Remote Sensing of Environment 331 (2025) 115051

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Breaking the limitations of scenes and sensors variability: A novel unsupervised domain adaptive instance segmentation framework for agricultural field extraction

Ren Wei [a], Lin Yang [a,*], Xiang Li [a], Chenxu Zhu [a], Lei Zhang [c], Jie Wang [a], Jie Liu [a], Liming Zhu [d], Chenghu Zhou [a,b]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[d] College of Hydraulic Science and Engineering, Yangzhou University, Yangzhou 225009, China

ARTICLE INFO

Editor: Marie Weiss

Keywords:
Agricultural field parcel extraction
Unsupervised domain adaptation
Instance segmentation
Image adaptation
Instance adaptation
Consistency mutual learning

ABSTRACT

Extraction of agricultural field parcels is of great importance for agricultural condition monitoring, farm management, and food security. Several methods have been developed to map the distribution of agricultural field parcels, among which deep learning-based supervised learning is increasingly employed. Nevertheless, advanced deep learning models face two major limitations: limited ability to generalize across different spatial , temporal and sensor contexts with varying scene and object characteristics, and high requirement for annotated datasets to support training and validation. To address this challenge, we introduce a novel unsupervised domain adaptation (UDA) framework (UDA-Field Teacher, UDA-FT) for agricultural field parcel instance segmentation, which is designed to transfer knowledge from labeled source domains to unlabeled target domains. UDA-FT is based on the Mask R-CNN framework and incorporates a target-oriented teacher model and a cross-domain student model. This cross-domain student model embeds an image adaptation module and an instance adaptation module, employing adversarial learning strategies to mitigate cross-domain distribution differences. Additionally, we propose a consistency mutual learning module based on soft pseudo-label technology, overcoming the limitations of traditional hard pseudo-labeling in confidence threshold selection and improving model robustness in the target domain. Furthermore, to address the difficulty in generating independent instance labels for densely packed agricultural field parcels and capturing spatial contextual relationships during soft pseudo-label generation, we propose two data augmentation methods, namely CutMatch (CM) and LeakyMask (LM). We adopted the proposed framework on cross-scene and cross-sensor datasets to evaluate its effectiveness and robustness under different scenes. Quantification and visualization results demonstrate our UDA-FT outperforms existing domain adaptation methods for cross-scene and cross-sensor agricultural field parcels across all metrics. Ablation studies highlight the substantial impact of strong data augmentation on model performance, emphasizing the importance of learning from out-of-distribution data. As an innovative application of unsupervised domain adaptation in agricultural field parcel instance segmentation, this research provides a novel method for domain shift in agricultural remote sensing imagery, enabling more accurate field instance segmentation with significant implications for global agriculture.

1. Introduction

Agricultural fields, as a pivotal resource, play a key role in food security (Weiss et al., 2020) and support environmental sustainability through biodiversity conservation and resource management (Pande and Moharir, 2023). Accurate information on the position and extent of agricultural fields is essential for resource and environmental management, such as farmland area estimation, crop yield prediction, and

* Corresponding author.
E-mail address: yanglin@nju.edu.cn (L. Yang).

https://doi.org/10.1016/j.rse.2025.115051
Received 12 April 2025; Received in revised form 17 September 2025; Accepted 25 September 2025
Available online 29 September 2025

farmland change pattern recognition (Corgne et al., 2016; Duveiller, 2010). Currently, acquisition of agricultural field data still relies heavily on manual work, such as field investigation and agricultural data statistics collection, requiring high labor costs and time expenditures, which often makes large-scale agricultural field survey impractical. Furthermore, on-site investigation and sampling of these fields are costly and time-consuming, typically allowing for the investigation of only a small number of locations and failing to capture data for the entire surface of the scene.

With the development of remote sensing technology, extraction of agricultural field parcels from satellite images has become a popular technology (Huang et al., 2018; Triantafyllou et al., 2019). Visual interpretation with ground observation is commonly needed to create detailed and accurate agricultural field maps at earlier time (Atzberger, 2013; Wachowiak et al., 2017), but those methods are labor-intensive and expensive. Automated methods based on gray scale, color, texture or spectral and spatial information of remote sensing images have then been proposed (Cheng et al., 2020; Graesser and Ramankutty, 2017; Rydberg and Borgefors, 2001). These methods usually need to set parameters carefully for each area, which limits model transferability and generalization.

Deep learning techniques, with great capacity in complex feature extraction and representation, have been applied in agricultural field extraction (Yuan et al., 2020). Specifically, semantic segmentation has shown great effectiveness in agricultural field parcel extraction. This type of approaches, including single-task models (Li et al., 2022; Xie et al., 2023; Zhang et al., 2020), multi-task models (Li et al., 2023; Long et al., 2022; Xu et al., 2023) and divided and stratified models (Wei et al., 2021; Xu et al., 2024; Xu et al., 2019) has greatly improved the accuracy and efficiency of remote sensing image interpretation. However, the distribution of agricultural field is often closely arranged, as single-task models built for semantic segmentation shows poor performance in delineating closely arranged agricultural field parcels (Diakogiannis et al., 2020; Kampffmeyer et al., 2016; Kemker et al., 2018). Multi-task models combining networks like semantic segmentation and edge detection, can improve boundary accuracy but require complex post-processing and lacks adaptability (Li et al., 2023; Papadomanolaki et al., 2021). Divided stratified models, tailored to specific scenes, require extensive labeled data, which increases training costs. Furthermore, scene-specific stratification and classification necessitate a well-defined manual partitioning approach, the accuracy of which ultimately dictates the algorithm's performance. (Mao et al., 2022; Xu et al., 2019; Zhou et al., 2019).

Given that semantic segmentation methods has the above-mentioned shortcomings, instance segmentation or panoptic segmentation has been adopted to detect individual instances for agricultural field parcel extraction (Garnot and Landrieu, 2021; Mei et al., 2022; Zhao et al., 2022). Garnot and Landrieu (2021) proposed an end-to-end, single-stage method for field parcel panoptic segmentation based on time series remote sensing imagery. Mei et al. (2022) used instance segmentation framework Mask R-CNN (He et al., 2017) based on WorldView-3 satellite imagery to delineate smallholder field boundaries in Northeast India. Zhao et al. (2022) utilized the Hybrid Task Cascade (HTC) (Chen et al., 2019) model to segment agricultural field parcels from the high-resolution remote sensing images of JiLin-1, securing the championship of the iFLYTEK Challenge 2021. Although the above methods achieves high accuracy and do not necessarily require complex post-processing, they are not able to adequately fit domains out of the training domain distribution, thus cannot be transferred to other scenes. The target study area often lacks instance annotations and exhibits a different data distribution compared to areas with available instance annotations, which limits the model's generalization capability and its applicability to real-world scenarios. Furthermore, annotating data for instance segmentation is label-intensive task, making it costly and time-consuming to acquire sufficient training data, especially for large-scale applications.

To address the challenges of annotation cost and the inability of fully supervised learning to adapt to other domain, a potential strategy is to transfer the knowledge from densely labeled agricultural field parcel data of source domain to target domain data without annotations. Specifically, the source and target domains share the same feature space and label space, but have different marginal and/or conditional distributions. This can be achieved by training on the source domain and subsequently adapting the learned model parameters to the target domain through specific strategies. This unsupervised domain adaptation (UDA) technology has become a focus of research in the field of computer vision with the emergence of deep learning, particularly in image classification, object detection and semantic segmentation (Biasetton et al., 2019; Guan et al., 2022; Guizilini et al., 2021; Oza and Sindagi, 2024; Yu et al., 2022; Zou et al., 2018). These studies have demonstrated enhanced generalization and transferability capabilities by many advanced methodologies to reconcile feature distributions across various domains, including discrepancy-based (Chen et al., 2020), reconstruction-based (Hu et al., 2022) and adversarial-based (Biasetton et al., 2019; Ganin et al., 2016; Guan et al., 2022; Guizilini et al., 2021). As the first work to apply adversarial learning to domain adaptation, domain-adversarial neural network (DANN) has served as inspiration for a multitude of subsequent research efforts (Ganin et al., 2016). Adversarial learning is an effective strategy in unsupervised domain adaptation for deep learning, utilizing generative adversarial networks (GANs) to learn domain-invariant features (Goodfellow et al., 2020). Specifically, GANs employ a generator to produce features from both source and target domains, while a discriminator distinguishes between these features. Through this adversarial process, the source and target domain features gradually become more similar and achieve domain alignment. Besides adversarial learning strategies, another effective approach in UDA is the pseudo-labeling method (Cho et al., 2022; Litrico et al., 2023; Yuan et al., 2024). These methods typically employ a teacher-student model framework for UDA. The teacher model, trained on the source domain, generates pseudo-labels leveraging its acquired prior knowledge. The student model then learns from these pseudo-labels to adapt to the target domain, thereby enhancing the effectiveness of model transferability from source to target. Xiao et al. (2021) proposed a teacher-student competitive (TSC)-based unsupervised domain adaptation method to solve the source bias problem. An unsupervised domain adaptation method UDA-COPE was proposed by Lee et al. (2022), where a teacher-student framework were used to address annotation scarcity in category-level object pose estimation. Although teacher-student models achieve satisfying results in UDA by generating pseudo labels to guide student training, the teacher model operates independently without receiving feedback from the student model, which may lead to suboptimal or biased pseudo-label generation. On the basis of teacher-student architecture, mutual learning framework, firstly proposed by Zhang et al. (2018), allows both models to refine each other through collaborative training by enabling bidirectional knowledge exchange. Liu et al. (2020) proposed a cycle-consistent panoptic domain adaptation-based Mask R-CNN architecture, which enables unsupervised nuclear instance segmentation by learning from fluorescence microscopy images, significantly outperforming existing unsupervised domain adaptation methods. Li et al. (2022) developed a framework called Adaptive Teacher, which significantly improves cross-domain object detection performance through domain adversarial learning, weak-strong data augmentation, and teacher-student mutual learning. Yang et al. (2025) proposed the Versatile Teacher framework, which significantly improves cross-domain object detection performance through class-aware pseudo-label selection (CAPS) and instance-level alignment using saliency maps. These works have demonstrated that teacher-student mutual learning is a powerful paradigm for UDA (Liu et al., 2022). Although these SOTA methods have demonstrated impressive performance, they often rely on "hard" pseudo-labels, which are sensitive to thresholding and can discard valuable information. Employing a teacher-student mutual

learning approach with a soft pseudo-labeling strategy could be a potential solution to address these challenges (Xu et al., 2021).

Currently, adversarial learning and pseudo-labeling have also been widely applied in remote sensing (Chen et al., 2022a; Chen et al., 2022b; Luo et al., 2024; Zhang et al., 2022a). Peng et al. proposed full-level domain adaptation network (FDANet), which leverages adversarial learning and a self-training pseudo-labeling approach for domain-adaptive building extraction from very-high-resolution (VHR) imagery (Peng et al., 2022). Huang et al. proposed a spatial-spectral weighted adversarial domain adaptation (SSWADA) network, utilizing an adversarial learning strategy to extract domain-invariant features for hyperspectral wetland mapping (Huang et al., 2023). Luo et al. (2024) proposed the CDST framework for unsupervised domain adaptation (UDA) in object detection for remote sensing imagery. This framework comprises two stages: mitigating domain shift between source and target domains using Cycle-GAN, and facilitating cross-domain transfer via a pseudo-labeling strategy. Although UDA has been widely applied in remote sensing, research in the agricultural domain and particularly in agriculture extraction remains relatively scarce. For agricultural field extraction, Zhang et al. (2022b) employed semantic segmentation and adversarial training approaches to extract agricultural field parcels in target domain without labels. However, this research did not produce instance masks for individual agriculture field parcels.

Although UDA strategy can solve the cross-domain remote sensing segmentation and detection problems, few previous studies has applied it in cross-domain agricultural field parcel instance segmentation. The variation of shapes and sizes of agricultural field parcels across different domains, coupled with inconsistencies in their spectral and textural characteristics pose significant challenges for cross-domain field parcel instance segmentation. In this paper, we propose an unsupervised domain adaptation framework for instance segmentation to address the challenge of suboptimal transfer learning performance. To address the challenges of varying environmental contexts and sensor characteristics, we also collected and processed diverse remote sensing images of agricultural landscapes, create two different datasets (cross-scene dataset and cross-sensor dataset) to train and evaluate the proposed unsupervised agricultural field segmentation method. The cross-scene dataset comprises remote sensing images from diverse geographical regions, enabling evaluation of model performance across different agricultural practices and landscapes. The cross-sensor dataset includes images captured by various remote sensing sensors with different resolution and spectral bands, which is intended to assess the model's robustness and adaptability to the unique characteristics of each sensor in the detection and delineation of agricultural field parcels.

In summary, our contributions are primarily reflected in three aspects:

(1) We propose an unsupervised domain adaptation instance segmentation framework based on adversarial learning and consistency mutual learning, which we name as Unsupervised Domain Adaptation Field Teacher-Student network (UDA-FT). This framework includes an Image Adaptation Module, an Instance Adaptation Module, and a Consistency Mutual Learning Module for the segmentation of agricultural field parcels across scenes and sensors. To the best of our knowledge, we are the first to apply unsupervised domain adaptation to the extraction and mapping of agricultural field parcels across diverse scenes and sensors.

(2) We developed a data augmentation method called CutMatch (CM) to address the challenges of segmenting densely distributed agricultural field parcels across scenes or sensors, along with a data augmentation approach called LeakyMask (LM) to capture detailed contextual semantic information.

(3) To address the shortcomings of hard pseudo-labels commonly used in teacher-student models, we propose a soft pseudo-labeling strategy. This approach eliminates the need for manually setting threshold values and avoids issues such as noise introduction or the loss of valuable targets caused by inappropriate threshold selection, ensuring better training outcomes.

2. Study area and datasets

2.1. Study areas

We selected two types of datasets: cross-scene dataset and cross-sensor dataset to evaluate the applicability and effectiveness of our method facing the challenges of morphology, texture and spectral distribution discrepancy across several domains in agricultural field parcel extraction. To expand diverse geographical and agricultural landscapes, we selected multiple study areas spanning Europe and Asia, as illustrated in Fig. 1. The source domain in Denmark lies in a gentle, low-lying, and temperate landscape where parcels, though morphologically diverse, have a relatively large overall area. In contrast, the target domains in China, Japan, Cambodia, and Vietnam are dominated by smallholder farming practices. In these countries, farmland plots are characteristically small and densely distributed, which adds to the difficulty in the delineation of their fine-scale boundaries. This substantial appearance shift is useful for evaluating cross-domain generalization without leaking target-specific priors.

2.2. Data preparation

2.2.1. Cross-scene datasets

The proposed method is designed to adapt cross-scene domain-adaptive agricultural field parcel extraction. Towards cross-scene aspect, we selected Denmark as the source domain (referred to as the European dataset), and the Asian study areas as the target domain. For Denmark, we downloaded cloud-free (<2 % cloud cover) Esri Imagery (https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9), a global high-resolution imagery service provided by Esri, which aggregates data from multiple satellite and aerial sources including DigitalGlobe's WorldView-2 and WorldView-3 satellites, GeoEye-1, and other providers. The provided data has been pre-processed and orthorectified, ensuring it is analysis-ready. This service offers imagery across multiple zoom levels, with spatial resolutions ranging from coarse scales to ultra-high resolutions finer than 0.1 m. While the native resolution of the underlying imagery in the mosaic varies, the service allows users to request data at specific resolutions. This resolution of 1.5 m was chosen because it effectively captures the boundaries of small to medium-sized agricultural field parcels, particularly in the Asian dataset where fields are often smaller and densely packed, while also balancing computational efficiency for large-scale processing. We specifically used the RGB bands of the Esri World Imagery for the Denmark dataset at the beginning of the growing period in May 2016 to ensure optimal visibility of vegetation for agricultural field delineation.

Like the European dataset, the Asian dataset was constructed by mosaicking multiple cloud-less Esri images (<2 %). We also downloaded 1.5 m resolution Esri Imagery for the Asian study areas, which allows for the accurate delineation of fine-scale field boundaries in densely packed smallholder farming regions, while maintaining manageable data processing demands. We used the RGB bands of the Esri Imagery for the Asian dataset, with acquisition times focused on the peak growing season of June to July to align with optimal vegetation growth for field parcel extraction.

The ground-truth reference annotations for this study were compiled from multiple public datasets and supplemented by our own manual visual interpretation and digitization. Annotations of Denmark dataset are obtained from the 2016 Denmark "Market" dataset, which is part of the European Union Land Parcel Identification System (LPIS) initiative (Xu et al., 2023). The Denmark study area includes more than 550,000 crop field parcel polygons, each assigned with a unique identification number, a crop type from 293 classes, and its field area. For the Asian target domains, the labels for Japan were sourced from the fiboa project

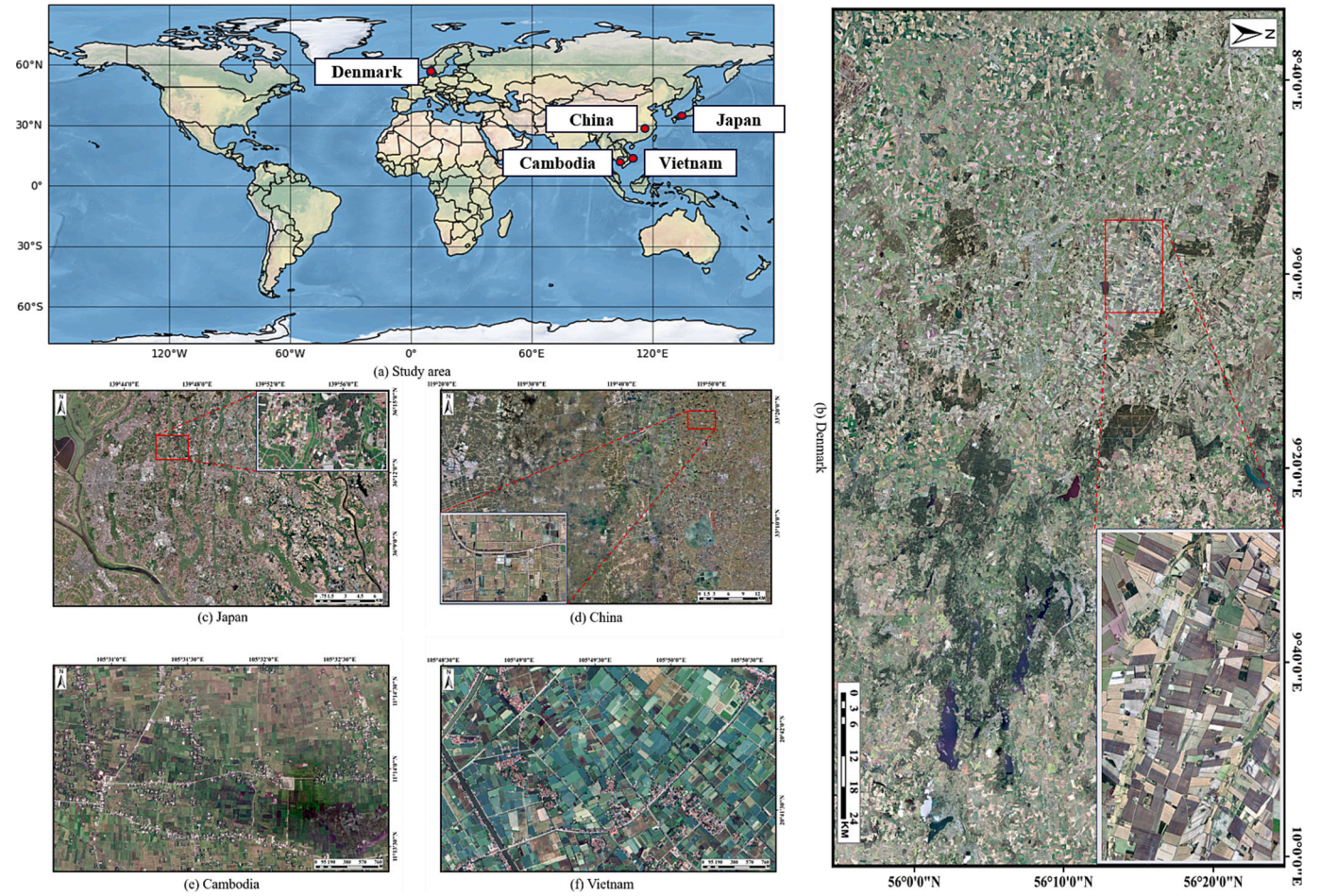

Fig. 1. Study area. Cross-scene dataset involves five countries: Denmark, China, Japan, Cambodia and Vietnam. Cross-sensor dataset involves more than 10 regions in above-mentioned Asian countries.

