scientific reports

Check for updates

OPEN

A few-shot high-resolution remote sensing image semantic segmentation method

Han-Lin Jiang[1], Ning Wang[1], Bo Geng[5], Zu-Kui Li[6], Rong-Hai Wu[1], Xiao-Wei Li[1], Ben-Hui Chen[4], En-Ming Zhao[3], Guo-Peng Ren[2], Mei Zhang[1]✉ & Deng-Qi Yang[1]✉

Semantic segmentation of high-resolution Unmanned Aerial Vehicle (UAV) remote sensing images plays a crucial role in environmental monitoring, urban planning, agricultural assessment, and disaster management. Semantic segmentation methods that are based on deep learning have demonstrated superior performance; however, they rely on large amounts of annotated data, and thus their performance significantly degrades in small-sample scenarios. To obtain better performance on small-scale remote sensing semantic segmentation datasets, methods combining knowledge distillation and semi-supervised learning are proposed. These methods use models pre-trained on large-scale natural image datasets (such as ImageNet) to guide the training of student models on target datasets directly, achieving significant performance gains. However, the feature distribution of natural image datasets differs significantly from that of remote sensing image datasets. Therefore, student models, directly guided by teacher models pre-trained on natural image datasets, often struggle to obtain the optimal performance, especially when few samples are labeled in the target domain. Whether introducing a medium-scale remote sensing dataset as an intermediate domain between natural image datasets and the target remote sensing dataset can further improve model performance is a question worth exploring. This study proposed a few-shot remote sensing image semantic segmentation method that combined multi-stage knowledge distillation (MKD) and semi-supervised learning (SSL) to progressively bridge domain gaps and leverage unlabeled data. The experimental results on the Erhai UAV dataset (EH) show that the proposed MKD + SSL method achieves a mean IoU of 77.05% with only 880 labeled samples, outperforming the widely used single-stage KD method by + 3.06% mIoU, with per-class IoU gains up to +(2.17% – 5.21%). On the Cityscapes benchmark, our framework further surpasses state-of-the-art methods such as UniMatch, achieving a + 1.5% and + 1.4% improvement in mIoU under 1/16 and 1/8 labeled settings, respectively. These results demonstrate that the proposed method effectively enhances segmentation accuracy in few-shot settings and generalizes well across diverse datasets, with wide practical value.

Keywords Semantic segmentation, Multi-stage knowledge distillation, Semi-supervised learning, UAV remote sensing image, Few-shot learning

Semantic segmentation of high-resolution UAV remote sensing images plays a crucial role in environmental monitoring[1], urban planning[2], and disaster management[3], attracting widespread attention. Traditional remote sensing image semantic segmentation methods typically rely on handcrafted features (e.g., texture, spectral indices, and shape descriptors manually designed by experts) combined with classical machine learning algorithms[4]. These methods generally necessitate the manual design of feature extractors. While such approaches can yield acceptable performance in simple or low-resolution scenarios, they face considerable challenges when applied to complex environments and high-resolution UAV imagery that is rich in detail[2–4]. With the rapid development of deep learning technology[5–8], deep learning-based remote sensing image semantic segmentation methods have demonstrated superior performance due to their powerful feature extraction and modeling capabilities, gradually becoming a research hotspot in this field[8–12].

[1]College of Mathematics and Computer Science, Dali University, Dali 671003, Yunnan, China. [2]College of Agricultural and Biological Sciences, Dali University, Dali 671003, Yunnan, China. [3]College of Engineering, Dali University, Dali 671003, Yunnan, China. [4]Department of Mathematics and Information Technology, Lijiang Teachers College, Lijiang 674100, Yunnan, China. [5]China Tower Co., LTD., Dali Branch, Dali 671003, Yunnan, China. [6]Yunnan Hualiang Data Group Co., LTD, Dali 671003, Yunnan, China. ✉email: meizi108hn@163.com; dqyang@dali.edu.cn

Deep learning-based remote sensing image semantic segmentation models have shown outstanding performance across various tasks[13]. However, their effectiveness heavily depends on large amounts of accurately annotated training data[14]. Although the widespread adoption of UAV technology has made remote sensing image acquisition more accessible, annotating pixel-level training samples remains a time-consuming and labor-intensive task[15], requiring expert annotators. On the other hand, reducing the number of training samples often leads to severe degradation in model performance[3,10–12]. As a result, research teams face significant challenges[16].

To address the issue of insufficient labeled data, researchers have proposed methods that combine semi-supervised learning with transfer learning[17,18]. Semi-supervised learning leverages a small amount of labeled data along with a large volume of unlabeled data, enabling the model to learn underlying patterns from unlabeled samples. This reduces the dependence on labeled samples and provides an effective approach to mitigating the scarcity of annotated data[19,20]. Transfer learning facilitates knowledge transfer from large-scale image datasets to target tasks[21], thereby reducing the demand for extensive training samples in the target domain. Existing transfer learning-based methods for remote sensing image semantic segmentation primarily rely on parameter transfer, in which pre-trained model parameters from large-scale natural image datasets such as ImageNet[22] are directly transferred and fine-tuned on remote sensing datasets[23–25]. Although this approach enhances model performance, the significant differences in data distribution and feature representation between remote sensing images and natural images limit the effectiveness of parameter transfer[25]. To overcome this challenge, researchers have adopted knowledge distillation as an alternative to parameter transfer, aiming to fully utilize the knowledge learned from large-scale natural image datasets[26,27].

Knowledge distillation is an extended approach to transfer learning. First, a teacher model is trained on a large-scale dataset. Then, its outputs are used to guide the training of a student model on a target dataset, which is often much smaller in scale. The student model learns from the teacher model by mimicking its soft labels (i.e., probability distribution vectors), thereby improving its own performance and generalization ability. Previous studies have shown that knowledge distillation often achieves better results than direct parameter transfer[26,27]. Currently, most semantic segmentation methods based on knowledge distillation employ models pre-trained on the ImageNet dataset as teacher models to guide the training of student models on small-scale remote sensing datasets[27–29]. For instance, Wang et al.[27] used knowledge distillation to transfer knowledge from a teacher model pre-trained on ImageNet to student models trained on the Cityscapes[15], CamVid[30], and Pascal VOC[31] semantic segmentation datasets. Their experimental results showed that, compared to parameter transfer, their method improved the mean Intersection over Union (mIoU) by 3.6%, 2.75%, and 2.27% on these three datasets, respectively. However, due to the significant differences between natural image datasets and remote sensing image datasets, when only a limited number of labeled samples are available in the target remote sensing dataset, directly using a teacher model trained on large-scale natural image datasets to guide the student model may not achieve optimal performance. A promising research question is whether introducing an intermediate-scale remote sensing dataset between the large-scale natural image dataset and the small-scale target remote sensing dataset—utilizing a multi-level knowledge distillation approach—could further enhance the performance of the student model.

