SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

August 2025, Vol. 68, Iss. 8, 182301:1–182301:16
https://doi.org/10.1007/s11432-024-4307-1

Mitigating representation bias for class-incremental semantic segmentation of remote sensing images

Xiaoqian SUN†, Xingxing WENG†, Chao PANG & Gui-Song XIA*

School of Computer Science, Wuhan University, Wuhan 430072, China

Received 15 July 2024/Revised 7 October 2024/Accepted 10 February 2025/Published online 4 July 2025

Abstract Class-incremental semantic segmentation methods for remote sensing images primarily rely on knowledge distillation to mitigate catastrophic forgetting. These methods overlook (1) the representation bias stemming from the conflict between cross-entropy loss and distillation loss, and (2) issues of interstep imbalance and interstep similarity that become particularly pronounced after multiple incremental learning steps, thus limiting their performance. To address these challenges, we propose MiR (mitigating representation bias), a novel framework that alleviates forgetting and facilitates learning new classes effectively. MiR replaces the traditional feature-classifier mode with the feature-SegToken interaction, leveraging implicit coarse-to-fine segmentation to mitigate representation bias. Furthermore, a weighting strategy is introduced to adaptively adjust cross-entropy and distillation losses, effectively tackling issues of interstep imbalance and interstep similarity. Extensive experiments conducted on the DeepGlobe, Potsdam, and Vaihingen datasets demonstrate the effectiveness of MiR in learning new knowledge while retaining old knowledge.

Keywords remote sensing image, semantic segmentation, class-incremental learning, knowledge distillation, representation bias

Citation Sun X Q, Weng X X, Pang C, et al. Mitigating representation bias for class-incremental semantic segmentation of remote sensing images. Sci China Inf Sci, 2025, 68(8): 182301, https://doi.org/10.1007/s11432-024-4307-1

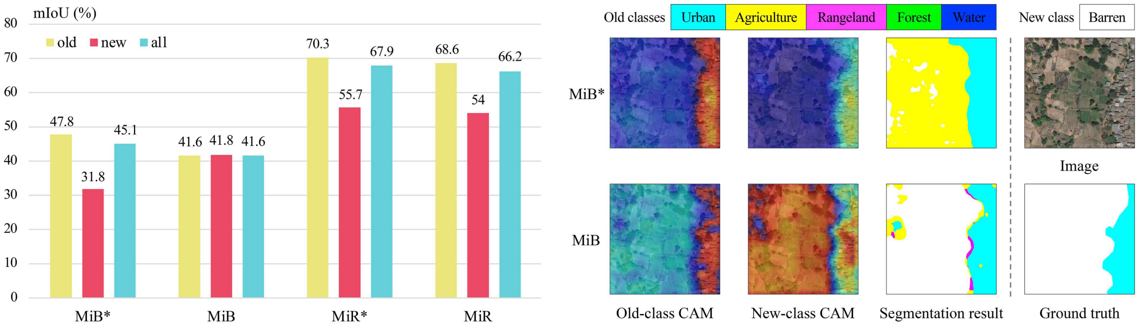
1 Introduction

Semantic segmentation of remote sensing images, which aims to categorize each pixel into a specific class or object, plays an important role in many applications such as natural resources survey [1] and urban planning [2]. With the remarkable advancements in deep learning, significant progress has been achieved in this field [3]. Nonetheless, a notable limitation persists in most semantic segmentation models: their inability to continually receive and learn new knowledge while retaining previously learned knowledge. This seriously hinders the widespread application of the remote-sensing segmentation model. As new images are acquired from diverse geographical locations, the need arises to identify new categories of land cover in addition to the preset classes. Therefore, segmentation models should be capable of continuously learning knowledge about new categories while preserving existing knowledge of previous classes.

A naive solution to achieve the goal is to fine-tune segmentation models using both previously and newly collected images. However, factors such as data privacy and the high cost of data storage, often result in the unavailability of previous remote sensing images. Fine-tuning segmentation models only on new images would overwrite model weights associated with previous classes, leading to significant performance degradation. This phenomenon is known as catastrophic forgetting [4]. To address the issue of forgetting, recent efforts have been dedicated to class-incremental semantic segmentation of remote sensing images [5–11].

Inspired by class-incremental image classification [12], knowledge distillation [13] is employed as a fundamental technique to maintain the predictions of the segmentation model on previous classes when updating its parameters with new images of novel categories. Building on this foundation, existing methods further enhance the protection of old-class knowledge by implementing designs such as adopting contrastive distillation [5], utilizing cross-image relationship distillation [9], trying multi-scale feature

* Corresponding author (email: guisong.xia@whu.edu.cn)
† These authors contributed equally to this work.



(a) (b)

Figure 1 (Color online) Illustration for representation bias. (a) Results (mIoU in %) on the DeepGlobe dataset (1-1s). (b) Visualization of class activation maps (CAM) and segmentation results extracted from MiB [15], applying different hyper-parameter settings on the DeepGlobe dataset (1-1s). Old, new, and all refer to mIoU computed on old, new, and all classes, respectively. * indicates models trained with $\lambda_{\rm kd}=10$, while others are trained with $\lambda_{\rm kd}=5$. Adjusting the hyperparameter $\lambda_{\rm kd}$ from 10 to 5 leads to a 10% performance improvement on new classes, suggesting that MiB, using the default parameter ($\lambda_{\rm kd}=10$), tends to maintain old classes. Furthermore, MiB ($\lambda_{\rm kd}=10$) can produce high responses in old-class regions but struggles in new-class regions. By reducing $\lambda_{\rm kd}$, MiB's tendency to retain old classes is alleviated, allowing it to generate sufficient responses in new-class regions. Hence, we conclude that the competition between cross-entropy and distillation losses leads to the issue of representation bias.

distillation [10] or combining with old-class feature memory replay [11, 14]. Despite the remarkable progress achieved by these methods, they overlook several challenges, resulting in limited performance.

Specifically, current methods typically employ cross-entropy loss and knowledge distillation loss to jointly optimize semantic segmentation models. Handpicked hyperparameters $\lambda_{\rm ce}$ and $\lambda_{\rm kd}$ are used to balance these loss terms, hoping that models can make a trade-off between learning new classes and not forgetting old classes. However, due to the lack of theoretical guarantees or standards, achieving this trade-off is challenging. Consequently, the model faces a dilemma in representation learning: it tends to bias towards either new or old classes. If biased towards new classes, there is a risk of forgetting old classes. Conversely, bias towards old classes hinders the model's ability to learn new ones. This phenomenon is termed representation bias in this paper. Figure 1(a) presents quantitative results (mIoU in %) of MiB [15] method applying different hyperparameter settings on the DeepGlobe dataset (1-1s) [16]. Adjusting the hyperparameter $\lambda_{\rm kd}$ from 10 to 5 leads to a 10% performance improvement on new classes, suggesting that MiB, using the default parameter ($\lambda_{\rm kd}=10$), tends to maintain old classes. Additionally, comparing the ground truth with class activation maps in Figure 1(b), it is evident that MiB ($\lambda_{\rm kd}=10$) can produce high responses in old-class regions but struggles in new-class regions, resulting in misclassification of new classes as old ones. By reducing $\lambda_{\rm kd}$, the model's tendency to retain old classes is alleviated, enabling it to generate sufficient responses in new-class regions. Therefore, we conclude that the competition between cross-entropy and distillation losses leads to representation bias.

In addition to representation bias, interstep imbalance and interstep similarity present challenges for class-incremental semantic segmentation of remote sensing images using knowledge distillation. After performing multiple incremental learning steps, the data volume and pixel proportion of new classes are typically lower than those of old classes, resulting in data imbalance and foreground-background imbalance. Interstep imbalance can make the model focus on old classes while overlooking new ones. Moreover, as the number of learned classes increases, visually similar classes (e.g., meadow and forest) may exist across different incremental steps. When new and old classes exhibit high similarity, the model faces challenges in assigning accurate labels. The cross-entropy loss guides the model to assign labels of new classes, while the distillation loss forces it to predict old ones. This confusion not only hinders the learning of new classes but also decreases the performance of old classes.