(Field Boundaries for Agriculture, https://source.coop/repositories/fiboa/japan/description), the data for Vietnam and Cambodia were derived from the AI4SmallFarms dataset (Persello et al., 2023), and the reference data for China were produced through a rigorous manual digitization process with ArcMap 10.6 software by our team. To ensure the reliability and consistency of all reference data from different sources, a final quality control process was performed. We manually inspected all polygons against the imagery and found that over 95 % correctly aligned with visible field boundaries. For the small fraction with deviations, we performed manual corrections to standardize the final dataset to the highest possible quality. During editing, we enforced quantitative QC by applying a positional tolerance of ≤ 1 pixel at edges, removed duplicates using IoU ≥ 0.50, and corrected or excluded invalid polygons (self-intersections/holes).

Ensuring temporal consistency between imagery and reference data was another critical aspect of our data preparation. For the Denmark domain, both the Esri imagery and the LPIS annotations are from 2016. For the Asian domains, we exported Esri basemap tiles in June–July 2024 (Wayback versions), and the tile-level Acquisition Date falls within May–August across 2015–2024. The corresponding annotations for Japan and China were also created in 2024. For Vietnam and Cambodia, the reference data were digitized in August 2021. To mitigate the potential impact of this temporal gap, a dedicated manual verification was conducted. We visually inspected these older annotations against the imagery and confirmed that the majority of parcel boundaries remained stable. Parcels that exhibited changes or mismatches were manually re-digitized to accurately reflect the imagery, thereby minimizing the impact of the temporal discrepancy.

In our study, for central Denmark, a total of 74,653 crop field parcels were available for training and validation. After preprocessing, which included converting multi-polygons to polygons for clarifying each field parcel, removing self-intersecting polygons and parcels categorized as unrelated to agricultural use (e.g., natural or permanently protected areas, forests), approximately 68,000 crop field parcels were retained for model training and validation. Due to GPU memory limitation, the large mosaicked remote sensing images could not be directly fed into training. To address this, we employed a sliding window approach with a size of 512 × 512 pixels to crop both the European dataset and Asian dataset into smaller image patches. Annotations for both the source domain (European dataset for training) and target domain (Asian dataset for validation) were segmented using identical sliding window coordinates, resulting in corresponding annotations for each 512 × 512 image. The cropped images are shown in Fig. 2. The first and second rows display the Esri aerial imagery and corresponding annotation visualizations from the European dataset, respectively. The third and fourth rows display the Esri Imagery and corresponding annotation visualizations from the Asian dataset, respectively. After cropping, the European dataset comprised 2432 images with corresponding COCO-format annotations, while the Asian dataset contained 11,086 images. After that, an 8:2 ratio was used to split each of the two cropped datasets into training and validation sets. This resulted in 2026 training images and 406 validation images for the European dataset, and 8585 training images and 2501 validation images for the Asian dataset. With Denmark as the source domain, and China, Japan, Cambodia, and Vietnam as target domains, four datasets were created. As shown in Table 1, the four datasets are Denmark-China, Denmark-Japan, Denmark-Cambodia, and

Fig. 2. Cross-scene dataset. (a) Imagery from the source domain (Denmark). (b) Source domain annotations. (c) Imagery from the target domain (various Asian regions). (d) Target domain annotations.

Table 1
Cross-scene datasets.

	Source domain	Target domain			
	Denmark	China	Japan	Cambodia	Vietnam
Number of imageries used	2	2	2	28	30
Image acquisition time	May 2016	July–August 2023	June–July 2024	June–July 2015	May–June 2021
Number of patches	2432	4714	3294	1678	1400
Train/Validation number	2026/406	3771/943	2634/660	1337/341	1103/297
Dataset	\	Denmark-China	Denmark-Japan	Denmark-Cambodia	Denmark-Vietnam
Annotations sources	LPIS	Ours	fiboa	AI4SmallFarms	AI4SmallFarms
Annotations acquisition time	May 2016	June 2024	July 2024	August 2021	August 2021

Denmark-Vietnam, with cropped sample sizes of 4714, 3294, 1678, and 1400, respectively. Each dataset was split into training and validation sets with an 8:2 ratio. The final sample sizes for training and validation are displayed in the fourth row of the table.

2.2.2. *Cross-sensor datasets*

Data from several different remote sensors, i.e. JiLin-1, PlanetScope, Esri Imagery and GF–2 (only for large-scale experiment) were used Fig. 3. Table 2 presents the sensors and study areas involved in the cross-sensor dataset. For Esri Imagery and PlanetScope, representative large-scale agricultural areas in Japan were selected as study areas. The Esri Imagery used here is the same Esri Imagery as in the cross-scene dataset, with a 1.5 m resolution and RGB bands, acquired during the growing season in June and July. The Jilin-1 public dataset consists of imagery collected during the vegetation growing season in June. GF-2 was selected to represent large-scale agricultural areas in Jiangsu Province in China. To investigate the practicality of our proposed method across different sensor data, we selected four cross-sensor datasets: JiLin-1-Esri Imagery, Esri Imagery-PlanetScope, Esri Imagery-JiLin-1, and PlanetScope-Esri Imagery. The JiLin-1 remote sensing images and

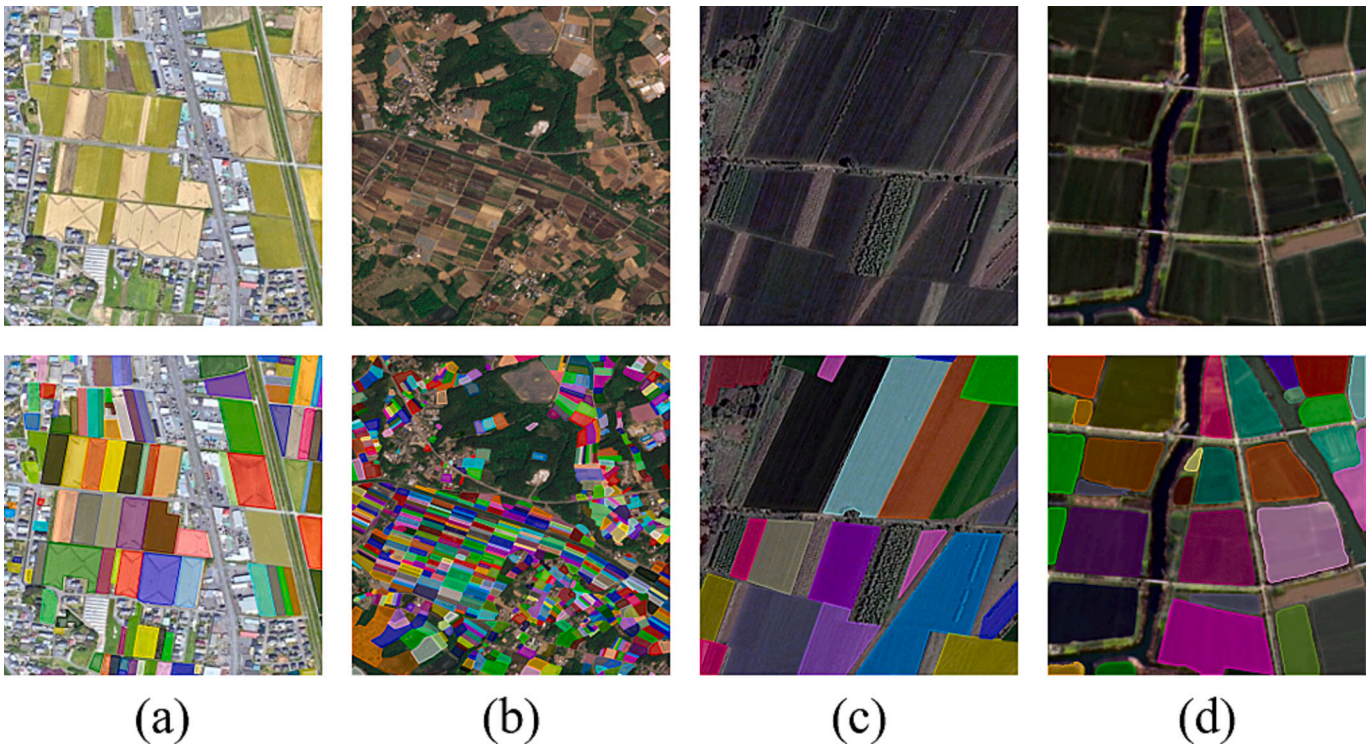

Fig. 3. Cross-sensor dataset. (a) Esri Imagery and its corresponding annotations. (b) PlanetScope imagery and its corresponding annotations. (c) JiLin-1 imagery and its corresponding annotations. (d) GF-2 imagery and its corresponding annotations.

Table 2
Cross-sensor datasets.

	Esri Imagery	JiLin-1	PlanetScope	GF-2
Number of imageries used	4	16	7	6
Image acquisition time	June–July 2024	June 2021	June–July 2024	June–August 2023
Samples' location	Japan	China	Japan	China
Number of patches	6372	4267	1792	8816
Resolution	1.5 m	0.75 m ~ 1.1 m	3 m	1 m
Train/Validation number	5074/1298	3550/717	1428/364	
Annotations sources	fiboa	iFLYTEK dataset	fiboa	Ours
Annotations acquisition time	July 2024	June 2021	July 2024	June 2024

annotations, sourced from the iFLYTEK Challenge 2021 Cultivated Land Extraction Competition dataset (Zhao et al., 2022), cover many agricultural areas in various regions in Southern China. Original data includes 16 JiLin-1 satellite images of varying sizes, with pixel counts ranging from 5000 to 18,000. For consistency in our analysis, we used only the RGB bands of the JiLin-1 imagery. Each image contains four spectral bands: blue, green, red, and near-infrared. The spatial resolution of the JiLin-1 multispectral images is between 0.75 and 1.1 m. The number of annotations varies for each JiLin-1 image, ranging from 149 crop field instance annotations of one image to 4821 (Zhao et al., 2022). The four-band PlanetScope imagery from Planet Lab Inc. (San Francisco, CA, USA) was also used in this study. We downloaded data from Planet Explorer (https://www.planet.com/) through a research and education license with Planet Lab Inc. This study utilized orthorectified and pre-processed PlanetScope Level 3B surface reflectance data, ensuring geometric, radiometric, and atmospheric accuracy. This 3 m resolution imagery, including red, green, blue, and near-infrared bands, was specifically downloaded to match the spatial extent of the Asian dataset used for cross-scene analysis. We used the RGB bands of the PlanetScope imagery, with acquisition times in June and July to align with the growing season for optimal field visibility. And the Esri Imagery utilized is the same as used in the cross-scene dataset.

To further validate the generalizability of our method, we additionally incorporated experiments using GaoFen-2 (GF-2) satellite data for large-scale agricultural field parcel extraction. The GF-2 images were selected for large-scale inference and evaluation purposes, and the annotations were created through manual visual interpretation and digitization of a small portion of these images. We only randomly selected multiple cloud-free (<2 % cloud cover) GF-2 images between June and August 2023 covering various regions within Jiangsu Province in China. These images contain both multispectral and panchromatic bands. For the multispectral data, we performed radiometric calibration, atmospheric correction, and orthorectification to obtain surface reflectance data. The panchromatic bands underwent radiometric calibration and orthorectification. Finally, we fused the processed multispectral and panchromatic data to generate GF-2 images with a spatial resolution better than 1 m. For consistency, we used the RGB bands of the GF-2 imagery for inference. After excluding images with limited agricultural field coverage, we retained six images for our study.

Similar to the cross-scene dataset, the large remote sensing images were cropped into smaller patches using a sliding window approach to create the final training datasets. For both the JiLin-1, PlanetScope, Esri Imagery and GF-2, we opted for a cropping size of 512 × 512 pixels to reduce GPU memory consumption. Finally, we divided each dataset into training and validation sets using an 8:2 ratio.

2.2.3. Datasets distribution analysis

Due to differences in geographic regions, imaging sensors, and

acquisition times, various remote sensing images often exhibit significant spectral shifts, as well as differences in target morphology and distribution. Fig. 4 illustrates the distribution differences within the cross-scene and cross-sensor datasets based on PCA algorithm. In Fig. 4 (a), it can be observed that in the cross-scene dataset, only the Japan dataset overlaps with the Denmark source domain, while the other three target domain datasets have substantial spatial distances from the Denmark dataset. For the cross-sensor dataset, Fig. 4(b) shows considerable overlap between the GF-2 and JiLin-1 datasets; apart from this, there are significant spatial distances between all other datasets.

3. Methodology

3.1. Domain adaptation field teacher framework

3.1.1. Problem formulation and framework overview

Traditional deep neural networks (DNNs) excel at associating input samples $x_i \in X$ with labels $y_i \in Y$ within a specific domain. In the context of agricultural field extraction, this often involves using spatial information and spectral bands as input features to segment agricultural field. However, real-world applications often encounter a significant challenge: the data used for training the model (source domain, $D_s = \{(x_i^S \in X_S, y_i^S \in Y_S)\}_{i=1}^n$) may differ significantly from the data encountered in the real-world application (target domain,$D_T = \{(x_i^T \in X_T)\}_{i=1}^n$). These differences can arise from various factors, including variations in sensor types, and even seasonal variations in data acquisition time. This mismatch in data distributions, denoted as $P_S(X_S) \neq P_T(X_T)$ and $Q_S(Y_S|X_S) \neq Q_T(Y_T|X_T)$, where P_s, P_T are marginal distribution of source domain and target domain, Q_s, Q_T are conditional distribution of source domain and target domain, respectively, can lead to a significant degradation in model performance when directly applying a model trained on D_s to D_T.

To address the aforementioned challenge, a novel model named Unsupervised Domain Adaptation Field Teacher-Student network (UDA-Field Teacher, UDA-FT) is proposed to extract individual agricultural field objects in a target domain using existing agricultural field labels from a source domain. The overall architecture of the model is shown in Fig. 5. FT consists of two main modules: target-oriented teacher model and cross-domain student model. Both the teacher and student model adopt the Mask R-CNN network as their base architecture. The teacher model is specifically designed to generate high-quality soft pseudo-labels for the target domain, while the student model is trained using a combination of strongly-augmented labeled source domain data and soft pseudo-labeled target domain data.

A two-stage training strategy is employed in UDA-FT. The first stage involves a robust pretraining process, where the model is trained on the source domain to learn inherent data representations using Mask R-CNN network with various strong data augmentation methods. The second stage introduces teacher model and student model, facilitating knowledge transfer by aligning the feature representations of both the teacher and student networks through consistency mutual learning. Unlike the fully supervised Mask R-CNN, FT incorporates an adversarial domain discriminator to minimize the domain shift between source and target domains at the image level and instance level. Additionally, during the second stage's mutual learning process, consistency learning between the teacher and student models ensures accurate knowledge transfer. In the inference phase, obtaining the final result only requires a single forward pass similar to the standard Mask R-CNN.

3.1.2. Mask R-CNN

Mask R-CNN is a two-stage instance segmentation architecture built upon the object detection framework Faster R-CNN (Ren et al., 2015). It not only performs object detection but also generates high-quality segmentation masks for each detected instance. In this study, we employ a Mask R-CNN based framework to design an unsupervised domain adaptation method for the precise extraction of crop field parcels. Mask R-CNN consists of three parts: the backbone network for feature extraction, the neck network for feature fusion and the RoI head for task-specific output.

The backbone of Mask R-CNN is ResNet101 (He et al., 2016), a deep residual network known for its exceptional performance in image recognition tasks. The neck component of the Mask R-CNN framework is the Feature Pyramid Network (FPN)(Lin et al., 2017), which is integrated with the ResNet101 backbone to enhance multi-scale feature representation. FPN constructs a feature pyramid by combining low-resolution, semantically strong features with high-resolution, semantically weak features. In addition, the detector head comprises multiple modules for extracting crucial information and generating task-specific objectives. First, the Region Proposal Network (RPN) generates candidate object proposals by analyzing the multi-scale feature map from FPN. These proposals are then refined and filtered by RoI head, resulting in a manageable set of Regions of Interest (RoIs). To ensure accurate spatial alignment, RoI Align is employed, addressing quantization misalignment issues inherent in Faster R-CNN's RoI pooling. The box head then processes these aligned RoIs to predict object categories and

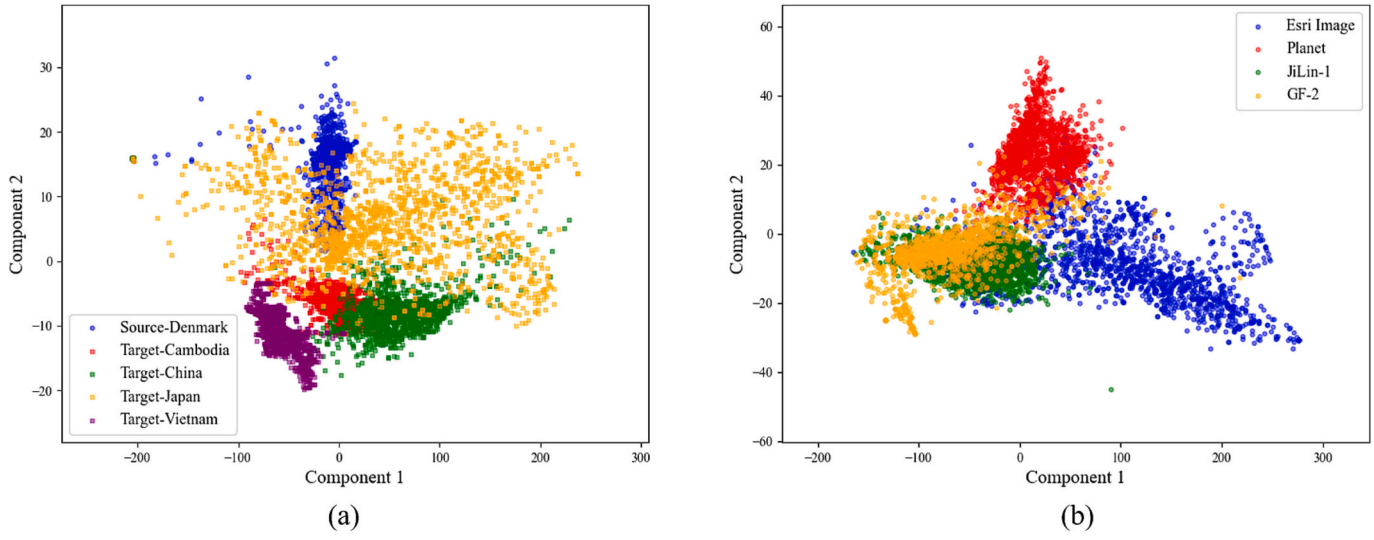

Fig. 4. Cross-scene dataset and cross-sensor distribution visualization. Component 1 is the first PCA component, and Component2 is the second PCA component. (a) Cross-scene dataset distribution visualization. (b) cross-sensor distribution visualization.