This study proposed a few-shot remote sensing image semantic segmentation method that combines multi-level knowledge distillation with semi-supervised learning. Our main contributions are as follows:

(1) We have optimized the encoder of the DeepLabV3 + model[32] by replacing the High-Resolution Network (HRNet) as the backbone network and introducing the Polarized Self-Attention (PSA) module, which enhanced the model's ability to restore fine details and improved pixel-level classification accuracy.
(2) We proposed a few-shot semantic segmentation method that combined multi-stage knowledge distillation and semi-supervised learning, achieving high performance with only a few labeled samples.

Related work

To improve the accuracy of semantic segmentation of high-resolution UAV images with small samples, researchers have explored various approaches. Early studies mainly relied on transfer learning (TL), where models pre-trained on large-scale natural image datasets (e.g., ImageNet) were fine-tuned on small-scale remote sensing datasets[33,34]. This strategy has been shown to enhance the performance and convergence speed of the semantic segmentation model, but its effectiveness is limited due to the large domain gap between natural and remote sensing images. To address this limitation, knowledge distillation (KD) has been proposed as an advanced scheme of transfer learning. KD transfers soft labels or intermediate features from a teacher model to a student model, enabling the latter to better adapt to the target task[35–39]. For instance, Tuia et al.[40] combined KD with unsupervised domain adaptation to improve cross-domain adaptability. Li et al.[41] extended this by introducing cross-domain and cross-modal KD to align 2D–3D features, and Zhou et al.[42] designed a graph-attention guided KD to capture land-cover relationships. More recently, Yuan et al.[43] proposed feature augmentation KD, improving student learning, though at a higher computational cost.

Semi-supervised learning (SSL) has emerged as another key approach to alleviate the scarcity of labeled data. By combining a small number of labeled samples with abundant unlabeled data, SSL methods have been shown to significantly improve model performance[44–53]. A variety of techniques have been developed, including self-training[48,52], consistency regularization[45], and generative adversarial networks (GANs)[44,53]. For instance, Zhang et al.[51], proposed a consistency-based network with multi-view augmentation and dynamic pseudo-labels; Yang et al.[47] developed ST + + to mitigate noisy pseudo-labels through strong augmentation and selective retraining; Ke et al.[54] proposed Structured Consistency Loss, enforcing consistency at both pixel and structural levels; and Chen et al.[55] presented Noise-Robust Consistency Regularization, which mitigates pseudo-label noise through feature perturbation and robust loss. These studies collectively demonstrate that SSL can effectively improve accuracy under limited labeled data.

Recent trends have focused on integrating transfer learning, knowledge distillation, and semi-supervised learning into unified frameworks, leveraging their complementary advantages. Transfer learning initializes the model with large-scale datasets, SSL enhances model performance by leveraging a large amount of unlabeled data, and KD enables the model to stably adapt across different domains. For example, in 2023, Yuan et al.[56] proposed a semi-supervised framework known as Mutual Knowledge Distillation. This framework utilizes dual mean teacher models and pseudo-labeling, along with data and feature augmentation, to enhance training diversity and improve segmentation performance in scenarios with limited labeled data. In the same year, Chen et al.[57] proposed a semi-supervised knowledge distillation framework for large-scale urban object mapping in remote sensing, transferring knowledge from pre-trained teacher models to student models with SSL, thereby improving learning under limited labeled data and boosting segmentation accuracy for man-made objects. In 2024, Ma et al.[58] introduced a semi-supervised road detection method based on knowledge distillation. By utilizing knowledge distillation to transfer feature knowledge from a teacher model to a student model and incorporating a semi-supervised learning strategy, this approach further optimizes segmentation performance in road detection tasks. In 2025, Song et al.[59] introduced RS-MTDF, which employs multiple frozen vision foundation models (e.g., DINOv2, CLIP) as teachers, achieving state-of-the-art results on ISPRS Potsdam, LoveDA, and DeepGlobe. These hybrid approaches highlight the potential of combining TL, KD, and SSL for robust remote sensing segmentation.

Nevertheless, most methods still rely on ImageNet-pretrained teachers, whose data distribution differs significantly from remote sensing imagery, limiting their effectiveness in small-sample scenarios. To bridge this gap, we proposed a semantic segmentation method that combines multi-level knowledge distillation with semi-supervised learning, introducing a medium-scale dataset as an intermediate bridge between large-scale natural image datasets and the target domain, thereby improving adaptation and segmentation accuracy.

Dataset and method

Dataset

This study utilized the ImageNet natural image dataset[22], Huawei Ascend Cup remote sensing dataset (HW)[60], and the land cover UAV remote sensing dataset of Erhai (EH) Valley in Yunnan Province. ImageNet, a benchmark dataset for image classification, consists of 14,197,122 images with 21,841 Synset indices, including approximately 12 million training images, 50,000 validation images, and 100,000 test images. It serves as a fundamental dataset for pre-training deep learning models, which are later fine-tuned for various tasks, including remote sensing image segmentation. The HW dataset comprises 100,000 semantic segmentation images of 256×256 pixels, covering 17 land cover categories, primarily used for land-use classification. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio. The EH dataset, captured using UAVs, boasts a spatial resolution of 3 cm and features images with dimensions of 512×512 pixels, offering detailed geographic information about the Erhai watershed (25.67°N, 100.16°E). The dataset comprised six land cover classes: Buildings, Vegetation, Roads, Water Bodies, Farmland, and Others. It included 11,000 labeled images and 168,987 unlabeled images. The labeled images were randomly selected, cropped, and labeled from various large-sized images to ensure the representativeness and diversity of the training set. This study employed a fixed data partitioning strategy[61], dividing the labeled images into a training set, a validation set, and a test set in a ratio of 8:1:1. A separate test set was reserved to ensure the reproducibility of the results and to balance the computational efficiency and reliability. The unlabeled images were utilized in the semi-supervised learning stage. Image examples and annotations for the three datasets are shown in Fig. 1.

High-resolution remote sensing image semantic segmentation method

Considering the significant differences in feature distributions between natural images and semantic segmentation images, we proposed a few-shot high-resolution remote sensing image semantic segmentation framework that integrates multi-stage knowledge distillation and semi-supervised learning (MKD + SSL) (Fig. 2). This framework introduced a moderate-scale remote sensing dataset as an intermediate domain between the ImageNet source domain and the target UAV remote sensing dataset, leveraging multi-stage knowledge distillation to model with limited target domain labeled samples. Furthermore, semi-supervised learning was employed to fully utilize the vast number of unlabeled samples in the target domain, thereby enhancing the model's overall performance.

Backbone model selection and optimization

Currently, DeepLabV3 + is one of the mainstream models widely recognized for its superior performance in the field of semantic segmentation[32]. It typically employs ResNet as its backbone network. However, ResNet struggles with effectively preserving details and accurately defining boundaries, particularly when processing high-resolution remote sensing images that have complex boundaries and rich details. To overcome these shortcomings, this study replaced the DeepLabV3 + backbone with the High-Resolution Network (HRNet) and further introduced the Polarized Self-Attention (PSA) module after the backbone's feature maps to enhance semantic segmentation performance. The applications of the HRNet enabled the DeepLabV3 + to maintain high-resolution features throughout the network, significantly improving multi-scale feature fusion capability, and making it especially suitable for high-resolution remote sensing image segmentation tasks with complex boundaries and fine details. The PSA module employed a polarized filtering strategy to maintain high internal resolution and utilizes SoftMax and Sigmoid functions to enhance the high dynamic range, refining and strengthening high-level semantic features.

