

Bridging the Gap between Labeled and Unlabeled Data via Unified Flow with Feature Memory Bank

Shanwen Wang, *Graduate Student Member, IEEE*, Xin Sun, *Senior Member, IEEE*, Danfeng Hong, *Senior Member, IEEE*, Junyu Dong, *Member, IEEE*, Patrick Le Callet, *Fellow, IEEE*

Abstract—Although semi-supervised semantic segmentation (S^4) utilizes abundant unlabeled data to reduce manual labeling burdens, independent training of labeled and unlabeled data causes the former to dominate, which severely degrades pseudo-label quality. To address this challenges, we propose a novel remote sensing (RS) S^4 method via unified flow with feature memory bank (UFFM). Specifically, UFFM comprises two key innovations: unified flow (UF) and feature memory bank (FMB). The UF is a new training flow that generates less biased pseudo-labels by combining an external visual foundation model (VFM) with an RS domain teacher, and jointly optimizes labeled and pseudo-labeled data under a unified training objective. The FMB is a novel memory module for S^4 that dynamically updates class-specific features during training and reduces the feature discrepancy between labeled and unlabeled data through class-feature alignment. To verify the effectiveness of our model, we conduct extensive experiments on RS datasets. The experimental results show the superiority of our method over SOTA S^4 methods. Moreover, the results demonstrate the effectiveness of our contributions in bridging the optimization and feature representation gap between labeled and unlabeled data. Our code is released at <https://github.com/wangshanwen001/RS-UFFM>.

Index Terms—Semi-supervised semantic segmentation, Visual foundation model, Remote sensing images, Feature memory.

I. INTRODUCTION

SEMANtic segmentation has shown great potential in applications such as flood monitoring [1], precision agriculture [2], ecosystem assessment [3], and urban planning [4]. However, manually labeling extensive RS datasets is highly time-consuming and labor-intensive [5]–[7]. Semi-supervised semantic segmentation (S^4) has attracted significant attention in the remote sensing (RS) community by leveraging a small set of labeled samples alongside a wealth of unlabeled data [8]–[10]. Several RS S^4 models address issues like high inter-class similarity, long-tailed distributions, and low-quality pseudo-labels [11]–[13]. Nevertheless, training labeled and unlabeled data independently prevents effective feature interaction, causing inconsistent semantic representations. Consequently, the explicitly annotated data dominates optimization, severely compromising pseudo-label quality and aggravating

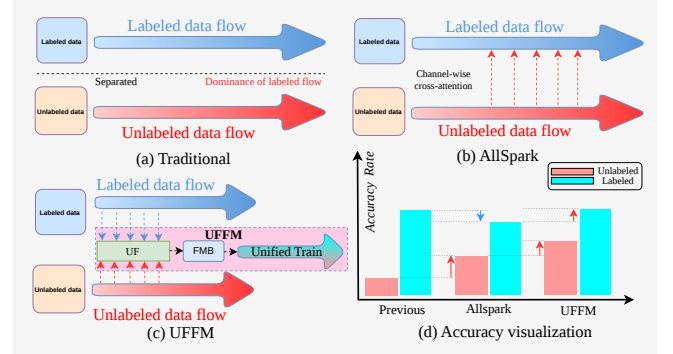


Fig. 1. (a), (b) and (c) Comparison between the training data flows of previous methods, AllSpark and ours UFFM. (d) Comparison between the accuracy rate of previous methods, AllSpark and ours UFFM.

confirmation bias. To address this, AllSpark [14] reconstructs labeled features from unlabeled features using a channel-level cross-attention mechanism. While this approach improves unlabeled data accuracy, it unfortunately degrades labeled data performance, as shown in Fig. 1(d). This occurs because reconstructing labeled samples from unlabeled features can introduce noise into the discriminative features of labeled data. Specifically, during training, these reconstructed samples align poorly with ground-truth annotations, compromising supervised learning effectiveness. This challenge is even more pronounced in the RS domain by extreme label scarcity.

To address the above challenge, we propose a unified flow with feature memory bank (UFFM). As shown in Fig. 1, UFFM bridges the gap between labeled and pseudo-labeled data, surpassing traditional S^4 methods. Specifically, UFFM introduces two key innovations: a unified flow (UF) paradigm and a feature memory bank (FMB). Instead of isolating labeled and unlabeled data, UF integrates them through collaborative pseudo-labeling strategy and a data fusion mechanism. It generates less biased pseudo-labels by combining the external knowledge of a vision foundation model (VFM) with the domain-specific expertise of an in-domain teacher. Then, UF mixes these pseudo-labels with labeled data under a unified loss function to establish a cohesive supervision framework. Furthermore, features of the same category across labeled and unlabeled data should exhibit high similarity. To achieve this, we propose a feature memory bank (FMB) to inject long-term category feature memory into S^4 training. The FMB creates and updates a feature prototype for each semantic category and leverages them to evaluate the reliability of pixel-level pseudo-labels. UFFM integrates UF and FMB to mitigate the dominance of labeled data during training. By optimizing

This work is supported by the Science and Technology Development Fund - International Collaborative Research, Macao SAR (0001/2025/AIJ), Science and Technology Development Fund, Macao SAR - Ministry of Science and Technology: National Key R&D Program of China (0007/2025/AMJ, 2025YFE0202900), and Science and Technology Development Fund, Macao SAR - Basic Research (0006/2024/RIA1)

S. Wang and X. Sun are with Faculty of Data Science, City University of Macau, 999078, SAR Macao, China. D. Hong is with School of Automation, Southeast University, Nanjing, 211189, China. J. Dong are with the Department of Computer Science and Technology, Ocean University of China, Qingdao, 266100 China. P.L. Callet is with Nantes Université, Ecole Centrale Nantes, CAPACITES SAS, CNRS, LS2N, UMR 6004, Nantes, France.

labeled and unlabeled samples within a shared feature space, it effectively bridges the gap between labeled and pseudo-labeled data. It yields superior performance across both labeled and unlabeled data compared to existing S^4 baselines (Fig. 1(d)). In summary, our contributions are as follows:

- 1) We propose UFFM, a novel S^4 model for RS, to address the challenge of separated traditional training paradigm, where explicitly annotated data dominates training and ultimately degrades pseudo-label quality.
- 2) We propose UF that generates less biased pseudo-labels by fusing external VFM knowledge with domain-teacher expertise, and performs a specific unified training by integrating pseudo-labels with ground-truth labels.
- 3) We propose the FMB that introduces long-term category feature memory into S^4 training. It optimizes labeled and unlabeled samples within a shared feature space, by updating a feature prototype for each category.
- 4) Extensive experiments demonstrate the effectiveness of UFFM over state-of-the-art methods. The comprehensive ablation studies highlight our success in bridging the optimization and feature representation gap between labeled and unlabeled data, significantly boosting performance across both data types.

The rest of this article is organized as follows: Section II provides an overview of existing related research. In Section III, we formally propose and analyze our UFFM model. Section IV presents comprehensive experimental results, including comparative analyses with SOTA methods and ablation studies. Finally, Section V concludes and discusses the article.

II. RELATED WORK

This section reviews relevant research on S^4 methods, and also summarizes recent progress in S^4 for RS domain.

A. Semi-supervised Semantic Segmentation

Semantic segmentation is a foundational image analysis task with widespread application in areas such as land cover classification [15], urban modeling [16], and environmental monitoring [17]. However, conventional supervised methods are constrained by the heavy demand and high cost of pixel-level annotations [18]–[21]. Consequently, S^4 has attracted growing interest by leveraging large-scale unlabeled data [22]–[27].

In computer vision, classical S^4 strategies predominantly rely on consistency regularization, pseudo-labeling, teacher–student frameworks, and adversarial learning [28]–[30]. A representative method is FixMatch [31], which generates high-confidence pseudo-labels from weakly augmented unlabeled images to supervise strongly augmented views via consistency regularization. UniMatch [32] extends this paradigm by introducing unified dual-stream perturbations across both image- and feature-level representations, effectively expanding the perturbation space to strengthen weak-to-strong consistency learning. Owing to its superior performance, UniMatch has become a widely adopted baseline. Then, UniMatch_v2 [33] upgrades the architecture by swapping traditional ResNet backbones for vision foundation

models like DINOv2 [34]. To prevent labeled data from dominating the training process, AllSpark [14] reconstructs labeled features from unlabeled representations using a channel-level cross-attention mechanism.

