Advancing Data-Efficient Exploitation for Semi-Supervised Remote Sensing Images Semantic Segmentation

Liang Lv and Lefei Zhang, *Senior Member, IEEE*

***Abstract*— To reduce the dependence of remote sensing (RS) image semantic segmentation models on extensive pixel-level annotated images, this article aims to address the issue of insufficient exploitation of RS images' potential within existing semi-supervised learning methods, introducing a novel semi-supervised RS image semantic segmentation method. Specifically, for unlabeled samples, the multiperturbation dynamic consistency (MDC) is proposed to align multiple predictions from diverse data augmentations; MDC leverages a dynamic decay threshold (DDT) instead of fixed thresholds to learn more reliable information, enriching the perturbation space and assisting the segmentation model in acquiring more discriminative feature representations. Furthermore, considering the rich contextual information in RS images, the class prototype memory (CPM) derived from labeled samples is maintained during the training stage, which is leveraged to guide the refinement of predictions from segmentation model at the inference stage. Extensive experiments are conducted on six RS image semantic segmentation datasets, including DFC22, iSAID, MER, MSL, GID-15, and Vaihingen. The experimental results demonstrate the superiority of the proposed method. The code is available at https://github.com/lvliang6879/MCSS.**

***Index Terms*— Class prototype memory (CPM), dynamic decay threshold (DDT), multiperturbation dynamic consistency (MDC), remote sensing (RS) images, semantic segmentation, semi-supervised learning.**

I. Introduction

REMOTE sensing (RS) images' semantic segmentation is an important branch of RS image processing [1], [2]; it involves precisely categorizing every pixel in RS images, enabling the automatic identification and classification of various features within them. In recent years, deep learning methods have achieved remarkable results in the field of RS image semantic segmentation [3], which automatically learn higher level feature representations and have stronger expressiveness and generalization ability. However, although deep learning methods perform well in RS image semantic segmentation, their characteristic of requiring pixel-level annotations also brings some problems. For massive and large-scale RS images, the cost of pixel-level annotation is very high [4]. This not only requires a lot of human resources but also involves the professional knowledge and experience requirements [5], which limits the promotion and application of deep learning methods in practical applications. Regarding the issue of limited labeled data, semi-supervised learning [6] has been proven to effectively enhance the performance of deep learning models by making full use of a large amount of unlabeled data. In the field of computer vision, the initial applications of semi-supervised learning were focused on image classification tasks, including methods such as self-training [7], co-training [8], semi-supervised dictionary learning [9], and label propagation algorithms [10]. As deep learning continues to advance, there have been further developments in methods based on self-training [11], consistency learning [12], [13], and adversarial techniques [14]. In particular, semi-supervised methods based on consistency learning have made significant progress. In the field of semi-supervised semantic segmentation, the paradigm of weak-to-strong consistency learning proposed by FixMatch [13], where the semantic segmentation network predicts pseudolabels for unlabeled images after weak data augmentation, supervising network predictions for unlabeled images after strong data augmentation, has become a current research focus. Methods like [15], [16], and [17] have achieved leading experimental results within this paradigm.

Deep learning has made significant advancements in land cover classification of RS images [18]. The high cost and scarcity of pixel-level annotations have promoted the development of semi-supervised learning methods within the field of RS image processing. Early studies such as [19], [20], and [21] have particularly focused on the semi-supervised classification of hyperspectral RS images. Works such as [22], [23], [24], [25], [26], [27], [28], and [29] have explored semi-supervised semantic segmentation in RS images. Particularly, LSST [26] provides standard datasets' protocols, and a method based on self-training has shown good results. WSCL [28] further improves accuracy based on the weak-to-strong consistency learning paradigm within the standard datasets' protocol. The core of semi-supervised semantic segmentation lies in how to effectively leverage limited labeled samples and a large number of unlabeled samples. However, current methods based on consistency learning significantly fall short in fully exploiting the potential of RS images. Specifically, for unlabeled images, the use of a single consistency constraint hinders the model's ability to explore a broader perturbation space,

Manuscript received 29 November 2023; revised 15 March 2024; accepted 2 April 2024. Date of publication 12 April 2024; date of current version 25 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62122060, Grant 42192580, and Grant 42192583; and in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100014. *(Corresponding author: Lefei Zhang.)*

The authors are with the Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: lianglyu@whu.edu.cn; zhanglefei@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3388199

and the fixed confidence threshold fails to adapt to changes during the model training phase, limiting the ability to extract reliable information from unlabeled data. At the same time, in terms of using labeled images, current methods rely solely on supervised training without further exploring how to make more effective use of these images.

To enhance the data-efficient exploitation of RS images, we propose a **m**ultiple perturbation dynamic consistency with **c**lass prototype memory for semi-supervised RS **s**emantic **s**egmentation method (MCSS). To broaden the initial perturbation space, multiperturbation dynamic consistency (MDC) is proposed to extend consistency from single strongly augmented (unlabeled images through strong data augmentation) predictions to multiple strongly augmented predictions, pseudolabels generated through weakly augmented images and enforce alignment between strongly augmented predictions, and pseudolabels to reinforce the constraints, thereby learning more discriminative representations. Meanwhile, to mitigate information loss and noise caused by a fixed threshold, MDC uses a dynamic decay threshold (DDT). It keeps a high threshold during the initial training stages when the model's performance is suboptimal and progressively lowers the threshold throughout training. This adaptive strategy enhances the efficiency of leveraging unlabeled data. For labeled data, we maintain a class prototype memory (CPM) to leverage the contextual information in RS images with limited computational burden. The class prototype features in CPM are extracted through mask average pooling during the training stage. During the inference stage, the cosine similarity is computed between the CPM and deep features obtained from the encoder, and the similarity information is subsequently integrated with the output of the network's decoder, rectifying the segmentation predictions. In summary, our work contributes in the following ways.

1) We propose an advanced method MCSS for semi-supervised RS images' semantic segmentation. Our framework leverages the MDC and CPM to further enhance the data-efficient exploitation of labeled and unlabeled RS images, thereby improving the performance of RS images' semi-supervised semantic segmentation.
2) We introduce the MDC to ensure alignment between predictions from multiple strongly augmented images (SAIs) and pseudolabels generated from weakly augmented images; meanwhile, MDC uses a DDT instead of a fixed threshold. This extends and enriches the perturbation space, enabling the model to learn more valuable information, thereby facilitating the learning of more distinctive feature representations.
3) Our method innovatively maintains a CPM from labeled samples and calculates the similarity information between CPM and test images to guide the refinement of segmentation model predictions.

The remainder of this article is organized as follows. Section II introduces related work. Section III provides a detailed description of the method proposed in this article. Section IV presents experimental datasets and implementations settings, along with an analysis of the experimental results. Finally, Section V concludes the article.

II. Related Work

A. Semi-Supervised Image Semantic Segmentation

Semi-supervised learning, situated between supervised and unsupervised learning, harnesses the power of both partially labeled and abundant unlabeled data to train machine learning models. With the rise of deep learning, a wave of innovative semi-supervised methods have developed, finding application across a spectrum of tasks, including object detection and semantic segmentation. Within the domain of semi-supervised semantic segmentation, various compelling approaches have been introduced.

1) *Adversarial Methods:* These techniques draw inspiration from the structures seen in generative adversarial networks (GANs) [31]. They leverage adversarial training, involving a generator and a discriminator, to amplify the model's learning from unlabeled data. Noteworthy methods in this category encompass GCT [32] and S4GAN [33].
2) *Self-Training:* Self-training methods typically rely on previous predictions made on unlabeled data. They use models trained on labeled data to generate pseudolabels for unlabeled instances. A notable method in this category is ST++ [11].
3) *Consistency Learning:* Methods falling into this category emphasize consistency. They encourage the model to produce consistent outputs for the same input undergoing various data augmentations, leading to enhancements in segmentation performance. Prominent examples include CCT [15] and CPS [16].
4) *Contrastive Learning:* This learning paradigm focuses on clustering similar elements together while separating them from dissimilar elements in a dedicated representation space. Methods such as ReCo [34] and UP2L [35] exemplify this approach.

B. Semi-Supervised RS Image Semantic Segmentation

Despite the relatively delayed progression of semi-supervised semantic segmentation in RS images, numerous advanced methodologies have developed recently. Wang et al. [22], based on the consistency regularization method, proposed an average update of pseudolabel to effectively harness unlabeled data, consequently enhancing segmentation performance on the RS datasets. LSST [26] presents a self-training framework; it uses the linear sampling to select reliable pseudolabels for training. Works based on the consistency learning method include RanPaste [23], ICNet [27], and WSCL [28]. Both RanPaste [23] and WSCL [28] focus on new data augmentation methods. RanPaste [23] proposed a new data augmentation method, which pastes part of the labeled image into the unlabeled image and mixes them, ultimately achieving improved outcomes on diverse RS datasets. WSCL [28] is proposed as an end-to-end semi-supervised approach for RS image segmentation; it introduced a novel sparse dual-view cross-sample image generation technique and uses an adaptive reweighting strategy based on entropy maps.

These enhancements expand training diversity and reduce noise in pseudolabels, resulting in improved segmentation performance. ICNet [27], on the other hand, focuses on structural improvements; it introduces a dual teacher–student structure with iterative training to achieve enhanced performance across various RS datasets.