Multi-stage knowledge distillation strategy

This study proposed a multi-stage knowledge distillation strategy to achieve efficient model adaptation from the source domain to the target domain through progressive transfer and knowledge transmission (Fig. 2). In

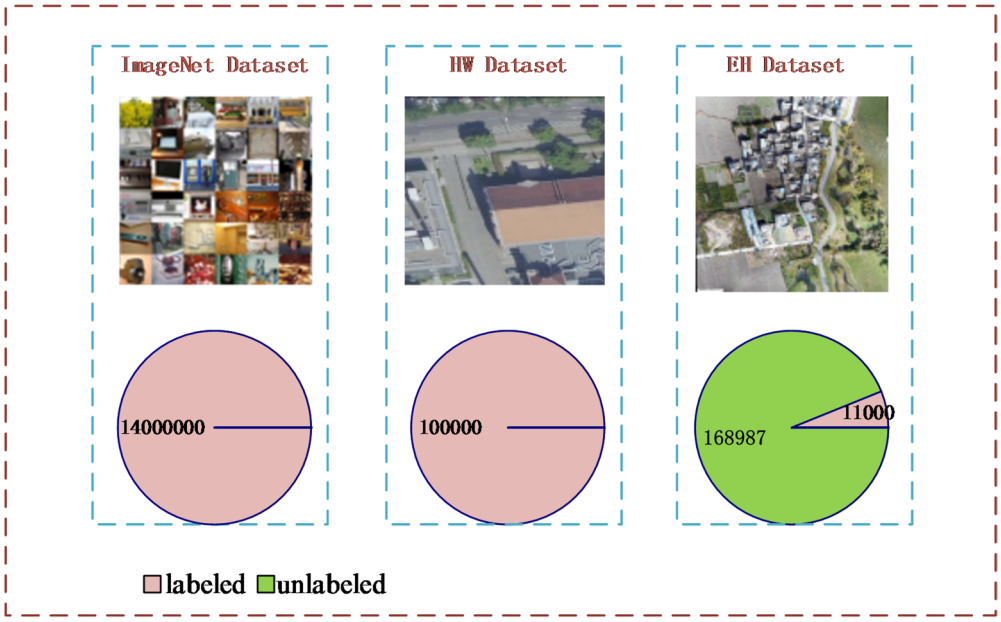

Fig. 1. Sample images and annotations of the three datasets.

the first stage of knowledge distillation, the pre-trained model (M) on ImageNet was fine-tuned using the HW data set by updating only the parameters of the fully connected layer. The fine-tuned model served as the initial teacher model (M_{T1}). We then randomly initialized the parameters of the student model (M_{S_init1}) and employed the initial teacher model M_{T1} to guide the student model to train on the HW dataset. Throughout the entire training process of the student model (M_{S_init1}), the parameters of the teacher model remained fixed to fully preserve the knowledge acquired from the source domain. The student model (M_{S_init1}) started with randomly initialized parameters and was progressively optimized during training. The optimization was guided by three types of loss functions. First, the supervised loss, based on cross-entropy, ensured that the student model learned accurate semantic segmentation from the labeled samples. Second, the knowledge distillation loss transferred knowledge from the teacher to the student by allowing the student to mimic the softened probability outputs of the teacher, which, compared to one-hot labels, revealed inter-class similarities and provided richer guidance. In this study, a temperature of three was used to generate the soft outputs, encouraging the student to inherit semantic knowledge from the teacher and enhance its generalization under limited labeled data. Third, the Elastic Weight Consolidation (EWC) loss imposed constraints on important parameters of the student model, thereby reducing the risk of forgetting critical source domain knowledge while adapting to new tasks. The final training objective was achieved by combining these three loss functions with corresponding weights. By incorporating EWC loss, the student model was optimized through the joint minimization of the supervised loss and knowledge distillation loss, yielding the well-trained student model ($M_{S\_trained1}$).

In the second stage, $M_{S\_trained1}$ was further fine-tuned on the EH dataset (updating only the full connection layer) to generate a new teacher model (M_{T2}), and a student model (M_{S_init2}) was randomly initialized. Then the new teacher model (M_{T2}) was used to guide the training of the student model (M_{S_init2}) on the EH dataset. This process also employed the supervised loss, knowledge distillation loss, and EWC loss described above, allowing $M_{S\_trained2}$ to adapt to the EH dataset while inheriting semantic knowledge from M_{T2} and maintaining stability of key parameters. Through this multi-stage distillation process, the model achieved efficient cross-domain adaptation from the ImageNet dataset to the HW dataset and then to the EH dataset, ultimately providing a high-quality initial model for subsequent semi-supervised learning.

Semi-supervised learning strategy

Annotating high-resolution remote sensing images is often costly and time-consuming, whereas unlabeled data is more abundant and easier to obtain in practical applications. To fully utilize these unlabeled samples and further enhance the accuracy of the segmentation model, this study combined semi-supervised learning (SSL) techniques with the training process (Fig. 2). At the beginning of training, the multi-level knowledge distillation (MKD) model served as the initial teacher model (M_{T3}), and a new student model (M_{S_init3}) was initialized with the parameters of the M_{T3} model. Given an unlabeled sample x_u, the teacher model M_{T3} generated a pseudo-label $\widehat{\boldsymbol{y}}_{\boldsymbol{u}}$. To ensure pseudo-label reliability, an entropy-based filtering mechanism was applied, retaining only low-entropy samples[62](x_u^{conf}, $\widehat{\boldsymbol{y}}_{\boldsymbol{u}}^{conf}$), which were then treated as pseudo-labeled data and combined with labeled data (x_l,y_l) for training the student model. During training, the total loss function consisted of two key components: (1) Supervised loss L_{sup}, which applied to labeled data (x_l, y_l) and measured the difference between the student model's predictions and the ground-truth labels; and (2) Consistency loss L_{cons}, which applied to pseudo-labeled data (x_u^{conf}, $\widehat{\boldsymbol{y}}_{\boldsymbol{u}}^{conf}$), encouraging the student model to produce predictions that were consistent with the high-confidence outputs of the teacher model. By jointly optimizing these losses,