These methods have achieved remarkable performance in natural-image segmentation and offer valuable design insights for RS S^4 approaches [9], [35], [36]. However, RS images present distinct challenges, including substantial scale variations, strong visual similarity across categories, complex backgrounds, and intricate textures [37], [38]. Consequently, directly transferring these general strategies to RS S^4 remains challenging [39].

B. Remote Sensing Semi-supervised Semantic Segmentation

To better accommodate the distinctive characteristics of RS imagery, recent studies have developed S^4 methods specifically tailored to the RS domain [40]–[43]. For example, Ni et al. [44] and Wang et al. [37] addressed multi-scale variations through contextual label refinement in the label space and multi-scale uncertainty consistency, respectively. To mitigate the high visual similarity among different classes, Xin et al. [45] combined contrastive learning with a cross-teacher–student attention network. Huang et al. [12] proposed decoupled weighting learning (DWL) to alleviate the adverse effects of inaccurate pseudo-labels and long-tailed class distributions. DWL decouples the predictions for labeled and unlabeled data during training and introduces a rank-based weighting module that adaptively weights pseudo-labels according to their relative confidence within each pseudo-class, thereby improving the reliability of pseudo-label learning. To address insufficient multimodal fusion and limited pixel-level annotations, Li et al. [42] proposed Semi-Mamba, which incorporates a semi-supervised Mamba-based cross-modality fusion module to facilitate cross-modal feature interaction in RS imagery. Furthermore, Geng et al. [46] proposed KGCSL to address the scarcity of labeled samples in hyperspectral image (HSI) classification. This semi-supervised method integrates multi-scale and multi-directional geometric features extracted using the contourlet transform, enabling accurate HSI classification under limited-label conditions.

With the rapid advancement of large-scale foundation models, vision–language models (VLMs) [47] and visual foundation models (VFMs) [48] have increasingly been incorporated into RS S^4 . For example, SemiEarth [13] is the first approach to leverage VLMs to improve pseudo-label quality, thereby enhancing the performance of S^4 on RS imagery. Similarly, Song et al. [49] employed multiple VFMs as teachers and introduced a distillation-and-fusion mechanism to guide the training of an RS S^4 framework.

However, existing RS S^4 methods remain largely confined to conventional architectural paradigms. These methods often prioritize increasingly complex framework designs while overlooking the optimization imbalance between labeled and unlabeled data. Consequently, the supervised pathway with labeled samples dominates the training process, resulting in low-quality of the pseudo-labels generated for unlabeled data. To address this limitation, we propose UFFM, which bridges the gap between labeled and pseudo-labeled data.

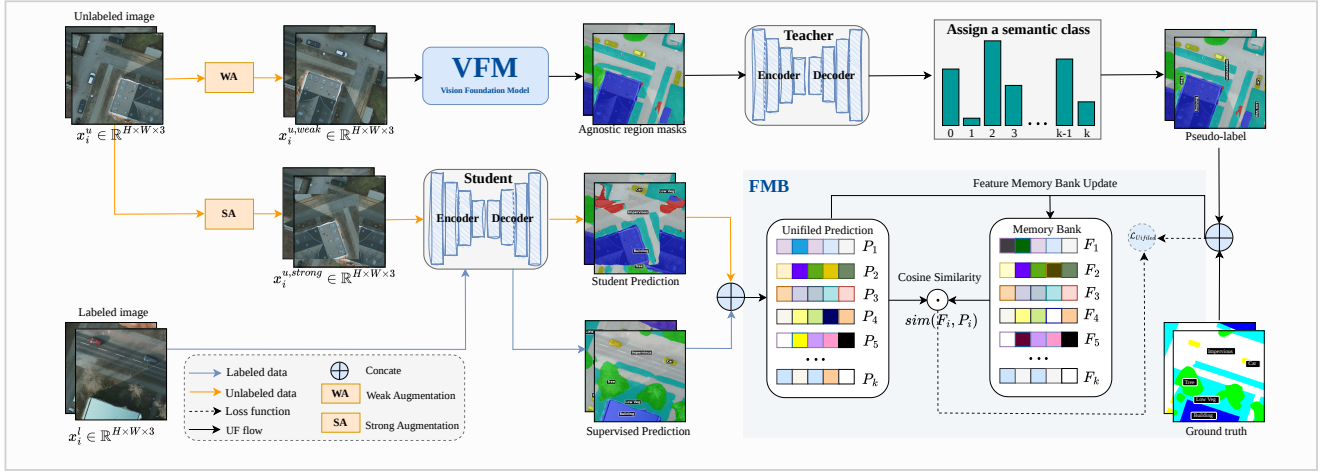


Fig. 2. Overall architecture of our UFFM model for RS images. Flow paths are color-coded: blue for labeled data, yellow for unlabeled data, black for UF processing, and dashed for loss functions. As shown by the black path in the top row, pseudo-labels in the UFFM model are initially generated by the VFM as class-agnostic masks and subsequently mapped to semantic classes by the teacher. The light blue-gray region highlights the FMB module, which comprises two stages: feature memory bank updating and loss computation, where loss weights are determined via cosine similarity.

III. METHODS

This section is organized as follows. Section III-A describes the main framework, Section III-B introduces the principles of the UF module, and Section III-C presents the basic principles of the FMB module.

A. Main Framework

Semi-supervised learning trains a model using a small set of labeled samples together with a large amount of unlabeled data. As discussed above, conventional strategies that process these two data sources separately fail to fully exploit the potential of unlabeled data. To overcome this limitation, we propose UFFM, a specific S^4 framework illustrated in Fig. 2. Unlike conventional approaches that treat labeled and unlabeled samples as separate training sets, UFFM integrates them through a novel pseudo-labeling strategy and data-fusion mechanism.

We denote the labeled set as $\mathcal{D}^L = \{(x_i^l, y_i^l)\}_{i=1}^{N_L}$ and the unlabeled set as $\mathcal{D}^U = \{x_i^u\}_{i=1}^{N_U}$, where $x_i^l \in \mathbb{R}^{H \times W \times 3}$ and $x_i^u \in \mathbb{R}^{H \times W \times 3}$ denote the labeled and unlabeled images, respectively, with height H and width W . $y_i^l \in \{0, 1\}^{H \times W \times K}$ denotes the ground-truth across K classes. N_L and N_U represent the total counts of labeled and unlabeled samples, where typically $N_U \gg N_L$. The standard S^4 loss function is given by:

$$\mathcal{L} = \frac{1}{N_L} \sum_{i=1}^{N_L} \mathcal{L}_{CE}(p_i^l, y_i^l) + \frac{1}{N_U} \sum_{i=1}^{N_U} \mathcal{L}_{CE}(p_i^{u,s}, y_i^{u,t}), \quad (1)$$

where the former component denotes the supervised loss on labeled samples and the latter corresponds to the unsupervised loss on unlabeled images. p_i^l is the prediction for labeled input x_i^l , and $\mathcal{L}_{CE}$ represents the cross-entropy loss. While standard S^4 methods process labeled and unlabeled data separately, our proposed UFFM framework trains them jointly through a specially designed weighting strategy. The main loss function is formulated as:

$$\mathcal{L} = \frac{1}{N_L + N_U} \sum_{i=1}^{N_L + N_U} \mathcal{L}_{CE}([w_i^l, w_i^u]([p_i^l, p_i^{u,s}], [y_i^l, y_i^{u,t}])). \quad (2)$$

where w_i^l and w_i^u are derived from the FMB module, the details of which are provided in Section III-C. As illustrated in Fig. 2, UFFM unifies supervised and unsupervised learning within a teacher–student framework. For the unsupervised branch, both weak and strong augmentations are applied to unlabeled data to facilitate reliable pseudo-label generation and enhance feature learning robustness. Specifically, weakly augmented samples provide stable representations for the VFM and teacher model to yield high-confidence pseudo-labels, whereas strongly augmented samples expose the student model to broader data variations. The student model is then optimized under the supervision of the pseudo-labels generated by the UF module. The teacher parameters θ^t are updated via Exponential Moving Average (EMA) from student parameters θ^s , i.e., $\theta_t^t = \alpha \theta_{t-1}^t + (1 - \alpha) \theta_t^s$ where α denotes the decay rate. However, conventional methods process labeled and unlabeled data independently, allowing labeled samples to dominate optimization and degrade pseudo-label quality. To resolve this limitation, we introduce the UF training flow and FMB module, detailed in the subsequent sections.