The aforementioned methods focus on different challenges in RS image semi-supervised semantic segmentation, making improvements in enhancing the quality of pseudolabels, introducing new data augmentation techniques, or developing novel consistency learning frameworks, and have achieved certain successes. However, they fail to further exploit the potential of RS images. We improve the limitations of the existing methods on RS images and propose a new semi-supervised semantic segmentation method that fully leverages limited labeled data and a large number of unlabeled data.

III. Methodology

A. Preliminaries

Given a limited labeled dataset $\{x_i^l, y_i^l\}_{i=1}^{B_l}$ and an unlabeled dataset $\{x_j^u\}_{j=1}^{B_u}$, where $x_i^l \in \mathbb{R}^{H\times W\times 3}$ represents the ith labeled input image, $y_i^l \in \mathbb{R}^{H\times W\times K}$ is the corresponding pixelwise ground truth, and K denotes the number of classes. Similarly, $x_j^u \in \mathbb{R}^{H\times W\times 3}$ denotes the jth unlabeled input image. B^l and B^u indicate the quantities of labeled and unlabeled images, respectively. The optimization objective function of the weak-to-strong consistency method is formulated as

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u. \tag{1}$$

$\mathcal{L}_s$ and $\mathcal{L}_u$ represent the supervised and unsupervised loss function, respectively, λ is a hyperparameter that adjusts the weight of $\mathcal{L}_u$, and $\mathcal{L}_s$ calculates the cross-entropy loss between the model's prediction $f_\theta(x_i^l)$ and the ground-truth label y_i^l

$$\mathcal{L}_s = \frac{1}{B_l}\sum_{i=1}^{B_l} \mathcal{L}_{\mathrm{CE}}\left(f_\theta\left(x_i^l\right), y_i^l\right). \tag{2}$$

$\mathcal{L}_u$ calculates the cross-entropy loss between predictions from the same image, obtained separately with strong data augmentation Ω and weak data augmentation ω, with the model predictions from weak data augmentation being assigned as pseudolabels y_j^u

$$x_j^{u\_\omega} = \omega\left(x_j^u\right) \tag{3}$$

$$p_j^u = f_\theta\left(x_j^{u\_\omega}\right) \tag{4}$$

$$y_j^u = \arg\max\left(p_j^u\right) \tag{5}$$

$$\mathcal{L}_u = \frac{1}{B_u}\sum_{j=1}^{B_u} \mathbb{1}\left(\max\left(p_j^u\right) \geq \tau\right)\mathcal{L}_{\mathrm{CE}}\left(f_\theta\left(\Omega\left(x_j^{u\_\omega}\right)\right), y_j^u\right) \tag{6}$$

where τ is a predefined confidence threshold, and $\mathbb{1}$ is an indicator function.

Our method, based on the weak-to-strong consistency learning, further explores the efficient exploitation of labeled and unlabeled images for semi-supervised RS semantic segmentation. An overview of the proposed approach is illustrated in Fig. 1; during the training stage, the MDC uses the same encoder and decoder as supervised learning; it takes predictions y_j^u from weakly augmented images as pseudolabels

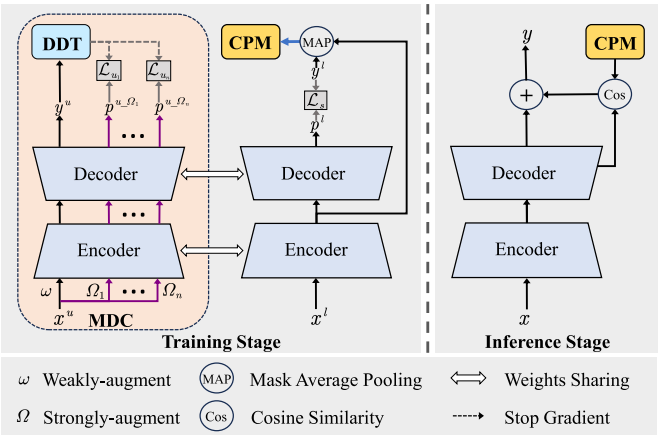

Fig. 1. Overview of the proposed MCSS, which builds on the consistency learning schema, incorporating novel MDC, DDT, and CPM. The purple solid arrows denote paths of strong data augmentation, the blue solid arrows signifies the EMA process, and the gray arrow indicates the loss computation between (pseudo) labels and model's output.

to supervise the predictions ($p^{u\_\Omega_1}$ to $p^{u\_\Omega_n}$) from multiple strongly augmented images and compute multiple unsupervised losses ($\mathcal{L}_{u_1}$ to $\mathcal{L}_{u_n}$), and the network parameters are jointly optimized by the supervised and unsupervised losses. A DDT mechanism is introduced for consistency learning; initially, a high threshold is used to prevent the network from learning excessive noise from pseudolabels; as training progresses, the threshold gradually decreases, enabling the network to progressively uncover reliable information within the unlabeled data. For labeled samples, the prototypes for each class are computed through mask average pooling from labeled samples; then, the EMA is used to dynamically maintain the CPM. In the inference stage, the cosine similarity between CPM and the deep features from test images is calculated to guide the refinement of network's predictions.

B. Multiperturbation Dynamic Consistency

The effectiveness of the weak-to-strong consistency learning lies in its ability to fully leverage a large-scale unlabeled image. It achieves this by aligning the predictions of a semantic segmentation model on the same sample after undergoing different levels of data augmentation, which pseudolabels obtained from weakly augmented images, supervise predictions obtained from SAIs; this compels the model to learn from different perspectives, helping it better understand the diversity of the data and thus improving generalization performance. In RS images, characterized by complex scenes, there is a need for an expanded perturbation space and more robust consistency constraints. Drawing inspiration from [36] and [37], the MDC regularization is proposed. As illustrated by the purple solid lines in Fig. 2, our method integrates multiple SAIs by applying strong data augmentation repeatedly to unlabeled images. The predictions from these SAIs are supervised using pseudolabels, and multiple losses are computed with the DDT. This strategy not only provides a richer array of perturbations but also imposes stronger consistency constraints. The methodology can be succinctly described as follows:

$$\mathcal{L}_u = \sum_{v=1}^{n} \mathcal{L}_{u_v}. \tag{7}$$

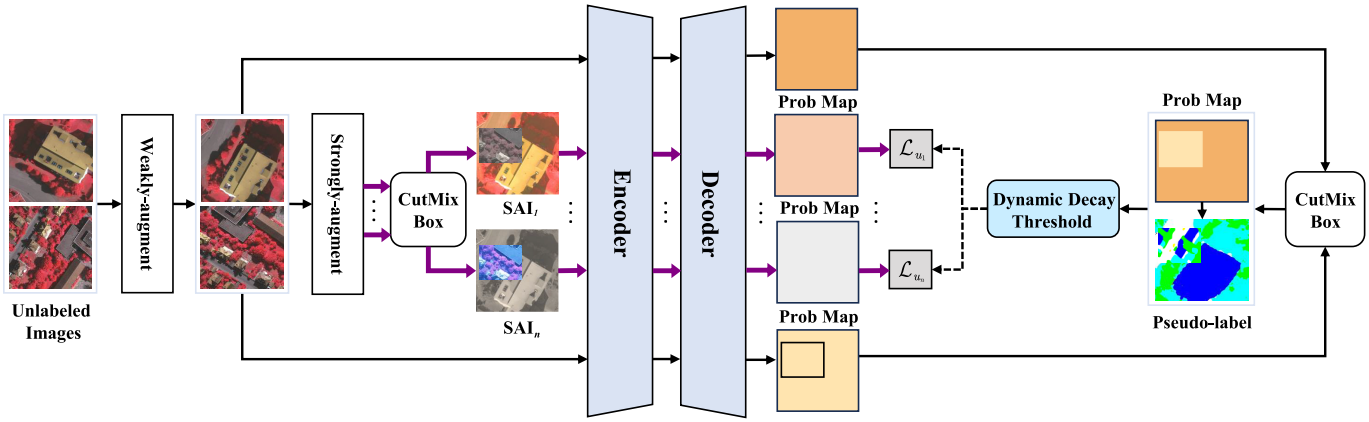

Fig. 2. Detail of MDC. Two unlabeled images first undergo weak augmentation and then go through several distinct rounds of strong augmentation. Subsequently, they are combined through CutMix augmentation to obtain new SAI$_i$. To acquire more reliable pseudolabels, it is necessary to perform CutMix on the predictions from two weakly augmented images. The Prob Map represents the probability maps generated by the network. Confidence for each pixel can be obtained from probability maps, and a DDT is used to select pixels that meet the real-time threshold requirements for loss calculation.

TABLE I
LIST OF DIVERSE DATA AUGMENTATION METHODS USED IN MCSS

Weak Data Augmentation	
Random scale	Randomly resizes the image by [0.5, 2.0].
Random flip	Horizontally flips the image with a probability of 0.5.
Random crop	Randomly crops a region from the image (320 × 320).
Strong Data Augmentation	
Grayscale	Randomly converts the image to grayscale with a probability of 0.5.
Gaussian Blur	Blurs the image with a Gaussian kernel.
Brightness	Adjusts the brightness of the image by [0.5, 1.5].
Contrast	Adjusts the contrast of the image by [0.5, 1.5].
Saturation	Adjusts the saturation of the image by [0.5, 1.5].
Hue	Adjusts the hue of the image by [-0.25, 0.25].
Cutmix Augmentation	
CutMix [30]	Combines regions of two images to create training samples.