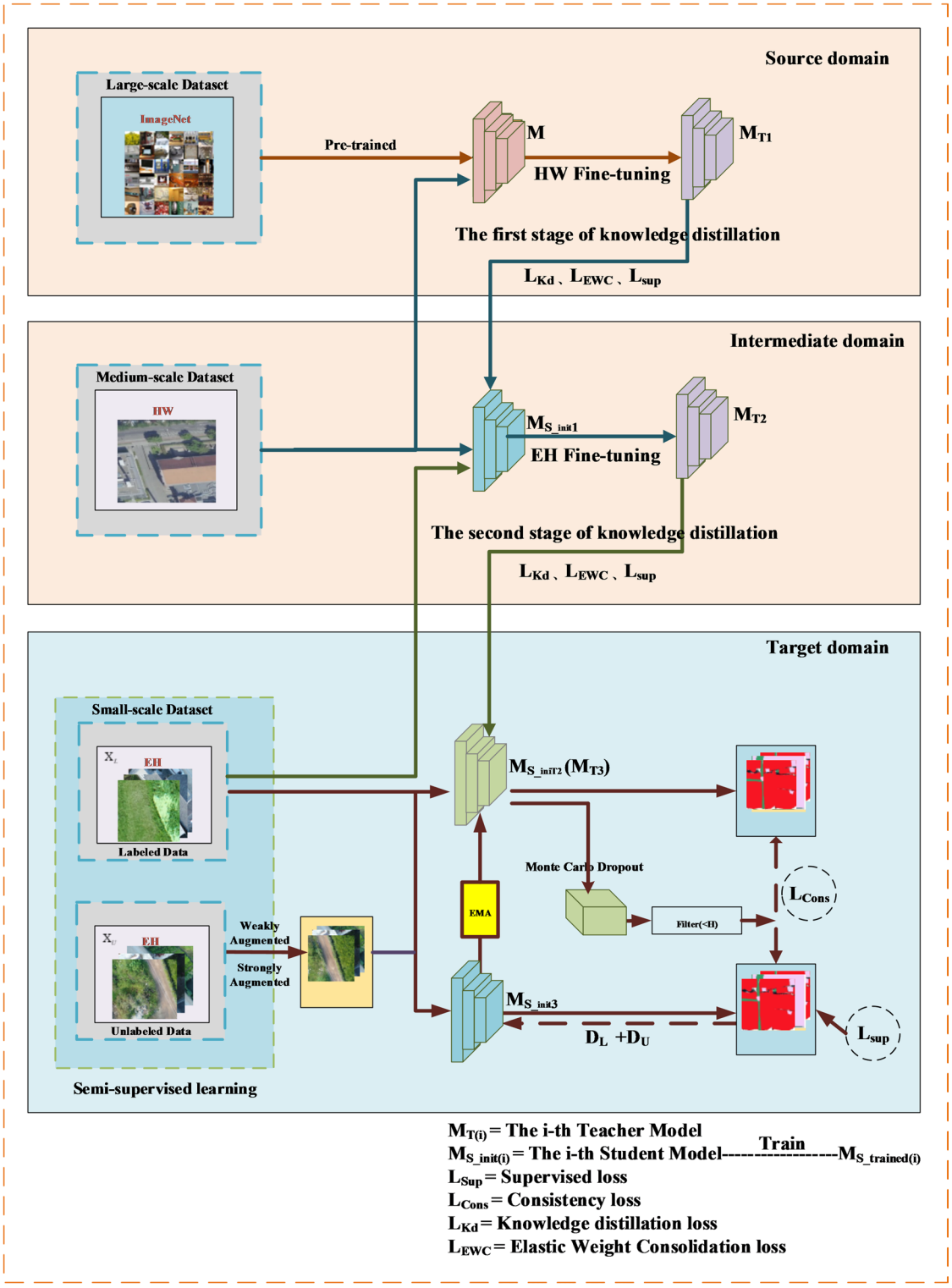

Fig. 2. A few-shot remote sensing semantic segmentation framework combining multi-stage knowledge distillation and semi-supervised learning.

the student model received strong supervision from labeled data while learning features from unlabeled data, enhancing its adaptation to the target domain. Additionally, the teacher model's parameters were dynamically updated through an Exponential Moving Average (EMA) mechanism[63] rather than being fixed. If the student model's and teacher model's parameters at step t^{th} were respectively denoted as $W_t^{(S)}$ and $W_t^{(T)}$, the EMA update formula was: $\boldsymbol{W}_t^{(T)} = \boldsymbol{\alpha} \cdot \boldsymbol{W}_{t-1}^{(T)} + (1 - \boldsymbol{\alpha}) \boldsymbol{W}_t^{(S)}$ where α was typically set close to 1 (e.g., 0.999) to ensure

smooth updates, allowing the teacher model to gradually absorb new information from the student model. This dynamic update mechanism enabled the teacher model to continuously generate higher-quality pseudo-labels, progressively improving training effectiveness. Through an iterative cycle of pseudo-label generation, supervised learning, and consistency learning, the student model ($M_{S\_trained3}$) drove improvements in the teacher model, which in turn provided increasingly refined pseudo-labels. This mutual enhancement process ensured that the teacher and student models collaborated effectively.

Model evaluation and parameter settings

Model evaluation metrics

This study primarily used GFLOPs (Giga Floating Point Operations) and IoU (Intersection over Union) as evaluation metrics to comprehensively assess the model's performance and efficiency in remote sensing image semantic segmentation tasks. GFLOPs measured the computational complexity and resource consumption of the model, where a lower value indicated reduced computational requirements during inference. IoU was one of the most commonly used evaluation metrics in semantic segmentation, assessing the segmentation accuracy for each category by calculating the overlap between predicted and ground truth regions. The mean IoU (mIoU) averaged the IoU values across all categories, providing a more holistic measure of the model's performance in segmentation tasks. Additionally, FPS (Frames Per Second) was used to quantify the model's inference speed in practical applications, indicating the number of images the model can process per second under a given hardware setup.

Model training and hyperparameters

In this study, deep learning experiments were conducted using an NVIDIA GeForce RTX 4090 GPU. The software environment was configured with PyTorch 1.7.0 as the deep learning framework, accelerated by CUDA 12.0 for efficient computation. The hyperparameter settings for the model were as follows: the input image size was set to 512 × 512 pixels with 3 channels, the batch size was 8, and the total number of training epochs was 100. The choice of 100 training epochs was intended to ensure sufficient model convergence while avoiding overfitting, especially when handling large datasets and complex models, which aligns with common practices in relevant studies[33,43,48]. During training, the initial learning rate was set to 0.001, and Stochastic Gradient Descent (SGD) was used as the optimizer. The loss function employed was Online Hard Example Mining (OHEM)[64], which improves segmentation performance by prioritizing difficult samples. Additionally, the momentum value was set to 0.9, the confidence threshold was 0.95, and weight decay was 0.001 to enhance model stability and prevent overfitting. In the semi-supervised learning phase, data augmentation techniques such as random flipping and brightness adjustment were applied during pseudo-label generation by the teacher model to improve model robustness. These configurations and parameter settings provided the necessary technical support and optimization conditions for achieving high-accuracy semantic segmentation in high-resolution remote sensing images.

Results

Experimental results of model improvement

To evaluate the effectiveness of the proposed model improvements, we conducted a comparative experiment on the EH remote sensing image dataset (Table 1). In these experiments, the DeepLabV3 + model utilized various networks as the backbone, and we employed traditional knowledge distillation (single-stage) instead of multi-stage knowledge distillation. All experiments utilized a pre-trained model on ImageNet to guide the training of the student model on the EH dataset.

To ensure consistency and reliability of the method, all models were trained using the entire labeled training set of the EH dataset, which consists of 8,800 images. For a fair comparison, all models used identical parameter configurations, data augmentation strategies, and evaluation datasets.

