomposed into

$$\phi = \left[\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_c\right]^\top, \boldsymbol{\phi}_i \in \mathbb{R}^m \tag{11}$$

where c is the number of all seen foreground classes. Similarly, the bias can be decomposed into $b = [b_0, b_1, \ldots, b_c]^\top$. Therefore, the generated instances of old classes in the deep feature space $\mathcal{Z}$ can be fed to the classifier, and the corresponding cross-entropy loss is

$$\mathcal{L}_{\text{gen}} = \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} \frac{1}{K_C} \sum_{i=1}^{K_C} \frac{1}{s} \sum_{j=1}^{s} -\log\left(\frac{e^{\boldsymbol{\phi}_k^\top \hat{Z}_{k,ij} + b_k}}{\sum_{q=0}^{c} e^{\boldsymbol{\phi}_q^\top \hat{Z}_{q,ij} + b_q}}\right) \tag{12}$$

where c_{old} is the number of total old classes upon step t (include the background class), K_C is the number of the instances of each class we generated which is set to be 128, and $c = c_{\text{old}} - 1 + c_{\text{new}}$ is the number of all seen foreground classes at step t.

Since the term in the last equation, $\mathcal{L}_{\text{gen}}$, is computationally inefficient when K_C and c_{old} are large. Inspired by IL2A [32], we try to optimize an upper bound of this term based on Jensen's inequality. Formally, when $K_C \to \infty$, the term in last equation is

$$\begin{aligned} \mathcal{L}_{\text{gen}} &= \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} \frac{1}{K_C} \sum_{i=1}^{K_C} -\log\left(\frac{e^{\boldsymbol{\phi}_k^\top \boldsymbol{p}_{k,i} + b_k}}{\sum_{q=0}^{c} e^{\boldsymbol{\phi}_q^\top \boldsymbol{p}_{q,i} + b_q}}\right) \\ &= \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} \mathbb{E}_{\boldsymbol{p}_k}\left[\log\left(\sum_{q=0}^{c} e^{(\boldsymbol{\phi}_q^\top - \boldsymbol{\phi}_k^\top)\boldsymbol{p}_k + (b_q - b_k)}\right)\right] \\ &\le \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} \log\left(\mathbb{E}_{\boldsymbol{p}_k}\left[\sum_{q=0}^{c} e^{(\boldsymbol{\phi}_q^\top - \boldsymbol{\phi}_k^\top)\boldsymbol{p}_k + (b_q - b_k)}\right]\right) \\ &= \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} \log\left(\sum_{q=0}^{c} e^{\boldsymbol{v}_{q,k}^\top \mu_k + (b_q - b_k) + \frac{\gamma}{2} \boldsymbol{v}_{q,k}^\top \Sigma_k \boldsymbol{v}_{q,k}}\right). \end{aligned} \tag{13}$$

In above equation, $\boldsymbol{v}_{q,k} = \boldsymbol{\phi}_q - \boldsymbol{\phi}_k$ and we set $\gamma = 2$ in our experiments. We can write the upper bound of $\mathcal{L}_{\text{gen}}$ as a common cross-entropy form

$$\mathcal{L}_{CSR} = \frac{1}{c_{\text{old}}} \sum_{k=1}^{c_{\text{old}}} -\log\left(\frac{e^{\boldsymbol{\phi}_k^\top \boldsymbol{\mu}_k + b_k}}{\sum_{q=0}^{c} e^{\boldsymbol{\phi}_q^\top \mu_k + b_q + \frac{\gamma}{2} \boldsymbol{v}_{q,k}^\top \Sigma_k \boldsymbol{v}_{q,k}}}\right). \tag{14}$$

The mean and covariance matrix of each class are computed and stored simultaneously with the local semantic memory, eliminating the need for further inference steps. In addition, the mean is an m-dimensional vector, while the corresponding variance is an $m \times m$ matrix, so the class-specific features require minimal memory usage.

E. Loss Function

Our method is based on PLOP [45] which applies knowledge distillation to different layers of feature maps. Specifically, for an input $I \in \mathcal{I}^t$ and every layer $l \in \{1, \ldots, L\}$ of the network, the feature maps of the new and old network are $Z_l^t = f_l^t(I)$ and $Z_l^{t-1} = f_l^{t-1}(I)$, respectively. Based on Z_l^t and Z_l^{t-1}, local POD embeddings are collected at several scales and concatenated to be feature vectors $\boldsymbol{Z}_l^t$ and $\boldsymbol{Z}_l^{t-1}$. Therefore, the local POD loss is

$$\mathcal{L}_{\text{pod}} = \frac{1}{L} \sum_{l=1}^{L} \left\| \boldsymbol{Z}_l^t - \boldsymbol{Z}_l^{t-1} \right\|^2. \tag{15}$$

Moreover, PLOP introduces a pseudolabeling strategy which will copy the real label of the foreground classes and, with a certain degree of confidence, select the class labels predicted by the old model. With the pseudolabel P, the classification loss is

$$\mathcal{L}_{\text{pseudo}} = -\frac{\nu}{H \times W} \sum_{i,j}^{H,W} \sum_{C \in \mathcal{C}^t} P(i, j, C) \log \hat{P}(i, j, C) \tag{16}$$

where $\hat{P}$ is the predictions of the current model, H is the height, W is the width, and ν is the ratio of accepted old class pixels over the total number of such pixels.