To tackle these challenges, we propose a novel class-incremental semantic segmentation framework for remote sensing images, named MiR (mitigating representation bias). Drawing inspiration from recent work on supervised semantic segmentation [17], MiR incorporates tokens to model category-wise information of both old and new classes, termed SegToken. Subsequently, it performs multiple interactions between image features and SegToken. These interactions progressively update category-wise information according to pixel features while leveraging the updated category-wise information to mine more feature regions corresponding to each semantic category. The SegToken derived from the final interaction serves as the segmentation results. Through such a design, MiR effectively mitigates representation bias. In

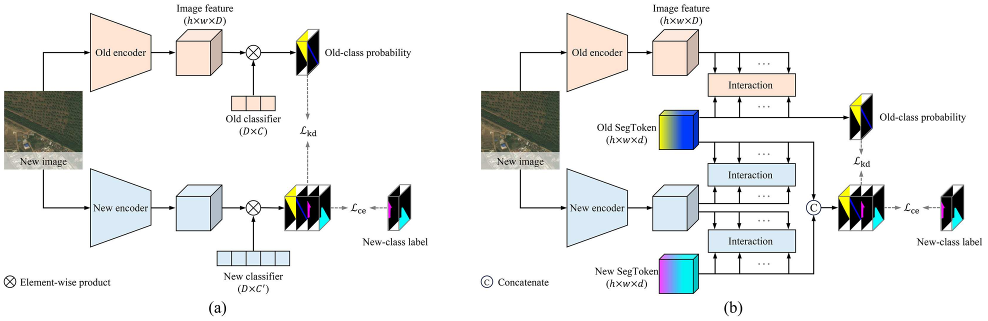


Figure 2 (Color online) Framework comparison between prior methods based on knowledge distillation (a), and our proposed MiR (b) for class-incremental semantic segmentation of remote sensing images. $\mathcal{L}_{\rm ce}$, $\mathcal{L}_{\rm kd}$ are cross-entropy loss and distillation loss, respectively. The SegToken is a learnable tensor to model category-wise information. MiR replaces the typical feature-classifier mode with the feature-SegToken interaction mode, aiming to utilize implicit coarse-to-fine segmentation to mitigate representation bias stemming from the competition between cross-entropy and distillation losses.

cases where the model is biased towards new classes, image features may have weak associations with old classes. The interactions between image features and SegToken are similar to achieve coarse segmentation using general information, and subsequently refine the segmentation using customized information, thus alleviating the forgetting issue induced by representation bias. Conversely, the features may have weak associations with new classes. The coarse-to-fine segmentation can also facilitate the learning of new classes. Figure 2 compares the class-incremental semantic segmentation framework between our work and prior methods.

Furthermore, MiR introduces a simple yet effective weighting strategy to address the issues of interstep imbalance and interstep similarity. It adaptively increases the cross-entropy loss according to the current loss of new class pixels at each training iteration. This ensures that the model pays attention to the smaller number of new classes. Since the negative effect of interstep similarity becomes prominent after multiple learning steps, MiR weights the knowledge distillation loss depending on the confidence in old knowledge, at the last learning step. This weighting strategy improves both stability (not forgetting old classes) and plasticity (effectively learning new classes).

The proposed MiR has been evaluated on three commonly used datasets for class-incremental semantic segmentation of remote sensing images. Compared to state-of-the-art methods, such as MiSSNet [11] and CoMFormer [18], MiR improves over the best baseline by 10.3% in the most challenging setting of the DeepGlobe dataset (1-1s) [16], in terms of mIoU computed on all classes. On the Potsdam dataset (2-3)[1] and the Vaihingen dataset (2-2-1)[2], MiR approaches the upper bounds set by joint training, with gaps of −0.1% and 1.1%, respectively. To summarize, the contributions of this paper are as follows.

(1) We present a novel class-incremental semantic segmentation framework for remote sensing images, dubbed MiR, which provides strong performance baselines on three widely used datasets: DeepGlobe, Potsdam, and Vaihingen.

(2) We propose a feature-SegToken interaction mode to replace the traditional feature-classifier mode. This design leverages implicit coarse-to-fine segmentation to mitigate representation bias stemming from the competition between cross-entropy and distillation losses.

(3) We introduce a simple yet effective weighting strategy for class-incremental semantic segmentation based on knowledge distillation. This strategy adaptively adjusts cross-entropy and distillation losses, thereby alleviating issues of interstep imbalance and interstep similarity that arise after multiple incremental steps.

2 Related work

2.1 Deep learning-based semantic segmentation of remote sensing images

Deep learning-based semantic segmentation of remote sensing images has been a research hotspot in the field of remote sensing information mining. Due to their distinctive imaging perspectives and large field of view, remote sensing images present characteristics such as complex backgrounds, rich content, and

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx.
[2] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.

varied object sizes. These attributes pose challenges to popular semantic segmentation networks (e.g., FCN [19] and Deeplab-v3 [20]) in computer vision. Consequently, numerous studies have emerged aiming at developing accurate semantic segmentation models tailored for remote sensing images. Recognizing these unique characteristics, many studies explore innovative feature extraction strategies, or network architectures to obtain more discriminative feature representations.

For instance, addressing the challenge of varied object sizes inspires research into multi-scale feature extraction methods. These include dense connections [21], pyramid structures [22–24], attention mechanisms [25,26], and size-aware sub-segmentation networks [27]. Global contextual information is beneficial to object recognition within complex scenes. Transformers, known for their proficiency in modeling long-range dependencies, have inspired a series of segmentation models utilizing transformer architectures, such as GlOTS [28], STDSNet [29], and MMT [30]. Traditional transformers, however, typically model long-range dependencies at a single scale, potentially struggling with varied object sizes. To address this, some studies explore hierarchical transformer designs to extract multi-scale representations effectively [31]. Furthermore, efforts have been made to integrate the strengths of both convolutional neural networks (CNNs) and transformers, with studies combining global and local features to enhance the performance of semantic segmentation for remote sensing images [32–34].

Despite significant advancements in semantic segmentation of remote sensing images, existing models exhibit limited capabilities for continual learning, which makes them inadequate for effectively meeting the dynamic demands in diverse environments [35].

2.2 Class-incremental semantic segmentation for remote sensing images

Class-incremental semantic segmentation aims to enable segmentation models to identify a growing number of semantic classes. The intuitive solution to achieve this is storing a small portion of old images and updating the model together with new images [36,37]. However, such methods incur additional costs for data storage and repeatedly replay sparse old data, which easily leads to overfitting old knowledge. Since images used to learn new classes typically contain objects belonging to old classes (labeled as background), a consequence of the large field of view in remote sensing images, alternative strategies to replay old knowledge for the model are employing old models to generate pseudo labels [8,10] for old classes. Nevertheless, due to the distribution shift between old and new data, predictions by old models on new images may be inaccurate. To alleviate this issue, probability difference [8] or information entropy [10] are used to quantify the confidence of old models' predictions, selectively utilizing only highly confident predictions to preserve old knowledge.

The filtering mechanism of old models' prediction may exclude a large number of pixels, resulting in inadequate information for preserving old knowledge. Consequently, more efforts transfer knowledge from old models to new models by adding regularization terms on logits [7] or features [5,9,10]. Building on this foundation, specific designs are developed to consider the traits of semantic segmentation and remote sensing images, thus improving the model's ability to mitigate forgetting. For instance, techniques such as contrastive distillation [5], old-class feature perception [8], and multi-scale feature distillation [10] aim to mitigate conflicts between new and old classes at the feature level. To tackle forgetting of edges and small objects, diversity distillation [7] and cross-image relationship distillation [9] are developed. Additionally, some methods integrate knowledge distillation with memory replay, including old-class feature storage [11] and old-class feature generation [14].

Although these knowledge distillation-based methods report impressive results in class-incremental semantic segmentation of remote sensing images, they overlook the representation bias stemming from the competition between cross-entropy and distillation losses, thus resulting in limited performance.

3 Method

3.1 Problem definition

Class-incremental semantic segmentation involves training the model over multiple learning steps $t \in \{1, 2, \ldots, T\}$, with a set of new classes introduced at each step. In step t, the model Φ^t is trained on the dataset $S^t = \{x_n^t, y_n^t\}_{n=1}^{N^t}$, where $x_n^t \in \mathbb{R}^{H\times W\times 3}$ is a remote sensing image of height H, width W, and y_n^t is the corresponding ground truth. y_n^t only has labels for new classes C^t. If objects belonging to old classes $C^{1:t-1}$ or future classes $C^{t+1:T}$ exist in the image x_n^t, they are classified as the background