The experiment results indicated that both HRNetV2-W48 and its improvement HRNetV2-W48_PSA achieved superior segmentation performance while maintaining relatively low computational complexity (1,160.1 GFLOPs and 1,262.1 GFLOPs, respectively). Their mIoU scores reached 81.31% and 81.93%, respectively. Compared to the traditional ResNet-101 backbone, HRNetV2-W48_PSA not only exhibited lower computational complexity (1,262.1 GFLOPs) but also significantly improved segmentation performance, with mIoU increasing from 79.04% to 81.93%.

The experimental results indicated that the HRNetV2-W48_PSA model, which integrated the PSA module, achieved a 0.6% increase in mIoU compared to the HRNetV2-W48 model. Additionally, the computational

	Weights	Backbone	GFLOPs	mIoU
DeepLabv3+	ImageNet	ResNet-101	1,661.6	79.04
DeepLabv3+	ImageNet	Resnet101_ibn_a	1,778.7	79.83
DeepLabv3+	ImageNet	ResNext-101	1,788.2	80.01
DeepLabv3+	ImageNet	Xception71	1,344.6	80.22
DeepLabv3+	ImageNet	HRNetV2-W48	1,160.1	81.31
DeepLabv3+	ImageNet	HRNetV2-W48_PSA	1,262.1	**81.93**

Table 1. Performance of DeepLabV3 + with different backbone networks on the EH dataset. Significant values are in bold.

cost rose by over 100 GFLOPs. However, the tests indicated that the actual inference speed of these two models was nearly equivalent (HRNetV2-W48_PSA achieves 42 FPS, while HRNetV2-W48 reaches 44 FPS). Therefore, we recommend choosing HRNetV2-W48_PSA, as it maintains efficient inference while enhancing feature representation and segmentation accuracy, thus achieving the best balance between performance and computational cost.

Experimental results of semi-supervised learning and multi-stage knowledge distillation

To validate the effectiveness of the framework that combines multi-stage knowledge distillation and semi-supervised learning in small-sample target tasks, this study conducted experiments using 10% of the labeled samples (880 images) from the EH labeled training set and all unlabeled images (168,987 images). To ensure the consistency and reliability of the method, the experiments maintained the same parameter configurations, data augmentation strategies, and evaluation dataset. We compared the performance differences with and without semi-supervised learning under different knowledge distillation strategies (Table 2). The experimental results indicated that the introduction of an intermediate domain consistently enhanced performance across all categories.

When semi-supervised learning was not employed, the multi-stage knowledge distillation (MKD) model (ImageNet_HW/No) improved the mIoU by 5.58% over the single-stage knowledge distillation (SKD) model from ImageNet (ImageNet/No). This enhancement led to IoU increases of 8.4% for Road, 9.98% for Farmland, 5.48% for Building, and 5.04% for Grassland. Similarly, in comparison with the SKD method used in HW (HW/No), the mIoU of the MKD method rose by 3.58%, and the IoU for each category saw improvements ranging from 1.01% to 5.05%, demonstrating the efficacy of intermediate region distillation. When combined with semi-supervised learning (SSL), the mIoU of the MKD + SSL model (ImageNet_HW/Yes) was still 3.06% higher than that of the SKD model of ImageNet (ImageNet/Yes). This improvement resulted in IoU gains of 5.21% for Road, 3.74% for Farmland, 2.17% for Building, and 3.19% for Grassland. Compared with the SKD model based on the HW dataset (HW/Yes), the MKD + SSL model (ImageNet_HW/Yes) also showed similar results. The mIoU of the MKD + SSL method has increased by 2.42%, while the IoU of each category has increased by 0.16% to 4.15%.

The experimental results also indicated that semi-supervised learning significantly enhanced the performance of both single-stage and multi-stage KD models. Specifically, the mIoU of the MKD + SSL model has increased by 3.01% compared to that of the MKD model. The IoU values for all categories have also improved, with the improvement range ranging from 1.10% to 4.96%.

To provide an intuitive visualization of the segmentation performance, we randomly selected several unseen samples from the EH test set for qualitative analysis. The visualization results in the semi-supervised learning setting demonstrated that the ImageNet_HW17 strategy achieved the best performance (Fig. 3).

Effect of unlabeled data quantity

Upon utilizing various proportions of unlabeled data from the EH dataset, we examined the effect of different quantities of unlabeled samples on the performance of the MKD + SSL framework. This dataset contained a total of 168,987 unlabeled images. We conducted semi-supervised learning using 10% and 50% of the unlabeled images randomly selected, as well as all the unlabeled images (Table 3). The experimental results indicated that, in comparison to the baseline without SSL (74.04% mIoU), including only 10% of the unlabeled data can increase the model's mIoU by 1.16%. When the amount of unlabeled data was increased to 50%, the mIoU further improved to 76.40% (+2.36%), and utilizing the complete set achieved 77.05% (+3.01%). The results indicated that the mIoU of the proposed MKD + SSL framework was positively correlated with the total amount of unlabeled data used in semi-supervised learning. However, as the number of unlabeled samples participating in the semi-supervised training increases, the improvement in the model's mIoU tends to plateau after reaching a certain threshold. Therefore, when computing resources are abundant, it is recommended to utilize as much unlabeled data as possible to enhance the performance of the model. Conversely, when computing resources are limited, it is advised to use approximately half of the available unlabeled data for semi-supervised learning.

Knowledge Distillation Strategy/Semi-Supervised Learning	IOU per category (%)						mIoU
	Background	Road	Farmland	Building	Water	Grassland	
No Distillation/No	71.20	39.17	60.69	70.40	70.81	54.10	61.07
No Distillation/Yes	74.11	56.37	78.90	79.25	77.53	65.88	72.03
ImageNet/No	75.27	51.94	70.63	76.09	73.91	62.89	68.46
ImageNet/Yes	74.82	60.09	81.08	81.34	78.96	67.68	73.99
HW/No	74.29	55.29	74.75	78.26	75.15	65.01	70.46
HW/Yes	76.24	62.17	80.67	83.36	78.04	67.29	74.63
ImageNet_HW/No	**75.30**	**60.34**	**80.61**	**81.57**	**78.51**	**67.93**	**74.04**
ImageNet_HW/Yes	**76.40**	**65.30**	**84.82**	**83.51**	**81.41**	**70.87**	**77.05**

Table 2. Comparison of multi-stage Knowledge Distillation Strategies with and Without Semi-Supervised Learning. Significant values are in bold.