Algorithm 1 MiSSNet at Step t

Input: f^{t-1}, g^{t-1}, T, $\mathcal{I}_t$, $\mathcal{C}^t$, $\mathcal{C}^{B^t}$
Result: f^t, g^t, R_{t+1}, $\{\mu, \Sigma\}_t$
if $t = 1$ **then**
  $f^1, g^1 \leftarrow \text{FineTune}(f^0, g^0)$
  if $1 < T$ **then**
    $\mathcal{R}_{0:1} \leftarrow \text{SaveSemanticFeature}(f^1, \mathcal{I}_1, \mathcal{C}^1, \mathcal{C}^{B^1})$
    $R_2 \leftarrow \mathcal{R}_0 \cup \mathcal{R}_1$
    $\{\mu, \Sigma\}_1 \leftarrow \text{SaveClassFeature}(\mathcal{R}_0, \mathcal{R}_1)$
  end
else
  Input: R_t, $\{\mu, \Sigma\}_{1:t-1}$, λ, α, β
  Compute $\mathcal{L}_{LSD}$ with R_t
  Compute $\mathcal{L}_{CSR}$ with $\{\mu, \Sigma\}_1, \cdots, \{\mu, \Sigma\}_{t-1}$
  Compute $\mathcal{L}_{pseudo}$, $\mathcal{L}_{\text{pod}}$
  $\mathcal{L} \leftarrow \mathcal{L}_{pseudo} + \lambda\mathcal{L}_{\text{pod}} + \alpha\mathcal{L}_{LSD} + \beta\mathcal{L}_{CSR}$
  $f^t, g^t \leftarrow \text{Update}(f^{t-1}, g^{t-1}; \mathcal{L})$
  if $t < T$ **then**
    $\mathcal{R}_t \leftarrow \text{SaveSemanticFeature}(f^t, \mathcal{I}_t, \mathcal{C}^t)$
    $R_{t+1} \leftarrow R_{t-1} \cup \mathcal{R}_t$
    $\{\mu, \Sigma\}_t \leftarrow \text{SaveClassFeature}(\mathcal{R}_t)$
  end
end

The overall loss of MiSSNet at step $t > 1$ is the following:

$$\mathcal{L} = \mathcal{L}_{\text{pseudo}} + \lambda\mathcal{L}_{\text{pod}} + \alpha\mathcal{L}_{\text{LSD}} + \beta\mathcal{L}_{\text{CSR}} \tag{17}$$

where λ, α, and β are the hyperparameters. With this loss function, we can update the old model and overcome the catastrophic forgetting problem. It is important to emphasize that the two loss functions, $\mathcal{L}_{\text{LSD}}$ and $\mathcal{L}_{\text{CSR}}$, we construct during the training process only involve the utilization of memory and do not add any new parameters to the model. From our practice in Section III, although MiSSNet causes more extra memory consumption when the number of steps increases, the upper of the extra memory size is limited by the total number of classes and the size of the dataset, which ensures that the memory required is kept to about 10–20 MB. The pseudocode at step t is shown in Algorithm 1.

III. Experiment

We introduce the datasets used to test MiSSNet in Section III-A. Next, we cover CISS protocols, implementation details, and baselines in Section III-B. In Sections III-C–III-E, we present the experimental results. Furthermore, we demonstrate the validity of the two proposed modules by showing the results of the ablation experiment. And, the visualization results are presented in Section III-G. We conducted experiments on the effects of hyperparameters in the last of the experimental part. We will publish our code online.[1]

A. Datasets

MiSSNet is evaluated on three segmentation datasets for remote sensing domains, namely, Aeroscapes [56], WHDLD [57], [58], and Mars-seg [59].

[1]https://github.com/Lab-PANbin/

1) Aeroscapes: The Aeroscapes benchmark includes 3269 images captured using commercial drones within an altitude range of 5–50 m, with a uniform image size of 1280×720 pixels. The annotations for these images contain a background category and 11 object categories in total. The categories are listed alphabetically as follows: animal, bike, boat, car, construction, drone, obstacle, person, road, sky, and vegetation. This dataset is novel in terms of viewpoint, scene composition, and object scale compared to other datasets that concentrate solely on indoor scene domains or ground-level views.

2) WHDLD: The WHDLD is a densely labeled dataset suitable for multilabel tasks, such as remote sensing image retrieval (RSIR), classification, and other pixel-based semantic segmentation. The six class categories in this dataset are bare soil, building, road, pavement, vegetation, and water, wherein each image pixel is labeled. The dataset does not contain a background class, and a total of 4940 images, each measuring 256×256 pixels, are included.

3) Mars-seg: The Mars-seg dataset contains rich high-resolution images of Mars scenes, including single-channel grayscale images and RGB images. In order to be consistent with the previous experimental scene, we use all RGB images for the added experiment. The dataset contains 4134 images and nine classes: background, tracks, Martian soil, sands, unknown, bedrock, gravel, shadows, and rocks.

B. Protocol and Implementation Details

1) CISS Protocols: Two different CISS settings are available: *Disjoint* and *Overlapped*. The main difference between these two settings is that they use different data at the same step. In the Disjoint setting, at step t, images include only pixels of old and current classes ($\mathcal{C}^{1:t-1} \cup \mathcal{C}^t$), while the pixels of old and background classes are assigned as the background class of step t. Meanwhile, in the Overlapped setting, at step t, the images include pixels of old, current, and future classes ($\mathcal{C}^{1:t-1} \cup \mathcal{C}^t \cup \mathcal{C}^{t+1:T}$) and the pixels of old, background, and future classes are assigned as the new background class. Since images with future classes cannot be excluded in real-life situations, the overlapped setting is more realistic and challenging. For this reason, we conduct all experiments in the overlapped setting. The training images are only labeled for current classes ($\mathcal{C}^B \cup \mathcal{C}^t$), while the testing images are labeled for all previously seen classes ($\mathcal{C}^B \cup \mathcal{C}^{1:t-1} \cup \mathcal{C}^t$). Following previous works, the principle of our experiment is to have the same number of new categories in different steps as much as possible. We evaluate numerous CISS protocols for each dataset so that the number of new categories is similar across each step. On aeroscapes, the tasks follow two strategies: 1) adding five classes all at once in the new task (6-5, $T = 2$ steps) and 2) adding four classes in step two and three classes in the third step (4s-3, $T = 3$ steps). On WHDLD, there are two strategies: 1) learning three classes in the first step and the following three classes in the second step ($T = 2$ steps) and 2) learning two classes in the first step, followed by two additional steps each having two new classes ($T = 3$ steps). As for the Mars-seg dataset, we divide the eight categories into old categories and new categories for