B. Unified Flow

This section details the principle of the proposed UF module. Unlike traditional paradigms that process labeled and unlabeled data independently, UF introduces a joint pseudo-labeling and data-fusion strategy. To generate less biased pseudo-labels, UF leverages cross-domain collaboration between a VFM and an in-domain teacher network. This strategy leverages the external knowledge encoded in the VFM to mitigate biases inherited from the labeled data while effectively exploiting the teacher model’s domain-specific knowledge. Specifically, for each unlabeled image, the VFM first generates a set of class-agnostic regions. Each region is then assigned

the dominant class predicted by the teacher model, and its reliability is assessed based on class consistency and the teacher's mean confidence. Only regions that satisfy both reliability thresholds are retained as VFM-assisted pseudo-labels. For pixels not covered by reliable VFM regions, we use the teacher predictions that satisfy the conventional confidence threshold. Combining these two sources yields the final pseudo-label map for each unlabeled training sample, thereby reducing its dependence on biases inherited from the labeled data. This process is formally defined as follows. First, the weakly augmented unlabeled image $x_i^{u,weak}$ is fed into the VFM to generate a set of class-agnostic region masks:

$$M_i = \{m_i^1, m_i^2, \dots, m_i^j\} = \text{VFM}(x_i^{u,weak}), \quad (3)$$

where m_i^j denotes the j^{th} class-agnostic mask region generated by the VFM for unlabeled image i . The teacher model yields the class probability distribution for each pixel:

$$p_i^{u,t}(x_i^{u,weak}) = \text{Softmax}(\text{Teacher}(x_i^{u,weak})). \quad (4)$$

The predicted class $y_i^{u,t}$ and confidence score c_i^k of the teacher for each pixel are given by:

$$y_i^{u,t} = \arg \max_{k \in [1, K]} (p_i^{u,t}(x_i^{u,weak})), \quad (5)$$

$$c_i^k = \max_{k \in [1, K]} (p_i^{u,t}(x_i^{u,weak})), \quad (6)$$

where k indexes the classes. For each region m generated by the VFM, a semantic class is assigned based on the teacher's pixel-level predictions within that region. Specifically, the majority class occupying the largest proportion of the region is determined as:

$$y_i^V(m) = \arg \max_{k \in [1, K]} \sum_{m \in M} (y_i^{u,t}(m) = k). \quad (7)$$

We calculate both the class consistency and the average confidence within the target region. This design stems from the rationale that significant discrepancies between the segmentation predictions of the VFM and the domain-specific teacher model signify low inter-model consistency. It indicates that either the VFM produces inaccurate boundaries or the teacher model predicts incorrect semantic categories. Consequently, a region is regarded as a reliable VFM pseudo-label only when both the consistency and confidence thresholds are satisfied.

$$\gamma_i^m = \frac{\sum_{m \in M} (y_i^{u,t}(m) = y_i^V(m))}{|M|}, \quad (8)$$

$$c_i^m = \frac{1}{|M|} \sum_{m \in M} c_i^k, \quad (9)$$

$$Q_i^m = (1 * (\gamma_i^m \geq \tau_{\text{consistency}})) * (1 * (c_i^m \geq \tau_{\text{conf}})), \quad (10)$$

where $\tau_{\text{consistency}}$ represents the consistency threshold between the RS teacher and the VFM, and τ_{conf} denotes the pseudo-label confidence threshold. Consequently, the pixel-level pseudo-labels are defined as follows.

$$y_i^V = \begin{cases} y_i^V(m), & m \in M, Q_i^m = 1 \\ 0, & \text{otherwise} \end{cases}, \quad (11)$$

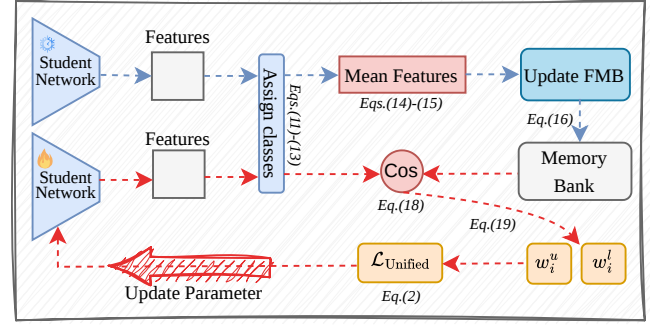


Fig. 3. The overall workflow of our FMB. It mainly consists of two stages: the memory bank update stage (the red lines) and the loss computation stage (the blue lines).

where 0 denotes unreliable pixels. If a VFM region is classified as unreliable, the method reverts to standard teacher pseudo-labels.

$$y_i^{u,t} = \begin{cases} y_i^{u,t}(m), & m \in M, c_i^k \geq \tau_{\text{conf}} \\ 0, & \text{otherwise} \end{cases}. \quad (12)$$

Unreliable pixels from the teacher model are simply discarded during training. By combining Equations (11) and (12), the final UF pseudo-labels y_i^u are unified as follows:

$$y_i^u = \begin{cases} y_i^V(m), & m \in M, Q_i^m = 1 \\ y_i^{u,t}(m), & m \in M, c_i^k \geq \tau_{\text{conf}} \\ 0, & \text{otherwise} \end{cases}. \quad (13)$$

We separately weight the generated pseudo-labels y_i^u and the ground truth p_i^l using the FMB module. Finally, the weighted pseudo-labels and ground-truth labels are combined to compute the loss according to Eq. 2, completing the UF pipeline. UF combines a VFM with the teacher model to yield less biased pseudo-labels, establishing a unified form of supervision with the ground truth. The next section details the principles and implementation of the FMB module.

C. Feature Memory Bank

The FMB is a novel memory module for updating class feature prototypes in S^4 , reducing the feature discrepancy between labeled and unlabeled data through class-feature alignment. The core principle of FMB is that samples from the same category, whether labeled or unlabeled, should exhibit similar feature representations. The overall workflow of FMB is illustrated in Fig. 3. FMB consists of two main stages: memory bank updating (indicated by red lines) and loss computation (indicated by blue lines). The update stage maintains a set of class prototypes within the memory bank to capture and preserve shared characteristics across both labeled and unlabeled data. In the loss computation stage, the model's learned features are compared against these prototypes. By aligning representations with prototypes that encode common characteristics, the proposed approach optimizes labeled and unlabeled samples within a unified feature space, ultimately preventing labeled data from dominating the training process.

During the FMB update stage, which updates the class prototypes, the student encoder is kept frozen to encode image

features. FMB applies the same update mechanism to both labeled and unlabeled data. For each class $k \in \{1, \dots, K\}$, we update a class prototype F_k . Let Ω_k denote the set of pixels belonging to class k within a training batch. To determine class membership, unlabeled pixels rely on the pseudo-labels y_i^u provided by UF, whereas labeled pixels use ground-truth annotations. Accordingly, the mean feature vector of class k within a training batch is given by:

$$f_k = \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} f_i, \quad (14)$$

where f_i denotes the feature vector of pixel i , while f_k represents the mean feature vector of all pixels belonging to class k . Before updating the FMB prototypes, we perform normalization on f_k as follows:

$$f_k = \frac{f_k}{\|f_k\|_2}. \quad (15)$$

The class prototypes in FMB are updated using an EMA mechanism as follows:

$$F_k^t = \alpha F_k^{t-1} + (1 - \alpha) f_k, \quad (16)$$

where α is the EMA decay and t is training step. Through this mechanism, the class prototypes in FMB are updated to preserve the shared characteristics between labeled and unlabeled data. Next, we describe how FMB derives the loss weights w_i^l and w_i^u in Eq. 2 for labeled and unlabeled pixels, respectively.