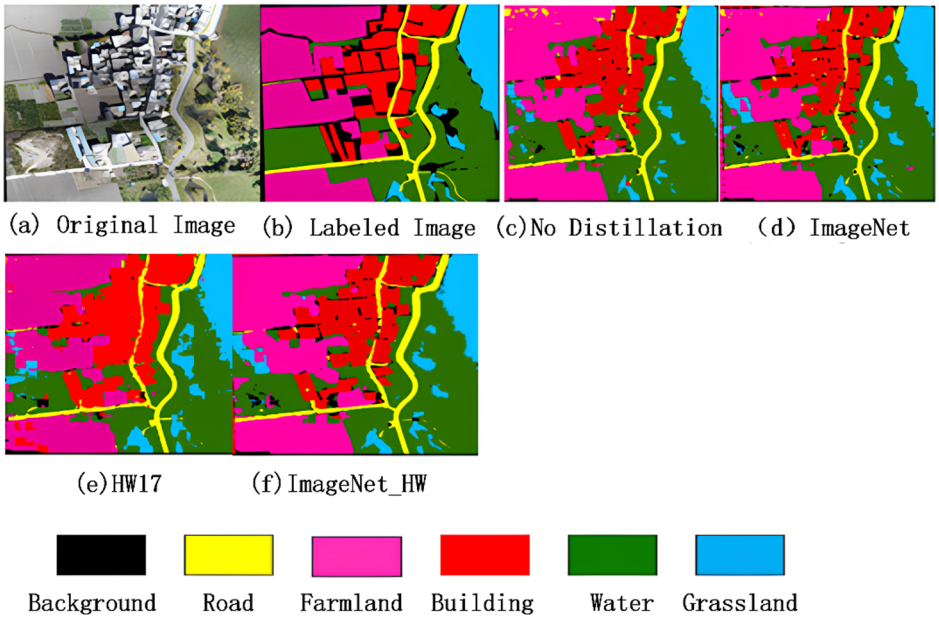

Fig. 3. Visualization of segmentation results on EH dataset samples.

Unlabeled Ratio	mIoU (%)
Baseline (No SSL)	74.04
10%	75.20
50%	76.40
100% (Full SSL)	77.05

Table 3. Impact of unlabeled data proportion on segmentation accuracy (mIoU).

Method	Pre-trained weights	Backbone	1/16	1/8	1/4	1/2
SupBaseline	ImageNet	ResNet-101	66.3	72.8	75.0	78.0
MT[46]	ImageNet	ResNet-101	68.08	73.71	76.53	78.59
CCT[66]	ImageNet	ResNet-101	69.64	74.48	76.35	78.29
GCT[67]	ImageNet	ResNet-101	66.90	72.96	76.45	78.58
U2PL[39]	ImageNet	ResNet-101	74.9	76.5	78.5	79.1
UniMatch[65]	ImageNet	ResNet-101	76.6	77.9	79.2	79.5
MKD + SSL(Ours)	Imagenet_GTAV	HRNetV2-W48_PSA	**78.1**	**79.3**	**79.4**	**80.3**

Table 4. Comparison with the latest SOTA methods on the Cityscapes Dataset.

Discussion

Comparative analysis of experimental results

To further validate the effectiveness of the proposed semantic segmentation method, we conducted experiments on the Cityscapes dataset, which contains 2,975 training, 500 validation, and 1,525 test images[65]. The training set was used to construct labeled and unlabeled subsets: specifically, 1/16, 1/8, 1/4, and 1/2 of the training images were randomly selected as labeled data, while the remaining images were treated as unlabeled data. The resulting mIoU was then compared with that of existing approaches (Table 4).

In the comparative experiments, we employed our proposed MKD + SSL method with HRNetV2-W48_PSA as the backbone network, a pre-trained model on ImageNet as the teacher model, GTAV as the intermediate domain, and Cityscapes as the target domain. All methods evaluated in Table 4 were semantic segmentation approaches, ensuring consistency with our task. The Cityscapes dataset is a widely used benchmark in this field, and conducting experiments on this dataset provides a unified and fair evaluation for comparison. The results demonstrated that the proposed MKD + SSL framework consistently outperforms existing methods when using the same number of labeled samples. Moreover, as the size of the labeled dataset in the target domain decreases, the performance improvement of the proposed method becomes more significant. Compared to existing

methods, our model achieved mIoU improvements of 1.5% and 1.4% when using 1/16 and 1/8 labeled samples, respectively. When the labeled sample size increased to 1/4 and 1/2, the mIoU gain decreased, but was still superior to existing methods. These results indicated that our MKD + SSL framework offers notable advantages in small-sample scenarios, making it particularly well-suited for few-shot semantic segmentation tasks.

Impact of intermediate domain selection

We observed that the selection of the intermediate domain in multi-stage knowledge distillation has a significant impact on model performance. A well-chosen intermediate domain effectively bridges the gap between the source and target domains, facilitating more efficient knowledge transfer. However, if the intermediate domain is inappropriately selected, it may lead to negative transfer, causing a decline in model performance on the target domain. In the aforementioned comparative experiments, we attempted to use the HW dataset as the intermediate domain but encountered negative transfer effects. We hypothesized that this occurred due to significant differences in scene distribution between the intermediate domain (HW) and the target domain (Cityscapes).

To evaluate the impact of intermediate domain selection on the model performance, we conducted a t-SNE analysis[68] to visualize the feature distribution across different datasets (Fig. 4). The results showed that the HW dataset lay between ImageNet and EH in the feature representation space (Fig. 4a), with partial overlap with both domains. Specifically, HW included categories such as water, transportation, buildings, farmland, grassland, forest, and bare soil, many of which (e.g., water, buildings, farmland, grassland) overlap with the categories in the EH dataset (Background, Road, Farmland, Building, Water, Grassland). This semantic consistency explained why HW serves as an effective intermediate domain to bridge ImageNet and EH, thereby improving transfer performance. In contrast, when we used GATV as the intermediate domain, the results were even worse than the single-stage KD method. We attributed this to the fact that GATV mainly contained urban street scenes, which differed substantially from the EH dataset, a finding that was consistent with our t-SNE visualization. For the transfer from ImageNet to Cityscapes, however, GATV proved to be a more suitable intermediate domain. GATV features were closer to Cityscapes both visually and quantitatively, which facilitates smoother domain adaptation (Fig. 4b).

We further conducted experiments comparing GATV and HW as intermediate domains for Cityscapes, and the results showed that GATV yields better performance. This can be attributed to the fact that both GATV and Cityscapes were urban scene datasets with many overlapping categories, making their feature distributions more consistent. These observations were also supported by our t-SNE visualization. Therefore, when selecting an intermediate domain dataset, we recommend choosing one with a similar feature distribution or overlapping categories with the target domain dataset to ensure a smooth domain transformation.

Ablation experiments

To validate the contributions of each component, we conducted a comprehensive ablation analysis evaluating (i) multi-stage KD without SSL, (ii) SSL without multi-stage KD, and (iii) the impact of removing EWC and PSA. The results, summarized in Table 5, showed that removing SSL decreases mIoU by 3.01% (77.05% → 74.04%), removing multi-stage KD while keeping SSL yields 73.99% (vs. 77.05% for the full model), and removing PSA slightly reduces performance (74.04%→73.66%). These observations confirm that the components of the MKD + SSL framework are complementary, and EWC and PSA provide additional robustness to the model.