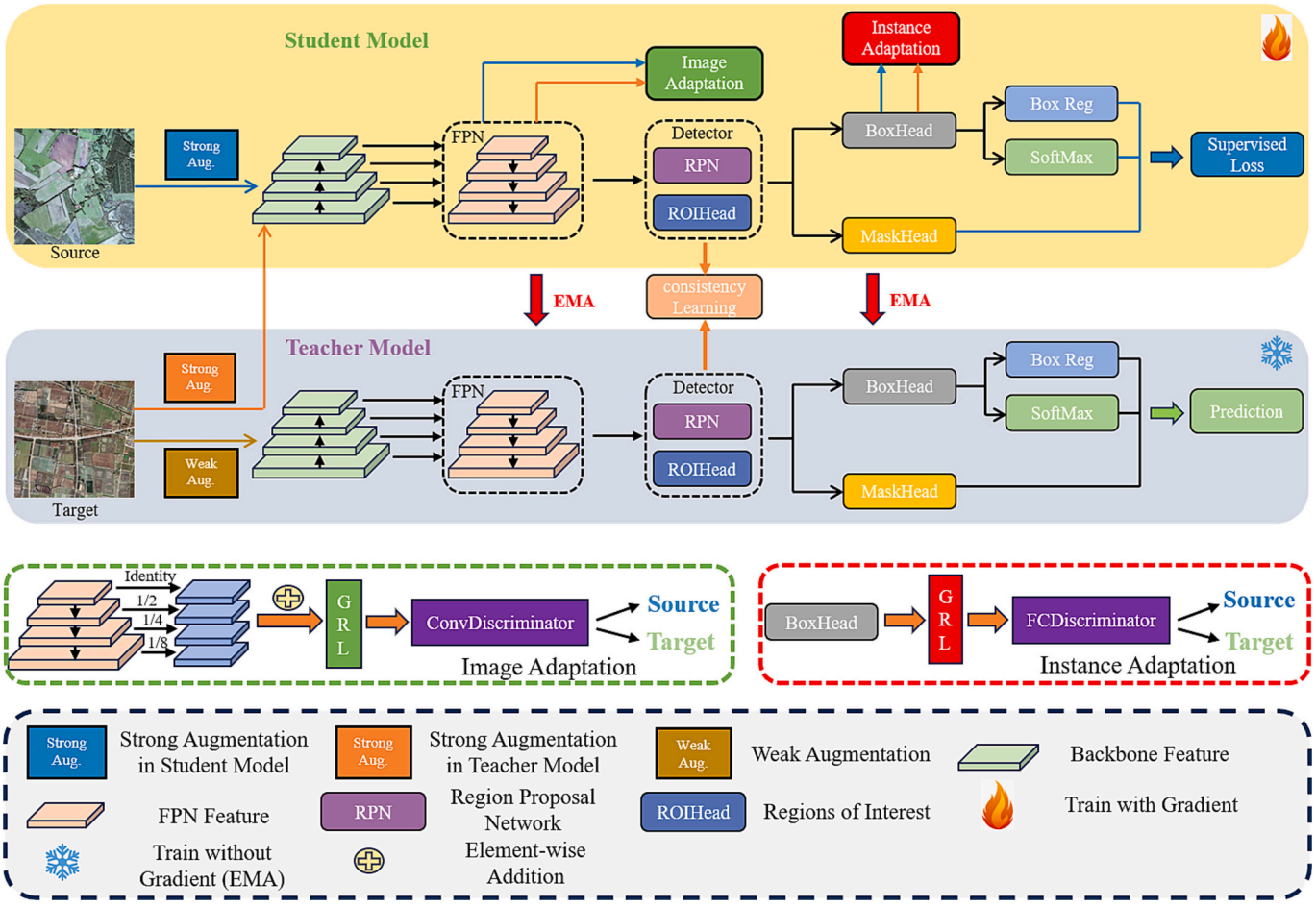

Fig. 5. The framework of the proposed UDA-Field Teacher (UDA-FT).

bounding box offsets. During the training phase, the mask head consists of fully convolutional network (FCN) (Long et al., 2025) that generates a binary mask for each RoI which is from the RPN by RoI Align in parallel. This parallel processing enables the framework to simultaneously predict object classes, bounding boxes, and segmentation masks. But in inference stage, the mask head generates a binary mask for each RoI from the box head output.

3.1.3. Consistency mutual learning module

Mutual learning is widely applied in knowledge distillation and semi-supervised learning. As shown in Fig. 6, the proposed consistency mutual learning module employs the teacher-student architecture. While both the teacher and student models employ the Mask R-CNN architecture, they differ in their parameter update mechanisms. The student model updates its parameters using standard gradient descent, while the teacher model updates its weights by distilling knowledge from the student model through an Exponential Moving Average (EMA) method. The EMA is defined as:

$$\theta_t = (1-\alpha)\theta_s + \alpha\theta_{t-1} \tag{1}$$

where θ_t and θ_{t-1} represent the parameters of the teacher model at the current and previous time steps respectively, θ_s represents the parameters of the student model at the current time step, and a represents the decay rate of the EMA. In this study, a is set to 0.9996.

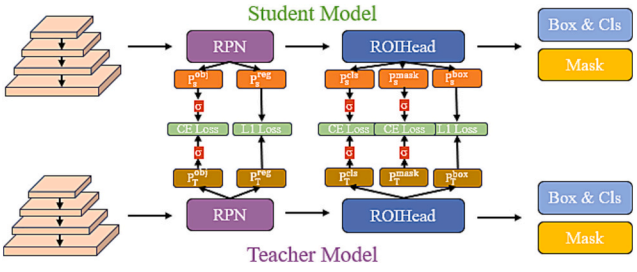

Fig. 6. Consistency mutual learning module.

In the mutual learning process of teacher-student models, the teacher model typically generates pseudo-labels for target domain images to facilitate student model learning.

Consequently, the model's performance is intrinsically linked to the quality of these pseudo-labels. In domain adaptive object detection and instance segmentation tasks, it is necessary to employ manually set confidence thresholds and non-maximum suppression (NMS) to filter the object bounding boxes predicted by the teacher model, thereby generating "hard" pseudo-labels. This approach has three notable disadvantages: 1)Pseudo-label quality depends heavily on confidence thresholds, with poor choices leading to errors that harm training. 2) Low-confidence bounding boxes often provide valuable semantic information and directly eliminating these boxes results in the loss of potential supervisory signals. 3) Since the teacher model's predictions often have low confidence scores in the early training stage, a high confidence threshold for pseudo-label selection risks discarding valuable information, while a low threshold can introduce substantial noise and hindering training effectiveness. To address the issues associated with hard pseudo-labels, we propose a soft pseudo-labeling approach for updating the student model, which eliminates the need for manually set confidence thresholds. This method facilitates mutual learning between the teacher and student models, specifically focusing on the RPN and RoI

Head components.

To be specific, after the features output by the FPN are fed into the RPN, both the teacher and student models generate coordinate offset matrices for bounding box localization and logits vectors for foreground-background classification. The RPN mutual learning loss function used to update the student model is:

$$L_{rpn} = L_{rpn}^{loc}\left(\theta_s^{loc}\left(A, x_s^{strong}\right), \theta_t^{loc}\left(A, x_t^{weak}\right)\right) + L_{rpn}^{cls}\left(\sigma\left(\theta_s^{cls}\left(A, x_s^{strong}\right)\right), \sigma\left(\theta_t^{cls}\left(A, x_t^{weak}\right)\right)\right) \quad (2)$$

where L_{rpn}^{loc} and L_{rpn}^{cls} represent the localization loss and classification loss of the RPN, θ_s^{loc} and θ_t^{loc} represent the parameters of the student and teacher models' localization networks, A denotes the anchors, x_s^{strong} and x_t^{weak} signify strongly and weakly augmented data inputs , and σ denotes the sigmoid activation function. In this study, the localization loss L_{rpn}^{loc} is smooth L1 loss and the classification loss L_{rpn}^{cls} is cross-entropy loss.

The output from RPN is fed into RoI Head to generate the bounding box feature, categorical feature and segmentation feature. Both the teacher and student models have same structure. The RoI Head mutual learning loss function used to update the student model is:

$$L_{roi} = L_{roi}^{box}\left(\theta_s^{box}\left(p_s^{loc}, x_s^{strong}\right), \theta_t^{box}\left(p_t^{loc}, x_t^{weak}\right)\right) + L_{roi}^{cls}\left(\sigma\left(\theta_s^{cls\prime}\left(p_s^{cls}, x_s^{strong}\right)\right), \sigma\left(\theta_t^{cls\prime}\left(p_t^{cls}, x_t^{weak}\right)\right)\right) + L_{roi}^{seg}\left(\sigma\left(\theta_s^{seg}\left(p_s^{loc}, x_s^{strong}\right)\right), \sigma\left(\theta_t^{seg}\left(p_t^{loc}, x_t^{weak}\right)\right)\right) \quad (3)$$

where L_{roi}^{box}, L_{roi}^{cls} and L_{roi}^{seg} represent the bounding box localization loss, classification loss and segmentation loss of the RoI Head. θ_s^{box}, $\theta_s^{cls\prime}$ and θ_s^{seg} represent the parameters of the student models' bounding box localization networks, classification networks and segmentation networks. Parameters with a subscript 't' denote the corresponding parameters in the teacher model, and p_s^{loc} and p_s^{cls} denote the RPN output. Consistent with the loss function employed for RPN mutual learning, the loss functions utilized for bounding box localization and classification are the smooth L1 loss and cross-entropy loss, respectively. Furthermore, the supplementary loss function incorporated into the RoI Head mutual learning framework for segmentation purposes also leverages the cross-entropy loss.

3.1.4. Image adaptation module & instance adaptation module

In the context of domain adaptation for instance segmentation, a significant challenge arises from the fact that annotations are typically only available for the source domain. When the teacher model generates pseudo labels for target domain images, it relies heavily on knowledge acquired from labeled source domain data. This inherent bias can result in noisy and perhaps inaccurate pseudo labels for the target domain, potentially causing the learning process to deteriorate or collapse. The discrepancy between domains exacerbates this issue, leading to a degradation in the overall performance of the adaptation framework.

To mitigate this critical challenge, it is essential to narrow the domain discrepancy between source and target domains. A particularly efficacious strategy to achieve this objective involves the integration of adversarial learning techniques into the framework. The adversarial learning as a powerful mechanism for aligning the statistical distributions across the two domains. This alignment process effectively diminishes the domain shift, thereby enhancing the model's generalization capacity. From the perspective of image features and target instances, image-level adaptation and instance-level adaptation are employed to reduce the shift between the source and target domains.

As illustrated in Fig. 7, image-level adaptation is implemented to align multi-scale features in the domain space subsequent to the FPN processing. Initially, feature maps of varying resolutions obtained from the FPN are downscaled to 1/32 of the original image size through average pooling. These downscaled features are then summed together to create a composite multi-scale information feature. This fused image feature is subsequently fed into a Gradient Reversal Layer (GRL)(Ganin and Lempitsky, 2025) for adversarial training, which serves to obfuscate inter-domain class distinctions. The mechanism of GRL is depicted in Fig. 7. $G_f(x;\theta_f)$ represents the feature extractor and $G_d(x;\theta_d)$ represents the domain classification network, both feature vectors derived from source and target domain inputs are passed through $G_d(x;\theta_d)$ to obtain domain classification results. To adapt the model parameters trained on the source domain to the target domain distribution, we aim to confuse the source and target domain distributions during training. This confusion makes the discriminator unable to determine whether the prediction originates from the source or target domain, effectively maximizing the domain classification loss. The domain classification loss can be defined as:

$$L_{dis} = -d\log G_d\left(G_f\left(x;\theta_f\right);\theta_d\right) - (1-d)\log\left(1 - G_d\left(G_f\left(x;\theta_f\right);\theta_d\right)\right) \quad (4)$$

During network training, confusing the source and target domains requires maximizing the domain classification loss function, which means:

$$L_{adv} = argmax L_{dis} \quad (5)$$

When using gradient optimization methods to search for optimal parameters, the domain classifier parameters are updated as follows:

$$\theta_d = \theta_d - \mu\frac{\partial L_{dis}}{\partial \theta_d} \quad (6)$$

The GRL as defined above is inserted between the feature extractor and the domain classifier, when passing through the GRL, the gradient is expected to be reversed by multiplying by $-\lambda$, which is formally defined as:

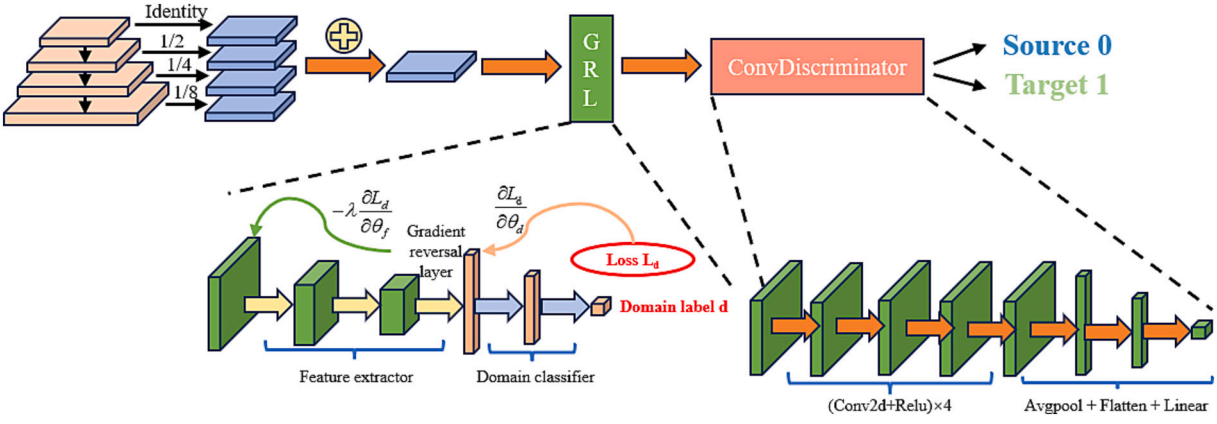

Fig. 7. Image adaptation module.

$$\theta_f = -\lambda \frac{\partial L_{dis}}{\partial \theta_f} \tag{7}$$

In this study, the value of λ is set to 1.

The output from the GRL is then fed to a convolutional discriminator (ConvDiscriminator) for image-level domain classification. The ConvDiscriminator consists of four Convolution+ReLU layers, followed by an adaptive pooling layer and a flattening operation. Finally, a Linear layer is used for domain classification. To facilitate this process, we employ a binary encoding scheme for domain categorization: the source domain is designated as 0, while the target domain is assigned a value of 1. The network parameters are then optimized using a binary cross-entropy loss function, which effectively quantifies the discrepancy between the predicted and actual domain classifications.

As shown in Fig. 8, instance-level adaptation aims to align the feature distributions of specific target instances between the source and target domains. It focuses on aligning the features of individual objects in images (crop field parcel) to reduce the differences in specific target instances across different domains. The output after the boxhead is used as the input for instance-level adaptation. Similar to image-level adaptation, the features are passed through a GRL. These features are then fed into a fully connection discriminator (FCDiscriminator) for instance-level domain classification. The FCDiscriminator consists of two Linear+ReLU layers, followed by a Linear layer for classification. Similarly, the source domain label is set to 0 and the target domain label is set to 1 for training.

3.2. Two-stage training strategy

3.2.1. Stage1: Robust pretraining on source domain

First, we trained a teacher model on the source domain data using a robust pre-training strategy. The key to unsupervised domain adaptation learning within the teacher-student model paradigm lies in improving the quality of target-domain pseudo-labels. During the initial phase of mutual learning, the effectiveness of pseudo-labels is predominantly influenced by the teacher model's capacity to generalize beyond its training dataset distribution (Xu et al., 2021). To enhance this capability, we employ a suite of comprehensive data transformation techniques. These include random resizing of images, random erasing and color jitter (Zhong et al., 2020). Such methods serve to diversify the input data, encouraging the model to develop more robust and adaptable feature representations. These techniques alter the distribution of the original data, enabling models trained on the source domain to achieve superior generalization capabilities.

Moreover, the student model of adversarial training is inherently unstable, particularly when a significant discrepancy exists between the source and target domain distributions, often leading to training collapse. Models trained with robust data augmentation strategies demonstrate enhanced generalization, which effectively mitigates the occurrence of training collapse. It is noteworthy that the random augmentation strategies continuously alter the data distribution, making it more challenging for the model to fit. Consequently, if the network in the first stage is not initialized with pre-trained weights, it is susceptible to early collapse. To avoid this, we initialize the network with weights pre-trained on ImageNet.

3.2.2. Stage2: Teacher-student mutual learning

The second stage of training integrates both labeled source domain data and unlabeled target domain data in a comprehensive approach. Source domain data undergoes strong augmentation before entering the student model, while target domain data is subjected to both strong and weak augmentations, feed into the student and teacher models respectively. While strong data augmentations can enable models to learn more generalized knowledge, they often focus on semantic understanding of the images and seldom consider the difficulties in knowledge transfer caused by differences in object morphology during domain adaptation.

As the morphological difference is particularly pronounced at different field parcel size in cross-scene task and different resolution in cross-sensor task, blurring the boundaries of agricultural field often leads to inaccurate predictions that hinder the training process. To address this, we propose a straightforward yet effective data augmentation technique named "CutMatch" (CM) as illustrated in Fig. 9 (a). This method involves randomly cropping a section from the target domain data (in this study, we used a range size cropping window) and subsequently resizing it to a larger or smaller scale. After the target domain data undergoes "CutMatch" preprocessing, it needs to be further processed through weak augmentation methods such as random flipping and random resizing before being input into the teacher model to generate predictions. For the student model, however, the process to enhance the capability of consistency learning is different. In addition to the same weak augmentation methods applied to the teacher model input, the data input to the student model needs to undergo a series of strong augmentation steps. These methods include conventional techniques such as color jittering and random blurring.