Having established the updated class prototypes in the FMB module, we next detail how they are utilized during loss computation, as highlighted by the red paths in Fig. 3. Taking unlabeled data as an example, FMB receives both the UF pseudo-labels and the student network's predictions as input. We first use the UF pseudo-label y_i^u to identify the corresponding class prototype. Then, we compute the cosine similarity between the student network's predictions and their corresponding class prototypes as follows:

$$p_i^{u,s} = \text{Student}(x_i^{u,strong}), \quad (17)$$

$$\cos_i^u = \frac{p_i^{u,s} F_k}{\|p_i^{u,s}\|_2 \|F_k\|_2}. \quad (18)$$

The resulting cosine similarity values, $\cos_i^u$, fall within the interval $[-1, 1]$. To guarantee that final loss function yields non-negative values, $\cos_i^u$ is linearly transformed into w_i^u according to:

$$w_i^u = \frac{\cos_i^u + 1}{2}. \quad (19)$$

The value of w_i^u lies in the range $[0, 1]$. Similarly, for labeled data, the ground-truth labels y_i^l and the student model predictions are fed to FMB. Following the same procedure as unlabeled data, we compute the corresponding weight w_i^l . Finally, these FMB weights are integrated into the UF loss function, as expressed in Eq. 2.

Algorithm 1 summarizes the core training procedure of UFFM to clarify the execution workflow, with the primary logic outlined in lines 7–17. Here, D^U and D^L denote the unlabeled and labeled datasets for a given epoch, containing

N_U and N_L images, respectively. $model_s$ and $model_t$ represent the student and teacher networks. Furthermore, $p_i^{u,s}$ and p_i^l denote the student's output predictions for unlabeled and labeled samples, respectively. Here, y_i^u denotes the final UF pseudo-label, y_i^l denotes the ground-truth label, and $\mathcal{L}_{CE}$ represents the cross-entropy loss function.

Algorithm 1 Training procedure of UFFM

Input:

$$D^U = \{(x_i^u)\}_{i=1}^{N_U}, D^L = \{(x_i^l, y_i^l)\}_{i=1}^{N_L}$$

Output:

Θ : optimal model parameters

1: **while** until converge:

2: **for** x_i^l, x_i^u in D^L, D^U :

3: $x_i^{u,weak} = \text{WeakAugment}(x_i^u)$

4: $x_i^{u,strong} = \text{StrongAugment}(x_i^u)$

5: $p_i^l = model_s(x_i^l)$

6: $p_i^{u,s} = model_s(x_i^{u,strong})$

7: $y_i^{u,t} = model_t(x_i^{u,weak})$

8: $M_i = \{m_i^1, m_i^2, \dots, m_i^j\} = \text{VFM}(x_i^{u,weak})$

9: **for** pixels in m_i^j :

10: **if** $Q_i^m = 1$:

11: $y_i^u = y_i^V(m)$

12: **else if** $Q_i^m \neq 1$ and $c_i^k > \tau_{\text{conf}}$:

13: $y_i^u = y_i^{u,t}(m)$

14: **else**:

15: $y_i^u = 0$

16: **end for**

17: Update FMB by Eq. 14-16.

18: Use FMB to compute w_i^l and w_i^u via Eq. 19.

19: $\mathcal{L} = \frac{1}{N_L + N_U} \sum_{i=1}^{N_L + N_U} \mathcal{L}_{CE}([w_i^l, w_i^u]([p_i^l, p_i^{u,s}], [y_i^l, y_i^u]))$

20: Update Θ via gradient descent on $\mathcal{L}$

21: Save the best checkpoint Θ_{best}

22: **end for**

23: **return** Θ

24: **end**

In this section, we presented the core principles of UF and FMB, providing a thorough analysis of how they bridge the optimization and feature representation gap between labeled and unlabeled data. In the following section, we conduct extensive experiments to demonstrate the effectiveness of our proposed methods.

IV. EXPERIMENTS

In this section, we evaluate the proposed UFFM model on benchmark remote sensing datasets. We first conduct comparative experiments against several state-of-the-art methods. Subsequently, we perform comprehensive ablation studies to validate the effectiveness of each proposed module. Our code is released at <https://github.com/wangshanwen001/RS-UFFM>.

A. Datasets

DeepGlobe: The DeepGlobe land cover dataset [50] is widely used for remote sensing image semantic segmentation

and land cover mapping. It features a spatial resolution of 0.5 meters and comprises 803 high-resolution satellite images of 2448×2448 pixels across seven categories, including Urban land, Agriculture land, Rangeland, Forest land, Water, Barren land, and Unknown. To facilitate model training, we cropped the original images into 512×512 patches, yielding 20,075 images in total. The cropped dataset is split into training, validation, and test sets containing 12,045, 4,015, and 4,015 patches, respectively.

ISPRS-Potsdam: The ISPRS Potsdam benchmark dataset is widely used for semantic segmentation on high-resolution remote sensing images [51]. It features a spatial resolution of 0.05 meters and consists of 38 large-scale satellite images, each with a size of 6000×6000 pixels. The dataset covers six land cover categories including impervious surfaces, buildings, low vegetation, trees, cars, and background. For efficient training, the original images are cropped into patches of 512×512 pixels, yielding 5,472 image patches. These patches are partitioned into training, validation, and test sets using a 6:2:2 ratio, containing 3,283, 1,094, and 1,095 images, respectively.

B. Data Augmentation and Experiment Settings

We first apply data augmentation to the labeled images using geometric transformations (including scaling, horizontal flipping, vertical flipping, and aspect ratio warping) along with Gaussian blurring. For unlabeled images, weak augmentation consists exclusively of geometric transformations, whereas strong augmentation incorporates photometric transformations, Gaussian blur, and CutMix [52].

All experiments are conducted on a single NVIDIA RTX H100 GPU using CUDA v11.7. Within our UFFM framework, SAM 3 [53] serves as the external-knowledge VFM model, while both the student and teacher adopt a DINOv2-small [34] backbone. To evaluate performance across varying supervision regimes, models are trained for 50 epochs on the ISPRS Potsdam and DeepGlobe datasets using 1%, 5%, and 10% labeled data partitions alongside the remaining unlabeled samples.

C. Evaluation metrics

Following the standard evaluation protocol of previous RS S^4 methods [12], we use mean Intersection-over-Union (mIoU) as the primary metric to assess model performance. For DeepGlobe and ISPRS-Potsdam, unknown and background pixels are excluded from evaluation, and mIoU is averaged over the six and five foreground classes, respectively. The mIoU is calculated as follows.

$$IoU_k = \frac{TP_k}{TP_k + FP_k + FN_k}, \quad (20)$$

$$mIoU = \frac{1}{K} \sum_{k=1}^K IoU_k, \quad (21)$$

where TP_k , FP_k , and FN_k represent true positives, false positives, and false negatives for class k , and K is the total number of classes.

TABLE I
COMPARISON RESULTS WITH SOTA METHODS ON DEEPGLOBE DATASET.
THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. IOU AND MIOU ARE
REPRESENTED AS PERCENTAGES.

Ratio	Model	IoU						mIoU
		Urban	Agr.	Range.	Forest	Water	Barren	
1%	CCT [54]	70.86	70.64	11.03	62.44	28.76	27.66	45.23
	CPS [55]	80.94	70.66	1.16	63.91	27.45	0.79	40.82
	LSST [56]	79.35	73.41	21.60	60.76	30.40	25.64	48.53
	FixMatch [31]	82.51	74.10	18.79	67.65	44.72	32.74	53.42
	UniMatch [32]	80.54	70.72	20.71	65.48	34.09	9.24	46.80
	DWL [12]	81.66	75.40	21.82	67.10	63.04	35.27	57.38
	AllSpark [14]	80.69	71.15	19.91	65.60	62.47	31.10	55.15
	SemiVL [57]	81.45	72.72	22.44	68.24	64.40	35.76	57.50
	UniMatch_v2 [33]	81.34	72.35	22.63	68.14	65.16	36.34	57.66
	SemiEarth [13]	81.63	71.43	23.43	68.74	66.99	43.16	59.23
	Our (UFFM)	82.26	73.41	20.62	70.97	71.52	50.49	61.55
5%	CCT [54]	81.20	76.14	12.38	64.05	49.88	42.97	54.44
	CPS [55]	84.15	78.67	11.31	71.15	57.49	43.23	57.67
	LSST [56]	84.26	81.67	30.71	68.25	65.62	55.16	64.28
	FixMatch [31]	85.31	82.96	32.22	67.47	69.76	59.09	66.13
	UniMatch [32]	84.13	81.36	30.69	69.83	65.84	54.38	64.37
	DWL [12]	86.08	83.43	36.62	70.22	70.77	59.86	67.83
	AllSpark [14]	80.88	81.23	31.28	66.65	67.46	57.14	64.11
	SemiVL [57]	85.96	73.94	28.48	76.39	78.58	64.16	67.92
	UniMatch_v2 [33]	84.24	73.12	29.13	76.36	78.44	62.28	67.26
	SemiEarth [13]	86.20	70.60	30.44	77.40	80.61	64.33	68.26
	Our (UFFM)	86.54	72.69	31.92	74.98	81.45	64.54	68.69
10%	CCT [54]	83.22	80.80	29.47	70.37	63.16	49.08	62.68
	CPS [55]	85.97	82.82	28.20	72.03	66.97	53.76	64.96
	LSST [56]	85.53	83.14	36.67	71.34	70.78	57.99	67.58
	FixMatch [31]	86.53	84.01	36.57	71.26	69.88	57.38	67.60
	UniMatch [32]	84.88	82.75	34.36	69.87	66.61	53.03	65.25
	DWL [12]	85.46	83.63	38.95	72.40	70.76	60.33	68.59
	AllSpark [14]	83.66	82.21	33.41	67.80	68.24	57.22	65.42
	SemiVL [57]	85.25	80.73	38.70	76.14	73.64	63.33	69.63
	UniMatch_v2 [33]	83.16	81.53	37.11	75.94	75.39	66.57	69.95
	SemiEarth [13]	86.70	73.26	40.20	78.92	76.58	67.24	70.48
	Our (UFFM)	87.27	76.65	41.15	79.32	77.60	67.25	71.54