Assuming we apply strong data augmentation to unsupervised images n times, at this point, $\mathcal{L}_u$ becomes the sum of unsupervised losses from n different predictions, and the vth strong augmentation is Ω_v, $\mathcal{L}_{u_v}$ can be formulated as

$$\mathcal{L}_{u_v} = \frac{1}{B_u}\sum_{j=1}^{B_u} \mathbb{1}\left(\max\left(p_j^u\right) \geq \tau\right)\mathcal{L}_{\mathrm{CE}}\left(f_\theta\left(\Omega_v\left(x_j^{u\_\omega}\right)\right), y_j^u\right). \tag{8}$$

To achieve a diverse perturbation space, we use a variety of data augmentation techniques. Weak data augmentation methods encompass random rotations, random scaling, and random cropping, while strong augmentation methods include grayscale transformation and random adjustments in brightness, contrast, and hue. The specific augmentation methods are detailed in Table I. French et al. [38] signify that CutMix is highly effective in semi-supervised semantic segmentation. Therefore, we incorporate CutMix augmentation to enrich our perturbation space. Given two images, x_1 and x_2, CutMix takes a portion from each of them and combines them to create a new image, which can be expressed by the following formula:

$$\tilde{x} = M \odot x_1 + (1 - M) \odot x_2. \tag{9}$$

As the CutMix Box in Fig. 2 shows, M is a generated rectangular mask smaller than the original image size, and its acquisition method is as follows:

$$M_{\mathrm{area}} = p_{\mathrm{area}} \times H \times W \tag{10}$$

where M_{area} is the mask area's proportion relative to the original image, and it is determined by randomly selecting a value between p_{area1} and p_{area2}. Within the range of [ratio$_1$, ratio$_2$], we randomly select the ratio of the mask's height to width

$$w_{\mathrm{cutmix}} = \left\lfloor \sqrt{\frac{M_{\mathrm{area}}}{\mathrm{ratio}}} \right\rfloor \tag{11}$$

$$h_{\mathrm{cutmix}} = \left\lfloor \sqrt{M_{\mathrm{area}} \times \mathrm{ratio}} \right\rfloor. \tag{12}$$

As shown in (11) and (12), using the area ratio M_{area} and ratio, we can compute the height h_{cutmix} and width w_{cutmix} of M. Finally, M is randomly selected with a region of the corresponding size in the original image.

In our study, to obtain a more diverse set of augmented samples, we first apply two RS images to weak augmentation. Subsequently, we independently apply strong augmentation to the weakly augmented images, resulting in enhanced images

$$\tilde{x}^u = M \odot \Omega_1\left(x_1^{u\_\omega}\right) + (1 - M) \odot \Omega_2\left(x_2^{u\_\omega}\right). \tag{13}$$

Correspondingly, the prediction results of the two weakly augmented images need to be merged using a mask to obtain the final pseudolabel

$$\tilde{y}^u = M \odot y_1^u + (1 - M) \odot y_2^u. \tag{14}$$

Semi-supervised algorithms based on the FixMatch [13] paradigm often use a fixed threshold to calculate unsupervised loss, considering only unlabeled data with prediction confidence exceeding the threshold. While this strategy ensures the involvement of high-quality unlabeled data in model training, it can lead to the exclusion of a considerable number of pixels in semantic segmentation tasks, resulting in the loss of valuable information. Setting a very low threshold, on the other hand, introduces a significant amount of pseudolabel noise, negatively impacting model performance.

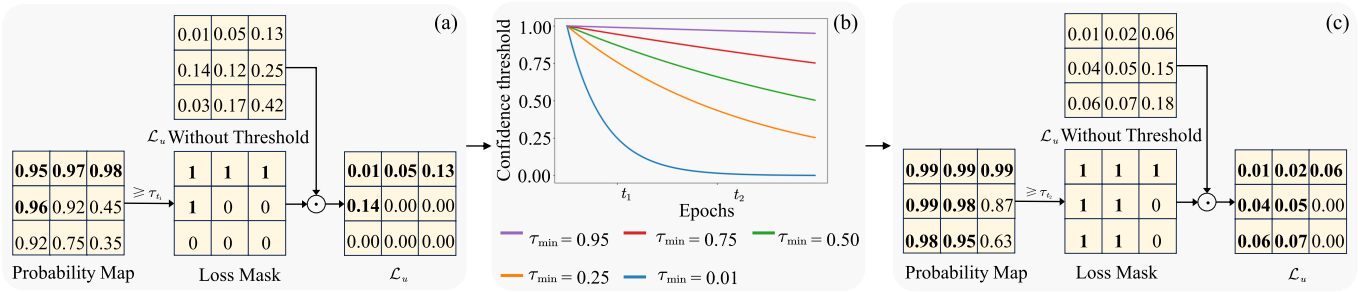

Fig. 3. Schematic of the calculation process of the unsupervised loss $\mathcal{L}_u$ with DDT at epochs t_1 and t_2, where $t_1 < t_2$. (a) Epoch $= t_1$ and $\tau_{t_1} = 0.95$. (b) DDT with different minimum values $\tau_{\min}$. (c) Epoch $= t_2$, $\tau_{t_2} = 0.89$, and $\odot$ is the elementwise multiplication.

In the initial stages of model training, especially when the model's performance is poor, a low threshold introduces significant noise that interferes with the learning process. However, as the model undergoes several epochs of supervised training, it becomes capable of generating high-quality pseudolabels. At this point, lowering the threshold becomes advantageous, allowing more pixels to participate in the model training process.

Based on the above analysis, to achieve a dynamic threshold while avoiding excessive changes, we design an exponentially decaying method to flexibly adjust the threshold, as represented by the following formula:

$$d_r = \left(\frac{\tau_{\min}}{\tau_0}\right)^{\frac{1}{t_{\text{sum}}}} \tag{15}$$

$$\tau_t = \tau_0 \cdot {d_r}^t \tag{16}$$

where d_r represents the decay rate, τ_0 denotes the initial threshold, $\tau_{\min}$ stands for the minimum threshold, t_{sum} indicates the total number of training epochs, and τ_t signifies the threshold at epoch t. In Fig. 3, we present a detailed schematic of the unsupervised loss L_u calculation using the dynamically decaying threshold, as well as the decay curve of τ_t relative to various $\tau_{\min}$ values. Notably, at epoch t_1, the probability maps, predicted from weakly augmented images, serve as the basis for deriving a loss mask, which is then generated using a specified threshold τ_{t_1}. This loss mask allows for the calculation of loss exclusively for pixels with a confidence level at or above τ_{t_1}. Consequently, at epoch t_1, owing to a relatively higher threshold, a limited number of pixels contribute to the loss calculation. As we progress to epoch t_2 and the threshold diminishes, the spectrum of pixels participating in training broadens, enabling the network to learn a richer set of informative pixels.

Theoretical Analysis for MDC: MDC intuitively expands the perturbation space to improve performance; based on the semi-supervised consistency learning theory proposed by Wei et al. [39], we delve deeper into understanding how MDC manages to make this improvement. This theory, assuming classwise expansion and interclass separation, shows semi-supervised methods have a lower error bound than traditional supervised learning.

We first introduce key concepts and symbols of this theory. Let P represent the distribution of unlabeled samples across input space $\mathcal{X}$, with P_i being the class-conditional distribution for class i. The neighborhood of x, denoted as $N(x)$, is defined as the set of points whose transformation set overlaps with the transformation set of x: $N(x) = \{x' : B(x) \cap B(x') \neq \emptyset\}$, and $B(x)$ denotes the set including class-invariant data augmentations. For $S \subseteq \mathcal{X}$, the neighborhood of S is defined as the union of the neighborhoods of its elements: $N(S) \triangleq \cup_{x \in S} N(x)$. Expansion means for a subset S of samples for class i, the probability within the neighborhood of S exceeds c times that of S itself, with $c > 1$ signifying the expansion factor. Specifically, the class-conditional distribution P_i satisfies (a, c)-expansion if for all $V \subseteq \mathcal{X}$ with $P_i(V) \leq a$, the following holds:

$$P_i(N(V)) \geq \min\{c P_i(V), 1\}. \tag{17}$$

If P_i satisfies (a, c)-expansion for all $\forall i \in [K]$, P satisfies (a, c)-expansion. The separation assumption refers to the situation where P can be correctly classified by an ideal classifier with a high probability $(1 - \mu)$, where μ represents an extremely small probability. The expansion assumption indicates that if a pseudolabel generator F_{pl} satisfies a baseline level of accuracy, i.e., the maximum misclassification probability $\bar{a} < 1/3$, and if P meets the $(\bar{a}, \bar{c})$-expansion for $\bar{c} > 3$, then it follows that $c \triangleq \min\{(1/\bar{a}), \bar{c}\}$. Suppose separation and expansion assumptions hold, then for any minimizer $\hat{F}$, Wei et al. [39] provide an upper bound for $\text{Err}(\hat{F})$

$$\text{Err}\left(\hat{F}\right) \leq \frac{2}{c-1}\text{Err}\left(F_{\text{pl}}\right) + \frac{2c}{c-1}\mu \tag{18}$$

where $\text{Err}(\hat{F})$ is the standard 0-1 loss on ground-truth labels: $\text{Err}(\hat{F}) \triangleq \mathcal{L}_{0-1}(\hat{F}, F^*)$. Inequality (18) shows that with the assumptions of expansion and separation, a higher semi-supervised learning coefficient c results in a classifier that outperforms the initial pseudolabeler.