Additionally, inspired by the MIC method(Hoyer et al., 2023), a Random LeakyMask (LM) data augmentation strategy was adopted to the second stage training. The LM method is illustrated in Fig. 9 (b). This methodology employs random masking for partial image occlusion, with the distinctive feature of incorporating a pixel "leakage" mechanism. This mechanism permits a controlled leak of original pixel information through the mask, and provides the model with latent contextual cues. Through this approach, the model can more effectively capture local and global semantic information, thereby enhancing its perception of the surrounding visual environment.

After processing through the backbone and FPN, image adaptation is performed between the features of the strongly augmented source domain data and the strongly augmented target domain data. Subsequently, as the data flows through the RPN and RoI Head, consistency mutual learning is applied to the features derived from the strongly and weakly augmented target domain data. The process continues with instance adaptation, conducted on the strongly augmented source and target domain boxhead processing. The total loss function for the second training stage is:

$$L_{total} = L_{sup} + \lambda_1 L_{img-dis} + \lambda_2 L_{ins-dis} + L_{rpn} + L_{roi} \tag{8}$$

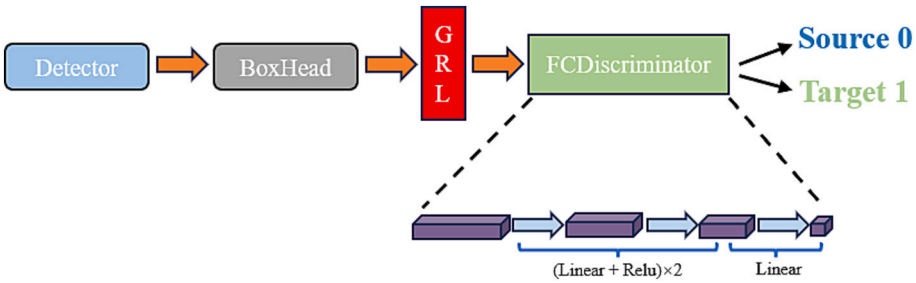

Fig. 8. Instance adaptation module.

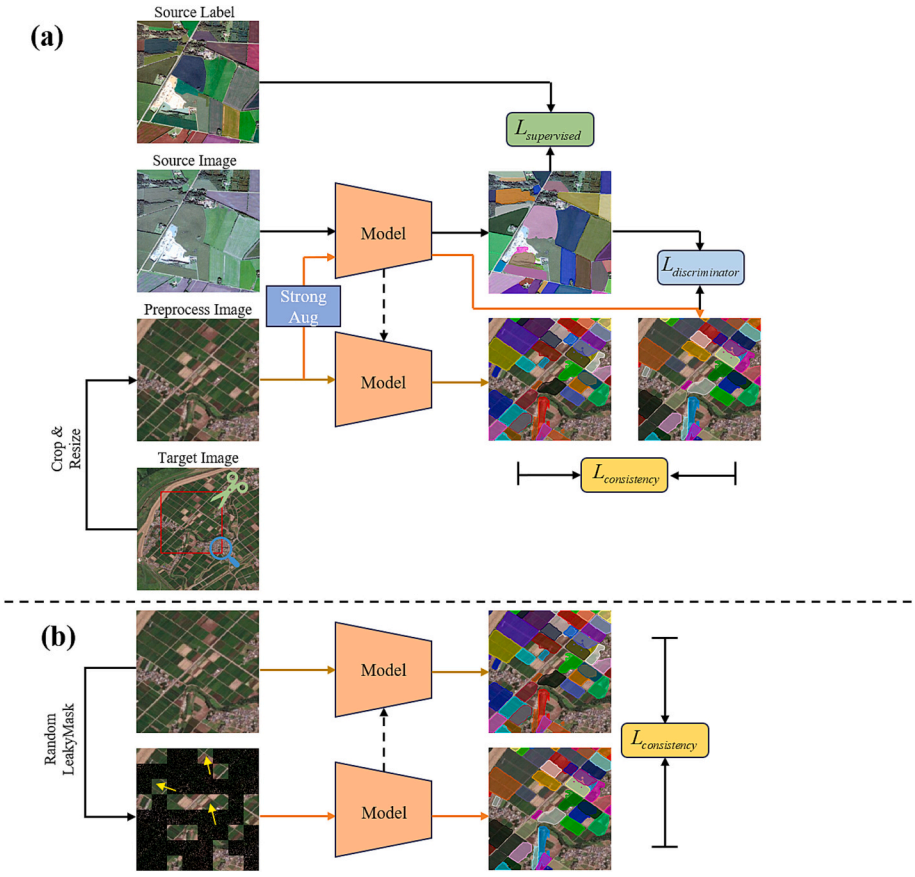

Fig. 9. CutMatch and LeakyMask. (a) CutMatch method. (b) LeakyMask method.

Where L_{sup} denotes the source domain supervised learning loss, $L_{img-dis}$ denotes the image-level adaptation loss and $L_{ins-dis}$ denotes the instance-level adaptation loss. λ_1 and λ_2 are the hyper-parameters used to control the weighting of the corresponding adaptation losses.

This training methodology results in the preservation of the student model's parameters as the definitive training outcome, effectively synthesizing domain adaptation techniques with mutual learning strategies to enhance the model's generalization capabilities and target domain performance.

4. Experiments

4.1. Implementation details

All models were compiled based on the PyTorch framework. We trained all models using two NVIDIA GeForce RTX 3090Ti GPU with 24 GB of memory. The first stage of training involves conventional supervised learning on the source domain data. The training parameters and protocol are as follows: the maximum number of iterations is set to 30,000 steps; the initial learning rate is established at 0.02, with a warm-up strategy implemented for the first 1000 steps of training. Subsequently, a step-wise decay strategy is adopted for learning rate adjustment. Specifically, at the 25,000th step, the learning rate is reduced to 0.002. The batch size is fixed at 16 samples. For parameter optimization, the Stochastic Gradient Descent (SGD) algorithm is employed.

The second stage of training initializes with the optimal parameters obtained from the first stage. The training configuration for this phase is as follows: the maximum iteration count is set to 55,000 steps. The initial learning rate is established at 0.002, with a warm-up strategy implemented for the first 100 steps. Subsequently, a step-wise decay strategy is adopted for learning rate adjustment. Specifically, at the 50,000th step, the learning rate is reduced to 0.0002. The batch size is maintained at 16 samples. The SGD algorithm is employed for parameter optimization. The total loss function L_{total} is a weighted sum of its components. The supervised loss L_{sup} and the consistency losses L_{rpn}, L_{roi} are applied with an implicit weight of 1.0, consistent with standard practice. Notably, the weight λ_1 and λ_2 assigned to the domain adaptation loss function are set to 0.01. This relatively small value is crucial, as excessively large weights for this component can lead to training collapse.

4.2. Description of comparative methods

Given the current absence of unsupervised domain adaptation algorithms specifically designed for agricultural field instance segmentation, direct comparison with other methods in this domain is not feasible. Consequently, we have opted to benchmark our approach against several baselines and other research domain SOTA methods. The source-only method serves as our primary baseline, which refers to a model trained exclusively on the source domain with weakly augmentation and directly applied to the target domain without any adversarial training or consistency learning mechanisms (He et al., 2017). We also introduce the strong augmentation method, which utilizes the stage 1 strong augmentation techniques for training on the source domain data before application to the target domain (Zhong et al., 2020). Additionally, the Adaptive Teacher (AT) method (Li et al., 2022), a state-of-the-art approach for domain adaptive object detection that employs a teacher-student structure serves as another baseline for comparison. To evaluate instance segmentation results, this study incorporates a mask head branch into the AT method for learning agricultural field masks. In addition, the Panoptic Domain Adaptive Mask R-CNN (PDAM) (Liu et al., 2020), the first unsupervised domain adaptation framework for

instance segmentation, is used as a baseline for comparison too. While originally proposed for medical image analysis, PDAM's generalizable architecture allows for its application to our task, maintaining its original structure. Lastly, the Oracle method, representing results obtained by training the model on the target domain under fully supervised conditions, serves as an upper bound for performance. This comparison serves as a critical baseline to evaluate the effectiveness of our proposed unsupervised domain adaptation framework.

4.3. Evaluation metrics

To quantify the model performance, five quantitative object-based evaluation metrics including mean average precision at *IoU* 0.5 to 0.95 (*mAP*, we abbreviate it as *AP* in the table and the text below), mean average precision at *IoU* 0.5 (mAP_{50}, we abbreviate it as AP_{50} in the table and the text below), precision, recall and F1-score were employed. These metrics evaluate the model's ability to correctly identify and segment field parcels as objects, considering both their spatial overlap (via *IoU*) and detection confidence. These object-based metrics are widely used in instance segmentation tasks, such as the COCO evaluation protocol, and particularly suitable for evaluating field parcel extraction in remote sensing imagery (Persello et al., 2023; Chen et al., 2023a; Chen et al., 2023b). While the COCO metric serves as the standard benchmark for comparing instance segmentation models, it is primarily designed for evaluating raster segmentation, which cannot fully reflect and evaluate the effectiveness and advantages of vector-based prediction methods. To address this gap, we introduce the polygons and line segments (PoLiS) metric, which is designed to assess the differences in both the shape and accuracy of polygons (Avbelj et al., 2015). To enable this vector-based evaluation, we perform a post-processing step where the raster-based instance masks generated by the model are converted into vector polygon format using the Suzuki–Abe border-following algorithm (Suzuki, 1985). Specifically, the soft probability masks are binarized, and the object contours are extracted via a standard contour-finding algorithm. These boundaries form the vertices of the predicted polygons, which are then used to compute PoLiS. This vectorization step ensures compatibility with shape-sensitive metrics and supports further GIS-based analysis. For the large-scale experiment, we additionally employed two object-based metrics S^{under} and S^{over}, to quantify the degree of under-segmentation and over-segmentation of extracted field parcels, respectively (Yan and Roy, 2014; Persello and Bruzzone, 2010). S^{under} measures the rate at which an extracted field parcel is larger than the corresponding ground truth parcel (under-segmentation), while S^{over} measures the rate at which an extracted field parcel is smaller than the ground truth parcel (over-segmentation). The formulas for calculating these metrics are as follows:

$$mAP = \frac{1}{c}\sum_{i=1}^{c}\left(\int_0^1 P_i(R)dR\right) \tag{9}$$

$$precision = \frac{TP}{TP+FP} \tag{10}$$

$$recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{12}$$

$$PoLiS = \frac{1}{2q}\sum_{p_j\in P_o}\min_{g\in\partial GT_o}\|p_j-g\| + \frac{1}{2r}\sum_{g_k\in GT_o}\min_{p\in\partial P_o}\|g_k-p\| \tag{13}$$

$$S^{under}(P_j;A_i) = \left(1-\frac{|P_j\cap A_i|}{P_j}\right) \tag{14}$$

$$S^{over}(P_j;A_i) = \left(1-\frac{|P_j\cap A_i|}{A_i}\right) \tag{15}$$

where P and R in formula (9) represent the Precision and Recall, c represents the number of classes (here the value of c is 1), TP represents the number of true positive samples, FN represents the number of false negative samples, and FP represents the number of false positive samples. Formula (10), (11) and (12) present the conventional instance segmentation evaluation metrics: Precision, Recall, and F1-score. Formula (13) presents PoLiS metric. In this equation, P_o is the predicted polygon and GT_o is the ground-truth polygon. The variables p_j and g_k represent the vertices of P_o and GT_o, respectively, while q and r are their corresponding total number of vertices. The metric computes a symmetric distance by summing the minimum distances from each vertex p_j to the boundary of the ground-truth polygon (GT_o), normalized by a factor of $(1/2q)$, and adding this to the sum of minimum distances from each vertex g_k to the boundary of the predicted polygon (P_o), which is normalized by $(1/2r)$. P_j and A_i in formula (14) and (15) represent the prediction field parcel and corresponding annotation. When P_j intersects with multiple fields in A_i, or A_i intersects with multiple fields in P_j, the values of Eq. (14) and (15) are determined by summing the S values for each intersecting field, weighted by their respective intersection areas. These metrics yield percentage values between 0 and 1, with the exception of PoLiS. To enhance the interpretability of the PoLiS metric, we convert the pixel-based spatial distances into actual ground distances, measured in meters.

5. Results

5.1. Benchmark comparison on cross-scene dataset

5.1.1. Results on the Denmark-Japan dataset

Fig. 10 presents the visualization results of several methods in the Denmark-Japan experiment. Overall, almost all methods identified most of agricultural field parcels in the target domain. This suggests a relatively small domain gap between the Japan dataset and Denmark dataset, indicating that features learned from the Denmark dataset transfer relatively well to the Japan target domain. This is consistent with the observation from Fig. 4 that the feature distributions of the Denmark dataset and Japan dataset are the most spatial distance similar. A closer examination of the visualized results reveals that the UDA-FT method exhibits the fewest false negatives, demonstrating its best overall performance. This is particularly evident for small and densely packed agricultural field parcels, where UDA-FT can identify more targets and exhibits the lowest number of missed detections. Specifically, the results in the second row highlighted by yellow squares of the Fig. 10 clearly show that only UDA-FT can detect almost all the small agricultural field parcels completely. The source-only method performs poorly in the second and third rows of the Fig. 10, often merging many small farmland parcels into a single large block. The strong-augmentation method performs better than the source-only method but still missed a lot of small targets. AT and PDAM can effectively identify larger agricultural field parcels and recognize some small ones, but they still miss a considerable number of small parcels, resulting in a worse overall performance than that of UDA-FT.

Table 3 presents a quantitative comparison of all models' performance on the Japan test dataset. The results highlight that the UDA-FT method consistently outperforms other approaches across all evaluation metrics, achieving the highest scores in AP, AP_{50}, Precision, Recall, and F1-score. This performance demonstrates the effectiveness of UDA-FT in mitigating the domain shift and accurately identifying and delineating agricultural field parcels. Notably, while all models exhibit satisfying performance on AP_{50}, a metric often used to assess object detection capabilities, the F1-score, which considers both Precision and Recall, reveals a more substantial performance gap. This difference emphasizes

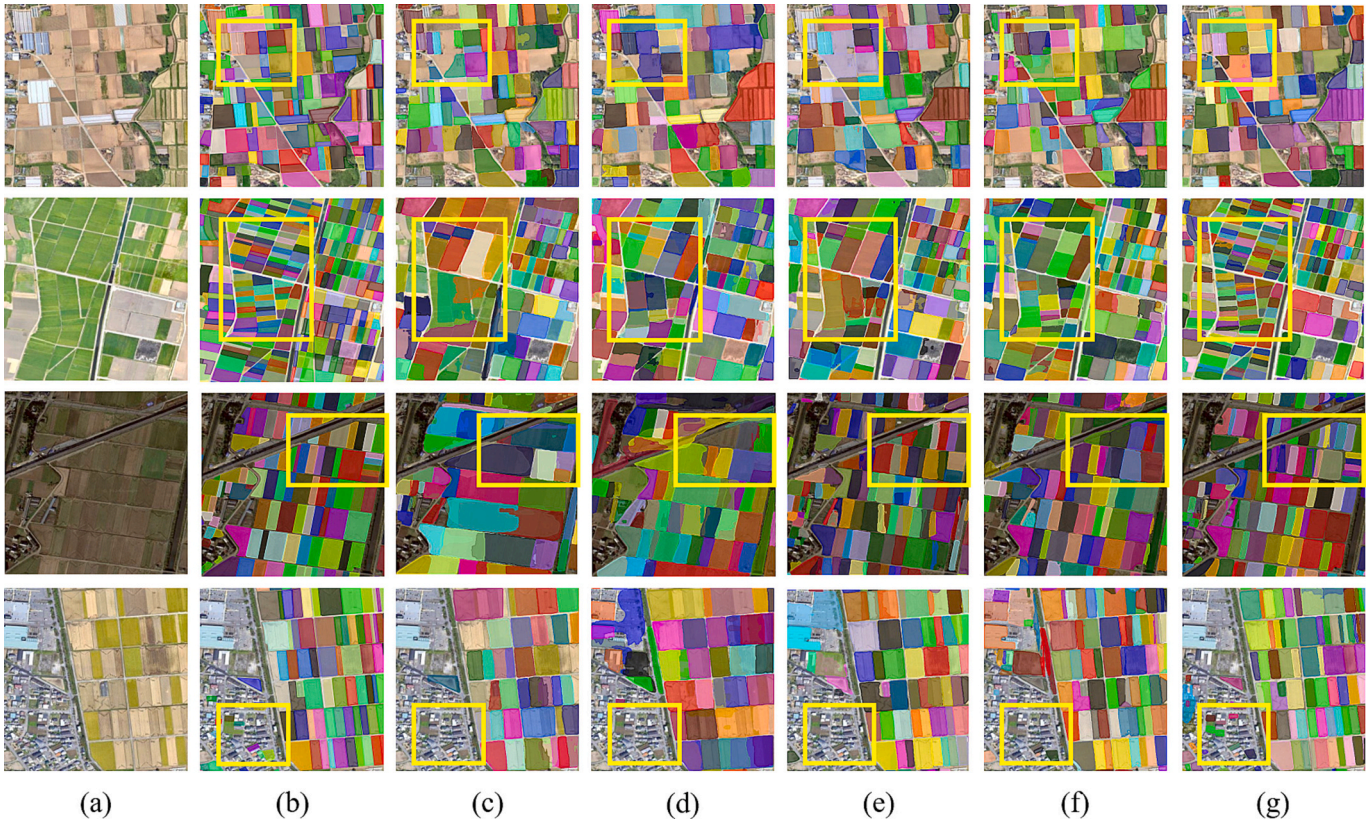

Fig. 10. Visualization results of the Denmark-Japan dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 3
Quantitative evaluation results of the Denmark-Japan dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.216	0.337	0.282	0.251	0.266	19.38
Strong-aug	0.238	0.372	0.557	0.430	0.485	14.15
AT	0.255	0.406	0.591	0.502	0.543	14.21
PDAM	0.267	0.427	0.682	0.621	0.650	13.43
UDA-FT	**0.284**	**0.450**	**0.715**	**0.722**	**0.708**	**10.17**
Oracle	0.348	0.524	0.861	0.787	0.822	6.44

the superior ability of UDA-FT to not only detect field parcel instances but also achieve precise segmentation. Specifically, UDA-FT achieves a Precision of approximately 0.71 and a Recall of 0.72, indicating its proficiency in both minimizing false positives and false negatives. These quantitative results align with the visual results depicted in Fig. 10. In addition, the PoLiS metric further demonstrates UDA-FT's advantage in polygon shape accuracy. UDA-FT achieves a PoLiS of only 10.17 m, which is about 24.4 % and 28.5 % lower than PDAM (13.43 m) and AT (14.21 m), respectively. This highlights UDA-FT's ability to more accurately reconstruct real field boundaries.

5.1.2. Results on the Denmark-China dataset

Compared to the Japan dataset, the China dataset has a more complex background, with less uniform agricultural field parcels. Fig. 4 also shows that the China dataset's feature distribution is further from the Denmark dataset than the Japan dataset's, presenting significant challenges for model learning. As shown in Fig. 11, the source-only method suffers from a significant number of missed detections, particularly in the images shown in the second and third rows. While the strong-augmentation method demonstrates some improvement, it still exhibits a considerable number of false negatives. Although AT and PDAM can identify most agricultural field parcel instances, they are prone to misclassification errors. For example, in the top-right corner of the image in the third row, highlighted by a yellow solid square, both AT and PDAM mistakenly classify the river as small field parcels and fragment some complete parcels into multiple smaller segments, leading to an increased false positive rate. In contrast, the UDA-FT method exhibits fewer misclassifications and can accurately identify and delineate entire farmland parcels without fragmentation, achieving the best overall performance.