TABLE I
MEAN IOU (%) ON THE AEROSCAPES DATASET FOR DIFFERENT IL AND CISS METHODS IN THE SETUP OF 6-5

Setting		1-6 (1st step)						7-11 (2nd step)					Overall
Method/Class	Background	Person	Bike	Car	Drone	Boat	Animal	Obstacle	Construction	Vegetation	Road	Sky	mIoU
FT	**73.67**	0.0	0.0	0.0	0.0	0.0	0.0	12.82	71.12	**93.47**	**88.59**	**94.98**	36.29
SI [30]	70.45	0.0	0.0	0.0	0.0	0.0	0.0	13.08	68.98	92.83	85.72	94.17	35.44
EWC [29]	71.13	0.0	0.0	0.0	0.0	0.0	0.0	9.53	**71.14**	92.27	85.41	93.29	35.23
ILT [46]	73.03	39.20	9.34	**87.14**	**60.71**	49.20	49.79	11.59	69.09	91.80	85.43	93.27	59.97
PLOP [45]	72.61	**40.50**	**13.18**	85.25	57.67	21.62	**59.13**	10.56	66.70	91.81	84.03	89.35	57.70
CLWSI [47]	70.27	40.24	12.50	84.82	57.50	55.37	49.61	4.45	67.55	90.57	83.23	91.98	59.01
Ours	72.27	39.82	7.42	86.15	58.06	**60.63**	57.47	**13.69**	67.35	91.93	84.31	90.30	**60.78**
Joint	82.19	43.60	15.46	86.81	57.64	71.70	47.99	15.04	75.40	93.38	93.14	95.37	64.81

TABLE II
MEAN IOU (%) ON THE AEROSCAPES DATASET FOR DIFFERENT IL AND CISS METHODS IN THE SETUP OF 4S-3

Setting		1-4 (1st step)				5-8 (2nd step)				9-11 (3rd step)			Overall
Method/Class	Background	Person	Bike	Car	Drone	Boat	Animal	Obstacle	Construction	Vegetation	Road	Sky	mIoU
FT	59.17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	91.00	85.87	**94.82**	27.57
SI [30]	60.22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**92.33**	**86.02**	94.16	27.73
EWC [29]	54.86	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	88.57	83.93	93.48	26.74
ILT [46]	59.76	35.67	7.68	83.95	27.80	0.25	38.08	2.43	58.18	86.42	82.86	91.03	47.84
PLOP [45]	64.53	45.59	**18.75**	86.00	**61.49**	0.0	19.57	6.93	**69.61**	88.56	82.81	91.41	52.94
CLWSI [47]	60.90	36.61	10.35	84.45	14.70	0.0	17.46	2.26	64.14	86.97	83.79	93.05	46.22
Ours	**64.78**	**49.27**	16.48	**86.12**	54.85	**19.75**	**38.96**	**7.25**	69.03	88.72	83.33	89.68	**55.69**
Joint	82.19	43.60	15.46	86.81	57.64	71.70	47.99	15.04	75.40	93.38	93.14	95.37	64.81

two steps, namely, 4s. As in previous cases, we report final results in mean Intersection over Union (mIoU).

2) Implementation Details: The experiments are conducted using PyTorch and a single Nvidia RTX 3090 GPU for both training and testing. The ResNet-101 pretrained on ImageNet is employed as the backbone network, and features are decoded using the Deeplab-v3 [55] architecture. In order to verify the complexity of the model, we calculated floating point operations (FLOPs) and the number of model parameters with $4 \times 3 \times 512 \times 512$ input shape. Among them, FLOPs is 541.81 GB and the number of model parameters is 57.93 MB. To better adapt the pretrained network to the new dataset, we execute a larger epoch at the initial step of each experiment, followed by smaller epochs for subsequent steps to conserve computing resources in certain experiments. Specifically, for Aeroscapes, WHDLD, and Mars-seg datasets, we designate 100, 60, and 60 epochs, respectively, for the first step, while for tasks 6-5, 4s-3, 3s, 2s, and 4s, we set 60, 90, 30, 60, and 60 epochs for subsequent steps, respectively. The model is trained using a batch size of 4, while the input image is cropped to fit the 512×512 input dimension. In terms of hyperparameters, α is set to 0.1, whereas β is set to 0.25. We follow the PLOP [45] guidelines for other experimental setups, such as γ. And, we set the initial learning rate of $1e-2$ for the first CISS step and $1e-3$ for all the following ones.

3) Baselines: We benchmark our model against the latest state-of-the-art CISS methods, namely, incremental learning techniques (ILT) [46], PLOP [45], and continual learning with structured inheritance (CLWSI) [47]. Moreover, we evaluate general incremental learning models based on Elastic Weight Consolidation (EWC) [29] and synaptic intelligence (SI) [30] for prior-focused methods. EWC makes use of the empirical Fisher matrix to calculate the importance of each parameter for old classes, while SI employs the learning trajectory. Finally, in order to show the performance of each model more clearly, we add the "joint" experiment; that is, the precision of the semantic segmentation task under full supervision. This is also considered a performance ceiling for incremental learning.

C. Experimental Results on the Aeroscapes Dataset

We conduct comparative experiments on the Aeroscapes dataset regarding the 6-5 task and 4s-3 task, and the results are shown in Tables I and II. Both tables list the experimental results of different incremental learning methods and CISS methods under various experimental settings, and the best results among the benchmark methods are highlighted in bold.