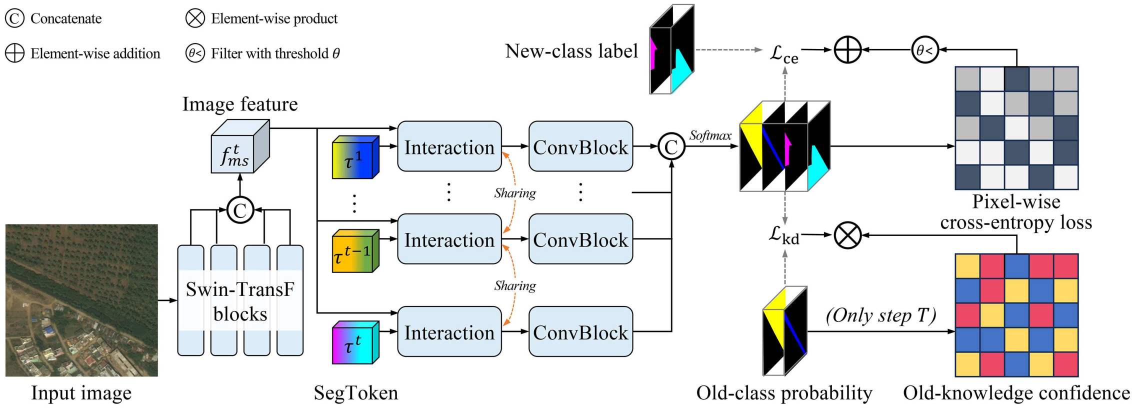


Figure 3 (Color online) Overall framework of our proposed MiR at learning step t. The old model Φ^{t-1} shares a similar architecture with the new model Φ^t, except for the absence of the new SegToken τ^t and a new ConvBlock. Thus, Φ^{t-1} is omitted for clarity. At step t, an image is processed by both Φ^{t-1} and Φ^t, mapping it to their respective output spaces. The cross-entropy loss $\mathcal{L}_{\rm ce}$ and distillation loss $\mathcal{L}_{\rm kd}$ from [15] are employed to train Φ^t. To mitigate representation bias arising from the conflict between $\mathcal{L}_{\rm ce}$ and $\mathcal{L}_{\rm kd}$, SegToken, which is a learnable tensor, is introduced to model category-wise information, with each category represented by distinct colors within the SegToken. Multiple interactions between SegToken $\tau^{1:t}$ and multi-scale image feature f^t_{ms} are then conducted, achieving an implicit coarse-to-fine segmentation. The detailed structure of the interaction module is shown in Figure 4. The ConvBlocks [39], composed of two 3×3 convolutions and a skip connection, are used for mask refinement. Additionally, we propose adaptively increasing $\mathcal{L}_{\rm ce}$ to alleviate interstep imbalance, while weighting $\mathcal{L}_{\rm kd}$ based on the confidence in old knowledge, thereby mitigating interstep similarity.

c_b in the ground truth y^t_n (i.e., $C^1 \cap C^2 \cdots \cap C^T = \emptyset$). The goal of learning step t is to enable the model to segment all seen classes $C^{1:t}$. However, since there is no supervised information related to old classes, the model is easily biased toward new classes and forgets previously learned classes. To alleviate the forgetting issue, we develop a novel class-incremental semantic segmentation framework for remote sensing images, dubbed MiR.

3.2 MiR framework

For semantic segmentation of remote sensing images, the model generally consists of two key components: a backbone for feature extraction and a classifier to generate the segmentation map. Since objects in the remote sensing images vary greatly in size, we adopt Swin Transformer [38] as the backbone to extract multi-scale features and model long-range dependencies. As depicted in Figure 3 [15, 39], given an input image $x \in \mathbb{R}^{H\times W\times 3}$, four Swin-Transformer blocks are applied for hierarchical feature extraction. Denote the feature maps as $f_i \in \mathbb{R}^{\frac{H}{r_i}\times\frac{W}{r_i}\times D_i}$ with downsampling rates from $r = [4, 8, 16, 32]$ and channel dimensions from $D = [96, 192, 384, 768]$. Following [38], hierarchical feature maps after downsampling or upsampling are concatenated along the channel dimension, to extract multi-scale context, i.e., $f_{ms} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1440}$.

After feature extraction, existing methods mostly feed f_{ms} to the classifier to perform pixel-wise classification. To address the forgetting issue, knowledge distillation is generally implemented by feeding new images x^t_n into the old model Φ^{t-1} to predict old-class segmentation maps, and then using them as pseudo labels to encourage the model Φ^t to produce similar predictions. Meanwhile, for learning new classes, the cross-entropy loss with the ground truth y^t_n is used to train the model Φ^t. The overall objective $\mathcal{L}$ of Φ^t can be defined as

$$\mathcal{L} = \lambda_{\rm ce}\cdot\mathcal{L}_{\rm ce} + \lambda_{\rm kd}\cdot\mathcal{L}_{\rm kd}, \tag{1}$$

where $\mathcal{L}_{\rm ce}$, $\mathcal{L}_{\rm kd}$ are cross-entropy loss and knowledge distillation loss, respectively. $\lambda_{\rm ce}$ and $\lambda_{\rm kd}$ are tunable hyperparameters to balance the loss terms, thereby making the trade-off between learning new classes and not forgetting old ones. However, due to the lack of theoretical guarantees or standards, the trade-off is still difficult to achieve. The trained model Φ^t is inevitably biased toward either new or old classes, consequently limiting the model's performance.

To mitigate representation bias, we design SegTokens with interaction modules, replacing the traditional classifier. The SegToken is a learnable tensor to model category-wise information, i.e., $\tau \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d}$, where $d = 7$ is the channel dimension. We feed the SegToken τ and the multi-scale feature f_{ms} into an interaction module to generate the low-resolution segmentation mask $M' \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 7}$.

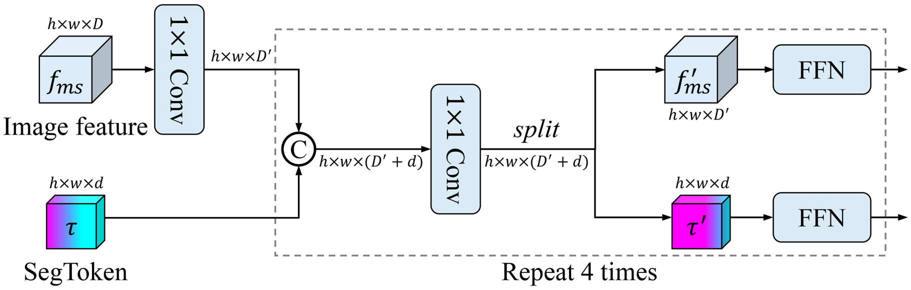


Figure 4 (Color online) Illustration about the architecture of interaction module. FFN refers to feed-forward networks [17], while the circular element represents the concatenation operation. Different colors within the SegToken represent information corresponding to various categories. This module tailors category-wise information based on the image feature, and leverages this customized information to mine more feature regions corresponding to semantic categories.

Then, a ConvBlock [39], composed of two 3×3 convolutions and a skip connection, is used to refine the mask, resulting in $M' \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times(|C|+1)}$. $|C|$ indicates the number of foreground classes. We upsample M' to the size of the input image and use the softmax function to get the final segmentation result $M \in \mathbb{R}^{H\times W\times(|C|+1)}$.

Figure 4 illustrates the interaction module's architecture. This module customizes category-wise information according to the image feature, and utilizes this customized information to mine more feature regions corresponding to semantic classes. To achieve this, SegToken τ and image feature f_{ms} are concatenated along the channel dimension, with the feature channels pre-adjusted to 512. Subsequently, a 1×1 convolution is employed for interaction, resulting in $\mathcal{F} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 519}$. $\mathcal{F}$ is then split along the channel dimension to yield updated SegToken τ' and image feature f'_{ms}. Two feed-forward networks (FFN) [17] are applied to further refine τ' and f'_{ms}, respectively. This process of concatenation, interaction, split, and refinement is repeated four times. Finally, the image feature is discarded, and the corresponding SegToken is considered the low-resolution segmentation mask M' for the subsequent process.

3.3 Class-incremental learning

In accordance with the proposed framework, we add a new SegToken to model information about new classes, thereby enabling the model to incrementally learn them. Specifically, at learning step t, we add a new SegToken τ^t while keeping the previous ones $\tau^{1:t-1}$. Given an image x^t where the subscript n is ignored for simplicity, the backbone processes it into the multi-scale feature f^t_{ms}. Subsequently, the forward pass of the interaction module is executed t times, with each forward pass involving f^t_{ms} and a distinct SegToken τ^i ($i \in \{1,\ldots,t\}$), producing t low-resolution segmentation masks. These masks undergo refinement through specialized ConvBlocks and are then concatenated along the channel dimension. Finally, an upsampling operation followed by a softmax function is applied to obtain the segmentation prediction for all learned classes $C^{1:t}$.

3.4 Loss weighting strategy

We adopt the formulations of $\mathcal{L}_{\rm ce}$ and $\mathcal{L}_{\rm kd}$ proposed by [15] to address the background shift problem unique to semantic segmentation, as follows:

$$\mathcal{L}_{\rm ce} = -\sum_{x^t\in S_t}\sum_{i\in x^t}\log p^t(i,c), \tag{2}$$

$$\mathcal{L}_{\rm kd} = -\sum_{x^t\in S_t}\sum_{i\in x^t} q^{t-1}(i,c)\log \hat{p}^t(i,c), \tag{3}$$

where

$$p^t(i,c) = \begin{cases} q^t(i,c), & \text{if } y^t(i) \neq c_b,\\ \sum_{k\in C^{1:t-1}} q^t(i,k), & \text{if } y^t(i) = c_b, \end{cases} \tag{4}$$

$$\hat{p}^t(i,c) = \begin{cases} q^t(i,c), & \text{if } c \neq c_b,\\ \sum_{k\in C^t} q^t(i,k), & \text{if } c = c_b, \end{cases} \tag{5}$$

and $q^t(i,c)$, $q^{t-1}(i,c)$ are the probabilities of class c for the i-th pixel of the image x^t. The superscript indicates the probability predicted by Φ^t or Φ^{t-1}. Forgetting is effectively mitigated through knowledge distillation and the proposed framework. However, after multiple incremental steps, issues of interstep imbalance and interstep similarity emerge, posing challenges for the current model to generate discriminative feature representations for new classes. Consequently, the model's plasticity is limited. To address these issues, we propose a weighting strategy that adaptively amplifies the cross-entropy loss and decreases the knowledge distillation loss. This enables the model to learn new classes effectively.