Table 4 reveals a noticeable performance decrease of all models compared to the Japan dataset results, indicating that the China dataset presents a more challenging domain adaptation task. This suggests that the diverse and irregular characteristics of agricultural field parcels in the China dataset, coupled with a larger domain shift from the source dataset, pose difficulties for models to effectively transfer knowledge and achieve accurate segmentation. Despite this challenge, UDA-FT consistently outperforms other methods, achieving the highest values for all metrics, including F1-score approaching 0.70. Moreover, the PoLiS metric verifies UDA-FT's superiority in polygon boundary reconstruction. UDA-FT achieves a PoLiS of 11.31 m, significantly lower than PDAM (15.01 m) and AT (16.46 m), with reductions of about 24.8 % and 31.3 %, indicating its high boundary recognition accuracy in complex shapes. In contrast, the source-only and strong-augmentation methods struggle to generalize to the target domain, as evidenced by their considerably lower F1-scores. While AT and PDAM show improvements over the baseline methods, their performance still lags behind UDA-FT in terms of overall performance, as reflected by the lower Precision and F1-score. Interestingly, PDAM achieves a higher Recall than UDA-FT, indicating its ability to identify a larger proportion of field parcel instances. However, this comes at the cost of a lower Precision, suggesting that PDAM may generate more false positive predictions compared to UDA-FT. This observation can be attributed to the trade-off between reducing missed detections (false negatives) and minimizing

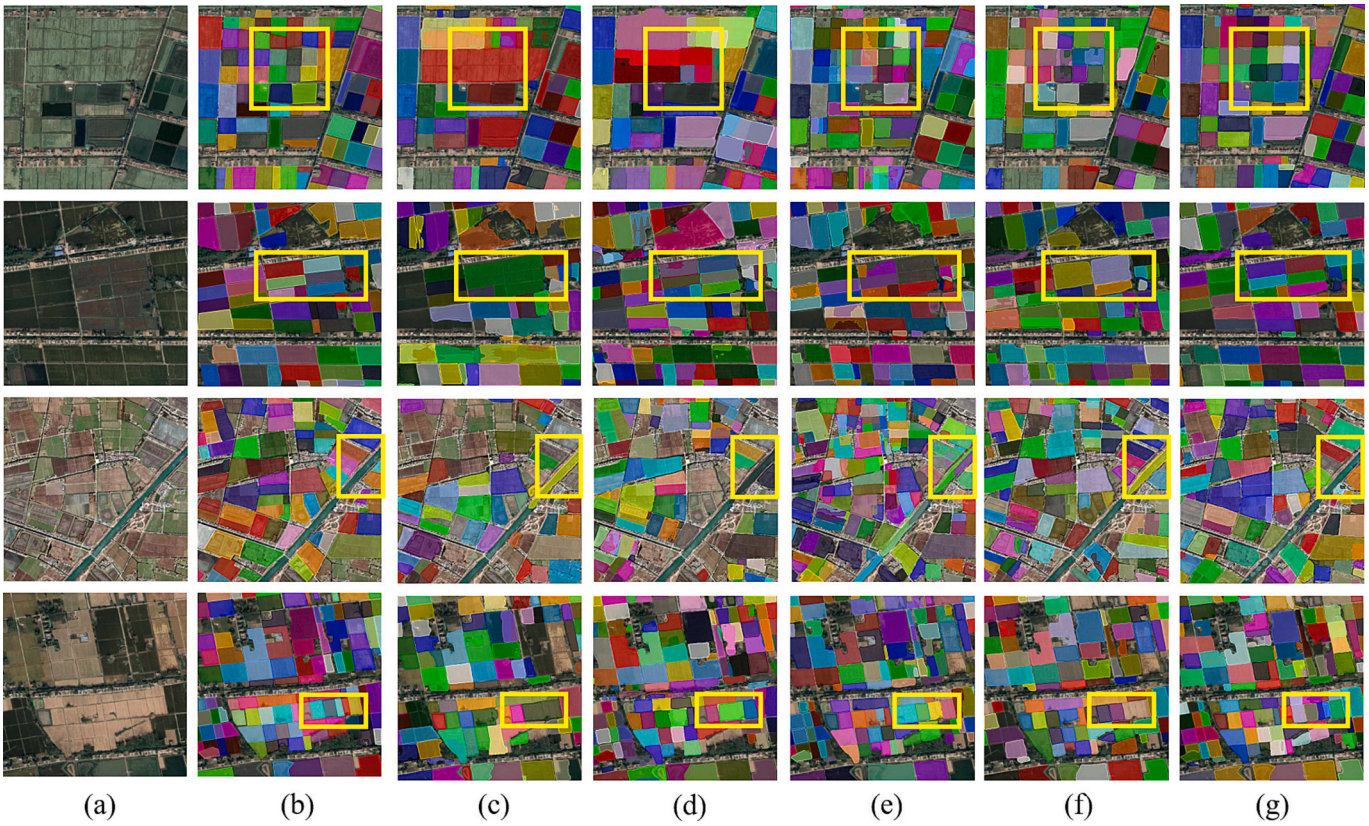

Fig. 11. Visualization results of the Denmark-China dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 4
Quantitative evaluation results of the Denmark-China dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.124	0.232	0.217	0.141	0.171	22.53
Strong-aug	0.179	0.289	0.372	0.288	0.325	18.72
AT	0.189	0.313	0.484	0.529	0.505	16.46
PDAM	0.200	0.342	0.531	**0.695**	0.602	15.01
UDA-FT	**0.227**	**0.373**	**0.698**	0.683	0.690	**11.31**
Oracle	0.305	0.451	0.812	0.775	0.793	6.95

misclassifications (false positives). PDAM's focus on maximizing Recall may lead to the inclusion of more background objects or the fragmentation of parcels, resulting in a higher number of false positive predictions. In contrast, UDA-FT achieves a better balance between Precision and Recall, demonstrating its ability to minimizing both missed detections and misclassifications. This balance contributes to UDA-FT's superior overall performance, as evidenced by its higher F1-score.

5.1.3. Results on the Denmark-Cambodia dataset

The Cambodia dataset exhibits significant differences in the shape and size distribution of agricultural field parcels compared to the Denmark dataset. Field parcels in Cambodia tend to be small and densely packed, posing considerable challenges for model learning. As shown in Fig. 12, the source-only has a high number of missed detections, particularly in the first three rows of images where the boundaries between parcels are unclear (highlighted by a yellow solid square in the figure) with almost no parcels detected. The strong-augmentation method demonstrates notable improvement over the source-only method, but still suffers from a significant number of missed detections and a tendency to merge multiple small parcels into single large blocks. The AT method can identify most farmland instances but still exhibits missed detections and merges small parcels. The PDAM method detects more detailed parcel objects with fewer misclassifications but still experiences some missed detections. In contrast, the UDA-FT method can comprehensively detect most parcels and avoids merging small parcels into larger blocks, achieving the best overall performance.

From Table 5, it can be found that the performance of source-only method does not compare with results obtained on the previous datasets, indicating that the Cambodia dataset presents the more challenging domain adaptation scenario. This suggests that the distinct characteristics of Cambodian agricultural field parcels, such as the prevalence of small, densely packed parcels and intricate field boundaries, coupled with a substantial domain shift from the source data, hinder the models' ability to effectively transfer knowledge and achieve accurate segmentation. Despite these challenges, UDA-FT continues to outperform other methods, achieving the highest values for all metrics, including an F1-score nearing 0.69. Additionally, UDA-FT excels in the PoLiS metric with a score of 8.07 m, which is approximately 15.5 % and 14.2 % lower than PDAM (9.55 m) and AT (9.41 m), respectively. This indicates the improved boundary reconstruction ability of UDA-FT in small and densely packed parcels. The source-only and strong-augmentation methods, on the other hand, struggle significantly to adapt to the target domain, as evidenced by their considerably lower F1-scores, especially for the source-only approach. Although AT and PDAM demonstrate improvements over the baseline methods, their performance still falls short of UDA-FT.

5.1.4. Results on the Denmark-Vietnam dataset

Fig. 13 illustrates the results of transferring models trained on the Denmark dataset to the Vietnam dataset. Compared to other datasets, some images in the Vietnam dataset exhibit blurred field parcel boundaries, as exemplified in the first row of images. Furthermore, the

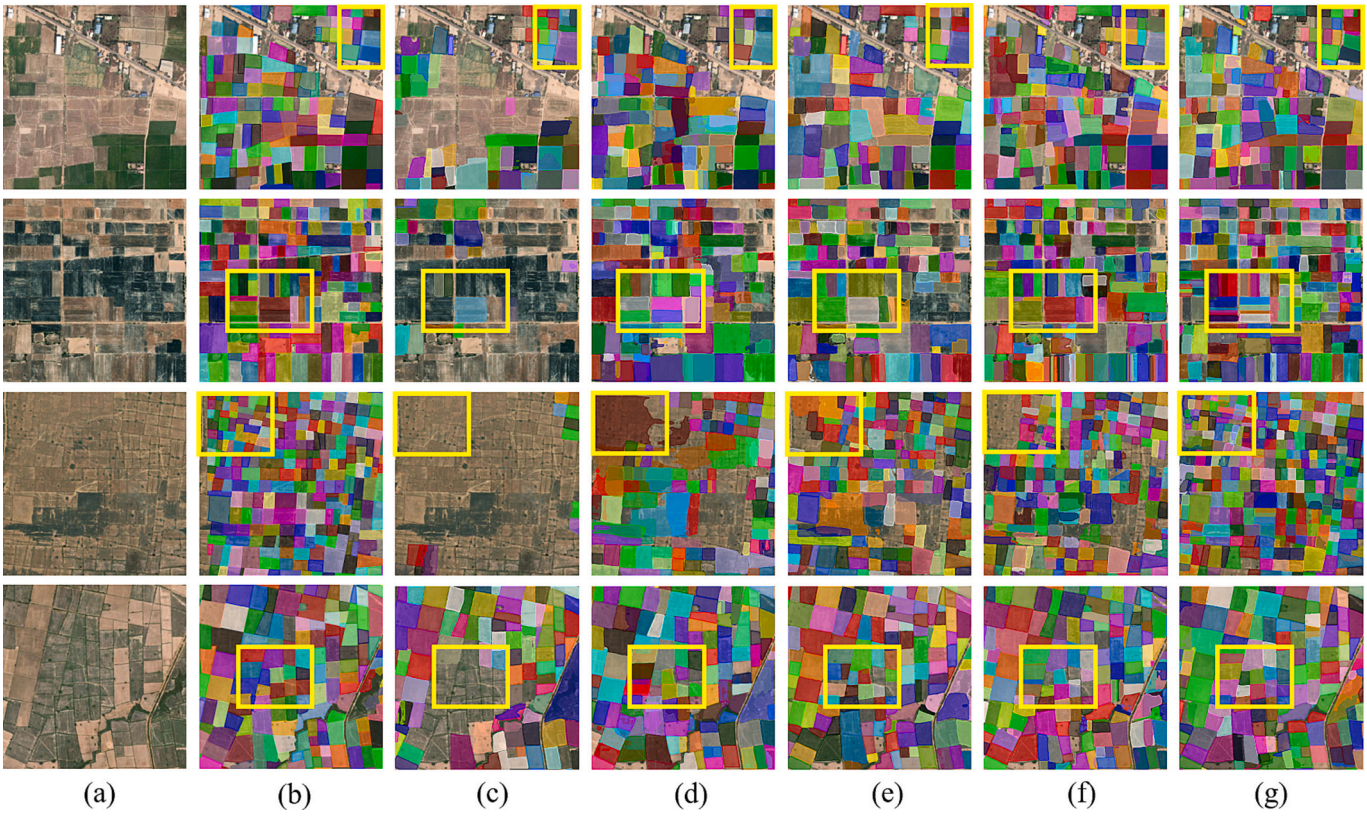

Fig. 12. Visualization results of the Denmark-Cambodia dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 5
Quantitative evaluation results of the Denmark-Cambodia dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.073	0.018	0.202	0.075	0.109	14.97
Strong-aug	0.131	0.258	0.398	0.311	0.349	12.14
AT	0.187	0.292	0.540	0.478	0.507	9.41
PDAM	0.195	0.322	0.664	0.561	0.608	9.55
UDA-FT	**0.219**	**0.349**	**0.719**	**0.659**	**0.688**	**8.07**
Oracle	0.312	0.443	0.828	0.753	0.789	5.01

Vietnam dataset contains a large number of elongated and densely packed parcels, a characteristic that differs significantly from the parcel distribution in the Denmark dataset. This observation is corroborated by Fig. 4, which shows that the Vietnam dataset's feature distribution is the furthest from that of the Denmark dataset. Consequently, the Vietnam dataset presents the most challenging domain adaptation scenario among all the datasets considered. As shown in the highlighted squares, the source-only method performs poorly, exhibiting a high number of misclassifications and missed detections. The strong-augmentation method demonstrates significant improvement over the source-only approach, suggesting that diverse data augmentation techniques enable the model to learn out-of-distribution features. The AT and PDAM methods outperform the first two methods considerably, but still suffer missed detections and a tendency to merge multiple small parcels into single large blocks. While the UDA-FT method also experiences some missed detections and misclassifications, it still outperforms other methods.

From Table 6, it is evident that the source-only method exhibits poor performance on the Denmark-Vietnam dataset, with an F1-score of only 0.056, which is considerably lower than the performance on other datasets. This indicates that the Vietnam dataset presents a particularly challenging domain adaptation scenario. The substantial domain shift between the source and target domains, along with the specific characteristics of Vietnamese agricultural fields, likely contribute to the difficulties in transferring knowledge effectively. Despite these challenges, UDA-FT demonstrates notable improvements, achieving the highest F1-score of 0.606 and outperforming all other methods. This suggests that UDA-FT is the most effective in adapting to the target domain, likely due to its ability to better generalize across the diverse and complex features of the Vietnamese fields. The PoLiS metric further confirms UDA-FT's robustness in shape preservation, with a score of 13.83 m—lower than PDAM (15.80 m) and AT (16.67 m) by 12.5 %–17.1 %—demonstrating better boundary fitting even on the most challenging dataset. The source-only and strong-augmentation methods continue to struggle, with strong-aug only showing moderate improvements, as reflected in their lower F1-scores. While AT and PDAM show some progress over the baseline methods, they still fall short of UDA-FT, which remains the best performing method on this dataset.

5.2. Benchmark comparison on the cross-sensor datasets

5.2.1. Results on the JiLin-1-Esri imagery dataset

Fig. 14 presents the experimental results on the JiLin-1-Esri Imagery dataset. It shows that directly transferring a model trained on the source domain JiLin-1 imagery to the Esri Imagery yields poor performance, characterized by a large number of misclassifications and missed detections. High-resolution images typically exhibit sharp and well-defined agricultural field parcel boundaries, whereas lower-resolution images often display blurred and compact boundaries. In this scenario, the source-only method performs poorly, while the strong augmentation approach shows a remarkable improvement compared to using only source data. However, limitations still exist. AT and PDAM demonstrate substantial enhancements over the first two methods, detecting more agricultural field parcels. However, they still suffer from missed detections and tend to merge multiple small parcels into single large

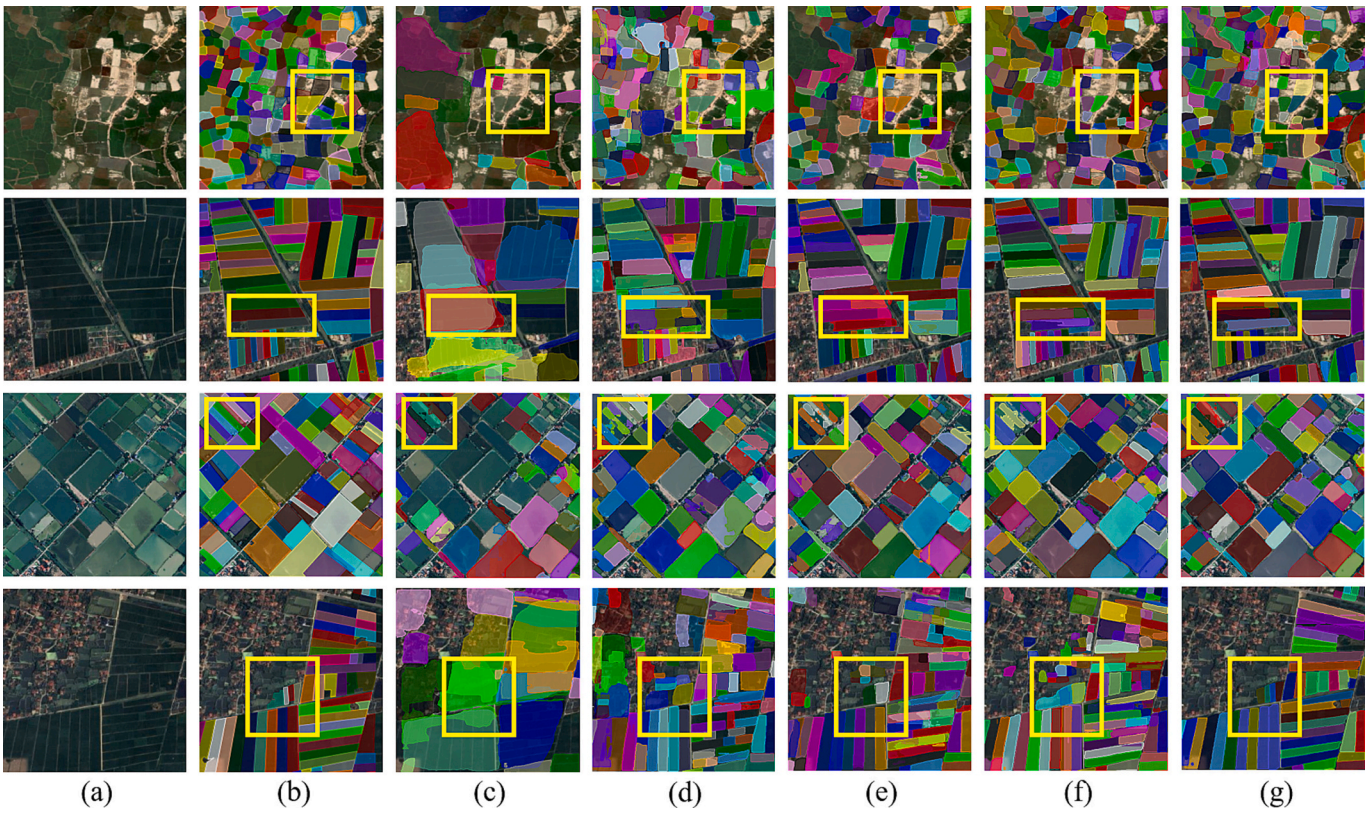

Fig. 13. Visualization results of the Denmark-Vietnam dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 6
Quantitative evaluation results of the Denmark-Vietnam dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.051	0.015	0.113	0.037	0.056	21.82
Strong-aug	0.128	0.210	0.320	0.405	0.358	17.98
AT	0.155	0.254	0.547	0.420	0.475	16.67
PDAM	0.173	0.289	0.521	0.479	0.499	15.80
UDA-FT	**0.192**	**0.337**	**0.626**	**0.587**	**0.606**	**13.83**
Oracle	0.245	0.392	0.762	0.684	0.721	6.64

blocks. In contrast, the UDA-FT method significantly outperforms other methods, especially in the fourth row of images, highlighted by a yellow solid square, where only UDA-FT detects the majority of parcels.