Specifically, we perform testing after completing all training steps and the categories used for training are arranged in the default order of the dataset. The specific values in the two tables include test IoU values for each class and mIoU for all classes, where the results for each class are listed in the order of the steps that introduced them during training.

For the 6-5 experimental setup, consisting of two learning steps with similar numbers of classes, we first learn the top six classes and then the remaining five classes. Traditional incremental learning algorithms cannot alleviate the catastrophic forgetting issue in the semantic segmentation task, as seen in the first three columns of Table I. They have no advantage over the most basic fine tuning (FT) for some classes. This outcome is primarily because traditional class increment learning algorithms are typically used for picture classification tasks rather than pixel-level tasks. They tend not to design specialized and efficient structures or regularizations that capture spatial or pixel-level information. Therefore, even if EWC and SI have well-designed regularization terms, they will not perform well. On the other hand, CISS algorithms are efficient at incremental learning due to specialized design, and they outperform standard IL baselines. Although in some classes, ILT, PLOP, and CLWSI have better IoU values than our method, our approach yields an mIoU of about 1% over ILT, 3% over PLOP, and 1% over CLWSI.

Regarding the 4s-3 experimental setup, where we obtain data of four classes for the first two steps, respectively, followed by learning the remaining three classes, Table II reflects that the traditional incremental learning paradigm does not help ease catastrophic forgetting issues, similar to the previous task. Notably, our proposed approach is significantly superior to the CISS paradigms of ILT, PLOP, and CLWSI. Furthermore, the results obtained using our method are not only superior to other CISS methods in terms of mIoU values but also in terms of IoU values for most classes, particularly for the boat category.

As the number of steps increases, the benefit from that MiSSNet only saves the memory of the background class in the first step and utilizes it in the following steps, MiSSNet can maintain better discrimination ability of the real background class. LSD and CSR are effective in keeping information of the real old categories because with the progress of incremental learning, the knowledge of the unknown categories stored by the memory module is gradually learned by the model, and the existing knowledge is overwritten.

D. Experimental Results on the WHDLD Dataset

The experimental results for the WHDLD dataset in the 3s and 2s settings are presented in Tables III and IV, respectively. Like the Aeroscapes dataset experiment, the results for each class are arranged in order, with IoU reported for each and mIoU reported at last. It is worth noting that there are no results for the background class since it is not included in the ground truth for WHDLD. Regrettably, conventional incremental learning methods are ineffective, even for datasets with a small number of classes.

The designed memory module, LSD module, and CSR module are mainly aimed at maintaining the classification ability of old categories, and our proposed MiSSNet maintains superiority in most of the old categories. Moreover, because LSD constructs a correlation matrix between all new and old categories, which is conducive to the balance of new and old information, MiSSNet also has higher IoU on many new categories. On the other hand, CSR does not impose any constraints on the learning of new categories, and it can release the learning ability of the model.

Our approach is only slightly superior to ILT, PLOP, and CLWSI when three classes are added in a single step (3s), and it remains superior to ILT, PLOP, and CLWSI in most new classes. This further illustrates the effectiveness of our approach in mitigating the problem of semantic distribution shift of the background class and catastrophic forgetting.

In the 2s setting, which is more challenging, most methods exhibit a reduced mIoU value, whereas our approach yields better results. This can be attributed to our approach's superiority in not only retaining past knowledge but also being capable of assimilating new knowledge compared with ILT, PLOP, and CLWSI. Therefore, our approach offers a more robust performance across all the learned classes and can tackle more complex tasks.

TABLE III
MEAN IOU (%) ON THE WHDLD DATASET FOR DIFFERENT IL AND CISS METHODS IN THE SETUP OF 3S

Settings	1-3 (1st step)			4-6 (2nd step)			Overall
Method/Class	building	road	pavement	vegetation	bare soil	water	mIoU
FT	0.0	0.0	0.0	79.73	39.11	**93.14**	35.33
SI [30]	0.0	0.0	0.0	**79.76**	39.57	93.09	35.40
EWC [29]	0.0	0.0	0.0	78.48	40.57	91.25	35.05
ILT [46]	56.36	61.07	38.79	77.90	40.68	91.36	61.03
PLOP [45]	56.06	60.84	**40.60**	78.33	38.48	91.43	60.96
CLWSI [47]	56.08	61.11	39.79	78.00	40.38	91.04	61.07
Ours	**56.58**	**61.31**	40.29	78.32	**41.22**	91.28	**61.50**
Joint	58.25	62.68	42.45	80.09	41.38	93.63	63.08

TABLE IV
MEAN IOU (%) ON THE WHDLD DATASET FOR DIFFERENT IL AND CISS METHODS IN THE SETUP OF 2S

Settings	1-2 (1st step)		3-4 (2nd step)		5-6 (3rd step)		Overall
Method/Class	building	road	pavement	vegetation	bare soil	water	mIoU
FT	0.0	0.0	0.0	0.0	**41.92**	**93.67**	22.60
SI [30]	0.0	0.0	0.0	0.0	40.92	93.52	22.41
EWC [29]	0.0	0.0	0.0	0.0	35.56	92.35	21.32
ILT [46]	53.18	57.70	40.10	76.90	40.51	90.85	59.87
PLOP [45]	**55.96**	60.56	39.74	69.61	36.85	88.59	58.55
CLWSI [47]	52.82	58.81	**40.74**	78.12	41.85	91.91	60.71
Ours	55.15	**61.15**	40.64	**78.29**	41.77	92.21	**61.53**
Joint	58.25	62.68	42.45	80.09	41.38	93.63	63.08