Model generalization analysis

To further assess the generalization capability of the proposed method, additional experiments were conducted on the Cityscapes dataset. This dataset differs significantly from the EH UAV remote sensing imagery in terms of resolution, imaging conditions, and target category distribution. Nevertheless, the results demonstrated that the proposed method still outperforms several representative semi-supervised approaches in urban scene segmentation (see Table 4). This indicates that the proposed MKD + SSL framework can adapt to semantic

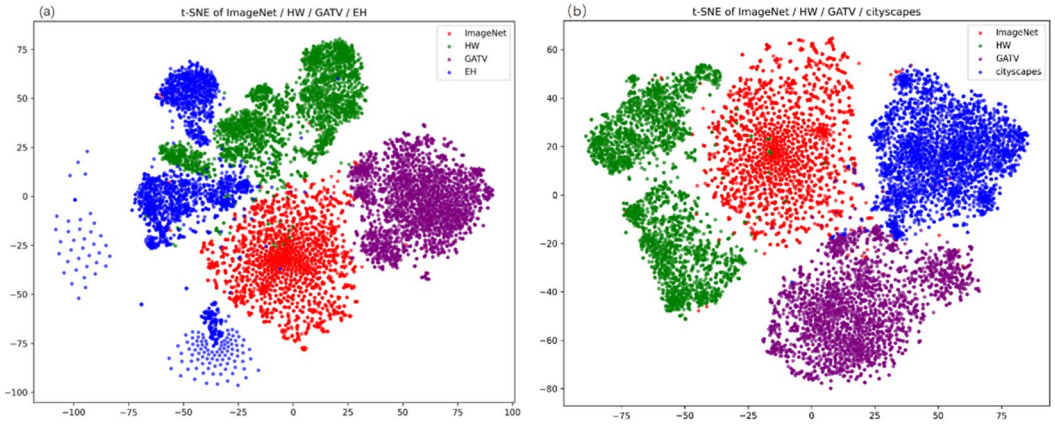

Fig. 4. t-SNE visualization of intermediate domain selection.

Group	Multi-stage KD	SSL	EWC	PSA	Test(mIoU)
1	√	√	√	√	77.05
2	√		√	√	74.04
3	√		√		73.66
4		√	√	√	73.99
5			√	√	68.46

Table 5. The effect of multi-stage KD, SSL, EWC, and PSA on the model's mIoU.

segmentation tasks in diverse scenarios, exhibiting strong cross-dataset generalization potential. In future work, we will conduct further experiments on other publicly available remote sensing datasets, such as UAVid and LoveDA, to further validate the robustness and practical applicability of the method in broader application contexts.

Broader impacts and ethical considerations

The proposed MKD + SSL framework has the potential to enhance the efficiency and accuracy of high-resolution remote sensing image semantic segmentation in various domains, such as urban planning, agricultural monitoring, ecological conservation, and disaster management. By reducing reliance on large-scale labeled datasets, this approach may facilitate the broader adoption of remote sensing applications. However, it is important to note that the method may also pose potential risks in practical use. For example, in urban surveillance or land-use regulation scenarios, remote sensing technology may involve issues of privacy and data security; in military or sensitive areas, its application without appropriate oversight could raise ethical concerns. Therefore, in future work, we will not only continue to improve the robustness and generalizability of the model but also place greater emphasis on addressing its ethical boundaries in real-world applications.

Limitations and future work

Although the proposed MKD + SSL framework demonstrates strong performance in few-shot remote sensing semantic segmentation, several limitations should be acknowledged. First, the effectiveness of multi-stage knowledge distillation largely depends on the selection of an appropriate intermediate domain. Although we have mentioned above that the selection of the intermediate domain should take into account the similarity with the feature distribution of the target domain dataset, further exploration is needed in future research on how to quantify this similarity to guide the selection of the intermediate domain. Second, the proposed framework introduces additional computational overhead due to multi-stage training and semi-supervised learning, which may limit its applicability in resource-constrained environments. Although the inference efficiency remains acceptable, the overall training cost is higher than that of single-stage approaches. Finally, the current method has been validated on UAV and urban scene datasets. Its performance on other remote sensing scenarios, such as multispectral or hyperspectral imagery, has not yet been explored. Future work will focus on reducing training complexity, improving intermediate domain selection strategies, and extending the framework to broader remote sensing applications.

Conclusion

This study addressed the high annotation cost associated with training semantic segmentation models for high-resolution remote sensing images by proposing a few-shot semantic segmentation method that combined multi-stage knowledge distillation and semi-supervised learning. By introducing an intermediate-domain dataset between the large-scale natural image dataset (ImageNet) and the small-scale UAV remote sensing image dataset, the proposed method gradually reduces the distribution gap between the source and target domains, effectively enhancing the model's transferability and adaptation in high-resolution remote sensing imagery. Meanwhile, the use of pseudo-labeling and consistency learning effectively leveraged large amounts of unlabeled data, further improving the segmentation accuracy of the model. The results validated the effectiveness of introducing an intermediate domain for multi-stage distillation, while also demonstrating the potential of pseudo-labeling and consistency learning in enhancing model performance. The proposed method was particularly valuable in real-world remote sensing scenarios where annotated data was extremely scarce and can be widely applied in fields such as environmental monitoring, urban planning, agricultural assessment, and disaster management. In summary, this study expands the scope of deep learning research in few-shot remote sensing semantic segmentation and provides a feasible technical framework for future large-scale, high-resolution remote sensing analysis and applications.

Data availability

The model codes and additional meta-data can be accessed on github (https://github.com/Mrjianghanlin/A-Few-Shot-High-Resolution-Remote-Sensing-Image-Semantic-Segmentation-Method3). The research utilizes three key datasets: (A) The UAV remote sensing data of the EH dataset can be accessed on Google Drive (https://drive.google.com/drive/folders/14YghGdWH-DzJy3sfTEQy0Tj3zhJtwd7Y?usp=sharing). (B) The high-resolution remote sensing images of the HW dataset are available on AI Studio (https://aistudio.baidu.com/datasetdetail/54302/0).C) Urban scene images of the Cityscapes dataset are freely accessible on the official website (https://

www.cityscapes-dataset.com/). If you are unable to access any of these datasets, please contact the Han-Lin Jiang author (email: 15691552855@163.com) for assistance.