D. Quantitative Results compared to SOTA

This section conducts experiments on DeepGlobe and ISPRS-Potsdam datasets, compared to the SOTA methods, including CCT [54], CPS [55], LSST [56], FixMatch [31], UniMatch [32], UniMatch_v2 [33], DWL [12], Allspark [14], SemiVL [57], and SemiEarth [13]. Specifically, we report the results under labeled data ratios of 1%, 5%, and 10% to comprehensively evaluate the effectiveness of our method. For all comparative methods, network configurations strictly follow the default settings reported in their respective papers and public code repositories. Some of the baseline results are reported directly from their original publications.

Tables I and II present the quantitative results on the DeepGlobe and ISPRS Potsdam datasets, respectively. Here, Agr. and Range. are abbreviations for Agriculture and Rangeland, respectively. We can observe that traditional S^4 methods, such as FixMatch and UniMatch, perform poorly. This is because, although they improve the accuracy of unlabeled data in various ways, they overlook the unique domain-specific challenges inherent to remote sensing imagery. Conversely, dedicated RS S^4 approaches like DWL achieves noticeable improvements by incorporating architectures tailored to handle distinct RS characteristics, such as rich multi-scale features. Although VLM-based frameworks like SemiVL and SemiEarth outperform standard S^4 baselines, they remain inferior to the proposed UFFM model. Furthermore, the results show that, on both

TABLE II
COMPARISON RESULTS WITH SOTA METHODS ON ISPRS-POTSDAM DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. IOU AND mIOU ARE REPRESENTED AS PERCENTAGES.

Ratio	Model	IoU					mIoU
		Building	Low vegetation	Tree	Car	Impervious surfaces	
1%	CCT [54]	54.48	61.28	48.56	52.95	60.71	55.59
	CPS [55]	59.35	69.16	62.89	59.88	66.33	63.52
	LSST [56]	68.74	75.24	54.74	62.09	68.80	65.92
	FixMatch [31]	76.95	71.59	64.71	65.85	72.81	70.38
	UniMatch [32]	76.52	70.99	65.44	66.62	72.64	70.44
	DWL [12]	72.34	77.08	62.74	62.57	72.22	69.39
	AllSpark [14]	83.70	65.92	59.64	69.77	75.31	70.87
	SemiVL [57]	84.73	67.28	58.87	72.95	77.16	72.20
	UniMatch_v2 [33]	84.75	67.83	65.88	75.50	77.15	74.22
	SemiEarth [13]	86.80	71.22	71.96	76.11	79.01	77.02
	Our (UFFM)	88.17	70.87	71.37	75.87	79.48	77.15
5%	CCT [54]	72.90	80.25	64.23	58.32	74.42	70.02
	CPS [55]	76.53	84.34	57.98	69.45	75.39	72.74
	LSST [56]	69.26	84.55	67.33	67.49	73.86	72.50
	FixMatch [31]	78.12	74.87	68.89	66.58	75.30	72.75
	UniMatch [32]	78.24	73.59	67.17	66.64	75.07	72.14
	DWL [12]	74.81	85.64	66.38	62.99	75.68	73.10
	AllSpark [14]	85.57	67.62	60.61	73.48	77.15	72.88
	SemiVL [57]	87.58	70.37	63.59	75.06	78.92	75.10
	UniMatch_v2 [33]	83.75	76.47	66.40	75.95	77.19	75.95
	SemiEarth [13]	88.51	74.45	74.06	78.14	79.87	79.01
	Our (UFFM)	88.68	73.48	74.85	78.72	80.44	79.23
10%	CCT [54]	73.09	83.94	61.12	60.45	73.06	70.33
	CPS [55]	77.80	87.15	61.12	68.48	75.89	74.09
	LSST [56]	70.92	86.06	68.91	70.22	74.89	74.20
	FixMatch [31]	77.97	76.17	70.09	70.97	76.14	74.27
	UniMatch [32]	77.34	87.75	70.79	56.65	76.46	73.80
	DWL [12]	76.37	88.42	66.54	64.37	77.14	74.57
	AllSpark [14]	86.29	69.83	64.17	75.23	78.31	74.76
	SemiVL [57]	88.62	72.23	65.40	76.34	80.71	76.66
	UniMatch_v2 [33]	84.93	78.59	68.97	75.54	79.83	77.57
	SemiEarth [13]	90.59	75.44	75.01	79.64	83.24	80.78
	Our (UFFM)	92.03	76.01	76.18	80.08	84.38	81.73

RS datasets, our proposed UFFM achieves the highest mean mIoU across almost all categories. This is because, although previous S^4 methods improve pseudo-label quality through various methods or filter out low-confidence pseudo-labels, they overlook the degradation of pseudo-label quality caused by the labeled-data dominated training process. In contrast, UFFM introduces a new strategy that effectively addresses this limitation.

E. Visualization Results compared to SOTA

To provide an intuitive comparison, we present visual segmentation results alongside state-of-the-art (SOTA) methods on the RS datasets. Fig. 4 illustrates qualitative comparisons on the ISPRS Potsdam dataset. Existing methods exhibit noticeable segmentation errors in complex regions, particularly where class boundaries overlap or intermix. Specifically, as highlighted within the black dashed boxes in Fig. 4, baseline models frequently misclassify the Low vegetation and Tree categories. In contrast, our proposed UFFM produces significantly cleaner boundaries and more accurate predictions.

The visualization results on the DeepGlobe are illustrated in Fig. 5. It can be observed that most models suffer from significant segmentation errors across various scenes. Specifically, in the first row, FixMatch and UniMatch misclassify water and rangeland areas as Agriculture, whereas AllSpark and DWL falsely predict non-existent Barren land. In the

TABLE III
ABLATION OF THE COMPONENTS.

Dataset	Network	Labeled mIoU	Unlabeled mIoU	mIoU
DeepGlobe	Baseline	91.32	56.89	57.35
	UF	97.83	59.74	60.94
	UF+FMB	98.49	60.97	61.55
ISPRS-Potsdam	Baseline	88.09	70.82	71.51
	UF	94.50	74.48	75.74
	UF+FMB	96.92	77.01	77.15

second row, FixMatch, AllSpark, and DWL misidentify large regions of Barren land as Water, while UniMatch misclassifies large areas as Forest. Although SemiEarth yields relatively better predictions, it still suffers from large-scale regional errors. In contrast, UFFM achieves the best segmentation performance, producing precise predictions in fine-grained regions while better preserving structural boundary details. Overall, our proposed model achieves the most robust and accurate performance among all evaluated methods.

F. Ablation Study

In this subsection, we conduct detailed ablation experiments on UFFM to validate the rationality of our model. We also provide an in-depth analysis and explanation of how our method bridges the optimization and feature representation gap between labeled and unlabeled data.