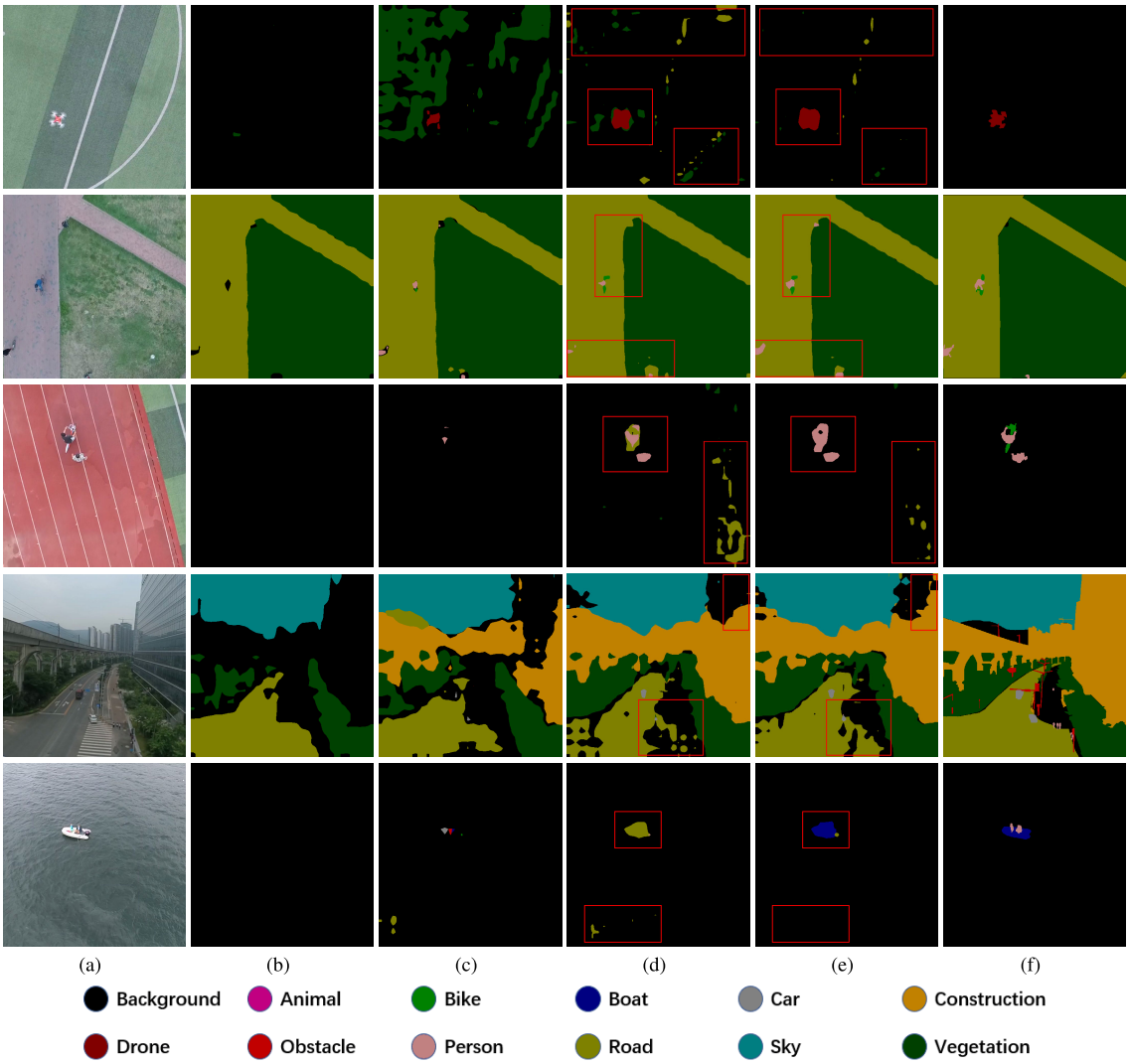

Fig. 2. Visualization of incremental semantic segmentation results on the 4s-3 (three steps) setting of the Aeroscapes dataset. Different columns of photographs from left to right indicate (a) input image, (b) EWC, (c) ILT, (d) PLOP, (e) ours, and (f) GT. The categories represented by the colors in this figure are shown in the legend. The red boxes indicate the important areas, where MiSSNet and PLOP differ significantly.

E. Experimental Results on the Mars-Seg Dataset

Our experiments on Mars-seg datasets also demonstrate the superiority of our algorithm. The Mars-seg dataset changes the perspective of remote sensing images from the traditional Earth to the universe, which brings more possibilities for the exploration of the universe. At the same time, our validation on this dataset also demonstrates the prospect of CISS algorithms. Furthermore, it also represents that our proposed algorithm MiSSNet can be applied in more application fields. The results of 4s experiment on the Mars-seg dataset are shown in Table V. The CLWSI achieves high accuracy on new categories. However, the performance of the proposed MiSSNet algorithm still has certain advantages on all metrics except the new categories of the second step.

TABLE V
MEAN IOU (%) ON THE MARS-SEG DATASET FOR DIFFERENT IL AND CISS METHODS IN THE SETUP OF 4S

Method/Class	Background	1-4 (1st step)	5-8 (2nd step)	mIoU
FT	19.55	0.0	61.71	29.60
SI [30]	19.84	0.0	60.74	29.20
EWC [29]	19.49	0.0	54.49	26.38
ILT [46]	52.33	52.89	57.21	54.75
PLOP [45]	82.10	54.21	49.21	55.09
CLWSI [47]	70.79	47.72	**58.76**	55.19
Ours	**82.51**	**55.77**	49.56	**55.98**
Joint	86.52	65.82	60.42	65.72

F. Ablation Experiment

In this section, we demonstrate the contributions of each component of our model through ablation experiments under the 6-5 experimental setting of the Aeroscapes dataset. First, we perform experiments on PLOP, which employs strip pooling and employs knowledge distillation in intermediate layers.