Specifically, during a training iteration with N_B images in the mini-batch, the model performs a forward pass and calculates the cross-entropy loss for all pixels. Then, we sort the new-class pixels based on their loss and select the top-K pixels with the highest loss. The average loss of these top-K pixels is combined with the cross-entropy loss of the mini-batch to perform a backward pass. In case any new-class pixels were overlooked, their loss would increase until they are likely to be included in the top-K selection. Consequently, the cross-entropy loss would increase, thereby prompting the model to focus on learning new classes.

Additionally, to alleviate the problem of interstep similarity, we propose to decrease knowledge distillation loss according to the confidence in old knowledge. Note that our focus is on the similarity across different incremental steps rather than within a learning step. When C^t has similar classes to $C^{1:t-1}$, the model Φ^{t-1} may struggle with distinguishing old classes and the background class, resulting in similar prediction scores. Motivated by this observation, we utilize the variance of pixel-wise classification probability to estimate the confidence of the old model in old knowledge, and then adjust the weighting of the knowledge distillation loss accordingly. Importantly, the negative impact of interstep similarity becomes pronounced after multiple incremental steps. Therefore, we propose to weight knowledge distillation loss only in the last learning step.

4 Experiments

4.1 Datasets

We conduct experiments on three publicly available datasets for class-incremental semantic segmentation of remote sensing images: DeepGlobe [16], Potsdam, and Vaihingen.

DeepGlobe is a large-scale satellite image dataset for land cover segmentation, which contains 803 high-resolution images with size 2448 × 2448. As done by [14], we split the DeepGlobe dataset into 455 images for training and 348 images for testing. 20% of the training images are used as the validation set for adjusting hyperparameters. This dataset focuses on 7 classes of land cover, including urban, agriculture, rangeland, forest, water, barren, and unknown/background.

Potsdam consists of 38 high-resolution aerial images with a size of 6000 × 6000 pixels. Following [7], we use 17 images as the training set, 7 images for the validation set, and 14 images as the testing set. The Postdam dataset contains 6 classes, namely impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

Vaihingen contains 33 high-resolution aerial images with an average size of 2494 × 2064 pixels. We follow the data split in [7] to select 11 images for training, 5 images for validation, and 17 images for testing. Like the Potsdam dataset, the Vaihingen dataset also has 6 classes.

4.2 Incremental settings and comparative methods

In computer vision [15, 40], there are two incremental settings: disjoint and overlapped. In the disjoint setting, each learning step comprises images where pixels belong to classes seen either in the current or previous steps. In contrast, the overlapped setting includes all images containing at least one pixel of new classes in each learning step. In both setups, new-class pixels are labeled and old-class pixels are considered background. The key distinction is that in the overlapped setting, the background class may encompass future classes as well. Unlike natural images, which typically have a narrow field of view, remote sensing images, captured from aerial vehicles or satellites, often contain pixels from previous, current, or future classes. Thus, we adopt the overlapped setting for our experiments. Following [14], we conduct experiments on the settings of DeepGlobe (3-3, 2-2-2, 1-1s), Potsdam (4-1, 3-2, 2-3) and Vaihingen (4-1, 3-2, 2-3, 2-2-1). We report the mean intersection-over-union (mIoU) as a quantitative measure to evaluate model performance.

Since more limited studies can be found for class-incremental semantic segmentation of remote sensing images, we compare MiR with popular class-incremental semantic segmentation methods in computer vision [12,15,18,41–45], and methods tailored for remote sensing images [7,10,11,14,36]. For fairness, the results are re-implemented by us using the code from the authors' GitHub repository and models are trained using the hyperparameters listed in their studies. As done by [14], we also report the result of simple fine-tuning (FT) on each S^t and joint optimization (Joint) on all classes at once. The latter can be regarded as an upper bound.

4.3 Implementation details

Network parameters. Most methods for class-incremental semantic segmentation of remote sensing images are built upon Deeplab-v3 (with 58 million parameters) [20]. For fairness, we employ the Swin-S backbone (with 50 million parameters) [38] as the feature extractor due to the comparable amount of model parameters. We empirically set $\lambda_{\text{ce}} = 0.5$, $\lambda_{\text{kd}} = 10$, and keep them unchanged in all experiments. To address the interstep imbalance issue, each training iteration by default selects top-131072 (i.e., $K = 131072$) pixels that have high cross-entropy loss. Additionally, if the loss of the Kth pixel is still greater than a given threshold, we would select all pixels with a loss greater than a given threshold. We experimentally use the threshold of 0.9 for the first step and 0.5 for the subsequent steps.

Training parameter. For all datasets, original images are cropped into size 512×512. In particular, we choose sequential cropping (with no overlap)/random cropping for the DeepGlobe dataset, and sequential cropping (with overlap) for the Potsdam and Vaihingen datasets. Standard data augmentation methods, such as random flip and random rotation, are used to enrich the training sets. During training, model parameters pretrained on ImageNet [46] are used to initialize the backbone at the first learning step. The SegToken added at each step is randomly initialized with a mean of 0 and a variance of 0.02. We adopt the AdamW optimizer with an initial learning rate of 0.00006, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. The batch size is set to 4. We train the model for 10 epochs on the DeepGlobe dataset, 30 epochs on the Potsdam dataset, and 100 epochs on the Vaihingen dataset. Swin-S backbone with Uper-Head [47] is used for joint training as the upper bound. All experiments are conducted on 4 NVIDIA TITAN RTX GPUs with Pytorch implementation and MMSegmentation[3] toolbox.

4.4 Comparison on the DeepGlobe dataset

Table 1 [10–12,14–16,18,41–45] reports the comparison with state-of-the-art methods on the DeepGlobe dataset with different incremental settings.

3-3 (2 steps). 3-3 indicates training the model initially on 3 classes, followed by training on 3 classes. Without employing any constraints, FT completely forgets old classes. PI, EWC, and RW employ different strategies to evaluate the importance of model parameters for old classes and then limit parameter updates, which do not provide benefits. Their performance on old classes is inferior to 4%. Based on knowledge distillation, LwF, LwF-MC, ILT, and MiB significantly improve the performance on old classes, with LwF-MC achieving the highest performance at 57.2%. This confirms the effectiveness of knowledge distillation on preventing catastrophic forgetting. Methods like DisALL, PFG-TKD, and MiSSNet target remote sensing images, outperforming generic methods on old classes. Particularly, DisALL gains 3.5% mIoU, PFG-TKD 8.4%, and MiSSNet 6.4%. It is interesting to compare the above methods with FT on learning new classes. These methods (excluding MiSSNet) avoid forgetting old classes at the cost of giving worse results on new classes (inferior to 59%). In contrast, CoMFormer and MiR can learn new classes well while maintaining good performance on old classes. MiR stands out, exceeding CoMFormer (sub-optimal performance) on both old (4.7%) and new (6.5%) classes. Overall, it obtains 71.5% mIoU, which is inferior to the upper bound of 1.2%. This demonstrates the superiority of our method to handle the forgetting issue.

2-2-2 (3 steps). 2-2-2 signifies three learning steps, each involving the sequential learning of 2 classes. As expected, FT, PI, EWC, and RW obtain good results on new classes (nearly 66%) but perform poorly on maintaining old knowledge (close to 0%). LwF, ILT, LwF-MC, and MiB improve the results in old classes: LwF obtains 39.1% mIoU, LwF-MC 48.0%, ILT 45.1%, and MiB 53.0%. Regarding methods customized for remote sensing images, MiSSNet is the best one, achieving mIoU of 61.0% on old classes, 62.8% on new ones, and 61.6% on all ones. While MiSSNet surpasses other models, CoMFormer obtains

[3] https://github.com/open-mmlab/mmsegmentation.

Table 1 Results (mIoU in %) on the DeepGlobe [16] dataset with different incremental settings. † indicates results re-implemented by us, while results of other methods are directly taken from [14]. Old, new, and all refer to mIoU computed on old, new, and all classes, respectively. The best results are in bold.

Method	3-3			2-2-2			1-1s		
	Old	New	All	Old	New	All	Old	New	All
FT	0.0	61.3	30.6	0.0	65.5	21.8	0.0	86.7	14.4
PI [41]	3.2	61.0	30.5	2.7	67.4	24.3	0.4	71.7	12.4
EWC [42]	3.0	59.0	32.1	2.2	65.9	23.5	0.4	73.0	12.5
RW [43]	3.2	60.0	31.0	3.3	63.6	23.4	0.7	74.9	13.0
LwF [12]	42.9	40.7	41.7	39.1	38.5	38.9	32.3	28.5	31.7
LwF-MC [44]	57.2	46.1	51.6	48.0	43.8	46.6	42.5	44.4	42.8
ILT [45]	49.9	45.0	47.4	45.1	46.7	45.6	42.2	49.0	43.3
MiB† [15]	56.5	57.5	57.0	53.0	53.2	53.0	47.8	31.8	45.1
DisALL [10]	60.7	50.1	55.4	54.1	51.6	53.2	51.9	55.1	52.4
PFG-TKD [14]	65.6	54.7	60.1	56.3	65.6	59.4	55.1	**66.4**	57.0
MiSSNet† [11]	63.6	62.0	62.6	61.0	62.8	61.6	56.6	43.3	54.4
CoMFormer† [18]	64.1	67.6	65.8	63.6	64.7	63.9	61.0	40.3	57.6
MiR	**68.8**	**74.1**	**71.5**	**70.8**	**72.4**	**71.3**	**70.3**	55.7	**67.9**
Joint†	69.7	75.6	72.7	71.7	74.5	72.7	73.9	66.4	72.7

a further boost. Compared to CoMFormer, MiR achieves an average improvement of about 7.4% on old, new, and all classes.