Viewing each strong augmentation as an independent event, for the jth augmentation, the class-conditional distribution P_{ij} meets the (a, c)-expansion for all $V_j \subseteq \mathcal{X}$ with $P_{ij}(V_j) \leq a$, we have $P_{ij}(N(V_j)) \geq c_j P_{ij}(V_j)$, through equivalent transformation, and it can be obtained

$$1 - \prod_{j=1}^{n}\left(1 - P_{ij}\left(N\left(V_j\right)\right)\right) \geq 1 - \prod_{j=1}^{n}\left(1 - c_j P_{ij}\left(V_j\right)\right). \tag{19}$$

For the overall V' with n times augmentations of MDC, we have $V' = \bigcup_{j=1}^{n} V_j$. Since each strong augmentation is mutually independent, it follows that $N(\bigcup_{j=1}^{n} V_j) = \bigcup_{j=1}^{n} N(V_j)$. According to the (a, c)-expansion in inequality (17), $P_i(\bigcup_{j=1}^{n} N(V_j)) \geq c' P_i(\bigcup_{j=1}^{n} V_j)$ holds; by applying

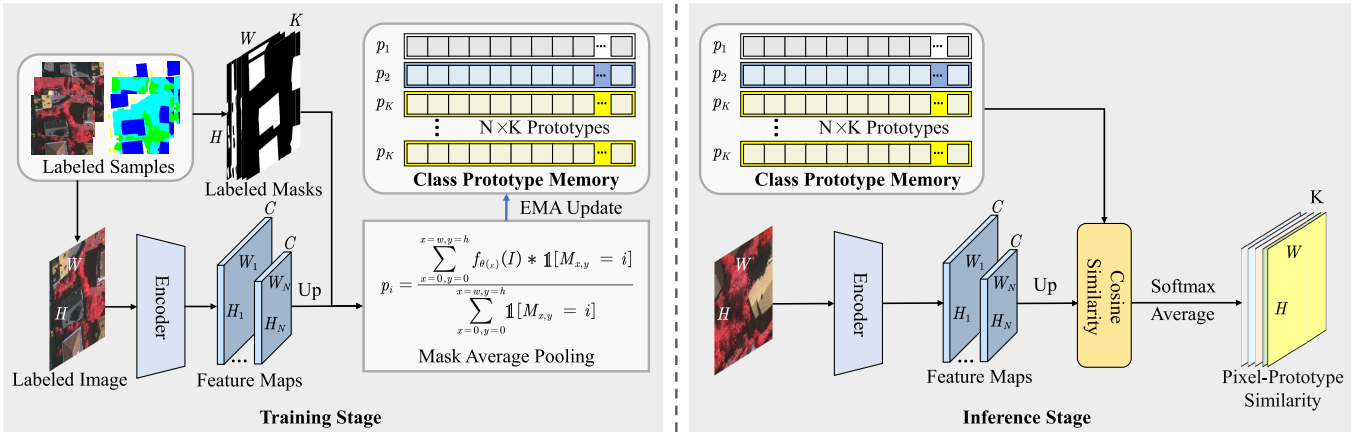

Fig. 4. Illustration of the CPM. Up means upsample by bilinear interpolation. During the training stage, the CPM is maintained through EMA using labeled samples. In the testing stage, pixel-prototype similarity between CPM and deep features from test image is computed to guide the refinement of segmentation result.

equivalent transformations, we can obtain

$$1-\prod_{j=1}^{n}\left(1-P_{ij}\left(N\left(V_j\right)\right)\right)\geq c'\left(1-\prod_{j=1}^{n}\left(1-P_{ij}\left(V_j\right)\right)\right). \tag{20}$$

Since the left sides of inequalities (19) and (20) are equivalent, we take a c' such that it satisfies: $c'(1-\prod_{j=1}^{n}(1-P_{ij}(V_j)))\leq 1-\prod_{j=1}^{n}(1-c_jP_{ij}(V_j))$, and therefore, we have

$$c'\leq\frac{1-\prod_{j=1}^{n}\left(1-c_jP_{ij}\left(V_j\right)\right)}{1-\prod_{j=1}^{n}\left(1-P_{ij}\left(V_j\right)\right)}. \tag{21}$$

Let $c'=\max_{j\in[n]}c_j$, the inequality clearly holds when $n=1$. When $n>1$, assuming it holds for $n=k$, through mathematical induction, it also holds for $n=k+1$. Therefore, $c'=\max_{j\in[k+1]}c_j$ holds. Considering the coefficient c' as a distribution, let the set of c' values that satisfy MDC be denoted as C'; there exists a subset $D\subseteq C'$ such that for any $c'\in D$, we have $c'\geq\max_{j\in[n]}c_j\geq c$; according to the upper bound of $\mathrm{Err}(\hat{F})$ in inequality (18), the following holds:

$$\mathrm{Err}\left(\hat{F}'\right)\leq\frac{2}{c'-1}\mathrm{Err}\left(F_{\mathrm{pl}}\right)+\frac{2c'}{c'-1}\mu. \tag{22}$$

Since MDC possesses a higher expansion factor c', it yields a smaller $\mathrm{Err}(\hat{F}')$. Consequently, compared with FixMatch, MDC demonstrates a lower error upper bound.

C. Class Prototype Memory

Current weak-to-strong consistency semi-supervised learning paradigms concentrate on extracting information solely from unlabeled images, often overlooking the intricate interplay between labeled and unlabeled data. Given the inherent richness in contextual information and global similarity within RS images, we propose a novel approach that exploits the underlying similarity between labeled and unlabeled data to rectify and refine the segmentation model predictions. To alleviate computational complexities associated with computing the similarity across all the labeled and unlabeled images, we draw inspiration from few-shot learning strategies, as exemplified by [40] and SG-One [41]. Our approach involves the extraction of prototypes for each class, facilitating the computation of the similarity between unlabeled data and prototypes. The subsequent processes involved in storing and using the CPM are detailed below.

Fig. 4 provides a visual representation of the CPM computation process. During the training phase, a labeled image $I\in\mathbb{R}^{H\times W\times 3}$ and its corresponding labeled mask $M\in\mathbb{R}^{H\times W\times K}$ (where K is the number of classes) are selected from the labeled samples. A set of N feature maps of varying scales are then extracted from the encoder

$$F=f_{\theta(E)}(I). \tag{23}$$

Here, $F=\{F_1,F_2,\ldots,F_N\}$, and $F_i\in\mathbb{R}^{H_i\times W_i\times C}$ represents the ith feature map. Subsequently, F is restored to the same size as M through bilinear interpolation. The prototypes $p=\{p_1,p_2,\ldots,p_K\}$, where $p_j\in\mathbb{R}^{1\times C}$, are then extracted through mask average pooling

$$p_j=\frac{\sum_{x=0,y=0}^{x=w,y=h}F_i*\mathbb{1}\left[M_{x,y}=j\right]}{\sum_{x=0,y=0}^{x=w,y=h}\mathbb{1}\left[M_{x,y}=j\right]}. \tag{24}$$

This process ensures that each pixel is assigned to the prototype of its corresponding class j based on the mask value $M_{x,y}=j$. For N feature maps, we obtain the overall prototypes P, which includes $N\times K$ class prototypes. To optimize the prototypes, we dynamically allocate features of labeled images from each batch to P through EMA updates, resulting in CPM (P_m). The CPM is initialized with the labeled images from the first batch, by calculating $P^{(i)}$ for each image in the first batch and then obtaining the initial P_m by averaging them. After the first batch, for the tth image, P_m is updated using the exponential moving average (EMA) method

$$\begin{cases}P_m^t=\lambda_m\cdot P_m^{t-1}+\left(1-\lambda_m\right)P, & t>B_l\\ P_m^t=\dfrac{1}{B_l}\displaystyle\sum_{i=1}^{B_l}P^{(i)}, & 0<t<B_l.\end{cases} \tag{25}$$

With the obtained P_m, the inference stage involves inputting unlabeled image I^u is input to the encoder $f_{\theta_{(E)}}$ to obtain deep features F^u of different scales. These features are then

interpolated back to the original image size. Next, the similarity between each pixel feature and the prototypes of each class in P_m is computed. Calculating the cosine similarity and normalizing the results using softmax yields the pixel-prototype similarity

$$S = \frac{\exp(\text{sim}(F^u, P_m))}{\sum_{j=1}^{K} \exp(\text{sim}(F^u, P_m))}. \tag{26}$$

Here, $S = \{S_1, S_2, \ldots, S_N\}$, $S_k \in \mathbb{R}^{H \times W \times K}$, and $\text{sim}(u, v) = \boldsymbol{u}^T \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$. Averaging the similarity results across classes for these N similarity results yields the final S_{average}

$$S_{\text{average}} = \frac{1}{N} \sum_{k=1}^{N} S_k. \tag{27}$$

The resulting S_{average} is then used to guide the rectification of predictions from segmentation model by fusing it with the output of the decoder

$$\text{output} = S_{\text{average}} + f_{\theta_{(D)}}(F_u). \tag{28}$$

In practice, CPM stores class-relevant features from all the labeled samples. During the inference stage, the similarity between each pixel and prototypes of each class in CPM is calculated. This approach refines segmentation results through probability fusion with the output of the semantic segmentation network. The proposed CPM not only establishes a connection between labeled and unlabeled data but also facilitates a dynamic and efficient utilization of prototype features to enhance semantic segmentation performance for RS images.