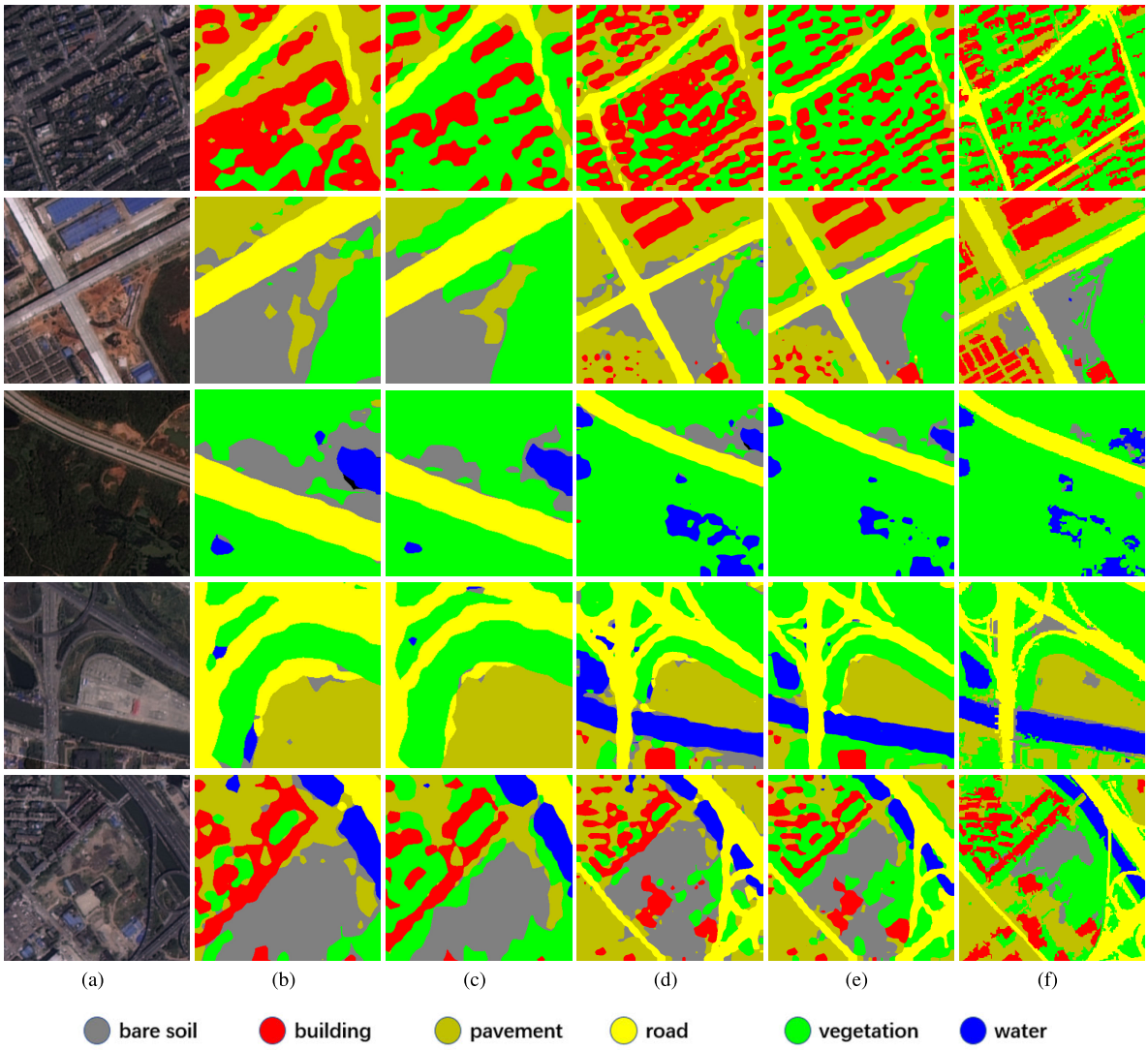

Fig. 3. Visualization of incremental semantic segmentation results on the 2s (three steps) setting of the WHDLD dataset. Different columns of photographs from left to right indicate (a) input image, (b) PLOP-local, (c) ours-local, (d) PLOP, (e) ours, and (f) GT. The categories represented by the colors in this figure are shown in the legend, and the black represents pixels that have been misclassified into nonexistent background classes.

TABLE VI
RESULT (%) OF ABLATION EXPERIMENT ON 6-5 SETTING

Method/Class	Background	1-6 (1st step)	7-11 (2nd step)	mIoU
baseline	**72.61**	46.22	68.49	57.70
+ LSD	72.58	51.39	68.05	60.10
+ CSR	72.18	**51.72**	68.69	60.50
+ LSD + CSR	72.27	51.59	**69.52**	**60.78**

Followed by this, the LSD module and the CSR module are incrementally included and evaluated.

As Table VI indicates, the results of the ablation on different components of our approach are presented, with IoU of the background class and mIoU for the classes belonging to different steps. The last column of the table demonstrates the mIoU of all classes after all training steps. The baseline model used here is PLOP, and the results are presented in the first row of Table VI. In the subsequent rows, we added two additional components of our approach onto the baseline, assessing the performance effect of each component. LSD and CSR may unintentionally impair the ability to distinguish real background classes; however, as depicted in the second column of Table VI, they can effectively enhance the ability to solve the catastrophic forgetting problem. Furthermore, our approach can also manage to improve the ability to learn new tasks when the two modules work together, as shown in the third column of the table.