1-1s (6 steps). This setting is the most challenging, as the model sequentially learns each new class one by one. PFG-TKD and CoMFormer significantly outperform other models in terms of mIoU on all classes. Notably, PFG-TKD's high performance is primarily attributed to its effective learning of new knowledge, while CoMFormer excels in maintaining old knowledge. MiR performs inferiorly compared to PFG-TKD on learning the last new class (55.7% vs. 66.4%), but it surpasses PFG-TKD by a significant margin, with a gap of 15.2% on five old classes. Moreover, MiR maintains a constant performance (around 70%) on preserving old classes from single-step to multi-step scenarios, while PFG-TKD's score decreases by more than 10%. Compared to CoMFormer, MiR exceeds it on all classes, with a particularly notable improvement of 15.4% mIoU on new classes. This confirms MiR's capability to effectively balance the stability and plasticity of the model.

Visualization. We present qualitative results from the most challenging 1-1s setting on the DeepGlobe dataset, to visually illustrate the superiority of our method. Different colors are used to indicate urban (cyan), agriculture (yellow), rangeland (purple), forest (green), water (blue), barren land (white), and background (black). Among the six foreground classes, only barran is a new class. As shown in Figure 5, MiSSNet and CoMFormer are biased toward the new class (see the 1st and 2nd rows). MiB is easily confused by similar old classes, such as rangeland, agriculture, and forest. Compared to these methods, MiR achieves a good trade-off between learning new classes and preserving old knowledge. This ultimately leads to segmentation results that closely match the ground truth, demonstrating the superiority of our method.

4.5 Comparison on the Potsdam and Vaihingen datasets

Table 2 [7, 10, 11, 14, 15, 18, 36, 42, 45] reports the comparisons on the Potsdam and Vaihingen datasets.

Results on Potsdam dataset. All models perform better on the Potsdam dataset compared to the DeepGlobe dataset. Even FT and EWC show performance improvements of over 40% from DeepGlobe (3-3) to Potsdam (2-3). One possible explanation is that while the DeepGlobe dataset comprises images from diverse geographic regions, the Potsdam dataset focuses solely on a specific geographic area. The varied features of old classes in the DeepGlobe dataset pose challenges in preserving old knowledge. MiB, ILT, and MiSSNet outperform EWC in the 4-1 setting but lag behind EWC in other settings. Essentially, these methods excel in preserving old knowledge but struggle with learning new classes. TANet, DisALL, PFG-TKD, and CoMFormer surpass the above methods in every setting, achieving average performances of 75.9%, 77.2%, 77.6%, and 84.0%, respectively. However, the best method is MiR, which surpasses the joint training upper bound. This is because the Potsdam dataset focuses on a specific geographical area, where the features of old-class objects in S^{t-1} and S^t exhibit significant similarities. Multi-step training of models actually reinforces old knowledge. Meanwhile, MiR addresses representation bias, interstep

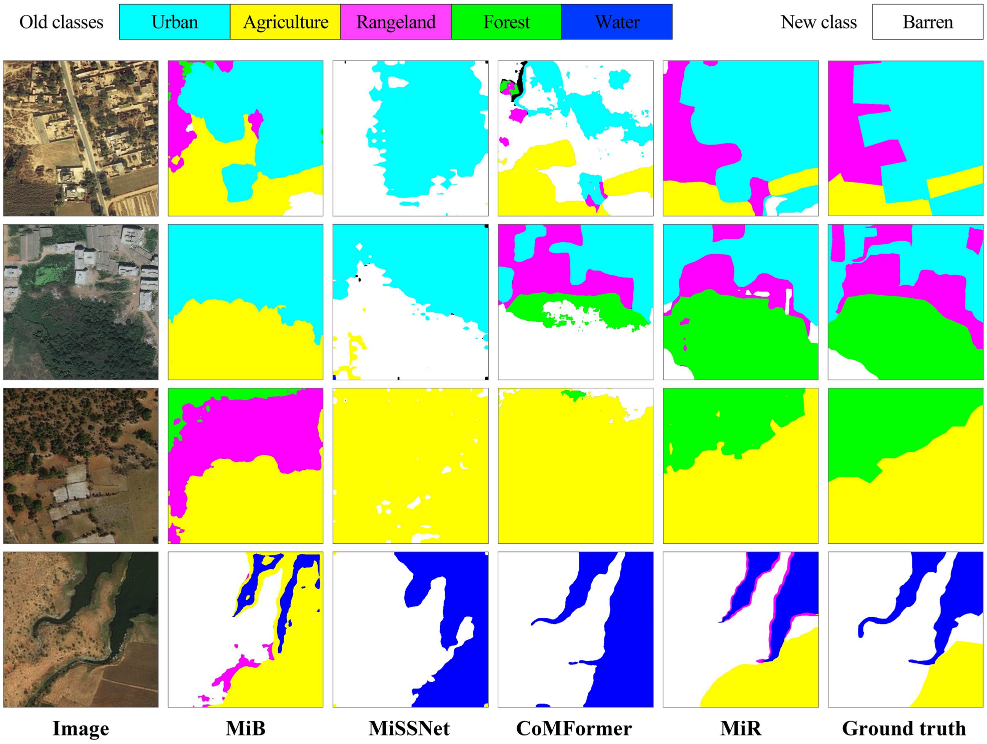


Figure 5 (Color online) Qualitative results on the DeepGlobe dataset with the setting of multi-step addition of five classes (1-1s). Different colors, i.e., cyan, yellow, purple, green, blue, white, and black, indicate urban, agriculture, rangeland, forest, water, barren, and background.

Table 2 Results (mIoU in %) on the Potsdam and Vaihingen datasets with different incremental settings. † indicates results re-implemented by us, while results of other methods are directly taken from [14]. Results refer to mIoU computed on all classes. The best results are in bold.

Method	Potsdam dataset			Vaihingen dataset			
	4-1	3-2	2-3	4-1	3-2	2-3	2-2-1
FT	59.3	70.6	72.1	17.6	57.0	55.2	10.5
EWC [42]	65.5	75.2	72.2	21.1	59.8	65.0	53.9
ILSS [36]	65.5	65.8	63.2	55.3	56.5	65.3	54.3
MiB [15]	69.1	70.6	72.2	74.0	73.2	73.7	72.8
ILT [45]	68.4	69.2	70.4	72.7	72.3	73.5	72.4
TANet [7]	76.7	76.0	75.1	75.7	74.8	74.4	74.0
DisALL [10]	77.6	77.2	76.7	74.9	74.6	74.5	74.1
PFG-TKD [14]	77.5	77.5	77.7	75.0	74.9	74.7	74.6
MiSSNet† [11]	75.1	71.8	72.1	64.8	58.3	56.9	53.1
CoMFormer† [18]	84.7	84.2	83.1	77.0	76.3	77.4	77.0
MiR	**85.5**	**85.6**	**85.4**	**79.2**	**78.7**	**78.6**	**78.4**
Joint†	85.3	85.3	85.3	79.5	79.5	79.5	79.5

imbalance, and interstep similarity, facilitating learning for new classes, ultimately achieving the highest mIoU of 85.5%.

Results on Vaihingen dataset. From Table 2, we can see that FT and EWC suffer a large performance drop, especially in the 4-1 and 2-2-1 settings. Comparative methods (except for ILSS and MiSSNet) outperform them by far in all settings. In particular, MiB achieves 73.4% mIoU averaged on four settings, ILT 72.7%, TANet 74.7%, DisALL 74.5%, PFG-TKD 74.8%, and CoMFormer 76.9%. Despite that, MiR provides a further boost in all settings, obtaining the average results of 78.7%.