Received: 3 June 2025; Accepted: 27 March 2026
Published online: 01 April 2026

References

1. Si, B. et al. ABNet: An aggregated backbone network architecture for fine landcover classification. *Remote Sens.* **16**(10), 1725 (2024).
2. Khan, B. A. & Jung, J.-W. Semantic segmentation of aerial imagery using U-Net with self-attention and separable convolutions. *Appl. Sci.* **14**(9), 3712 (2024).
3. Lu, Y. et al. Multi-dimensional manifolds consistency regularization for semi-supervised remote sensing semantic segmentation. *Knowl. Based Syst.* **299**, 112032 (2024).
4. Shen, X. et al. Multi-scale feature aggregation network for semantic segmentation of land cover. *Remote Sens.* **14**(23), 6156 (2022).
5. He, K. et al. Mask r-cnn. In *Proceedings of the Proceedings of the IEEE international conference on computer vision* (2017).
6. He, K. et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
7. Huang, G. et al. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).
8. Lin, T-Y. et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).
9. Zheng, G. et al. Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network. *Front. Earth Sci.* **11**, 1115805 (2023).
10. Li, G. et al. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021).
11. Xie, G-S. et al. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021).
12. Yang, B. et al. Prototype mixture models for few-shot semantic segmentation. In *Proceedings of the Computer Vision–ECCV. : 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, F, 2020 (Springer, 2020).
13. Yuan, X., Shi, J. & Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **169**, 114417 (2021).
14. Li, Y. et al. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **175**, 20–33 (2021).
15. Cordts, M. et al. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
16. Lyu, Y. et al. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **165**, 108–119 (2020).
17. Chen, Z. et al. Road extraction in remote sensing data: A survey. *Int. J. Appl. Earth Obs. Geoinf.* **112**, 102833 (2022).
18. Liu, P. et al. Survey of road extraction methods in remote sensing images based on deep learning. *PFG–J. Photogramm., Remote Sens. Geoinf. Sci.* **90**(2), 135–59 (2022).
19. Hung, W-C. et al. Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:180207934 (2018).
20. Chen, H. et al. SemiRoadExNet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning. *ISPRS J. Photogramm. Remote Sens.* **198**, 169–183 (2023).
21. Tan, B. et al. Transitive transfer learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015).
22. Recht, B. et al. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International conference on machine learning* (PMLR, 2019).
23. Marmanis, D. et al. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geosci. Remote Sens. Lett.* **13**(1), 105–109 (2015).
24. Hoyer, L. et al. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *Int. J. Comput. Vis.* **131**(8), 2070–96 (2023).
25. Gadiraju, K. K. & Vatsavai, R. R. Comparative analysis of deep transfer learning performance on crop classification. In *Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (2020).
26. Hinton, G. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:150302531 (2015).
27. Wang, Y. et al. Intra-class feature variation distillation for semantic segmentation. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, F, 2020. (Springer, 2020).
28. Park, S. & Heo, Y. S. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors* **20**(16), 4616 (2020).
29. Liu, Y. et al. Structured knowledge distillation for semantic segmentation. In *Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019).
30. Brostow, G. J., Fauqueur, J. & Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **30**(2), 88–97 (2009).
31. Everingham, M. et al. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).
32. Chen, L-C. et al. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:170605587 (2017).
33. Cui, B., Chen, X. & Lu, Y. Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection. *IEEE Access* **8**, 116744–55 (2020).
34. Vu, T-H. et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019).
35. Laine, S. & Aila, T. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:161002242 (2016).
36. Desai, S. & Ghose, D. Active learning for improved semi-supervised semantic segmentation in satellite images. Proceedings of the IEEE/CVF winter conference on applications of computer vision (2022).
37. Mittal, S., Tatarchenko, M. & Brox, T. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(4), 1369–79 (2019).
38. Zheng, Y. et al. Semi-supervised adversarial semantic segmentation network using transformer and multiscale convolution for high-resolution remote sensing imagery. *Remote Sens.* **14**(8), 1786 (2022).
39. Wang, Y. et al. Learning pseudo labels for semi-and-weakly supervised semantic segmentation. *Pattern Recogn.* **132**, 108925 (2022).
40. Tuia, D., Persello, C. & Bruzzone, L. Recent advances in domain adaptation for the classification of remote sensing data. arXiv preprint arXiv:210407778 (2021).
41. Li, M. et al. Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022).

42. Zhou, W. et al. Graph attention guidance network with knowledge distillation for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* https://doi.org/10.1109/TGRS.2023.3311480 (2023).
43. Yuan, J. et al. FAKD: Feature Augmented Knowledge Distillation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024).
44. Zhang, C. et al. Multitask GANs for semantic segmentation and depth completion with cycle consistency. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(12), 5404–15 (2021).
45. Sajjadi, M., Javanmardi, M. & Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Adv. Neural Inf. Process. Syst.* 29 (2016).
46. Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* 30 (2017).
47. Yang, L. et al. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022).
48. Zhai, X. et al. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019).
49. Zhou, W. et al. Graph attention guidance network with knowledge distillation for semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15 (2023).
50. Sohn, K. et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural. Inf. Process. Syst.* **33**, 596–608 (2020).
51. Zhang, B. et al. Semi-supervised semantic segmentation network via learning consistency for remote sensing land-cover classification. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2**, 609–15 (2020).
52. Li, J. et al. Semisupervised semantic segmentation of remote sensing images with consistency self-training. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–11 (2021).
53. Chen, C. et al. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *Proceedings of the International workshop on machine learning in medical imaging* (Springer, 2018).
54. Kim, J. et al. Structured consistency loss for semi-supervised semantic segmentation. arXiv preprint arXiv:200104647 (2020).
55. Zhang, H. et al. Noise-robust consistency regularization for semi-supervised semantic segmentation. *Neural Networks* **184**, 107041 (2025).
56. Yuan, J. et al. Semi-supervised semantic segmentation with mutual knowledge distillation. In *Proceedings of the Proceedings of the 31st ACM international conference on multimedia* (2023).
57. Chen, D., Ma, A. & Zhong, Y. Semi-supervised knowledge distillation framework for global-scale urban man-made object remote sensing mapping. *Int. J. Appl. Earth Obs. Geoinf.* **122**, 103439 (2023).
58. Ma, W., Karakuş, O. & Rosin, P. L. Knowledge distillation for road detection based on cross-model semi-supervised learning. In *Proceedings of the IGARSS 2024–2024 IEEE International Geoscience and Remote Sensing Symposium* (IEEE, 2024).
59. Song, J. et al. RS-MTDF: Multi-Teacher Distillation and Fusion for Remote Sensing Semi-Supervised Semantic Segmentation. arXiv preprint arXiv:250608772 (2025).
60. Datatang Huawei Cloud Cup 2019 China Internet+ College Student Innovation and Entrepreneurship Competition. Datatang (2023).
61. Lin, H. et al. A multi-task consistency enhancement network for semantic change detection in HR remote sensing images and application of non-agriculturalization. *Remote Sens.* **15**(21), 5106 (2023).
62. Zhu, J. & Gao, N. Entropy teacher: Entropy-guided pseudo label mining for semi-supervised small object detection in panoramic dental X-Rays. *Electronics* **14**(13), 2612 (2025).
63. Yang, Q. et al. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021).
64. Shrivastava, A., Gupta, A. & Girshick, R. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
65. Yang, L. et al. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023).
66. Ouali, Y., Hudelot, C. & Tami, M. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020).
67. Ke, Z. et al. Guided collaborative training for pixel-wise semi-supervised learning. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, F, 2020 (Springer, 2020).
68. Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–605 (2008).

Author contributions

H.-L.J. and N.W. designed the study and wrote the manuscript. B.G. and Z.-K.L. performed the experiments. R.-H.W. analyzed the data. X.-W.L. and M.Z. wrote the manuscript. B.-H.C. and E.-M.Z. prepared figures. G.-P.R. supervised the project. X.-W.L. E.-M.Z, and D.-Q.Y. Funding Acquisition. All authors reviewed and approved the final manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (32260131, 31960119, 62262001), the Yunnan Young and Middle-aged Academic and Technical Leaders Reserve Talent Project in China (202405AC350023, 202205AC160001), and the Scientific Research Fund project of the Education Department of Yunnan Province of China (2025Y1250, 2024Y850).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.Z. or D.-Q.Y.