LSD benefits from maintaining the correlation between the pixels of new and old classes, which can increase the model's classification accuracy for old classes while preserving the discrimination ability of the real background class. However, this approach limits the capacity to discriminate against new classes to some extent. On the other hand, CSR is designed to further manage the semantic distribution shift of the background class in CISS. As mentioned, in the incremental learning setting, the background classes of each step are always different from the real background class,

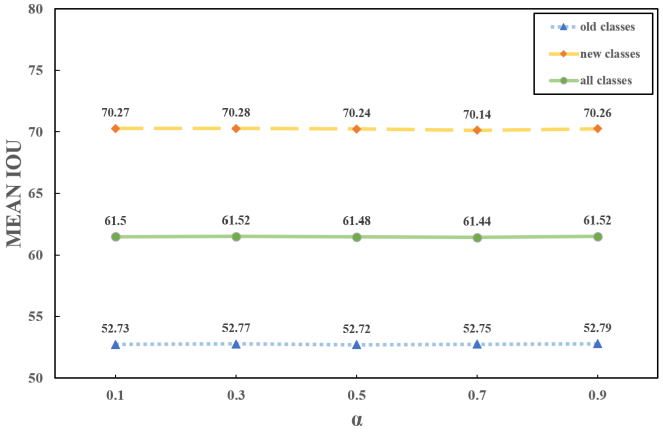

Fig. 4. Experimental results of parameter sensitivity in 3s settings of the WHDLD dataset. The pink line represents the average IoU of all the old categories that belong to the first step, and the green line represents the average IoU of all the new categories that belong to the second step. The black line shows the average IoU for all categories.

leading to a reduction in the classification accuracy of the real background class. Nevertheless, CSR significantly improves the segmentation accuracy of the model for classes other than the real background class. Taken together, LSD and CSR complement each other and ultimately improve our algorithm compared with PLOP.

G. Visual Results of Examples and Analysis

The final segmentation results of five example images are shown in Fig. 2. These input images are from the test dataset of Aeroscapes, and the semantic images are acquired after completing all training in the experimental setting of 4s-3. We show the results after parameter selection on the validation data. Our approach has obvious advantages for the semantic segmentation of smaller objects, such as boats, drones, and people. At the same time, our method is superior to other methods in terms of completeness and accuracy of segmented objects, and it is less likely to misclassify pixel points in large areas.

Our algorithm falls short in addressing the limitations of other methods. Compared with ground truth, it still has difficulty identifying classes that are very similar to the background or other categories, primarily in complex environments. These categories, such as obstacles and bikes, usually occupy only a small percentage of pixels in an image, further exacerbating the problem of misclassification or identification difficulties.

More visualization results of task 2s are presented in Fig. 3. To better demonstrate the advantages of our proposed algorithm on images with rich semantic information, we present the results of two algorithms, PLOP and MiSSNet, in this visualization image. And, the second and third columns are the locally enlarged images of the two algorithms. The results show the advanced nature of our algorithm, which can better segment detailed objects, and various shapes of the segmentation graph are more distinct.

H. Parameter Sensitivity

There are three hyperparameters in our loss function, and we keep the first hyperparameter γ consistent with PLOP. We adjust the third hyperparameter β to balance this loss with the local pod loss. As for the second hyperparameter α, to be fair, we have set it empirically to 0.1 in all comparison experiments.

In order to verify the influence of hyperparameters on our proposed model, we perform parameter sensitivity experiments on the second hyperparameter α. Especially, we divided the training set of the WHDLD dataset into a new training set and a validation set according to 4:1. In the 3s task scenario, we set the parameter values as 0.1, 0.3, 0.5, 0.7, and 0.9, respectively, and drew a line graph for the experimental results. The results are shown in Fig. 4. From the experimental results, our algorithm is not sensitive to the hyperparameters.

IV. Conclusion

We propose a memory-inspired CISS approach for remote sensing images by utilizing information saved and processed from old data, combining the distillation and regularization strategies. The saved memory includes the local semantic features and the class-specific features. Our method constructs a correlation matrix with local semantic features at the pixel level for distillation, greatly alleviating semantic distribution shift concerns of the background class. We further address the semantic distribution shift problem of the background class by using the class-specific features produced from the local semantic features processing for regulation, enabling the network to concentrate on both learned and new classes. Through extensive evaluation of various datasets (Aeroscapes, WHDLD, and Mars-seg) and settings (5 in total), our approach demonstrates a good performance. Our proposed algorithm provides more possibilities for the challenging semantic segmentation problems of remote sensing images, especially in the fields of urban planning, commercial drones, and space exploration. In the future, the research focus will be on developing incremental learning methods that store and utilize memory more efficiently, semantic segmentation methods coupled with weak supervision for incremental learning, and segmentation tasks related to domain generalization or domain adaptation.

References

[1] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[2] Y. Li, X. Lyu, A. C. Frery, and P. Ren, "Oil spill detection with multiscale conditional adversarial networks with small-data training," *Remote Sens.*, vol. 13, no. 12, p. 2378, Jun. 2021.

[3] J. Li, S. Du, R. Song, Y. Li, and Q. Du, "Progressive spatial information-guided deep aggregation convolutional network for hyperspectral spectral super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–15, Oct. 2023, doi: 10.1109/TNNLS.2023.3325682.

[4] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.

[5] X. Zheng, X. Chen, X. Lu, and B. Sun, "Unsupervised change detection by cross-resolution difference learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3079907.

[6] H. Zhou, F. Luo, H. Zhuang, Z. Weng, X. Gong, and Z. Lin, "Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508614.

[7] T. Guo, R. Wang, F. Luo, X. Gong, L. Zhang, and X. Gao, "Dual-view spectral and global spatial feature fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512913.

[8] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, "Multiscale diff-changed feature fusion network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3241097.

[9] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[10] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[11] W. Jing, Y. Yuan, and Q. Wang, "Dual-field-of-view context aggregation and boundary perception for airport runway extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702412.

[12] D. Zhao, C. Wang, Y. Gao, Z. Shi, and F. Xie, "Semantic segmentation of remote sensing image based on regional self-attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[13] J. Hou, Z. Guo, Y. Feng, Y. Wu, and W. Diao, "SPANet: Spatial adaptive convolution based content-aware network for aerial image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2192–2204, 2023.

[14] B. Li, P. Lv, Y. Zhong, and L. Zhang, "High resolution remote sensing image semantic segmentation based on ultra-lightweight fully convolution neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 3175–3178.