Visualization. Figure 6 shows qualitative results from the 3-2 setting on Potsdam and Vaihingen datasets. We use different colors to represent various classes: impervious surfaces (white), buildings (blue), low vegetation (cyan), tree (green), and car (yellow). Notably, tree and car are new classes. On the Potsdam dataset, MiR demonstrates superior segmentation results in terms of detail compared to

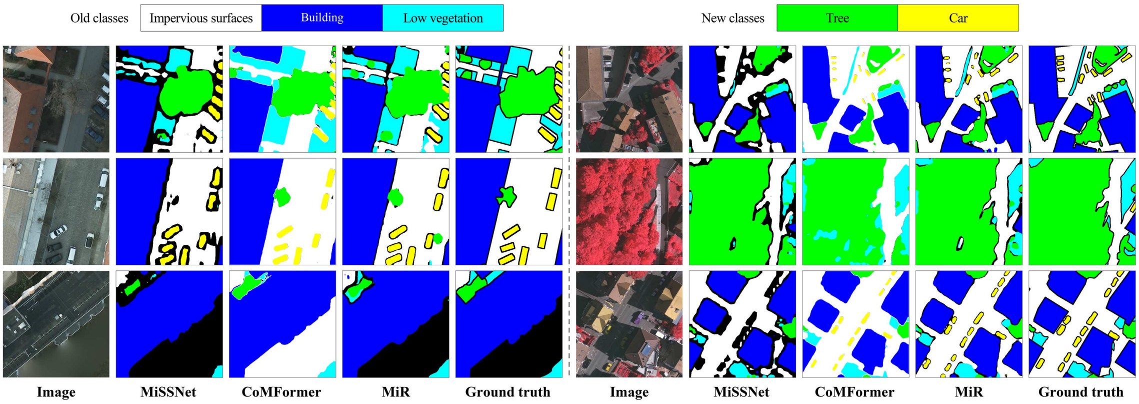


Figure 6 (Color online) Qualitative results on the Potsdam (left) and Vaihingen (right) datasets with the setting of single-step addition of two classes (3-2). Different colors, i.e., white, blue, cyan, green, yellow, and black, indicate impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

Table 3 Ablation study (mIoU in %) on the DeepGlobe [16] dataset with the setting of multi-step addition of five classes (1-1s). The subscripts, i.e., D and S, indicate MiB [15] method applying Deeplab-v3 and Swin-Transformer with Uper-Head, respectively. Old and all refer to mIoU computed on old and all classes. wce and wkd denote the weighting strategy for cross-entropy loss and knowledge distillation loss. The best results are in bold. The improvements are shown in brackets.

Method	Step 1	Step 2	Step 3	Step 4	Step 5	Old	Step 6	All
	urban	agriculture	rangeland	forest	water		barren	
MiB$_\mathrm{D}$	61.7	67.0	15.6	48.7	45.8	47.8	31.8	45.1
MiB$_\mathrm{S}$	78.3	83.3	34.0	64.3	75.3	67.0 (+19.2)	33.3 (+1.5)	61.4 (+16.3)
+SegToken	78.3	83.2	31.8	67.5	77.7	67.7 (+0.7)	53.0 (+19.7)	65.3 (+3.9)
+wce	78.5	85.3	33.6	**72.8**	80.3	70.1 (+2.4)	55.2 (+2.2)	67.6 (+2.3)
+wkd	**78.5**	**85.7**	**34.4**	72.6	**80.5**	**70.3** (+0.2)	**55.7** (+0.5)	**67.9** (+0.3)
MiB$_\mathrm{S}$	78.3	83.3	34.0	64.3	75.3	67.0	33.3	61.4
+wce-wkd	77.8	85.2	27.9	71.2	76.2	67.7 (+0.7)	55.4 (+22.1)	65.6 (+4.2)

MiSSNet and CoMFormer. On the Vaihingen dataset, it is apparent that MiSSNet fails to learn the new class car. Moreover, the similarity between the new class tree and the old class low vegetation poses a challenge for model learning. CoMFormer faces difficulties in accurately distinguishing between tree and low vegetation (see the 2nd row). In contrast, MiR is not influenced by the interstep similarity, producing satisfactory results.

4.6 Ablation studies

We conduct ablation experiments to verify the effect of the details of our method in the most challenging setting on the DeepGlobe dataset (1-1s). Specifically,

Effectiveness of SegToken and weighting strategy. As depicted in Table 3, we start from the baseline MiB$_\mathrm{D}$, which is built upon Deeplab-v3 and employs cross-entropy and distillation losses with modeling the background. We first use Swin-S with Uper-Head to replace Deeplab-v3 (MiB$_\mathrm{S}$): the transformer architecture provides a remarkable improvement of 19.2% on old classes, but it still struggles to learn new classes. This demonstrates that transformer architecture may excel in handling the forgetting issue. Second, we utilize the interaction module with SegToken in place of Uper-Head (+SegToken): this addresses the representation bias caused by competition between cross-entropy and distillation losses, providing benefits on both old and new classes. Remarkably, the improvement for new classes achieves 19.7% mIoU. Third, we modified the standard cross-entropy loss to handle the interstep imbalance issue (+wce). +wce outperforms +SegToken by 2.4% on old classes, 2.2% on new classes, and 2.3% on all classes. Finally, we account for the interstep similarity at the last learning step (+wkd). +wkd provides a slight improvement, where we gain 0.2% on old classes, 0.5% on new classes, and 0.3% on all classes. Though the improvement from the weighting strategy is small, it is not dispensable. We additionally train MiB$_\mathrm{S}$ with the strategy. It can be observed that wce and wkd provide an overall improvement of 22.1% on new classes, confirming their effectiveness. Overall, MiR obtains the best results of 70.3% on old classes, of 55.7% on new ones, and of 67.9% on all classes.

Table 4 Evaluation results (mIoU in %) of λ_{ce} and λ_{kd} on the DeepGlobe [16] dataset with the setting of multi-step addition of five classes (1-1s). (a) MiR. (b) MiB. λ_{ce}, λ_{kd} are tunable hyperparameters to balance cross-entropy loss and knowledge distillation loss. For all experiments in this paper, λ_{ce}, λ_{kd} of our method are 0.5 and 10, respectively. MiB sets $\lambda_{\mathrm{ce}} = 1$ and $\lambda_{\mathrm{kd}} = 10$ by default. Thus, in this experiment, we adjust hyperparameters at the same ratio, for example, doubling or halving λ_{ce}, and increasing or decreasing λ_{kd} by 5.

				(a)					
Weight setting		Step 1	Step 2	Step 3	Step 4	Step 5	Old	Step 6	All
		urban	agriculture	rangeland	forest	water		barren	
$\lambda_{\mathrm{ce}} = 1$	$\lambda_{\mathrm{kd}} = 10$	78.6	85.0	28.9	69.2	78.7	68.1	59.6	66.7
$\lambda_{\mathrm{ce}} = 0.5$		78.5	85.7	34.4	72.6	80.5	70.3	55.7	67.9
$\lambda_{\mathrm{ce}} = 0.25$		79.0	86.4	30.8	70.3	77.0	68.7	58.8	67.1
$\lambda_{\mathrm{ce}} = 0.5$	$\lambda_{\mathrm{kd}} = 15$	78.5	82.6	25.4	73.7	78.9	67.8	51.2	65.1
	$\lambda_{\mathrm{kd}} = 10$	78.5	85.7	34.4	72.6	80.5	70.3	55.7	67.9
	$\lambda_{\mathrm{kd}} = 5$	78.3	84.2	31.4	71.1	78.0	68.6	54.0	66.2

				(b)					
Weight setting		Step 1	Step 2	Step 3	Step 4	Step 5	Old	Step 6	All
		urban	agriculture	rangeland	forest	water		barren	
$\lambda_{\mathrm{ce}} = 2$	$\lambda_{\mathrm{kd}} = 10$	23.0	73.4	2.0	49.0	62.6	42.0	40.6	41.8
$\lambda_{\mathrm{ce}} = 1$		61.7	67.0	15.6	48.7	45.8	47.8	31.8	45.1
$\lambda_{\mathrm{ce}} = 0.5$		36.7	68.5	16.9	30.3	40.6	38.6	23.0	36.0
$\lambda_{\mathrm{ce}} = 1$	$\lambda_{\mathrm{kd}} = 15$	44.6	58.8	8.3	31.4	37.2	36.1	33.0	35.6
	$\lambda_{\mathrm{kd}} = 10$	61.7	67.0	15.6	48.7	45.8	47.8	31.8	45.1
	$\lambda_{\mathrm{kd}} = 5$	27.6	70.3	2.2	50.4	57.3	41.6	41.8	41.6

Table 5 Results (mIoU in %) of different learning orders on the DeepGlobe [16] dataset with the setting of multi-step of five classes (1-1s). Old and all refer to mIoU computed on old and all classes. In this experiment, we use the same hyperparameters setting for MiR.