The statistical results presented in Table 7 illustrate a significant performance gap between the models evaluated on the JiLin-1-Esri dataset. UDA-FT consistently outperforms other methods, achieving the highest scores across all metrics, including AP, AP_{50}, Precision, Recall, and F1-score. Specifically, UDA-FT attains an AP_{50} exceeding 0.32 and an impressive F1-score approaching 0.60, indicating its ability to effectively balance precision (minimizing false positives) and recall (minimizing false negatives). While some methods, such as PDAM, exhibit a marginally higher Precision, UDA-FT's Recall suggests its capability to detect a larger proportion of true field parcel instances. Additionally, the PoLiS metric further supports the superiority of UDA-FT in boundary precision. The source-only and strong-augmentation methods, in contrast, struggle to generalize to the target domain, particularly the source-only approach, which suffers from very low Precision and Recall, implying a high number of both missed detections and misclassifications. Although AT and PDAM demonstrate improvements over the baseline methods, their performance still lags behind UDA-FT.

5.2.2. *Results on the Esri Imagery-JiLin-1 dataset*

As depicted in Fig. 15, almost all methods can detect the majority of agricultural field parcels, and the source-only model demonstrates better performance compared to its counterpart in the JiLin-1-Esri Imagery experiment. This suggests that transferring knowledge learned from lower-resolution data to a higher-resolution domain is generally more feasible. Nevertheless, a closer examination reveals that the UDA-FT method exhibits the fewest missed detections and misclassifications, showcasing its superior overall performance. The source-only method suffers from a considerable number of missed detections, while the strong augmentation approach, despite having relatively fewer missed detections, exhibits some misclassifications. Although AT and PDAM detect more targets compared to the source-only and strong augmentation methods, they also experience misclassifications. For instance, the highlighted region in the third row shows that all methods except for source-only and UDA-FT incorrectly identify the central area as agricultural field parcels. Overall, UDA-FT achieves the best performance in terms of minimizing both missed detections and misclassifications, resulting in high object integrity.

Table 8 presents the quantitative evaluation results for the Esri Imagery-JiLin-1 dataset, a relatively easier cross-sensor domain adaptation scenario. UDA-FT consistently achieves the highest performance across all metrics, including a notable AP_{50} exceeding 0.42 and an F1-score approaching 0.72. This highlights its effectiveness in transferring knowledge from lower to higher resolution. While some methods like PDAM achieve a marginally higher Recall, UDA-FT maintains a superior balance between Precision and Recall, minimizing both false positives and false negatives, resulting in the highest F1-score. Meanwhile, the PoLiS metric also reflects shape reconstruction performance under cross-resolution transfer. UDA-FT achieved a PoLiS of 6.20 m, outperforming PDAM (7.34 m) and AT (7.96 m) by 15.5 % and 22.1 %, respectively, indicating robust boundary fitting even with significant resolution shifts. Performance of the source-only and strong-augmentation

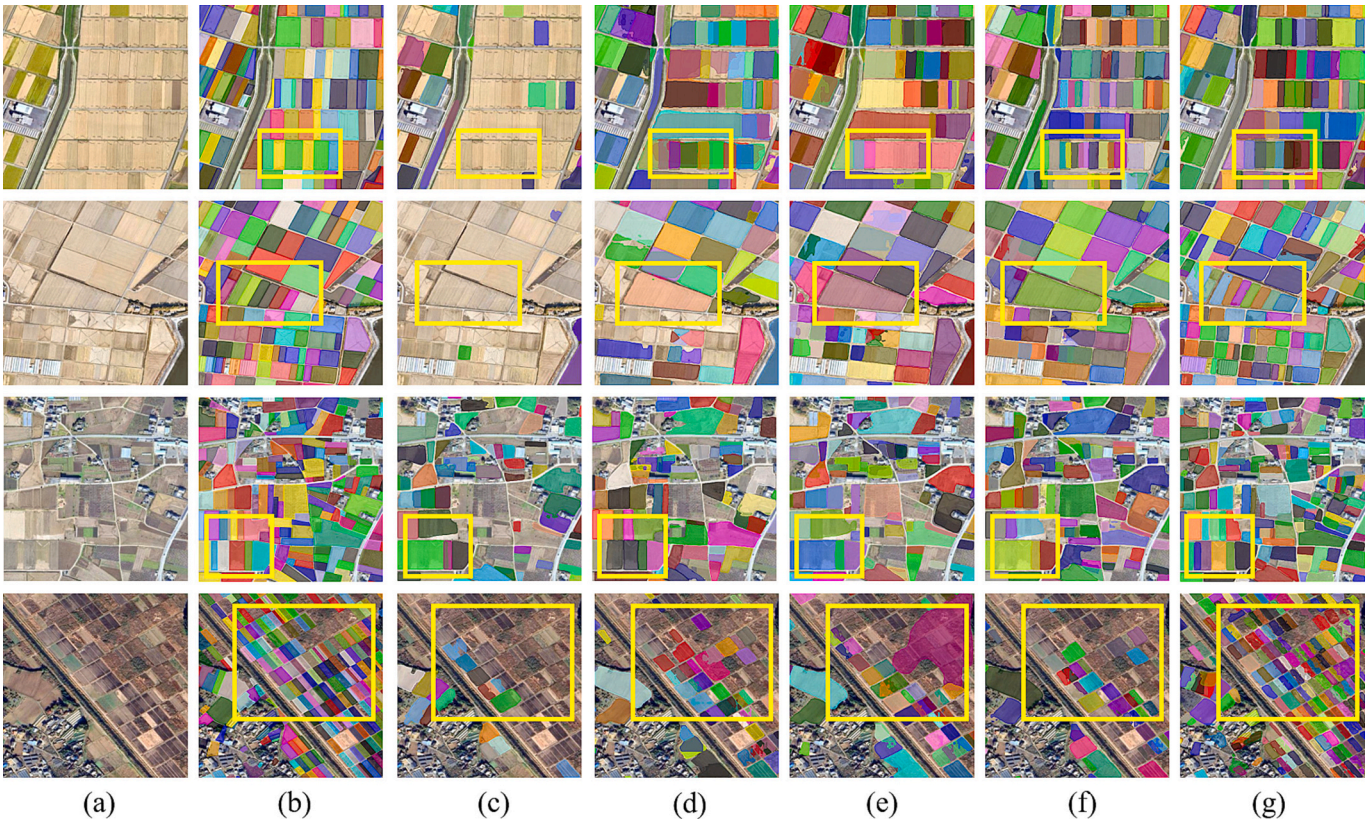

Fig. 14. Visualization results of the JinLin-1-Esri Imagery dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 7
Quantitative evaluation results of the JiLin-1-Esri Imagery dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.043	0.125	0.091	0.022	0.035	17.93
Strong-aug	0.120	0.198	0.287	0.234	0.258	10.85
AT	0.148	0.236	0.411	0.357	0.382	10.03
PDAM	0.151	0.257	0.481	0.413	0.444	11.14
UDA-FT	**0.194**	**0.321**	**0.575**	**0.601**	**0.588**	**9.38**
Oracle	0.348	0.524	0.861	0.787	0.822	6.44

methods improved compared with that in the case of experiments on the JiLin-1-Esri Imagery dataset, but still lag behind UDA-FT, especially in terms of Precision. AT and PDAM show improvements over the baselines but their overall performance, as indicated by the F1-score, remains lower than UDA-FT.

5.2.3. Results on the PlanetScope-Esri Imagery dataset

Because the PlanetScope dataset has the lowest resolution among all datasets involved in the comparison, transferring knowledge learned from the PlanetScope dataset to the Esri Imagery dataset is intuitively more challenging than transferring from the JiLin-1 dataset to the Esri Imagery dataset. Fig. 16 displays the results of transferring from the PlanetScope dataset to the Esri Imagery dataset. It can be observed that the predictions of all models tend to be fragmented and favor small agricultural field parcels, particularly for the source-only model. This is because field parcels in the low-resolution PlanetScope dataset are generally small in shape, leading the models to prioritize predicting small and fragmented parcels when transferred to the high-resolution Esri Imagery dataset. As highlighted in the squares, the UDA-FT method exhibits the fewest missed detections and misclassifications, demonstrating its superior overall performance. The source-only method suffers from a considerable number of missed detections and misclassifications, while the strong augmentation approach shows some improvement in terms of missed detections but still exhibits numerous omissions. Although AT and PDAM detect more targets compared to the source-only and strong-augmentation methods, they also experience misclassifications and over-segmentation.

Table 9 presents the quantitative results on the PlanetScope dataset to Esri Imagery dataset. UDA-FT consistently outperforms other methods across all evaluation metrics, achieving the highest AP, AP_{50}, Recall, and F1-score, although its Precision is marginally lower than that of PDAM method. Specifically, UDA-FT achieves an AP_{50} approaching 0.30 and an F1-score nearing 0.58, demonstrating its ability to effectively balance the trade-off between minimizing false positives (misclassifications) and false negatives (missed detections). While PDAM exhibits a slightly higher Precision, potentially indicating its ability to minimize misclassifications more effectively, UDA-FT's superior Recall suggests its capability to detect a larger proportion of true positive instances, which is crucial in this challenging scenario with potentially many small and fragmented parcels. In addition, the PoLiS metric confirms UDA-FT's shape reconstruction capability, achieving a PoLiS of 12.70 m, better than PDAM (13.66 m) and comparable to AT (12.15 m). The source-only method, as expected, performs poorly in this challenging scenario, particularly struggling with low Precision, implying a high number of misclassifications. Although AT and PDAM show improvements over the baseline methods, their overall performance, especially in terms of F1-score, still lags behind UDA-FT, emphasizing the importance of effectively addressing domain discrepancies and enabling precise feature alignment for successful cross-resolution domain adaptation.

5.2.4. Results on the Esri Imagery-PlanetScope dataset

Fig. 17 presents the experimental results on the Esri Imagery-PlanetScope dataset. It shows that the source-only method performs poorly, exhibiting a high number of misclassifications and missed

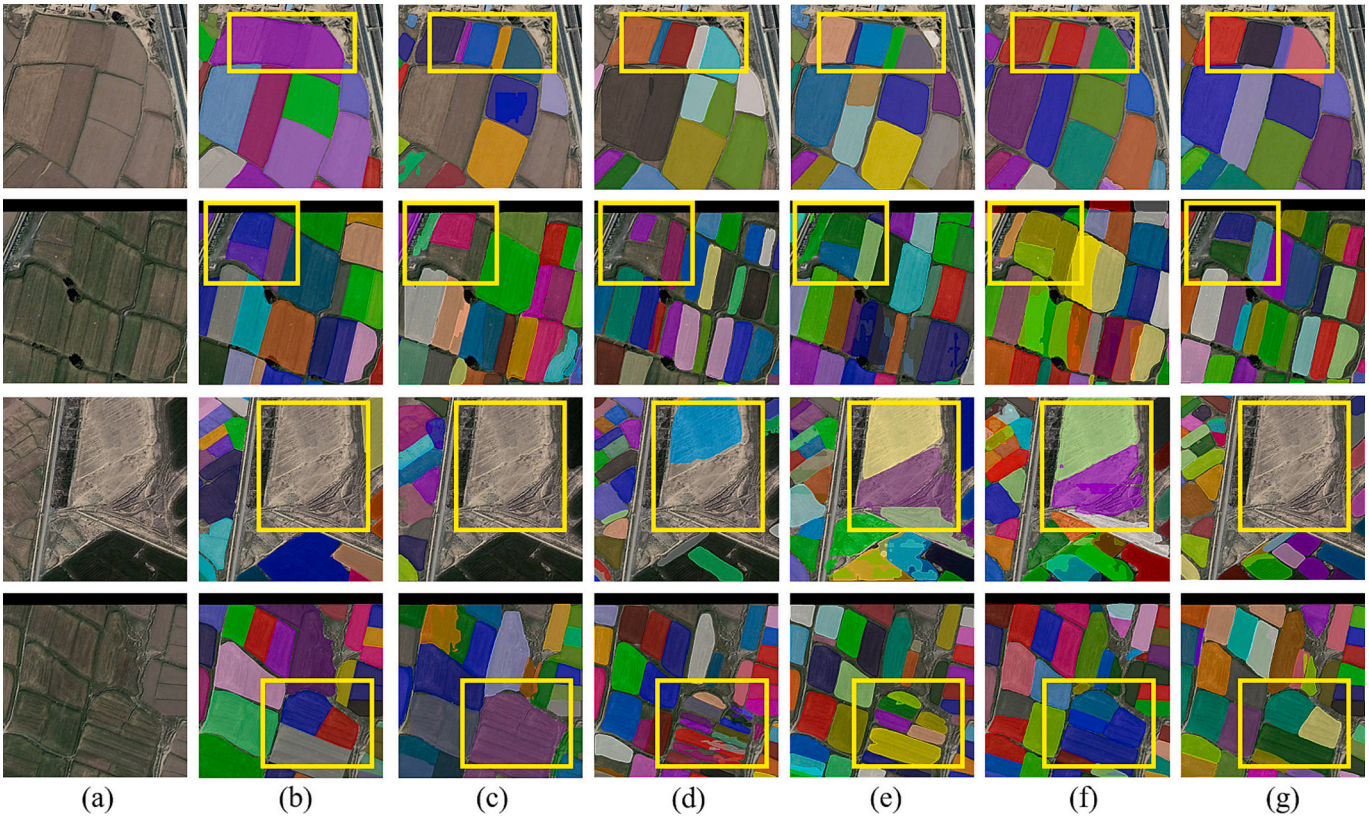

Fig. 15. Visualization results of the Esri Imagery-JinLin-1 dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 8
Quantitative evaluation results of the Esri Imagery-JiLin-1 dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.168	0.258	0.463	0.357	0.403	11.35
Strong-aug	0.204	0.322	0.511	0.482	0.496	9.21
AT	0.237	0.363	0.599	0.613	0.606	7.96
PDAM	0.257	0.393	0.586	0.629	0.607	7.34
UDA-FT	**0.272**	**0.424**	**0.742**	**0.698**	**0.719**	**6.20**
Oracle	0.351	0.551	0.887	0.800	0.841	1.80

detections. The strong-augmentation method shows a remarkable improvement over using only source data, detecting a larger number of targets. However, it tends to overlook fine-grained boundaries, leading to the merging of multiple small parcels into single large blocks. AT and PDAM demonstrate substantial enhancements over the first two methods, identifying more small targets. Nevertheless, they still suffer from missed detections and struggle to segment densely packed agricultural field parcels. In contrast, the UDA-FT method significantly outperforms other methods, especially in the highlighted areas in the third and fourth rows of images where only UDA-FT detects the majority of field parcels and can relatively accurately identify individual small parcels.

Table 10 presents the quantitative evaluation results on the Esri Imagery-Planet dataset, a challenging cross-sensor domain adaptation scenario involving transferring knowledge from higher to lower resolution imagery. UDA-FT consistently outperforms other methods, achieving the highest AP, AP_{50}, Precision, Recall and F1-score, demonstrating its effectiveness in handling this complex domain shift. Specifically, UDA-FT achieves an AP_{50} exceeding 0.23 and an F1-score approaching 0.54, indicating a superior ability to accurately detect and delineate farmland parcels while minimizing both false positives (misclassifications) and false negatives (missed detections). While PDAM exhibits a marginally higher Recall, suggesting a potential for capturing a larger proportion of true positive instances, UDA-FT maintains a superior balance between Precision and Recall, resulting in the highest F1-score. This balance is crucial in this challenging scenario, as it implies a better ability to minimize both types of errors (missed detections and misclassifications). The low value of PoLiS metric further illustrates UDA-FT's ability to preserve shape under low resolution. The source-only and strong-augmentation methods struggle significantly in this cross-resolution setting, exhibiting very low Precision and F1-scores, indicating a high number of both missed detections and misclassifications. Although AT and PDAM demonstrate improvements over the baseline methods, their overall performance, as reflected by the F1-score, still lags behind UDA-FT, emphasizing the efficacy of the proposed framework in achieving robust and accurate segmentation results even when transferring from higher to lower resolution imagery.

5.3. *Large-scale experiments results*

To verify the transferability and practicality of UDA-FT across diverse scenes and sensors at a large regional scale, we designed two large-scale experiments specifically addressing these two situations. For the cross-scene scenario, we trained UDA-FT on the Denmark-Japan dataset and randomly selected a large agricultural area in Japan from the Esri Imagery dataset for inference, predicting individual farmland parcels. Fig. 18a shows the large-scale distribution map of farmland parcels in the test area, while Fig. 18b and c display zoomed-in views of two selected smaller regions. The prediction results are the raw output of the model without any post-processing. From a large-scale perspective, the distribution of agricultural field parcel boundaries aligns well with the agricultural areas observed in the satellite imagery. At a smaller scale, even in densely packed farmland areas, UDA-FT can effectively identify and predict almost all agricultural field parcels without any training samples from the target domain.