IV. Experiments

In this section, we will describe a series of comparative experiments and ablation studies to evaluate the performance of our proposed method. We will compare it with FixMatch [13] and four open-source semi-supervised RS image segmentation methods across six RS image semantic segmentation datasets. Our evaluation will be based on intersection over union (IoU) for individual classes and mean IoU (mIoU) across all the classes. In Section IV-A, we introduce the experimental datasets and parameter settings. Section IV-B covers the ablation studies, while Section IV-C presents the experimental results on each dataset, comparing the performance of various methods under different proportions of labeled data.

A. Datasets

To comprehensively validate the proposed method and facilitate a fair comparison with the existing approaches, we follow the dataset protocols and partitioning conventions established by WSCL [28] and LSST [26]. We conduct experiments on six datasets: DFC22 [42], iSAID [43], MER [44], MSL [44], Vaihingen, and GID-15 [45]. Each dataset is divided such that the proportions of labeled images are 1/8 and 1/4, with the remaining images used as unlabeled.

The MiniFrance-DFC22 (DFC22) [42] is provided by the 2022 IEEE GRSS Data Fusion Contest. It consists of 766 pixel-level annotated images extracted from aerial imagery, each of size 2000 × 2000 pixels. This dataset is designed for training semantic segmentation models for land use mapping. It includes annotations for 12 land cover categories: urban, industrial, mine, artificial, arable, permanent, pastures, forests, herbaceous, open, wetlands, and water.

The iSAID [43] is a large-scale and densely annotated benchmark dataset for instance and semantic segmentation in aerial images. It comprises 655 451 object instances across 15 categories within 2806 high-resolution images. The 15 land cover categories are as follows: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, swimming pool, and soccer ball field.

Both the MER and MSL [44] datasets are derived from high-resolution Martian landscape images collected by Mars-Seg. They consist of 1023 high-resolution grayscale images and 4155 RGB images, respectively, encompassing nine land cover categories: martian soil, sand, gravel, bedrock, rocks, tracks, shadows, background, and unknowns.

The GID-15 [45] dataset is derived from high-resolution RS images captured by the GaoFfen-2 satellite. It comprises 15 land cover categories, including industrial land, urban residential areas, rural residential areas, traffic land, paddy fields, irrigated land, dry cropland, garden plots, arbor woodlands, shrub lands, natural grasslands, artificial grasslands, rivers, lakes, and ponds.

The Vaihingen dataset is provided by the International Society for Photogrammetry and RS Commission WG II/4; it consists of 33 RS images of varying sizes, with an average size of approximately 2494 × 2064 pixels and a spatial resolution of 9 cm. The dataset includes three channels: near-infrared, red, and green, and encompasses five land cover classes: impervious surfaces, buildings, low vegetation, trees, and cars.

B. Implementations

In this study, we use two RTX 3090 GPUs and construct a semantic segmentation model based on the PyTorch framework. The model uses ResNeXt-50 [46] as the backbone network and uses the UNet [47] architecture. The following are the key parameter settings for our model training: Optimizer, we use the AdamW [48] optimizer with an initial learning rate of 0.0001, and dynamic learning rate updates are performed using the OneCycleLR [49] scheduler. Each batch consists of eight labeled and unlabeled images, and the cropping size is set to 320 × 320 pixels. A series of data augmentation methods are applied following the strategies and ranges outlined in Table I. For a comprehensive evaluation of model performance, we set the training epochs to 1000.

C. Ablation Study

In this section, we conduct ablation experiments on the DFC22, MER, and Vaihingen datasets, all under the conditions where labeled images account for 1/8 of the training set. These experiments aimed to assess the effectiveness of the proposed MDC and CPM. Regarding MDC, three ablation studies are conducted. First, we examine the impact of varying

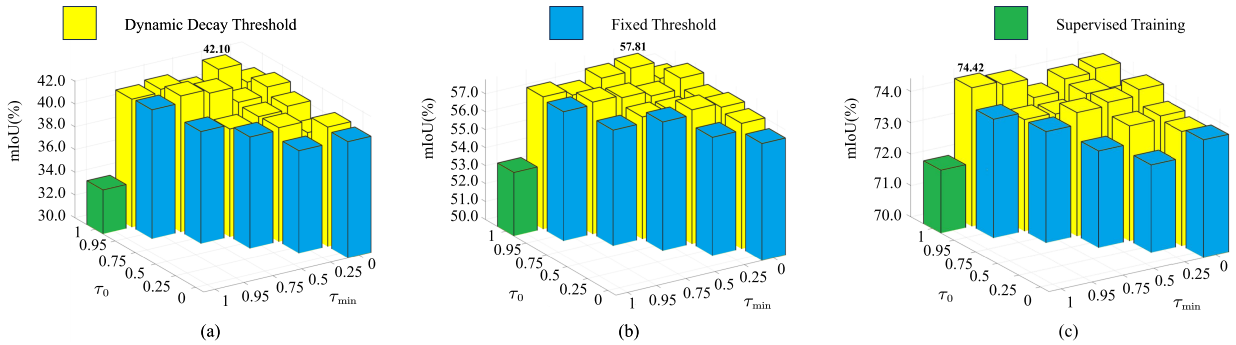

Fig. 5. Ablation study of confidence threshold τ in fixed threshold, τ_0 and $\tau_{\min}$ in DDT on three datasets. (a) DFC22. (b) MER. (c) Vaihingen.

TABLE II

COMPARISONS OF THE EFFECTS OF DIFFERENT AUGMENTATION TIMES (n) FOR MDC ON MIOU (%), TIME (S), AND MEMORY (M)

Method	DFC22		MER		Vaihingen		Memory
	mIoU	Time	mIoU	Time	mIoU	Time	
SupOnly	33.77	**11.5**	53.54	**13.6**	72.01	**10.5**	**8524**
FixMatch	38.52	15.4	55.31	19.4	73.39	13.3	10386
MDC(n=2)	41.23	17.6	57.20	21.5	**73.88**	15.5	12810
MDC(n=3)	39.99	19.5	**57.79**	23.7	73.58	16.5	15076
MDC(n=4)	**41.50**	20.6	57.52	25.5	73.79	17.5	16240
MDC(n=5)	40.31	21.7	57.03	26.7	73.29	18.6	18496

the number n of times strong data augmentations on model performance. Second, in the case of DDT, we compare the experimental results with different values of τ_0 and $\tau_{\min}$, and t_{sum} is the number of training epochs, set to 1000. Finally, we vary the value of the unsupervised loss weight λ, conduct experiments, and analyze the influence of λ on the experimental outcomes. For CPM, since we use EMA to update the CPM, we investigate the performance implications of different values of λ_m. These experiments are conducted to gain insights into how each component contributes to the overall performance of our method. In addition, we aim to provide a comprehensive analysis of the model's behavior under various settings, enhancing our understanding of the proposed approach.

1) Number of Strong Data Augmentations: To assess the performance of the number of strong data augmentations, we conduct ablation experiments on the DFC22, MER, and Vaihingen RS datasets, comparing the impact of different strong data augmentations times n ranging from 2 to 5 on model performance; these experiments are carried out under the fixed threshold of $\tau = 0.95$ and $\lambda = 0.5$. As shown in Table II, in comparison to the supervised training (SupOnly) approach that solely relies on labeled data, integrating FixMatch with unlabeled data training significantly enhances mIoU across three datasets. Specifically, improvements of 4.75%, 1.77%, and 1.38% are observed, respectively. Moreover, when compared with FixMatch's single strong augmentation technique, all the three datasets exhibit superior performance under the multistrong augmentation conditions. Notably, the DFC22 dataset manifests the most substantial improvement in mIoU, recording an increase of 2.98% when n is set to 4. Similarly, with n set to 3, the MER dataset achieves its maximal mIoU enhancement, at 2.48% above that of the standard FixMatch. In addition, when n is set to 2, the Vaihingen dataset still outperforms the conventional FixMatch approach, showing a 0.49% increase in mIoU. Under the same hyperparameter settings, increasing n results in a certain increase in training time and memory usage. For each increment of n by 1, the time increases by 1–2 s, and memory usage increases by approximately 2000 M.

It is evident that changes in the value of n have a certain impact on accuracy. The optimal value of n may vary depending on the dataset (DFC22: $n = 4$, mIoU = 41.50; MER: $n = 3$, mIoU = 57.79; Vaihingen: $n = 2$, mIoU: 73.88), and increasing n can sometimes even lead to a decrease in model performance, possibly due to the increased difficulty in learning from too many augmented views. It is noteworthy that when $n \geq 2$, all three datasets demonstrate significant improvements. This suggests that MDC offers richer perturbations and stronger constraints, thereby aiding semantic segmentation networks in better capturing the diversity of the data, ultimately enhancing generalization performance. Although increasing n enhances the model's performance, it also comes with an increase in computational costs. In addition, excessively large n values may increase the difficulty of learning, leading to a decline in performance. This suggests the need to adjust the extent of consistency learning constraints within a certain range based on different datasets and to select a tradeoff between efficiency and precision according to computational resources.

2) Dynamic Decay Threshold: We conduct separate ablation experiments for both fixed threshold and DDT to validate the effectiveness of the proposed method, and then compared the results. For ease of comparison, we conduct experiments with strong data augmentations numbers $n = 2$ and $\lambda = 0.5$.