1) *Ablation of Components*: To evaluate the individual contribution of each component, we conduct ablation experiments under a 1% labeled data setting using DINOv2-small as the backbone for both teacher and student networks and SAM 3 as the VFM. We systematically analyze our framework across three configurations: Baseline, UF, and UF + FMB. Specifically, the Baseline model operates without any of our proposed components.

The ablation results are presented in Table III. It can be seen that without the proposed UF and FMB, the performance is poor. Introducing the UF module improves mIoU over the baseline by 3.59% on the DeepGlobe dataset and 4.23% on the Potsdam dataset. Incorporating the FMB module provides additional gains, further boosting mIoU by 0.61% and 1.41% on DeepGlobe and Potsdam, respectively. These results confirm that UF and FMB significantly enhance S^4 in the RS domain. Furthermore, evaluating performance across labeled and unlabeled subsets reveals that adding UF improves mIoU for both data types, with the joint application of UF and FMB achieving the best overall performance. This validates the necessity and effectiveness of both proposed modules.

2) *Hyperparameters Analysis of UFFM*: We perform ablation studies on the key hyperparameters of UFFM, beginning with the consistency threshold, $\tau_{\text{consistency}}$, between the VFM and the domain-specific teacher model. As shown in Fig. 6, as $\tau_{\text{consistency}}$ increases, the mIoU initially improves before subsequently declining, achieving optimal performance at $\tau_{\text{consistency}} \approx 0.6$. This can be attributed to the fact that, when $\tau_{\text{consistency}}$ is too low, categories for which the VFM and in-domain teacher model disagree substantially are still accepted.

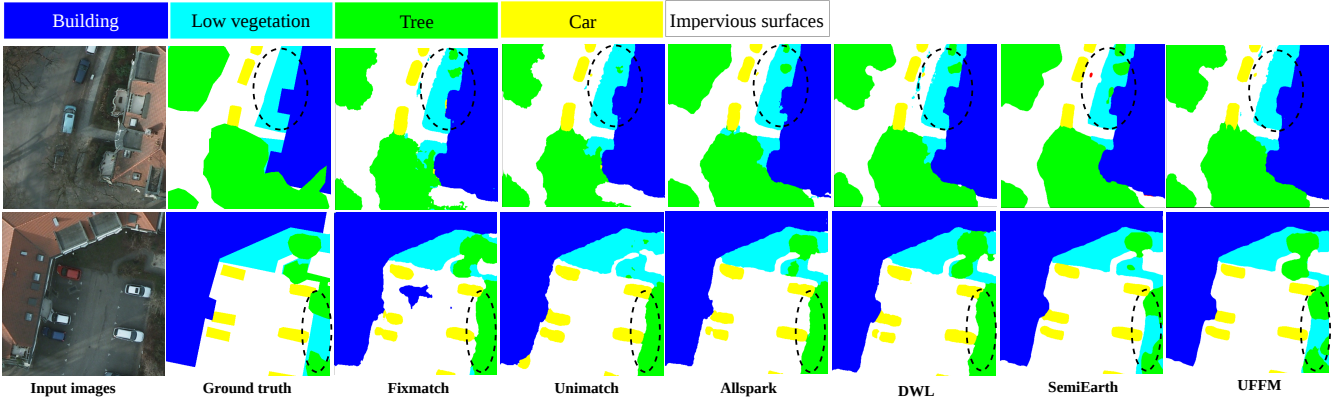


Fig. 4. Visual comparison of semantic segmentation results with different semisupervised methods on the ISPRS-Potsdam dataset.

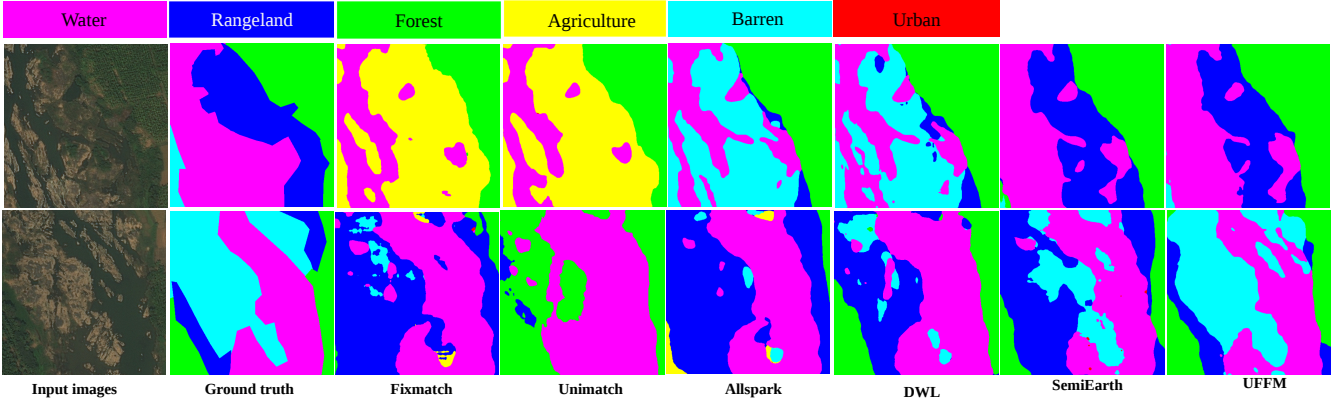
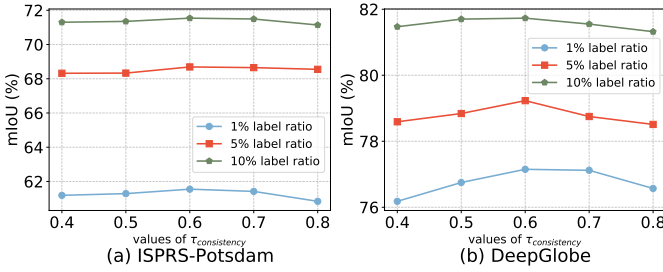
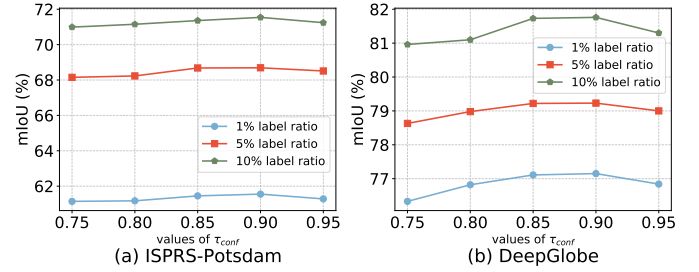


Fig. 5. Visual comparison of semantic segmentation results with different semisupervised methods on the DeepGlobe dataset.

Fig. 6. Hyperparameters Analysis $\tau_{consistency}$ for UF.Fig. 7. The ablation study of the hyperparameters τ_{conf} for UF.

Conversely, when $\tau_{consistency}$ is too high, the criterion becomes overly restrictive, causing most pixels to revert to the original teacher model's predictions.

We conducted experiments to evaluate the effect of the pseudo-label confidence threshold τ_{conf} . As the final safeguard for pseudo-label quality, τ_{conf} was set to a relatively high value in our experiments and varied from 0.75 to 0.95. As shown in Fig. 7, mIoU follows an inverted U-shape trend, i.e., first increasing and then decreasing as τ_{conf} rises. This occurs because a lower threshold allows a large number of low-quality pseudo-labels into training, thereby introducing noise and hindering student model optimization. Conversely, a high threshold τ_{conf} filters out too many pseudo-labels, starving the model of sufficient supervision. Peak mIoU is achieved when τ_{conf} is configured between 0.85 and 0.95.

3) Backbone Investigation: We further investigate the impact of backbone architecture by evaluating DINOv2 and DINOv3 backbones ranging from small to large for both the student and teacher models (Table IV). The results show that mIoU on the RS dataset consistently improves as model size increases. These findings confirm that UFFM effectively accommodates backbones of different parameter scales, highlighting its generalizability and versatility of the UFFM network architecture.

We further observe that DINOv3 consistently outperforms DINOv2 on DeepGlobe, while exhibiting degraded performance on Potsdam. We attribute this discrepancy to the larger patch size adopted by DINOv3. Given the high spatial resolution of Potsdam (0.05m), the larger patch size may discard fine-grained local information, leading to inferior segmentation performance. Considering the trade-off between model size

TABLE IV
THE BACKBONE INVESTIGATION.