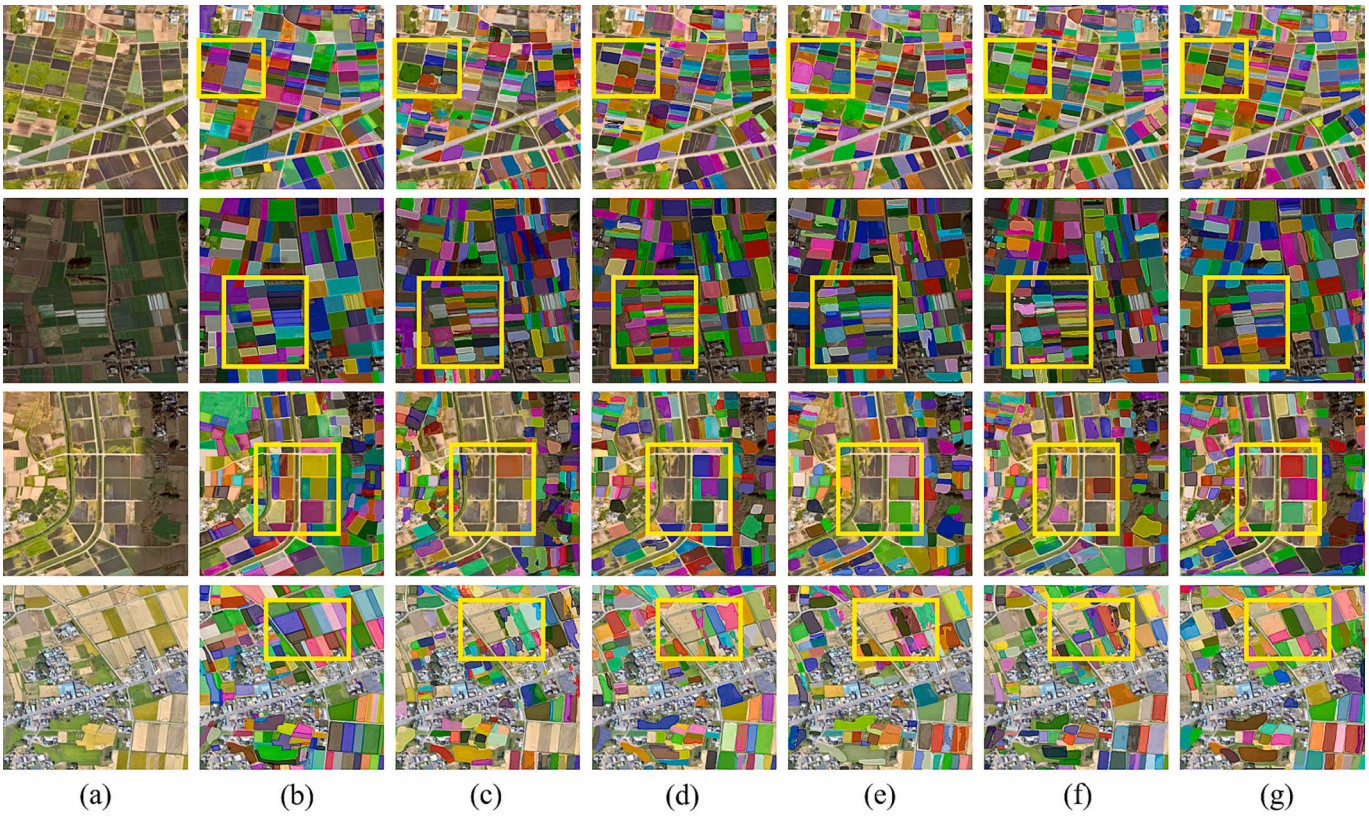

Fig. 16. Visualization results of the PlanetScope-Esri Imagery dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 9
Quantitative evaluation results of the PlanetScope-Esri Imagery dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.088	0.150	0.196	0.283	0.231	18.52
Strong-aug	0.140	0.221	0.387	0.439	0.411	15.21
AT	0.148	0.244	0.498	0.476	0.487	**12.15**
PDAM	0.167	0.272	**0.562**	0.523	0.542	13.66
UDA-FT	**0.182**	**0.297**	0.557	**0.589**	0.573	12.70
Oracle	0.348	0.524	0.861	0.787	0.822	7.16

Fig. 19 shows the large-scale distribution of predicted agricultural field parcels in the cross-sensor scenario. Specifically, we trained UDA-FT on the Esri Imagery-GF-2 dataset and randomly selected a large agricultural area in China from the GF-2 image for inference. Fig. 19(a) displays the large-scale distribution map of agricultural field parcels with GF-2 image in the test area, while Fig. 19b and c provide zoomed-in views of two selected smaller regions. The prediction results are the raw output of the model also without any post-processing. The zoomed-in results demonstrate that even without GF-2 ground truth, UDA-FT can still effectively align the predicted agricultural field parcel boundaries with the agricultural areas observed in the GF-2 image, especially in regions with densely packed agricultural field parcels.

Table 11 presents the quantitative evaluation results for the Denmark-Japan and Esri Imagery-GF-2 datasets. The improved performance in the Esri Imagery-GF-2 dataset can be attributed to the higher spatial resolution of GF-2 imagery (better than 1 m) compared to the 1.5 m resolution of Esri Imagery used in the Denmark-Japan dataset, which likely enhances the model's ability to capture fine-scale field boundaries. Additionally, the lower S^{under} and S^{over} values in the Esri Imagery-GF-2 dataset indicate better segmentation quality, with reduced under- and over-segmentation, consistent with the higher Precision and Recall values. These results further confirm UDA-FT's robustness and transferability across diverse scenes and sensors at a large regional scale.

5.4. Ablation study

To thoroughly evaluate the contributions of different components in our UDA-FT, we conducted a series of ablation experiments. These experiments include the image and instance adaptation, consistency mutual learning, and several data augmentation methods. Each experiment demonstrates the impact of these components on the overall performance of the model.

Table 12 demonstrates the effectiveness of the three above modules on UDA-FT. The first row shows the results of directly transferring the model trained with supervision on the source domain with weak-augmentation to the target domain. The second row displays the model's performance after applying strong data augmentation in stage 1 training. We can observe a significant improvement in the model's performance across both Denmark-Cambodia and PlanetScope-Esri Imagery datasets, with increases in AP_{50} of 0.071 and 0.069 respectively. The third and fourth rows show the results after adding the MIC (Hoyer et al., 2023) and our LM methods respectively. We can see that there is a considerable improvement in accuracy across both datasets for both methods. The MIC module increases AP_{50} by 0.011 and 0.007 for the Denmark-Cambodia dataset and PlanetScope-Esri Imagery dataset respectively. The LM methods, however, demonstrates superior performance, further improving AP_{50} by 0.015 and 0.013 respectively when compared with the strong augmentation baseline. This comparison clearly highlights the effectiveness of the LM method, which outperforms the MIC module in enhancing the model's contextual understanding capabilities. The fifth row shows the results after adding our CM method. We can observe a substantial improvement in the model's performance, with increases in AP_{50} of 0.027 and 0.022 respectively. This significant enhancement is further illustrated in Fig. 20, which provides a visual comparison of the model's output with and without the

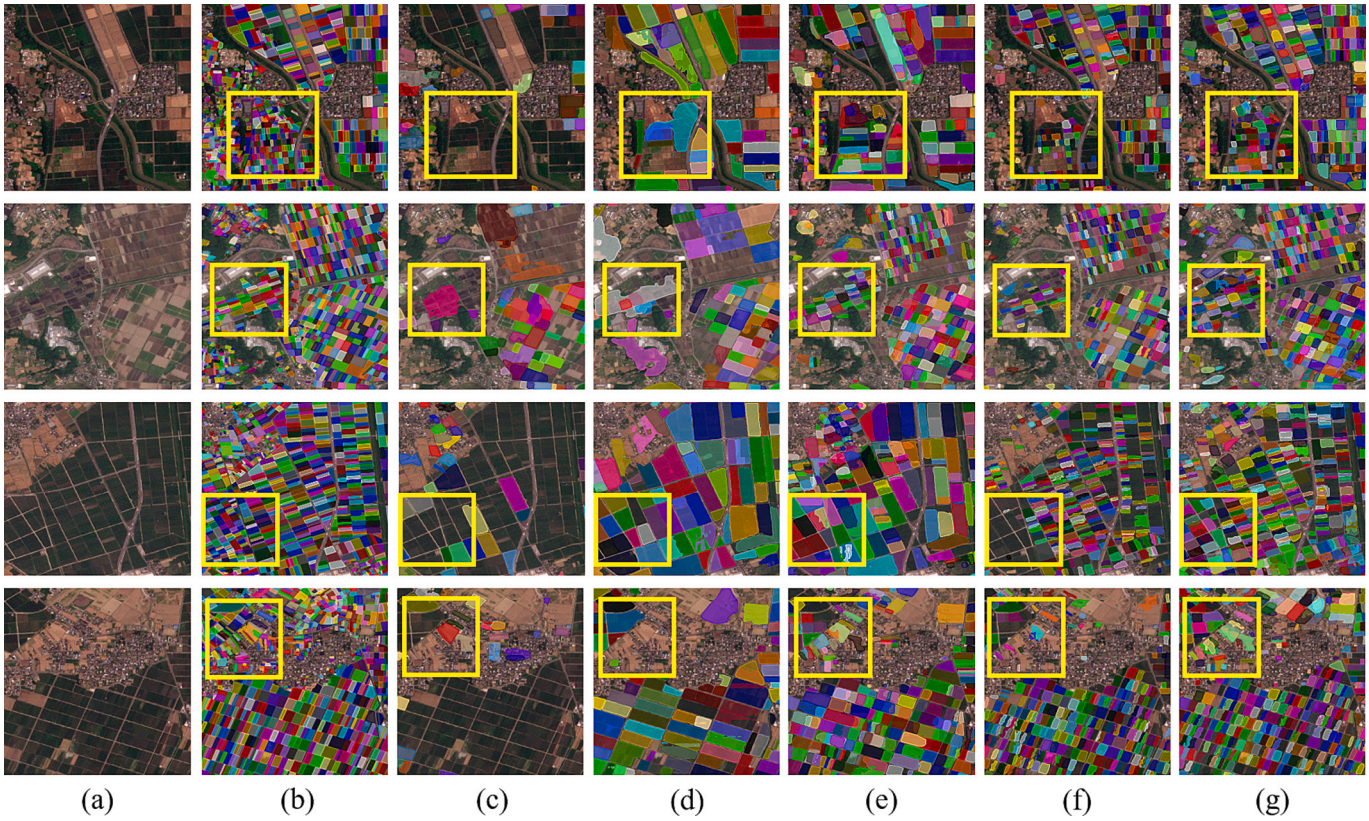

Fig. 17. Visualization results of the Esri Imagery-PlanetScope dataset. (a) Imagery. (b) Annotations. (c) Source-only model's results. (d) Strong-augmentation model's results. (e) AT's results. (f) PDAM's results. (g) Our UDA-FT's results.

Table 10
Quantitative evaluation results of the Esri Imagery-PlanetScope dataset.

Methods	AP	AP_{50}	Precision	Recall	F1	PoLiS (m)
Source-only	0.037	0.091	0.106	0.089	0.097	29.18
Strong-aug	0.055	0.125	0.287	0.261	0.273	23.61
AT	0.089	0.157	0.399	0.387	0.393	15.22
PDAM	0.124	0.191	0.432	0.467	0.449	11.54
UDA-FT	**0.156**	**0.232**	**0.553**	**0.518**	**0.535**	**9.36**
Oracle	0.189	0.321	0.761	0.583	0.660	5.43

CM module. The white boxes in Fig. 20 highlight specific areas where the CM module demonstrates its effectiveness. The sixth row displays the performance after incorporating the IA module. This addition improves the AP_{50} by 0.022 and 0.016 respectively. The last row demonstrates the results after adding the CML module, which further improves the AP_{50} by 0.027 and 0.025 respectively. It is evident that strong data augmentation in stage 1 and CML module in stage 2 provides the most significant boost to the model's performance, followed by CM method. The MIC, LM, and IA modules also contribute to the improvement, albeit to a lesser extent. This observation underscores the critical importance of enabling the model to learn from out-of-distribution data.

The substantial improvement from strong data augmentation in stage 1 suggests that exposing the model to a wider range of data variations during training enhances its ability to generalize across different domains. This is particularly crucial in the context of agricultural field parcels segmentation, where variations in imaging conditions, seasonal changes, and regional differences can significantly alter the appearance of agricultural fields. The contrasting contributions of the LM and MIC methods suggest that leaking some information to the mask is more beneficial for learning and understanding spatial context. In other words, providing the model with some guidance regarding the target object's location within the masked region proves more effective than completely masking it out. Furthermore, the significant improvement observed with the CM method highlights the necessity of further data augmentation specifically targeting size discrepancies between source and target domains. This approach encourages the reduction of morphological distribution differences between the source and target domains, thereby enhancing the model's ability to generalize across domains with varying agricultural field sizes. The contributions of the IA module underscore the importance of adversarial domain adaptation in mitigating the shift between different domains. By performing adversarial learning at both the image level and instance level, the IA module encourages the target domain data to align more closely with the source domain data in a high-dimensional feature space. This alignment further reduces the discrepancy between the two domains, contributing to the model's improved performance on the target domain. The notable contribution of CML module further reinforces the model's robustness to domain shifts. By encouraging consistent predictions across different augmented versions of the same input, this method helps the model to focus on invariant features that are likely to be relevant across domains.

6. Discussion

6.1. Source-target domain shift and alignment

A critical challenge in unsupervised domain adaptation for agricultural field instance segmentation is addressing the domain shift between source and target datasets. As shown in Fig. 21, our experiments with cross-scene and cross-sensor datasets transfers highlight the multifaceted nature of this shift. In the cross-scene scenario, we observed significant disparities in agricultural field parcel size, arrangement, color, and texture, reflecting the diverse agricultural practices and environmental conditions. In cross-sensor experiments, variations in resolution, color balance, and viewing angles of different sensors also pose

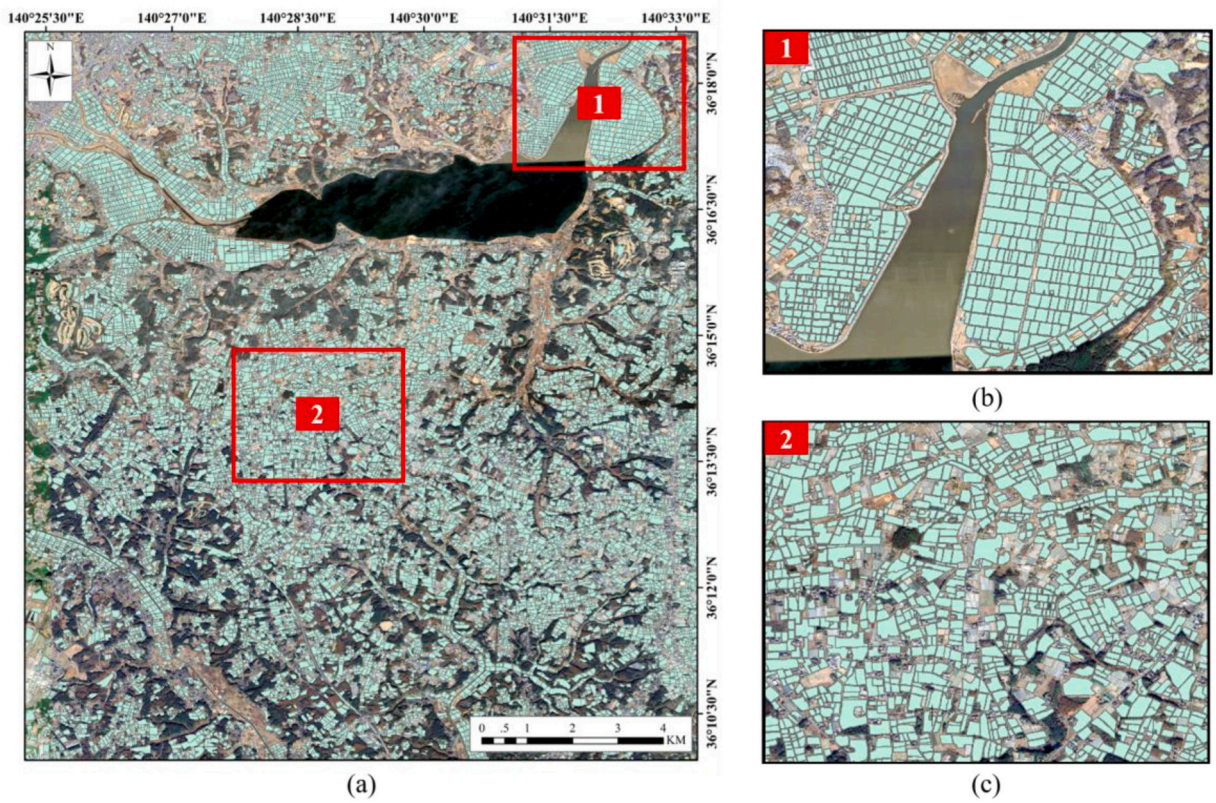

Fig. 18. Predicted distribution map of the agricultural field parcels with Esri Imagery. (a) Global visualization of study area in Japan. (b) and (c) Examples of local extraction results in (a).

challenges for model transfer. UDA-FT's success in both scenes can be attributed to its effective domain alignment strategies. The aligned domains after the image adaptation module and instance adaptation module are shown in Fig. 22. The adversarial learning component plays a crucial role in mitigating the domain shift by aligning the feature distributions between the source and target domains. It achieves this by training a discriminator network to distinguish between source and target features, while simultaneously training the feature extractor to fool the discriminator. This minimax game encourages the feature extractor to learn domain-invariant representations, effectively minimizing the discrepancy between the two domains. Our proposed method addresses the challenge of domain adaptation in transfer learning. By encouraging the model to learn domain-invariant features, it effectively mitigates the negative impact of differing domain distributions. This results in a more robust model that generalizes better to the target domain without requiring labeled target data.

6.2. Effect of hard pseudo-label & soft pseudo-label

The generation and utilization of pseudo-labels in the teacher-student mutual learning process are usually critical in unsupervised domain adaptation. Traditional approaches often rely on hard pseudo-labels, created by applying confidence thresholds and NMS to the teacher model's predictions. However, hard pseudo-labels in exists works are highly sensitive to threshold selection, often leading to false positives and negatives that can severely impact training quality. This is especially problematic in instance segmentation, where accurate detection boxes are crucial for segmentation precision. Moreover, hard thresholding discards low-confidence predictions, which, despite their uncertainty, often contain valuable semantic information. In this study, we implemented a soft pseudo-labeling approach in UDA-FT, eliminating the need for manually set confidence thresholds in pseudo-labels generation and preserving the full spectrum of the teacher model's predictions. This method retains the probability distribution of the model's output, allowing for an information-rich training signal. As shown in Fig. 23, within the red, yellow, and blue boxes, issues such as missed detections and incorrect shape recognition occur with hard pseudo-labels, whereas soft pseudo-labels help to avoid this problem. This indicates that soft pseudo-labels work effectively especially for small or irregularly shaped fields where hard thresholding might completely miss or misrepresent important features.

6.3. Effect of imagery acquisition timing and band combination

6.3.1. Imagery acquisition timing

The timing of imagery acquisition impacts cross-domain adaptation in agricultural remote sensing due to seasonal variations in crop growth stages. To minimize temporal variability, the cross-scene dataset used Esri Imagery for the Denmark source domain (Denmark-Summer, May 2016) and the Asian target domain (June–July), both during the growing season (see Section 2.2.1). The cross-sensor dataset used JiLin-1, GF-2 and PlanetScope imagery (June–July 2023). To investigate this impact, we conducted experiments using Sentinel-2 imagery for Japan-Summer (J^S, July, peak growing season) and Japan-Winter (J^W, December, post-harvest or dormant season), and used the Denmark-Summer Sentinel-2 imagery (D^S, July) as source domain. Fig. 25 illustrates the effect of acquisition timing on cross-domain adaptation. Fig. 25(a) shows Sentinel-2 imagery of two locations in J^S and J^W, highlighting seasonal differences—dense vegetation in J^S and sparse cover in J^W; and the results in Fig. 25(c) display UDA-FT can accurately capture individual agricultural field parcel despite seasonal variations. The model exhibits adaptability across both the summer vegetation growing season and the winter sparse vegetation period (bare soil phase). Notably, a model trained on source domain data from the summer season effectively adapts to winter conditions as well.