In Fig. 5, the green cube represents results obtained solely through supervised training ($\tau = \tau_0 = \tau_{\min} = 1$), while the blue cubes depict experiments with fixed threshold variations ($\tau = \tau_0 = \tau_{\min} = 0, 0.25, 0.5, 0.75, 0.95$). Clearly, the introduction of consistency loss significantly improves accuracy compared with using only supervised training, regardless of the fixed threshold value. This set of ablation experiments demonstrates that the variation in thresholds greatly influences the performance of consistency learning. Changing the fixed threshold yields the highest mIoU for three datasets as follows: DFC22 with $\tau = 0.95$, achieving an mIoU of 41.35; MER with $\tau = 0.50$, resulting in an mIoU of 57.20; and Vaihingen with

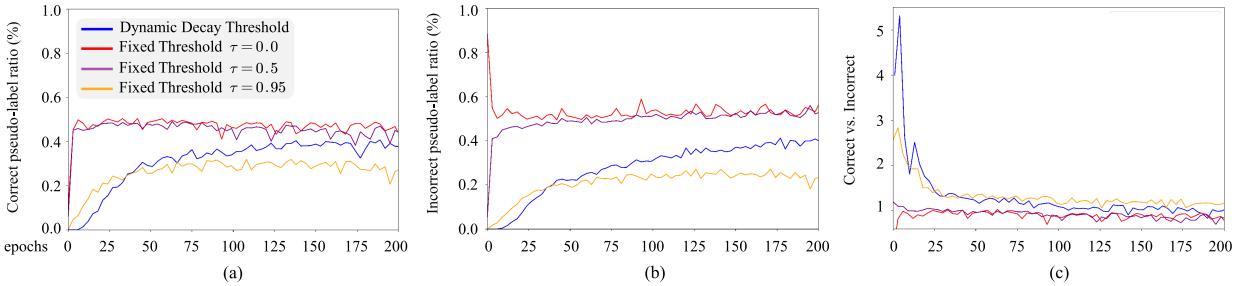

Fig. 6. During the training stage, the adoption of the fixed threshold versus the DDT leads to variations in (a) proportion of correct pseudolabels, (b) proportion of incorrect pseudolabels, and (c) ratio of correct-to-incorrect pseudolabels.

$\tau = 0.95$, reaching an mIoU of 73.85. It is observed that the highest accuracy is not always achieved at higher thresholds, indicating that the optimal threshold should be determined based on the specific dataset. The yellow cubes in Fig. 5 represent the results of DDTs, with the highest mIoU achieved for the three datasets with DDTs as follows: DFC22 with $\tau_0 = 1$ and $\tau_{\min} = 0.25$, recording an mIoU of 42.10; MER with $\tau_0 = 1$ and $\tau_{\min} = 0.25$, achieving an mIoU of 57.81; and Vaihingen with $\tau_0 = 1$ and $\tau_{\min} = 0.95$, achieving an mIoU of 74.42. Compared with fixed thresholds, this corresponds to an increase of 0.75%, 0.61%, and 0.57% for each dataset, respectively. Furthermore, it can be observed from Fig. 5 that DDTs evidently achieve higher accuracy compared with fixed thresholds. In addition, the experiment indicates that starting the decay of τ_0 from smaller values did not achieve higher accuracy, beginning the decay from $\tau_0 = 1$ and only adjusting $\tau_{\min}$ can reduce the complexity of hyperparameter selection and effectively control the decay rate to obtain the optimal threshold.

Fig. 6 illustrates the variations in pseudolabels during the training process using low, high, and DDTs. With a low threshold, although it enables a large number of correct pseudolabels to participate in training, it also introduces a significant amount of incorrect pseudolabels in the early stages of training which can interfere with model training. On the other hand, a high threshold maintains a smaller proportion of pseudolabels and also filters out many correct pseudolabels. The DDT achieves a better tradeoff between the proportions of correct and incorrect pseudolabels. Moreover, in the early stages of training, the number of correct pseudolabels is much higher than that of incorrect ones, which is beneficial for model training. It helps in learning more accurate label information while suppressing the interference of label noise.

3) Unsupervised Loss Weight: The unsupervised loss weight λ is a crucial hyperparameter in consistency learning. We explore the impact of four different λ values: 0.1, 0.5, 1.0, and 5.0 on the model's performance across the DFC22, MER, and Vaihingen datasets, with other primary hyperparameters set as follows: $n = 2$, and $\tau_{\min}$ values of 0.25, 0.25, and 0.95, respectively. Fig. 7 presents our experimental results. It is observed that the highest mIoU for the Vaihingen dataset, at 74.86%, is achieved at $\lambda = 0.1$. For the DFC22 and MER datasets, the peak mIoU, 42.10% and 57.81% respectively, is observed at $\lambda = 0.5$. However, as λ increased to 1.0 and 5.0, there is a significant decline in the mIoU across all three datasets. This suggests that an excessively high unsupervised loss weight λ can adversely affect the training process of the network, leading to a decrease in performance.

TABLE III

ABLATION STUDY OF λ_m IN THE CPM ON THE MIOU (%), STORAGE (KB), AND TIME (MM:SS) OF THREE DATASETS WITH 1/8 LABELED DATA

Method	DFC22			MER			Vaihingen		
	mIoU	Storage	Time	mIoU	Storage	Time	mIoU	Storage	Time
w/o CPM	42.10	–	–	57.81	–	–	74.86	–	–
$\lambda_m = 0.2$	42.45	49.29	7:52	58.28	37.29	6:45	75.16	21.29	5:17
$\lambda_m = 0.5$	42.67	49:29	9:35	58.31	37.29	5:27	75.15	21.29	5:25
$\lambda_m = 0.8$	42.36	49.29	4:32	58.18	37.29	5:23	75.10	21.29	4:43
$\lambda_m = 0.95$	42.30	49.29	7:21	58.16	37.29	6:17	75.08	21.29	8:22

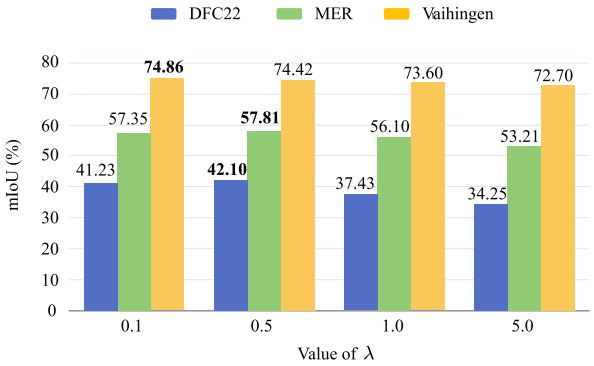

Fig. 7. Ablation study of unsupervised loss weight λ on the DFC22, MER, and Vaihingen datasets.

4) Class Prototype Memory: To validate the effectiveness of CPM in refining segmentation model predictions, we conduct experiments using different values of λ_m (0.2, 0.5, 0.8, 0.95), and the length of a single prototype is set to $C = 128$, with the number of different-scale feature maps denoted as $N = 4$, while keeping other hyperparameters consistent with the unsupervised loss weight. The experimental results are presented in Table III. Upon introducing CPM, the FC22 and MER datasets achieve their highest mIoU at $\lambda_m = 0.5$, recording 42.67% and 58.31%, respectively. For the Vaihingen dataset, the peak mIoU of 75.16% is observed at $\lambda_m = 0.2$. Compared with the scenarios without CPM, these represented increases of 0.57%, 0.50%, and 0.30%, respectively. Notably, the performance differences across the various λ_m values remained relatively stable, demonstrating the robustness of the proposed method to hyperparameter variations.

The experimental results indicate that rectifying the model predictions using the similarity between CPM from labeled samples and unlabeled samples can effectively enhance the

TABLE IV
COMPARISONS BETWEEN OUR PROPOSED METHOD WITH FIVE STATE-OF-THE-ART METHODS ON THE DFC22 DATASET

Method	Urban	Industrial	Mine	Artificial	Arable	Permanent	Pastures	Forests	Herbaceous	Open	Wetlands	Water	mIoU
1/8 Labeled Images													
SupOnly	63.50	36.46	nan	13.33	9.45	50.36	49.59	66.30	37.30	28.62	27.32	23.06	33.77
FixMatch [13]	66.02	37.68	nan	16.10	8.92	54.34	55.43	70.83	41.20	23.01	54.38	**34.35**	38.52
ICNet [27]	61.20	38.92	nan	12.80	9.75	55.62	52.35	68.59	40.20	20.73	53.80	28.49	36.87
RanPaste [23]	63.02	38.21	nan	14.72	10.20	52.34	52.10	69.55	40.05	25.70	55.25	28.65	37.48
LSST [26]	65.31	**45.81**	0.47	17.18	**13.32**	50.05	49.22	66.64	**41.28**	29.29	60.10	26.72	38.78
WSCL [28]	64.58	45.50	**3.49**	16.93	10.96	55.50	**57.23**	70.03	39.23	23.57	68.78	29.04	40.40
Ours	**67.74**	44.53	1.97	**17.54**	11.02	**56.14**	55.34	**70.97**	39.60	**46.75**	**73.01**	27.68	**42.67**
1/4 Labeled Images													
SupOnly	62.44	43.46	nan	4.48	5.39	49.37	52.57	65.72	38.47	40.60	51.81	31.67	37.17
FixMatch [13]	**63.30**	43.00	nan	3.98	12.33	61.36	51.63	70.66	40.76	37.00	59.40	27.77	39.27
ICNet [27]	61.80	39.74	1.05	4.52	9.98	45.12	52.63	69.80	39.77	46.63	54.38	28.96	37.87
RanPaste [23]	62.79	40.81	1.02	8.80	10.69	50.25	54.23	67.39	40.09	45.47	47.30	35.32	38.68
LSST [26]	62.64	42.30	nan	5.98	**12.35**	60.01	54.05	71.32	42.73	37.07	63.45	36.58	40.71
WSCL [28]	61.53	44.59	5.64	8.65	9.55	**62.84**	55.62	70.50	40.63	43.41	59.51	29.22	40.97
Ours	63.19	**44.93**	**5.76**	**9.72**	8.51	59.28	**56.52**	**71.53**	**43.39**	**47.59**	**64.67**	**38.76**	**42.82**