Dataset	Backbone	mIoU	Params
DeepGlobe	DINOv2-small	61.55	24.8M
	DINOv2-base	67.38	97.5M
	DINOv2-large	68.74	335.6M
	DINOv3-small	62.81	24.3M
	DINOv3-base	67.92	96.6M
	DINOv3-large	69.01	334.4M
ISPRS-Potsdam	DINOv2-small	77.15	24.8M
	DINOv2-base	80.03	97.5M
	DINOv2-large	80.86	335.6M
	DINOv3-small	73.60	24.3M
	DINOv3-base	75.67	96.6M
	DINOv3-large	78.06	334.4M

TABLE V
ABLATION STUDY OF UF.

Dataset	Network	Labeled mIoU	Unlabeled mIoU	mIoU
DeepGlobe	Baseline	91.32	56.89	57.35
	VFM	92.26	57.35	58.56
	Full UF	97.83	59.74	60.94
ISPRS-Potsdam	Baseline	88.09	70.82	71.51
	VFM	89.86	71.67	72.11
	Full UF	94.50	74.48	75.74

and effectiveness, DINOv2-small proves adequate for most cases with highly stable training. Therefore, it is selected as the default model in this study.

4) *Ablation Study of UF*: We perform an ablation study on the proposed UF to verify its design rationale. Specifically, we decouple the pseudo-label generation process to examine whether first generating anonymous labels with the VFM and then assigning semantic categories using an RS-domain teacher model produces less biased pseudo-labels. We compare three configurations: (1) pseudo-labels generated directly by a conventional RS teacher model (the baseline), (2) pseudo-labels generated directly by the VFM, and (3) pseudo-labels generated using the proposed UF strategy.

The experimental results are presented in Table V. The standalone RS teacher model yields the lowest performance. While pseudo-labels generated solely by the VFM marginally outperform those from the RS teacher model, the improvement is slight; despite its large-scale pre-training, the VFM lacks domain-specific RS knowledge. In contrast, our UF framework first leverages the VFM to generate anonymous pseudo-labels and then utilizes the RS teacher model to assign semantic categories, yielding a substantial performance gain. These results confirm that UF successfully capitalizes on the complementary strengths of VFM prior knowledge and RS domain expertise. Furthermore, pseudo-labels generated via UF exhibit reduced bias toward labeled data, ultimately enhancing performance across both labeled and unlabeled sets.

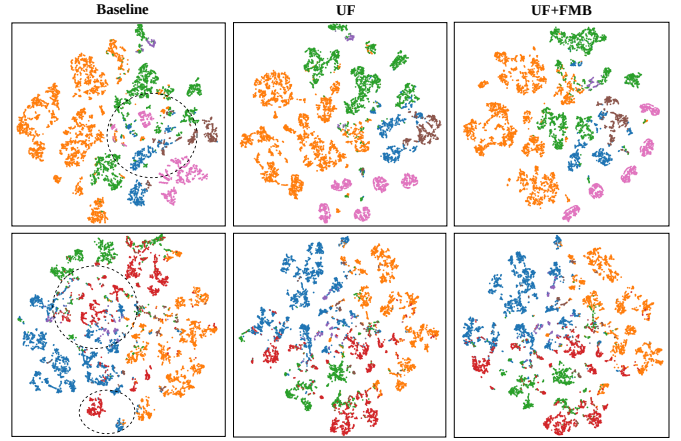


Fig. 8. t-SNE Visualizations of Learned Representations.

5) *t-SNE Visualizations of Learned Representations*: To evaluate the representations learned by different components of UFFM, we generate t-SNE visualizations across three configurations: the Baseline, Baseline with UF, and the complete model combining both UF and FMB.

The experimental results are visualized in Fig. 8, with different colors denoting features from different classes. The first and second rows show the results on the DeepGlobe and Potsdam datasets, respectively. In the baseline model, features from different classes overlap substantially, while features within the same class are loosely clustered, as highlighted by the black dashed ellipse. These patterns indicate the model’s limited ability to discriminate among classes. After incorporating UF and FMB, the inter-class separation becomes more distinct, and the intra-class feature clusters become more compact, demonstrating substantially improved feature discriminability.

V. CONCLUSION

In this work, we present UFFM to address the critical issue where independent training on labeled and unlabeled data causes labeled samples to dominate optimization, thereby severely degrading pseudo-label quality. Built upon a unified teacher–student framework, UFFM introduces two novel core innovations: Unified Flow and Feature Memory Bank. UF establishes an innovative training pipeline that synergizes Vision Foundation Models with domain-specific teacher models, successfully mitigating labeled data dominance and improving pseudo-label fidelity. Concurrently, FMB aligns feature representation discrepancies between labeled and unlabeled distributions by memorizing similar features of similar categories. Extensive experiments demonstrate that UFFM achieves state-of-the-art performance in mIoU across remote sensing benchmarks while offering strong interpretability. Comprehensive ablation studies further confirm that UFFM effectively alleviates data bias, enhancing segmentation accuracy simultaneously on both labeled and unlabeled data.

Nevertheless, current semi-supervised RS techniques still exhibit critical bottlenecks. Despite the promising performance achieved by our method in S^4 domain for RS imagery, its applicability is currently confined to a single downstream task,

i.e., semantic segmentation. More broadly, the RS community would benefit from a unified semi-supervised foundation model that can learn effective and transferable representations from limited labeled data while accommodating diverse downstream vision tasks, such as aerial object detection and land-use change detection. Accordingly, future work will investigate the development of a general-purpose semi-supervised foundation model for RS imagery.

Overall, to the best of our knowledge, UFFM is the first framework to address the optimization and feature representation gap between labeled and unlabeled data in the RS S^4 domain. We anticipate that UFFM will establish a foundational benchmark and inspire future exploration in this promising and emerging direction.

ACKNOWLEDGMENTS

We would like to express our sincere appreciation to the anonymous reviewers.