Method	Step 1	Step 2	Step 3	Step 4	Step 5	Old	Step 6	All
MiR	urban	agriculture	rangeland	forest	water		barren	
	78.5	85.7	34.4	72.6	80.5	70.3	55.7	67.9
	agriculture	barren	water	forest	rangeland		urban	
	77.9	53.4	81.7	72.9	33.9	64.0	79.2	66.5
	barren	water	forest	rangeland	agriculture		urban	
	48.6	83.5	73.0	35.3	82.4	64.6	79.0	67.0
	forest	urban	barren	agriculture	water		rangeland	
	74.3	78.9	47.3	87.2	78.4	73.2	40.1	67.7
Joint	urban	agriculture	rangeland	forest	water		barren	
	78.9	88.4	41.9	77.6	82.7	73.9	66.4	72.7

Effectiveness of mitigating representation bias. As mentioned previously, the competition between cross-entropy and distillation losses easily leads to the issue of representation bias, and empirical hyperparameter settings (e.g., $\lambda_{\mathrm{ce}} = 0.5$ and $\lambda_{\mathrm{kd}} = 10$) are difficult to solve this issue. To demonstrate the effectiveness of our method in mitigating representation bias, we test MiR with different loss weight settings. As shown in Table 4, our method outperforms the best comparative method across five weight settings, achieving an average improvement of 9% considering all classes. For $\lambda_{\mathrm{ce}} \in \{0.25, 0.5, 1\}$ and $\lambda_{\mathrm{kd}} \in \{5, 10, 15\}$, MiR achieves standard deviations of mIoU on old and new classes of 0.86 and 3.09, respectively, while MiB stands at 3.9 and 6.8. This shows that our method is more robust to the change in loss weights compared to the popular MiB method. In other words, MiR can effectively mitigate representation bias, leading to relatively consistent performance.

To intuitively illustrate the effect of our method, Figure 7(a) presents SegTokens extracted from MiR with different weight settings. We can observe that MiR ($\lambda_{\mathrm{ce}} = 1$, $\lambda_{\mathrm{kd}} = 10$) is biased toward new classes. The first interaction between image features and the new SegToken produces high responses in the regions of new classes. On the contrary, MiR ($\lambda_{\mathrm{ce}} = 0.5$, $\lambda_{\mathrm{kd}} = 15$) fails to produce sufficient responses in new-class regions after the first interaction. Through multiple interactions, MiR identifies more regions associated with old or new classes, thereby improving the final segmentation. In Figure 7(b), we present visualized examples of other old classes. These results highlight the importance of accounting for representation bias and demonstrate the effectiveness of our method.

Evaluation with different learning orders. We experiment with different learning orders to explore

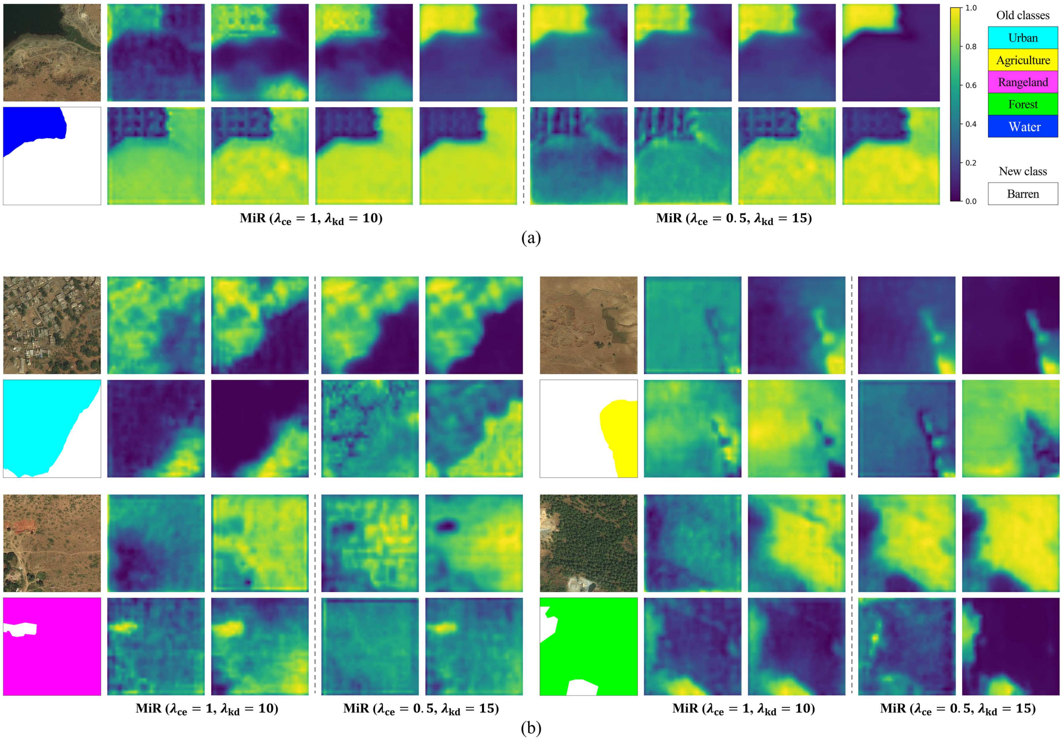

1.0
0.8
0.6
0.4
0.2
0.0
Old classes
Urban
Agriculture
Rangeland
Forest
Water
New class
Barren
MiR ($\lambda_{ce}=1, \lambda_{kd}=10$)
MiR ($\lambda_{ce}=0.5, \lambda_{kd}=15$)
(a)
MiR ($\lambda_{ce}=1, \lambda_{kd}=10$)
MiR ($\lambda_{ce}=0.5, \lambda_{kd}=15$)
MiR ($\lambda_{ce}=1, \lambda_{kd}=10$)
MiR ($\lambda_{ce}=0.5, \lambda_{kd}=15$)
(b)

Figure 7 (Color online) Visualization of SegTokens extracted from MiR, applying different loss weight settings on the DeepGlobe dataset with the setting of multi-step addition of five classes (1-1s). (a) SegTokens after four interactions; (b) SegTokens after the 1st and 4th interactions. In (a) and (b), odd rows represent old-class SegTokens, while even rows represent new-class SegTokens. The SegTokens of a model are visualized by accumulating pixel values along the channel dimension, resizing to match the image size, and then normalizing to the range [0, 1].

the stability of our method. Results are shown in Table 5. The mean and standard deviation of mIoU computed on all classes across the four learning orders are 67.3 and 0.6, respectively. This clearly demonstrates that our method is robust to the change of learning orders, and has significant stability. Additionally, when comparing the IoUs of the same semantic category as the new and old categories, we notice that the IoU gap for the urban class (79.1% vs. 78.7%) is only 0.4%, whereas for the rangeland (40.1% vs. 34.5%) and barren (55.7% vs. 49.8%) classes, the gaps exceed 5%. We argue that this is because the visual characteristics of these categories are complex and diverse, posing challenges for models to learn them effectively. See Table 5, even Joint achieves an IoU of only 41.9% on the rangeland class. Consequently, uncertain predictions from old models on these classes may be excluded during the subsequent knowledge distillation, leading to forgetting.

Analysis of computational cost. We compare the model size (number of parameters), computational complexity (in FLOPs), and inference time (in seconds per image) between MiR and the comparative methods with available code to assess its computational cost. All models are tested under identical settings: an image size of 512×512, a batch size of 1, and a single NVIDIA TITAN RTX GPU. As shown in Table 6 [11,15,18,38,48], the model size of MiR is comparable to that of other models, but the multiple interactions between image features and SegTokens significantly increase the FLOPs in MiR. However, as the interaction module only consists of a 1×1 convolution and two FFNs, it does not introduce substantial inference overhead. Consequently, MiR achieves the shortest inference time while delivering significant performance improvement on both old and new classes.

Evaluation on method scalability. Given that existing studies [7,14] have explored Deeplab-v3 and PSPNet [49] for class-incremental semantic segmentation of remote sensing images, we incorporate MiR with these two segmentation models, to evaluate its scalability. Specifically, following [7, 14], we select PSPNet with the ResNet-50 backbone and Deeplab-v3 with the ResNet-101 backbone. The original classifiers are replaced by the interaction module with SegToken. For PSPNet, the final pyramid pooling global feature is downsampled to a size of $[\frac{H}{16} \times \frac{W}{16}]$, and concatenated with SegToken before being input into the interaction module. For Deeplab-v3, feature maps from atrous spatial pyramid pooling are fused via 1×1 convolution, with size fixed to $[\frac{H}{16} \times \frac{W}{16}]$, and then fed into the interaction module with SegTokens.

Table 6 Evaluation results (mIoU in %) of computational costs across different methods. The comparative methods are implemented using code from the authors' GitHub. All models are tested under identical settings: an image size of 512×512, a batch size of 1, and a single NVIDIA TITAN RTX GPU.

Method	Backbone	#Parameters (M)	GFLOPs	Inference time (s/image)	DeepGlobe (1-1s) (%)		
					Old	New	All
MiB [15]	ResNet-101 [48]	57.9	67.7	0.079	47.8	31.8	45.1
MiSSNet [11]	ResNet-101 [48]	57.9	67.2	0.081	56.6	43.3	54.4
CoMFormer [18]	ResNet-101 [48]	63.0	81.6	0.102	61.0	40.3	57.6
MiR	Swin-S [38]	64.1	137.3	0.068	70.3	55.7	67.9

Table 7 Results (mIoU in %) of MiR with different segmentation models on the DeepGlobe [16] dataset with the setting of multi-step of five classes (1-1s). Old and all refer to mIoU computed on old and all classes. The improvements from CoMFormer to MiR are shown in brackets. The version of MiR implemented in this paper can be regarded as an integration with the segmentation model built on the Swin-S backbone and Uper-Head.