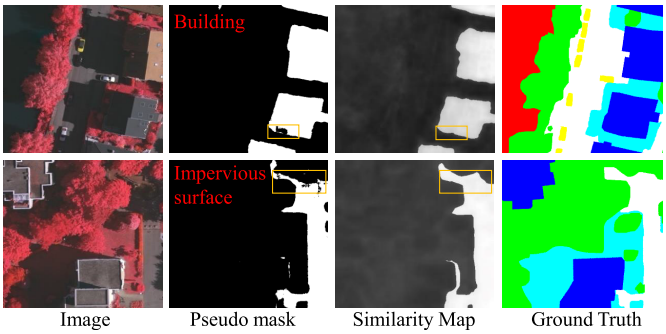

Fig. 8. Qualitative visualization results of the similarity map for the Vaihingen dataset. The categories of pseudo masks are indicated in red font. The improved areas are marked with orange boxes.

performance. This highlights the capacity of the proposed CPM to leverage information from labeled samples, thereby contributing to the improvement of semi-supervised RS semantic segmentation. In Table III, "Storage" refers to the storage space occupied by the CPM, while "Time" indicates the time required to obtain the appropriate CPM. The experimental results show that the CPM only occupies a few tens of kB of space (the specific size depends on the dataset), and an appropriate CPM can be obtained in a relatively short amount of time (approximately 4–10 min).

As shown in Fig. 8, we visualized the pixel-prototype similarity map. For clearer presentation, we only highlighted the pseudolabel mask and similarity map for individual categories, such as buildings and impervious surfaces. The generated pixel-prototype similarity map can identify the areas most likely belonging to these categories and filter out the incorrect regions.

D. Comparison With State-of-the-Art Methods

In this section, we use the results of fully supervised training with 1/8 or 1/4 of the labeled data as a reference. FixMatch is used as the baseline for the semi-supervised approach. We also replicate FixMatch along with four state-of-the-art RS image semi-supervised semantic segmentation methods: ICNet [27], RanPaste [23], LSST [26], and WSCL [28]. All the experiments are conducted on six RS image semantic segmentation datasets. Specifically, for the DFC22 dataset, we present a comparative and analytical result of the class IoU and mIoU by every method. The visual comparison results of each method are shown in Fig. 9. For the iSAID, MER, MSL, GID-15, and Vaihingen datasets, we conduct a comparative and analytical study of the mIoU results obtained by each method.

1) Comparison Results on the DFC22 Dataset: In the DFC22 dataset, MCSS is compared with other advanced methods. The experimental results in Table IV demonstrate that all the semi-supervised methods significantly improved accuracy compared with supervised training, with the proposed MCSS achieving the largest increase in mIoU. When the proportion of labeled images is 1/8, MCSS's performance improved by 8.9% compared with supervised training, with a notable increase in the IoU for all the categories, especially for Open and Wetlands classes, which increased by 18.13% and 45.69%, respectively. Compared with FixMatch, MCSS's mIoU improved by 4.15%, with significant IoU improvements in all the categories except the Water class. Against ICNet, RanPaste, LSST, and WSCL, MCSS gets the highest mIoU of 42.67%, with the highest IoU in most categories, especially in the Open class, where its IoU is significantly higher than other methods. In the first column of the visual comparison results in Fig. 9, image (g) shows MCSS's predictions, which more closely resemble the ground truth (b). When the proportion of labeled images is 1/4, there is an improvement in mIoU. At this ratio, compared with supervised training, MCSS's performance improved by 5.65%, and by 3.55% compared with FixMatch. Against ICNet, RanPaste, LSST, and WSCL, MCSS attains the highest mIoU of 42.82%, with the highest IoU in most categories.

2) Comparison Results on the iSAID Dataset: The data in the first column of Table V present the mIoU for various methods on the iSAID dataset when the proportion of labeled images is 1/8 and 1/4. Compared with supervised training,

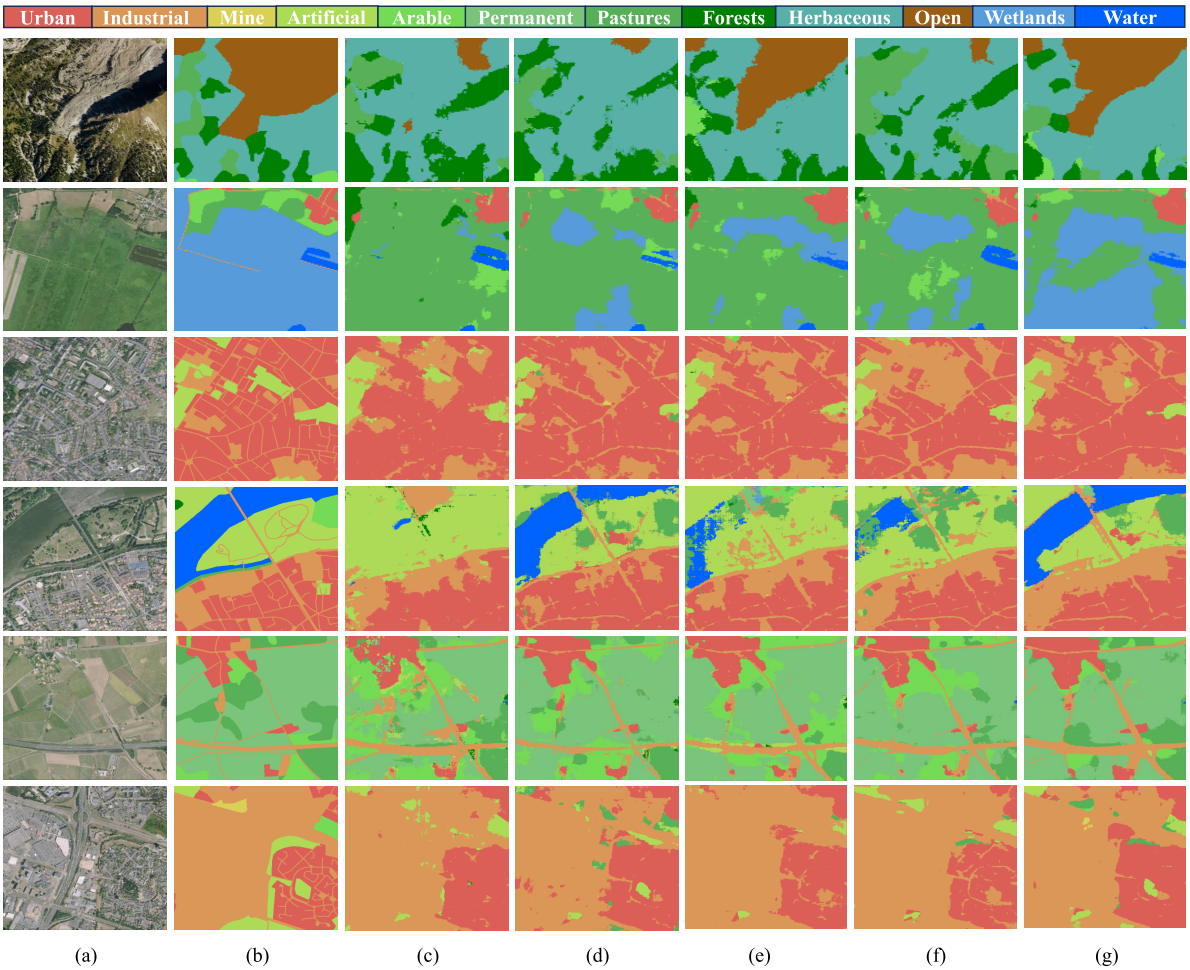

Fig. 9. Visual comparison of our proposed method with other state-of-the-art semi-supervised RS image semantic segmentation methods on the DFC22 dataset. The results are presented. (a) Input image. (b) Ground truth. (c) SupOnly. (d) FixMatch. (e) LSST. (f) WSCL. (g) Ours.

MCSS achieves the greatest increase in mIoU. Specifically, when the proportion of annotated images is 1/8 and 1/4, MCSS's performance improved by 11.64% and 6.04%, respectively. Compared with FixMatch, MCSS shows a performance enhancement of 4.88% and 4.09%, respectively. These experimental results indicate that the proposed semi-supervised learning method can effectively leverage data to significantly enhance network performance, although the accuracy gain brought by semi-supervised learning diminishes as the volume of annotated data increases. Against ICNet, RanPaste, LSST, and WSCL, MCSS achieved the highest mIoU of 70.12% and 80.17%, surpassing the latest WSCL by 1.96% and 2.14%, respectively. This demonstrates the advanced nature of the proposed method.