REFERENCES

- [1] X. X. Zhu, Z. Xiong, Y. Wang, A. J. Stewart, K. Heidler, Y. Wang, Z. Yuan, T. Dujardin, Q. Xu, and Y. Shi, "On the foundations of earth foundation models," *Communications Earth & Environment*, 2026.
- [2] D. Hong, C. Li, X. Li, G. Camps-Valls, and J. Chanussot, "Foundation models in remote sensing: Evolving from unimodality to multimodality," *IEEE Geoscience and Remote Sensing Magazine*, 2026.
- [3] Z. Wang, J. Yi, A. Chen, L. Chen, H. Lin, and K. Xu, "Accurate semantic segmentation of very high-resolution remote sensing images considering feature state sequences: From benchmark datasets to urban applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 824–840, 2025.
- [4] W. Zhou, C. Persello, D. Ming, S. Wang, and A. Stein, "A transformer based multi-task deep learning model for urban livability evaluation by fusing remote sensing and textual geospatial data," *Remote Sensing of Environment*, vol. 334, p. 115232, 2026.
- [5] X. Xue, H. Zhu, X. Li, J. Wang, L. Qu, and B. Hou, "Egpo: Enhanced guidance and pseudo-label optimization for semi-supervised semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.
- [6] W. Zhao, Z. Zhang, F. Zhao, H. Wang, Y. He, and H. Lu, "Remote sensing image generation via object text decoupling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 11 834–11 847, 2025.
- [7] W. Diao, H. Yu, K. Kang, T. Ling, D. Liu, Y. Feng, H. Bi, L. Ren, X. Li, Y. Mao, and X. Sun, "Ringmo-aerial: An aerial remote sensing foundation model with affine transformation contrastive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 10 900–10 913, 2025.
- [8] L. Lv, D. Wang, J. Zhang, and L. Zhang, "S5: Scalable semi-supervised semantic segmentation in remote sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 10, 2026, pp. 7726–7734.
- [9] W. Sun, Y. Lei, D. Hong, Z. Hu, Q. Li, and J. Zhang, "Rsprotosemiseg: Semi-supervised semantic segmentation of high spatial resolution remote sensing images with probabilistic distribution prototypes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 228, pp. 771–784, 2025.
- [10] W. Han, W. Jiang, J. Geng, and W. Miao, "Difference-complementary learning and label reassignment for multimodal semi-supervised semantic segmentation of remote sensing images," *IEEE Transactions on Image Processing*, vol. 34, pp. 566–580, 2025.
- [11] L. Zhang, Z. Tan, W. Zhang, Z. Li, G. Zhang, Z. Chen, and Z. Li, "More unlabeled data does matter: A full-cycle framework for semi-supervised semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [12] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "Decouple and weight semi-supervised semantic segmentation of remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 212, pp. 13–26, 2024.
- [13] S. Wang, X. Sun, D. Hong, and F. Zhou, "Vision-language model purified semi-supervised semantic segmentation for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2026.
- [14] H. Wang, Q. Zhang, Y. Li, and X. Li, "Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 3627–3636.
- [15] H. Chang, H. Bi, C. Xu, and F. Li, "Lsfmamba: Local-enhanced spiral fusion mamba for multi-modal land cover classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 5, pp. 6081–6094, 2026.
- [16] C. Wang, Y. Sun, J. Yang, and L. Luo, "Localized background-aware generative distillation for enhanced remote sensing object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 6, pp. 7663–7675, 2026.
- [17] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024.
- [18] M.-K. Xie, J.-H. Xiao, H.-Z. Liu, G. Niu, M. Sugiyama, and S.-J. Huang, "Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 7, pp. 8444–8455, 2026.
- [19] W. Suo, J. Ma, M. Sun, H. Zhang, P. Wang, Y. Zhang, and Q. Wu, "Semi-supervised vqa multi-modal explanation via self-critical learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 7, pp. 8361–8377, 2026.
- [20] Y. Fu, M. Wang, G. Vivone, Y. Ding, and L. Zhang, "An alternating guidance with cross-view teacher-student framework for remote sensing semi-supervised semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.
- [21] P. Liu and J. Liu, "When confidence fails: revisiting pseudo-label selection in semi-supervised semantic segmentation," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025, pp. 21 874–21 884.
- [22] B. Cheng, J. Lu, Y. Tian, H. Zhao, Y. Chang, and L. Du, "Cgmatch: A different perspective of semi-supervised learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 381–15 391.
- [23] K. Jing, H. Ma, C. Zhang, L. Wen, and Z. Zhang, "Recursive confidence training for pseudo-labeling calibration in semi-supervised few-shot learning," *IEEE Transactions on Image Processing*, 2025.
- [24] Q. Zeng, Z. Lu, Y. Xie, and Y. Xia, "Pick: Predict and mask for semi-supervised medical image segmentation," *International Journal of Computer Vision*, vol. 133, no. 6, pp. 3296–3311, 2025.
- [25] D. Chen, Z. Liu, C. Yang, D. Wang, Y. Yan, Y. Xu, and X. Ji, "Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24 045–24 055.
- [26] J. Yin, X. Jiang, T. Chen, G. Pei, Y. Yao, F. Shen, and H.-T. Shen, "Depmatch: Boosting semi-supervised semantic segmentation by exploring depth difference knowledge," *IEEE Transactions on Image Processing*, vol. 35, pp. 3256–3270, 2026.
- [27] J. Yin, T. Chen, G. Pei, H. Liu, Y. Yao, L. Nie, and X. Hua, "Semi-supervised semantic segmentation with multi-constraint consistency learning," *IEEE Transactions on Multimedia*, 2025.
- [28] L. Yan, F. Liu, L. Jiao, S. Li, J. Wang, L. Li, P. Chen, X. Liu, and X. Gou, "Language-guided modulation-update for semi-supervised semantic segmentation," *Pattern Recognition*, p. 113505, 2026.
- [29] Z. Lin, Y. Zhang, Z. Chen, Z. Fang, C. Wu, X. Chen, Y. Gao, and J. Luo, "Leo-split: A semi-supervised split learning framework over leo satellite networks," *IEEE Transactions on Mobile Computing*, 2025.
- [30] B. Dang, Y. Li, Y. Zhang, and J. Ma, "Progressive learning with cross-window consistency for semi-supervised semantic segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 5219–5231, 2024.
- [31] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.
- [32] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7236–7246.

- [33] L. Yang, Z. Zhao, and H. Zhao, “Unimatch v2: Pushing the limit of semi-supervised semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 3031–3048, 2025.
- [34] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khaidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “DINOv2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, 2024.
- [35] R. Sun, H. Mai, W. Li, Y. Chen, N. Luo, Y. Wang, and T. Zhang, “Beyond confidence: Exploiting homogeneous pattern for semi-supervised semantic segmentation,” in *Forty-second International Conference on Machine Learning*, 2025.
- [36] W. Xuan, H. Qi, and A. Xiao, “Tsg-seg: Temporal-selective guidance for semi-supervised semantic segmentation of 3d lidar point clouds,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 216, pp. 217–228, 2024.
- [37] S. Wang, X. Sun, C. Chen, D. Hong, and J. Han, “Semi-supervised semantic segmentation for remote sensing images via multiscale uncertainty consistency and cross-teacher–student attention,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.
- [38] X. Lu, L. Li, L. Jiao, X. Liu, F. Liu, W. Ma, and S. Yang, “Uncertainty-aware semi-supervised learning segmentation for remote sensing images,” *IEEE Transactions on Multimedia*, vol. 27, pp. 5548–5562, 2025.
- [39] L. Zhou, K. Duan, J. Dai, and Y. Ye, “Advancing perturbation space expansion based on information fusion for semi-supervised remote sensing image semantic segmentation,” *Information Fusion*, vol. 117, p. 102830, 2025.
- [40] X. Gan, W. Li, Y. Zhang, W. Long, Y. Lu, and Z. Chen, “Prior information guided semi-supervised semantic segmentation of remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [41] J. Chen, G. Chen, L. Zhang, M. Huang, J. Luo, M. Ding, and Y. Ge, “Category-sensitive semi-supervised semantic segmentation framework for land-use/land-cover mapping with optical remote sensing images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 134, p. 104160, 2024.
- [42] Y. Li, D. Li, W. Xie, J. Ma, S. He, and L. Fang, “Semi-mamba: Mamba-driven semi-supervised multimodal remote sensing feature classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 9837–9849, 2025.
- [43] J. Chen, B. Sun, L. Wang, B. Fang, Y. Chang, Y. Li, J. Zhang, X. Lyu, and G. Chen, “Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102881, 2022.
- [44] T. Ni, J. Wang, X. Zi, K. Thiyagarajan, S. Kodagoda, and M. Prasad, “Clr-dlr: A semi-supervised framework for high-fidelity remote sensing segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [45] Y. Xin, Z. Fan, X. Qi, Y. Zhang, and X. Li, “Confidence-weighted dual-teacher networks with biased contrastive learning for semi-supervised semantic segmentation in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.
- [46] X. Geng, L. Li, L. Jiao, X. Liu, F. Liu, and S. Yang, “Knowledge-aware geometric contourlet semantic learning for hyperspectral image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 698–712, 2025.
- [47] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, “Qwen3 technical report,” *arXiv preprint arXiv:2505.09388*, 2025.
- [48] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, “Sam 2: Segment anything in images and videos,” in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 28 085–28 128.
- [49] J. Song, K. Li, J. Yao, X. Cao, and D. Meng, “Rs-mtdf: Multiteacher distillation and fusion for remote sensing semi-supervised semantic segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 4411 415–4411 415, 2026.
- [50] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “Deepglobe 2018: A challenge to parse the earth through satellite images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 172–181.
- [51] ISPRS, “ISPRS Potsdam Dataset,” <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>, 2018, accessed: 2026-2-1.
- [52] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.
- [53] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, “Sam 3: Segment anything with concepts,” *arXiv preprint arXiv:2511.16719*, 2025.
- [54] Y. Ouali, C. Hudelot, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency training,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 674–12 684.
- [55] X. Chen, Y. Yuan, G. Zeng, and J. Wang, “Semi-supervised semantic segmentation with cross pseudo supervision,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2613–2622.
- [56] X. Lu, L. Jiao, F. Liu, S. Yang, X. Liu, Z. Feng, L. Li, and P. Chen, “Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.
- [57] L. Hoyer, D. J. Tan, M. F. Naeem, L. Van Gool, and F. Tombari, “Semivl: Semi-supervised semantic segmentation with vision-language guidance,” in *European Conference on Computer Vision*. Springer, 2024, pp. 257–275.