[15] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi, and M. He, "Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 7419–7434, 2022.

[16] R. Xiao, C. Zhong, W. Zeng, M. Cheng, and C. Wang, "Novel convolutions for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3265752.

[17] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, Mar. 2018.

[18] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[19] M. Yousefhussien, D. J. Kelbe, E. J. Ientilucci, and C. Salvaggio, "A multi-scale fully convolutional network for semantic labeling of 3D point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 191–204, Sep. 2018.

[20] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3256064.

[21] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[22] J. Zhang, Y. Liu, P. Wu, Z. Shi, and B. Pan, "Mining cross-domain structure affinity for refined building segmentation in weakly supervised constraints," *Remote Sens.*, vol. 14, no. 5, p. 1227, Mar. 2022.

[23] L. Bergamasco, F. Bovolo, and L. Bruzzone, "A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2147–2162, 2023.

[24] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[25] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 3–14, Sep. 2018.

[26] X. Pan, J. Zhao, and J. Xu, "Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7854–7870, Sep. 2021.

[27] R. French, "Catastrophic forgetting in connectionist networks," *Trends Cognit. Sci.*, vol. 3, no. 4, pp. 128–135, Apr. 1999.

[28] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.

[29] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. India A, Phys. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[30] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.

[31] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proc. Comput. Vis. (ECCV)*, 2018, pp. 556–572.

[32] F. Zhu, Z. Cheng, X.-y. Zhang, and C.-l. Liu, "Class-incremental learning via dual augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14306–14318.

[33] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Comput. Vis. (ECCV)*, 2018, pp. 144–161.

[34] M. Kang, J. Park, and B. Han, "Class-incremental learning by knowledge distillation with adaptive feature consolidation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16050–16059.

[35] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5533–5542.

[36] Q. Gu, D. Shim, and F. Shkurti, "Preserving linear separability in continual learning by backward feature projection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24286–24295.

[37] W. Liu, X. Nie, B. Zhang, and X. Sun, "Incremental learning with open-set recognition for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3173995.

[38] Y. Tai, Y. Tan, S. Xiong, and J. Tian, "Mine-Distill-Prototypes for complete few-shot class-incremental learning in image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5206013.

[39] S. D. Bhat, B. Banerjee, S. Chaudhuri, and A. Bhattacharya, "CILEA-NET: Curriculum-based incremental learning framework for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5879–5890, May 2021, doi: 10.1109/JSTARS.2021.3084408.

[40] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3102629.

[41] S. Dang, Z. Cao, Z. Cui, Y. Pi, and N. Liu, "Class boundary exemplar selection based incremental learning for automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5782–5792, Aug. 2020.

[42] L. Wang, X. Yang, H. Tan, X. Bai, and F. Zhou, "Few-shot class-incremental SAR target recognition based on hierarchical embedding and incremental evolutionary network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3248040.

[43] S. Dang, Z. Cao, Z. Cui, Y. Pi, and N. Liu, "Open set incremental learning for automatic target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4445–4456, Jan. 2019.

[44] J. Chen, S. Wang, L. Chen, H. Cai, and Y. Qian, "Incremental detection of remote sensing objects with feature pyramid and knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3042554.

[45] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "PLOP: Learning without forgetting for continual semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4039–4049.

[46] U. Michieli and P. Zanuttigh, "Incremental learning techniques for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3205–3212.

[47] Y. Feng, X. Sun, W. Diao, J. Li, X. Gao, and K. Fu, "Continual learning with structured inheritance for semantic segmentation in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3076664.

[48] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3524–3537, Sep. 2019.

[49] L. Shan, W. Wang, K. Lv, and B. Luo, "Class-incremental learning for semantic segmentation in aerial imagery via distillation in all aspects," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3135456.

[50] X. Rong, X. Sun, W. Diao, P. Wang, Z. Yuan, and H. Wang, "Historical information-guided class-incremental semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3170349.
[51] J. Li et al., "Class-incremental learning network for small objects enhancing of semantic segmentation in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3124303.
[52] N. Yang and H. Tang, "GeoBoost: An incremental deep learning approach toward global mapping of buildings from VHR remote sensing images," *Remote Sens.*, vol. 12, no. 11, p. 1794, Jun. 2020.
[53] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9230–9239.
[54] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12309–12318.
[55] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[56] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1499–1508.
[57] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, p. 964, Jun. 2018.
[58] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328 Jan. 2020, doi: 10.1109/JSTARS.2019.2961634.
[59] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 3152587.

Jiajun Xie received the B.S. degree from the School of Mathematics, Sichuan University, Chengdu, China, in 2022. He is currently pursuing the M.S. degree with the School of Statistics and Data Science, Nankai University, Tianjin, China.

His research interests include deep learning, incremental learning, and semantic segmentation.

Bin Pan (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2013 and 2019, respectively.

Since 2019, he has been an Associate Professor with the School of Statistics and Data Science, Nankai University, Tianjin, China. He has authored or coauthored over 40 scientific articles. His research interests include cross-domain remote sensing image processing, hyperspectral image classification, unmixing, and super resolution. Codes for his articles are published online https://github.com/Lab-PANbin/.

Dr. Pan serves as an Associate Editor for the *Infrared Physics and Technology*.

Xia Xu received the B.S. and M.S. degrees from the School of Electrical Engineering, Yanshan University, Qinhuangdao, China, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019.

She is currently an Assistant Professor with the College of Computer Science, Nankai University, Tianjin, China. Her research interests include hyperspectral unmixing, multiobjective optimization, and remote sensing image processing

Zhenwei Shi (Senior Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 200 scientific articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the *Pattern Recognition*, the *ISPRS Journal of Photogrammetry and Remote Sensing*, and the *Infrared Physics and Technology*.