3) Comparison Results on the MER Dataset: As delineated in the second column of Table V, the mIoU is evaluated on the MER dataset at labeled image ratios of 1/8 and 1/4. MCSS outperforms supervised training, evidenced by the most substantial increment in mIoU, the performance enhancements of 4.77% and 3.81% at labeled image ratios of 1/8 and 1/4, respectively. Relative to the benchmark set by FixMatch, MCSS displays superior efficacy, achieving increases of 3% and 2.56% for the corresponding labeled image ratios. These empirical outcomes validate the capability of MCSS to extensively leverage data, thereby substantially enhancing network performance. In comparison to other leading-edge techniques such as ICNet, RanPaste, LSST, and WSCL, MCSS attains the apex in mIoU, reaching 58.31% and 60.10%, surpassing

TABLE V
COMPARISONS OF OUR METHOD WITH FIVE STATE-OF-THE-ART METHODS ON FIVE DATASETS

Method	iSAID	MER	MSL	GID-15	Vaihingen
1/8 Labeled Images					
SupOnly	58.48	53.54	55.45	73.73	72.01
FixMatch [13]	65.24	55.31	56.27	70.01	73.39
ICNet [27]	64.74	55.12	55.95	69.59	73.38
RanPaste [23]	66.67	56.16	56.60	72.91	74.15
LSST [26]	66.16	55.89	57.80	74.78	73.47
WCSL [28]	68.16	56.44	59.31	76.01	74.39
Ours	**70.12**	**58.31**	**60.60**	**77.02**	**75.16**
1/4 Labeled Images					
SupOnly	74.31	56.29	56.65	76.27	73.60
FixMatch [13]	76.08	57.54	57.65	74.80	74.42
ICNet [27]	75.33	56.56	58.98	72.75	74.95
RanPaste [23]	76.16	57.25	58.96	74.63	74.93
LSST [26]	77.15	56.86	58.79	77.36	74.38
WSCL [28]	78.03	58.31	61.02	78.71	74.80
Ours	**80.17**	**60.10**	**63.21**	**80.44**	**75.55**

the state-of-the-art WSCL by margins of 1.96% and 2.14%, respectively. This highlights the progressive and innovative attributes of our proposed approach.

4) Comparison Results on the MSL Dataset: The data in the third column of Table V present the mIoU of various methods on the MSL dataset at labeled image ratios of 1/8 and 1/4. MCSS, our proposed method, exhibits the most substantial increase in mIoU compared with supervised training. Specifically, MCSS's performance improved by 5.15% and 6.56% at labeled image ratios of 1/8 and 1/4, respectively. In comparison to FixMatch, MCSS demonstrates a notable enhancement in performance, with increases of 4.33% and 5.56% for the respective ratios. These experimental outcomes indicate that our semi-supervised learning approach efficiently uses data to significantly boost network performance, effectively refining the FixMatch algorithm. Against advanced methodologies such as ICNet, RanPaste, LSST, and WSCL, MCSS achieves the highest mIoU of 60.60% and 63.21%, surpassing the latest WSCL by 1.29% and 2.19%, respectively.

5) Comparison Results on the GID-15 Dataset: The data in the fourth column of Table V display the mIoU for various methods on the GID-15 dataset at labeled image ratios of 1/8 and 1/4. Compared with supervised training, MCSS exhibits the most significant increase in mIoU. Specifically, at labeled image ratios of 1/8 and 1/4, MCSS's performance improved by 3.29% and 4.17%, respectively. Compared with ICNet, RanPaste, LSST, and WSCL, MCSS achieved the highest mIoU of 60.60% and 63.21%, outperforming the latest WSCL by 1.29% and 2.19%. Notably, results from FixMatch, ICNet, and RanPaste even show a decline compared with supervised training, indicating the superior robustness of our proposed MCSS, which demonstrates better adaptability across different datasets.

6) Comparison Results on the Vaihingen Dataset: The data in the last column of Table V demonstrate the mIoU for all the methods on the Vaihingen dataset with labeled image ratios of 1/8 and 1/4. It is observed that even under supervised training, the Vaihingen dataset achieves a relatively high mIoU, hence the marginal gains in mIoU brought by semi-supervised learning methods. However, the proposed MDAC still attains the highest mIoU, reaching 75.16% and 75.55% for labeled image ratios of 1/8 and 1/4, respectively. This represents an increase of 3.15% and 1.96% over supervised training, and it is 0.77% and 0.75% higher than the most recent WSCL, respectively.

V. Conclusion

To tackle the challenges posed by the time-consuming and labor-intensive task of pixel-level annotation for massive large-scale RS images, we present an advanced method MCSS for semi-supervised RS semantic segmentation. This approach is designed to make optimal use of limited labeled data in conjunction with abundant unlabeled data. The proposed method encompasses the introduction of MDC, which not only enriches the perturbation space but also reinforces alignment between predictions from multiple SAIs and pseudolabels generated from weakly augmented images, and it empowers the semantic segmentation network to acquire more discerning feature representations. To address the noise issue associated with pseudolabels, we introduce the DDT. Moreover, our method introduces a creative component known as the CPM, which is maintained and plays a pivotal role in rectifying the predictions of the semantic segmentation network by computing the similarity between deep features from unlabeled images and the CPM. Extensive comparative experiments and ablation studies, conducted across the DFC22, iSAID, MER, MSL, Vaihingen, and GID-15 datasets, provide compelling evidence of the advanced capabilities and effectiveness of our proposed method.

References

[1] 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.

[2] L. Jiao et al., "Brain-inspired remote sensing interpretation: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2992–3033, Feb. 2023.

[3] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.

[4] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Aug. 2021, Art. no. 5609413.

[5] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[6] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.

[7] Z. Zhao, L. Zhou, L. Wang, Y. Shi, and Y. Gao, "LaSSL: Label-guided self-training for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 9208–9216.

[8] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–152.

[9] D. Wang, X. Zhang, M. Fan, and X. Ye, "Semi-supervised dictionary learning via structural sparse preserving," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Mar. 2016, pp. 2137–2144.

[10] B. Wang, Z. Tu, and J. K. Tsotsos, "Dynamic label propagation for semi-supervised multi-class multi-label classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 425–432.

[11] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4268–4277.

[12] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1196–1205.

[13] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 596–608.

[14] J. Zhang, Z. Li, C. Zhang, and H. Ma, "Robust adversarial learning for semi-supervised semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, p. 65.

[15] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.

[16] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[17] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7236–7246.

[18] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, Aug. 2021, Art. no. 5509612.

[19] Q. Shi, L. Zhang, and B. Du, "Semisupervised discriminative locally enhanced alignment for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4800–4815, Sep. 2013.

[20] A. Ma, Y. Zhong, B. Zhao, H. Jiao, and L. Zhang, "Semisupervised subspace-based DNA encoding and matching classifier for hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4402–4418, Aug. 2016.

[21] Z. Wang, B. Du, L. Zhang, L. Zhang, and X. Jia, "A novel semisupervised active-learning algorithm for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3071–3083, Jun. 2017.

[22] J. Wang, C. H. Q. Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sens.*, vol. 12, no. 21, p. 3603, Nov. 2020.

[23] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 3102026.

[24] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10548–10559, Oct. 2021.

[25] Y. He, J. Wang, C. Liao, B. Shan, and X. Zhou, "ClassHyPer: ClassMix-based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 14, no. 4, p. 879, Feb. 2022.

[26] X. Lu et al., "Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543516.

[27] J.-X. Wang, S. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] X. Lu et al., "Weak-to-strong consistency learning for semisupervised image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3272552.

[29] Y. Xu, L. Yan, and J. Jiang, "EI-HCR: An efficient end-to-end hybrid consistency regularization algorithm for semisupervised remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3285752.

[30] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[31] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[32] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 429–445.

[33] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2019.

[34] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–18.

[35] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4248–4257.

[36] C. Zhang, J. Cheng, and Q. Tian, "Multi-view image classification with visual, semantic and view consistency," *IEEE Trans. Image Process.*, vol. 29, pp. 617–627, 2020.

[37] X. Liu et al., "Deep multiview union learning network for multisource image classification," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4534–4546, Jun. 2022.

[38] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2020, pp. 1–14.

[39] C. Wei, K. Shen, Y. Chen, and T. Ma, "Theoretical analysis of self-training with deep networks on unlabeled data," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–14.

[40] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9197–9206.

[41] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020.

[42] R. Hansch et al., "The 2022 IEEE GRSS data fusion contest: Semisupervised learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 334–337, Mar. 2022.

[43] S. W. Zamir et al., "ISAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2019, pp. 28–37.

[44] 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, 2022, Art. no. 3152587.

[45] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1492–1500.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–10.

[49] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," *Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.*, vol. 11006, pp. 369–386, May 2019.

Liang Lv received the B.S. degree from the Wuhan Institute of Technology, Wuhan, China, in 2016, and the M.S. degree from Shanghai Jiaotong University, Shanghai, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan.

His research interests include computer vision and remote sensing image analysis.

Lefei Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively.

He was a Big Data Institute Visitor with the Department of Statistical Science, University College London, London, U.K., and a Hong Kong Scholar with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. He is currently a Professor with the School of Computer Science, Wuhan University, and also with the Hubei Luojia Laboratory, Wuhan. His research interests include pattern recognition, image processing, and remote sensing.

Dr. Zhang serves as a Topical Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.