Method	Segmentation model	Step 1	Step 2	Step 3	Step 4	Step 5	Old	Step 6	All
		urban	agriculture	rangeland	forest	water		barren	
CoMFormer [18]	Mask2Former [50]	75.0	76.8	19.5	57.4	76.3	61.0	40.3	57.6
MiR	PSPNet [49]	76.1	80.5	24.0	64.3	57.7	60.5 (−0.5)	44.1 (+3.8)	57.8 (+0.2)
MiR	Deeplab-v3 [20]	65.0	77.1	23.7	61.0	74.8	60.3 (−0.7)	51.8 (+11.5)	58.9 (+1.3)
MiR	Swin-S [38]+Uper-Head [47]	78.5	85.7	34.4	72.6	80.5	70.3 (+9.3)	55.7 (+15.4)	67.9 (+10.3)

The hyperparameters for training PSPNet and Deeplab-v3, including the epoch and the learning rate, are kept consistent with those described in Subsection 4.3. The results are presented in Table 7 [16, 18, 20, 38, 47, 49, 50]. Compared to the current state-of-the-art method CoMFormer, which is specifically designed for Mask2Former [50], MiR achieves superior performance across various segmentation models. The highest gap reaches 10.3% when MiR is incorporated with the segmentation model composed of the Swin-S backbone and Uper-Head. Although MiR exhibits a degree of scalability, it is currently incompatible with architectures like Mask2Former, which can be studied in the future.

5 Conclusion

This paper focuses on the representation bias in class-incremental semantic segmentation of remote sensing images using knowledge distillation. We propose a novel framework, MiR to alleviate forgetting and enhance learning of new classes effectively. By replacing the feature-classifier mode with feature-SegToken interaction, MiR performs implicit coarse-to-fine segmentation to mitigate representation bias caused by the conflict between cross-entropy and distillation losses. On this basis, a weighting strategy is introduced to handle interstep imbalance and interstep similarity issues, thereby facilitating learning after multiple incremental steps. Extensive experiments on commonly used datasets demonstrate the superior performance of MiR over recent state-of-the-art methods such as CoMFormer and MiSSNet in various incremental settings. Besides, we conduct ablation experiments to further validate the efficacy of MiR's design details. Future studies may explore class-incremental semantic segmentation with long-tailed distributions [51], a common scenario in remote sensing images [52].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62325111, U22B2011). We would like to thank Shuguo JIANG for the insightful discussion.

References

1 Boonpook W, Tan Y, Nardkulpat A, et al. Deep learning semantic segmentation for land use and land cover types using landsat 8 imagery. ISPRS Int J Geo-Inf, 2023, 12: 14

2 He X, Zhou Y, Zhao J, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens, 2022, 60: 4408715

3 Huang L, Jiang B, Lv S, et al. Deep-learning-based semantic segmentation of remote sensing images: a survey. IEEE J Sel Top Appl Earth Observations Remote Sens, 2023, 17: 8370–8396

4 McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: the sequential learning problem. Psychol Learn Motiv, 1989, 24: 109–165

5 Arnaudo E, Cermelli F, Tavera A, et al. A contrastive distillation approach for incremental semantic segmentation in aerial images. In: Proceedings of International Conference on Image Analysis and Processing, 2022. 742–754

6 Li J, Diao W, Lu X, et al. SIL-LAND: segmentation incremental learning in aerial imagery via label number distribution consistency. IEEE Trans Geosci Remote Sens, 2022, 60: 5628820

7 Li J, Sun X, Diao W, et al. Class-incremental learning network for small objects enhancing of semantic segmentation in aerial imagery. IEEE Trans Geosci Remote Sens, 2022, 60: 5612920

8 Rong X, Sun X, Diao W, et al. Historical information-guided class-incremental semantic segmentation in remote sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 5622618

9 Rong X, Wang P, Diao W, et al. MiCro: modeling cross-image semantic relationship dependencies for class-incremental semantic segmentation in remote sensing images. IEEE Trans Geosci Remote Sens, 2023, 61: 5616218

10 Shan L, Wang W, Lv K, et al. Class-incremental learning for semantic segmentation in aerial imagery via distillation in all aspects. IEEE Trans Geosci Remote Sens, 2022, 60: 5615712

11 Xie J, Pan B, Xu X, et al. MiSSNet: memory-inspired semantic segmentation augmentation network for class-incremental learning in remote sensing images. IEEE Trans Geosci Remote Sens, 2024, 62: 5607913

12 Li Z, Hoiem D. Learning without forgetting. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 2935–2947

13 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015. ArXiv:1503.02531

14 Shan L, Wang W, Lv K, et al. Class-incremental semantic segmentation of aerial images via pixel-level feature generation and task-wise distillation. IEEE Trans Geosci Remote Sens, 2022, 60: 5635817

15 Cermelli F, Mancini M, Bulo S R, et al. Modeling the background for incremental and weakly-supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 10099–10113

16 Demir I, Koperski K, Lindenbaum D, et al. DeepGlobe 2018: a challenge to parse the Earth through satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2018

17 Lin F, Liang Z, Wu S, et al. StructToken: rethinking semantic segmentation with structural prior. IEEE Trans Circuits Syst Video Technol, 2023, 33: 5655–5663

18 Cermelli F, Cord M, Douillard A. CoMFormer: continual learning in semantic and panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3010–3020

19 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440

20 Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017. ArXiv:1706.05587

21 Peng C, Li Y, Jiao L, et al. Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 2612–2626

22 Nie J, Wang C, Yu S, et al. MIGN: multiscale image generation network for remote sensing image semantic segmentation. IEEE Trans Multimedia, 2022, 25: 5601–5613

23 Li Y S, Chen W, Huang X, et al. MFVNet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. Sci China Inf Sci, 2023, 66: 140305

24 Bai Q, Luo X, Wang Y, et al. DHRNet: a dual-branch hybrid reinforcement network for semantic segmentation of remote sensing images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 4176–4193

25 Yin P, Zhang D, Han W, et al. High-resolution remote sensing image semantic segmentation via multiscale context and linear self-attention. IEEE J Sel Top Appl Earth Observations Remote Sens, 2022, 15: 9174–9185

26 Zheng J, Shao A, Yan Y, et al. Remote sensing semantic segmentation via boundary supervision-aided multiscale channelwise cross attention network. IEEE Trans Geosci Remote Sens, 2023, 61: 4405814

27 Hang R, Yang P, Zhou F, et al. Multiscale progressive segmentation network for high-resolution remote sensing imagery. IEEE Trans Geosci Remote Sens, 2022, 60: 5412012

28 Liu Y, Zhang Y, Wang Y, et al. Rethinking transformers for semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens, 2023, 61: 5617515

29 Zhou X, Zhou L, Gong S, et al. Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery. IEEE J Sel Top Appl Earth Observations Remote Sens, 2023, 17: 175–189

30 Xu Z, Geng J, Jiang W. MMT: mixed-mask transformer for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens, 2023, 61: 5613415

31 Fu Y, Zhang X, Wang M. DSHNet: a semantic segmentation model of remote sensing images based on dual stream hybrid network. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 4164–4175

32 Xiao T, Liu Y, Huang Y, et al. Enhancing multiscale representations with transformer for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens, 2023, 61: 5605116

33 Yao M, Zhang Y, Liu G, et al. SSNet: a novel transformer and CNN hybrid network for remote sensing semantic segmentation. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 3023–3037

34 Xiang X, Gong W, Li S, et al. TCNet: multiscale fusion of transformer and CNN for semantic segmentation of remote sensing images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 3123–3136

35 Weng X, Pang C, Xu B, et al. Incremental deep learning for remote sensing image interpretation. J Electron Inf Technol, 2024, 46: 1–23

36 Tasar O, Tarabalka Y, Alliez P. Incremental learning for semantic segmentation of large-scale remote sensing data. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 3524–3537

37 Tasar O, Tarabalka Y, Alliez P. Continual learning for dense labeling of satellite images. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2019. 4943–4946

38 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022

39 He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks. In: Proceedings of the European Conference on Computer Vision, 2016. 630–645

40 Shang C, Li H, Meng F, et al. Incrementer: transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 7214–7224

41 Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: Proceedings of International Conference on Machine Learning, 2017. 3987–3995

42 Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA, 2017, 114: 3521–3526

43 Chaudhry A, Dokania P, Ajanthan T, et al. Riemannian walk for incremental learning: understanding forgetting and intransigence. In: Proceedings of the European Conference on Computer Vision, 2018. 532–547

44 Rebuffi S A, Kolesnikov A, Sperl G, et al. iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017. 2001–2010

45 Michieli U, Zanuttigh P. Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019. 1–8

46 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009. 248–255

47 Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision, 2018. 418–434

48 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016. 770–778

49 Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017. 2881–2890

50 Cheng B, Misra I, Schwing A, et al. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1290–1299

51 Liu X, Hu Y S, Cao X S, et al. Long-tailed class incremental learning. In: Proceedings of the European Conference on Computer Vision, 2022. 495–512

52 Bai Y, Shao S, Zhao S, et al. EME: energy-based multiexpert model for long-tailed remote sensing image classification. IEEE Trans Geosci Remote Sens, 2024, 62: 4701812