As shown in Table 13, quantitative analysis further confirms this

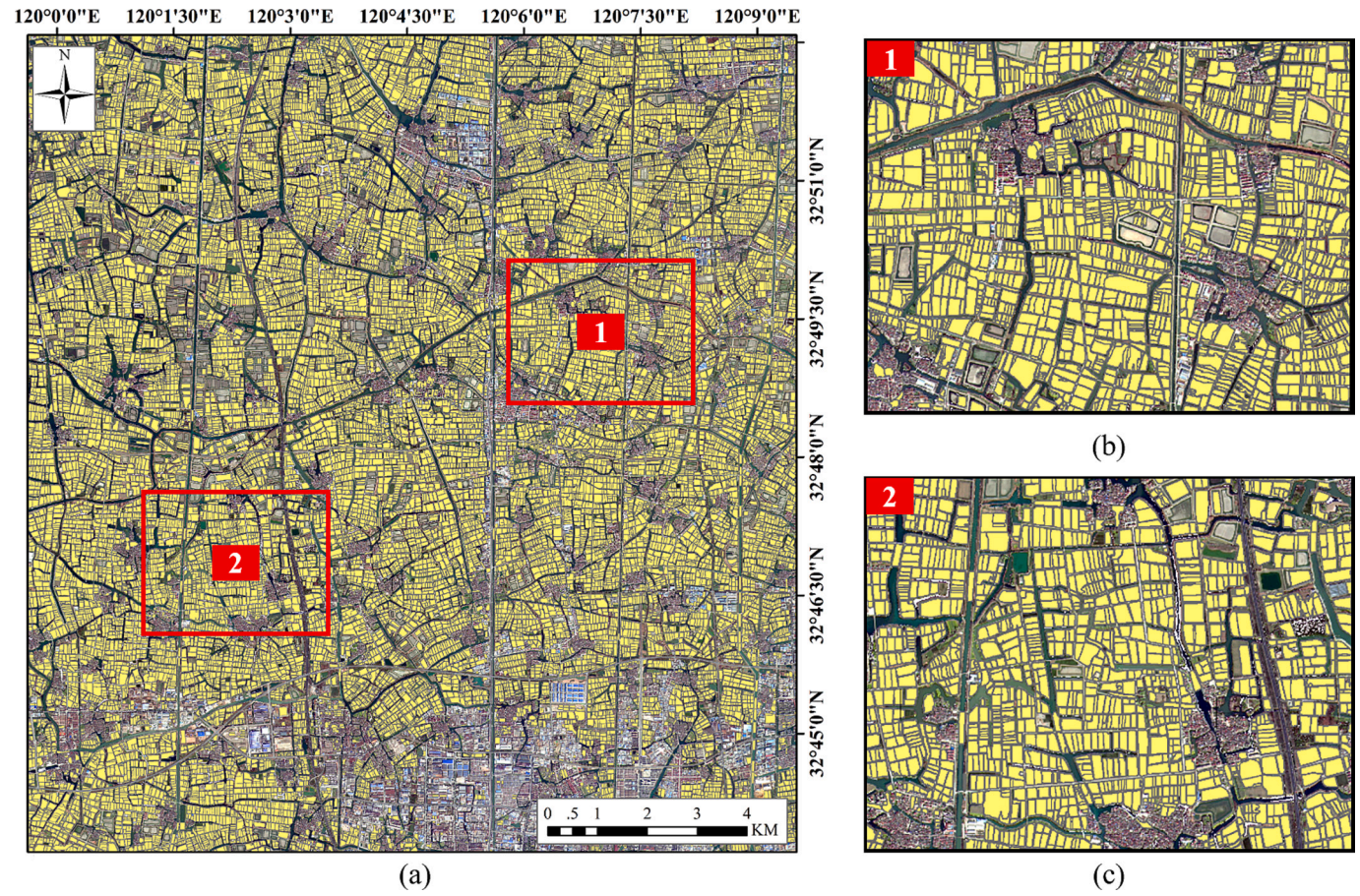

Fig. 19. Predicted distribution map of the agricultural field parcels in with GF-2 imagery. (a) Global visualization of study area in China. (b) and (c) Examples of local extraction results in (a).

Table 11
Quantitative evaluation results of the large-scale experiments across two distinct regions.

Metric	AP	AP_{50}	Precision	Recall	F1	S^{under}	S^{over}
Denmark-Japan	0.276	0.432	0.696	0.681	0.688	0.157	0.242
Esri Imagery-GF-2	0.294	0.460	0.782	0.735	0.758	0.125	0.197

impact. Direct transfer (source-only) from D^S to J^S achieves an F1-score of 0.072, while transfer to J^W drops to 0.054, a 25 % gap, due to increased domain shift in J^W. After applying UDA-FT, the F1-scores improve to 0.272 for $D^S \rightarrow J^S$ and 0.270 for $D^S \rightarrow J^W$, narrowing the gap to just 0.2 %, showcasing UDA-FT's ability to mitigate temporal domain shifts. This improvement is supported by the PCA-based distribution analysis in Fig. 24. Fig. 24(a) shows the feature distributions of D^S, J^S, and J^W before adaptation exhibit significant separation. D^S and J^S, both from the growing season, show relatively closer clustering due to similar vegetative states, while J^W, from the non-growing season, is markedly distant, reflecting the pronounced domain shift caused by sparse vegetation and exposed soil. After applying UDA-FT's image and instance adaptation modules, the feature distributions of D^S, J^S, and J^W becoming more compact and overlapping, particularly between D^S and J^W, indicating that UDA-FT effectively reduces the temporal-induced domain gap.

Table 12
Ablation experiments.

Methods	Weak	Strong*	MIC	LM	CM	IA	CML	APD- > C 50	APP- > E 50
Source-only (baseline)	✓							0.187	0.152
	✓	✓						0.258	0.221
	✓	✓	✓					0.269	0.228
	✓	✓		✓				0.273	0.234
	✓	✓		✓	✓			0.300	0.256
	✓	✓		✓	✓	✓		0.322	0.272
	✓	✓		✓	✓	✓	✓	0.349	0.297

Weak represents weak data augmentation, including random scaling and random flipping. Strong* represents strong data augmentation, including color jitter and random erase. MIC represents the SOTA data augmentation method in (Hoyer et al., 2023). LM represents the proposed LeakyMask method. CM represents the proposed CutMatch method. IA stands for image adaptation and instance adaptation, and CML stands for consistency mutual learning. APD- > C 50represents the metric AP_{50} in Denmark-Cambodia dataset, and APP- > E 50 represents the metric AP_{50} in Plant-Esri imagery dataset.

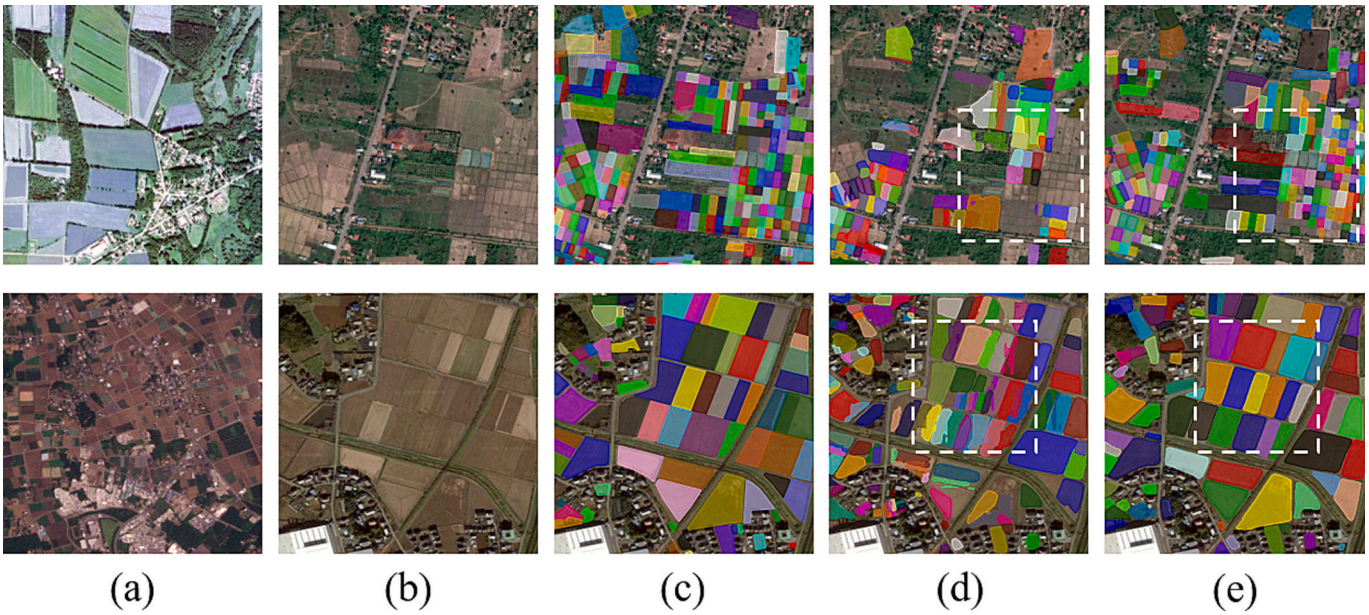

Fig. 20. Visualization results of the CutMatch method ablation experiment. (a) Source domain imagery. (b) Target domain imagery. (c) Annotations. (d) Results without CutMatch. (e) Results with CutMatch.

Table 13
Quantitative evaluation results of different seasons on the Japan Target Domain.

Methods	$D^S \rightarrow J^S$	$D^S \rightarrow J^W$	J^S Oracle	J^W Oracle	J^S source	J^W source
AP	0.105	0.104	0.193	0.189	0.022	0.014
AP_{50}	0.190	0.181	0.272	0.267	0.045	0.026
F1	0.272	0.270	0.361	0.355	0.072	0.054

6.3.2. Band combination

In our experiments, we utilized RGB bands to capture the visual characteristics of agricultural field parcels, such as color tones, morphology, texture, and size, which vary significantly across domains. RGB bands provide a straightforward representation of these features, making them suitable for unsupervised domain adaptation tasks where UDA-FT learns domain-invariant features based on visual appearance. To explore the impact of additional spectral information, we compared RGB with RGBN bands (including near-infrared, NIR) using Sentinel-2 imagery, consistent with the data in Section 6.3.1. The experiment involved multiple images from the Japan target domain, captured in July (summer growing season), with Denmark as the source domain. RGBN inputs were constructed by adding the NIR band, which enhances vegetation contrast, and false-color composites (NIR, red, green) were generated to highlight vegetation (Fig. 26b).

Table 14 shows that UDA-FT with RGBN bands achieves an AP of 0.128, AP_{50} of 0.245, and F1-score of 0.331, outperforming RGB by an improvement of 22.9 %, 28.9 %, and 21.7 %, respectively, confirming NIR's role in improving segmentation by better distinguishing vegetation. However, RGBN UDA-FT lags behind the RGBN Oracle (AP: 0.227, AP_{50}: 0.319, F1-score: 0.410), indicating spectral domain discrepancies as a challenge. Source-only models perform poorly (RGB Source: AP 0.022, AP_{50} 0.045, F1-score 0.072; RGBN Source: AP 0.031, AP_{50} 0.059, F1-score 0.084), underscoring UDA-FT's adaptation effectiveness. Fig. 26 shows RGBN predictions (column e) improve boundary delineation in dense regions (e.g., second and third rows), capturing more small-scale parcels with fewer over-segmentation errors than RGB (column d). However, in complex areas (e.g., second row), RGBN

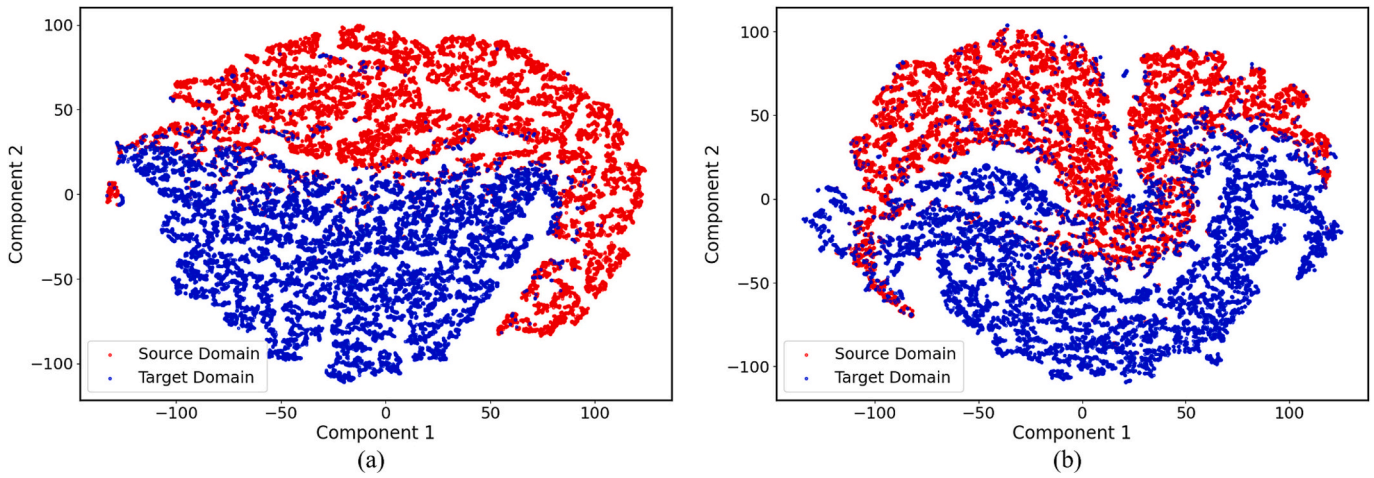

Fig. 21. The distribution of the different location and different sensor type by T-SNE algorithm. Red points represent the source domain and blue points represent the target domain. (a) is the distribution of the cross-scene dataset, (b) is the distribution of the cross-sensor dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

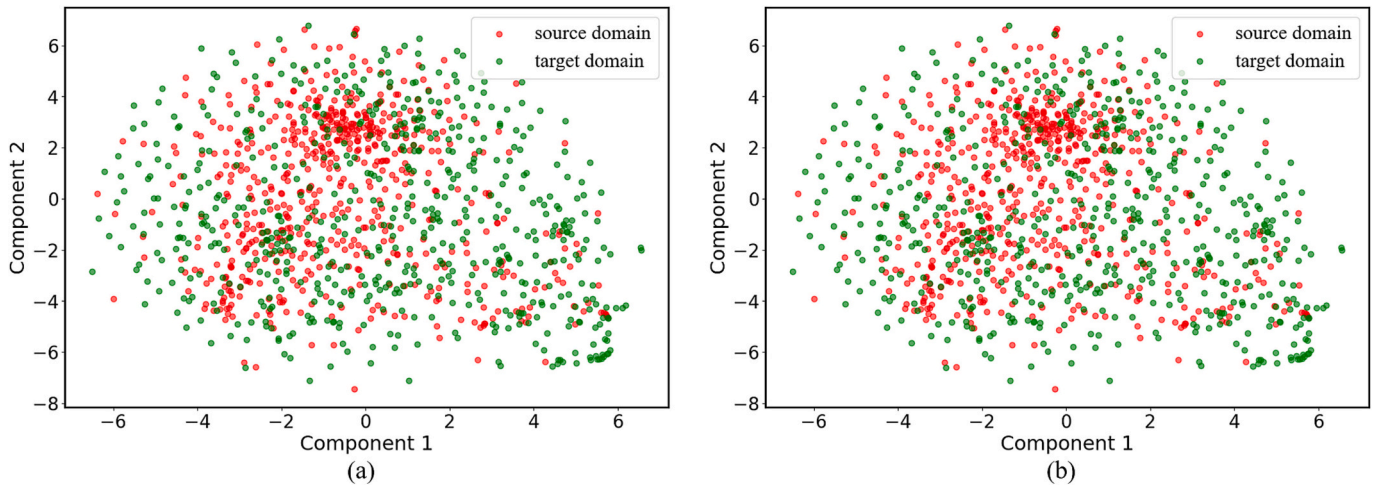

Fig. 22. The distribution of the cross-scene dataset and cross-sensor dataset after image adaptation and instance adaptation modules by T-SNE algorithm, based on the features from the final layer of the backbone network. Red points represent the source domain and green points represent the target domain. (a) is the distribution of the cross-scene dataset, (b) is the distribution of the cross-sensor dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

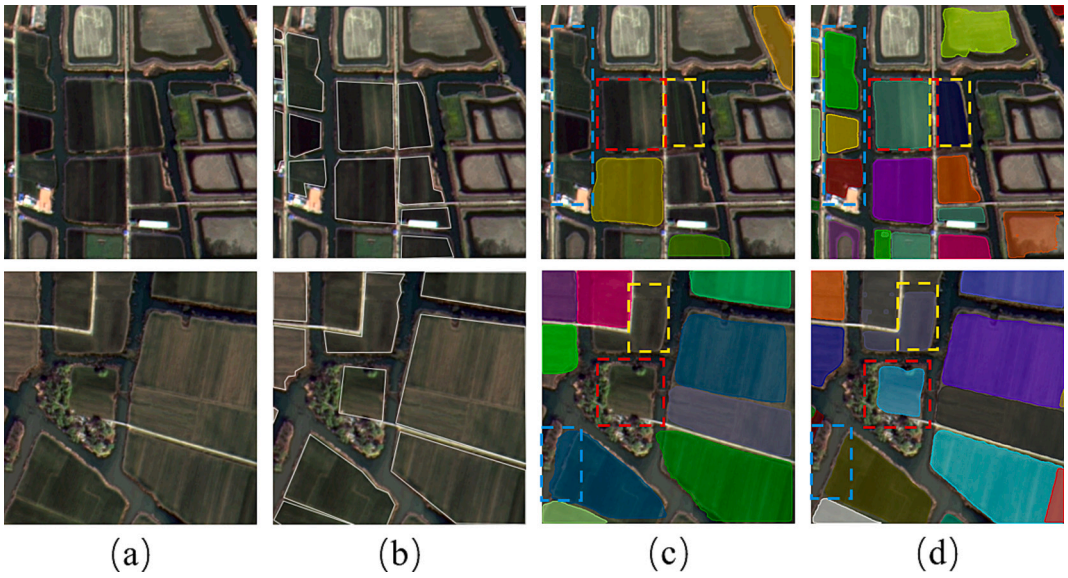

Fig. 23. Visualization results of hard pseudo-labels and soft pseudo-labels: (a) original image, (b) crop field annotations, (c) results of hard pseudo-labels, (d) visualization results of soft pseudo-labels.

introduces noise, misclassifying non-agricultural regions, likely due to spectral mismatches in NIR between domains. The larger performance gap between RGBN UDA-FT and RGBN Oracle (AP: 0.128 vs. 0.227) compared to RGB (AP: 0.105 vs. 0.193) supports this. While RGB-based predictions may lack precision, they offer greater robustness to spectral variations. This trade-off suggests that while RGBN enhances segmentation in vegetation-rich areas, future work should explore band-specific alignment, such as NIR-focused adversarial learning, to better leverage multispectral data.

6.4. *Potential applications, computing efficiency and limitations*

Our proposed UDA-FT framework directly addresses the challenge of the substantial annotation burden of manually digitizing field boundaries across diverse regions and sensors, thereby unlocking several significant applications. For data-scarce regions, particularly in developing countries, UDA-FT provides a pathway to generate accurate, fine-grained crop field maps where creating labeled datasets from scratch is often impractical. These maps can serve as a foundational data layer for developing field-level crop management systems, guiding precision agriculture, and supporting data-driven insurance services. Moreover, the UDA-FT framework could be a key component in establishing a dynamic, digitized global land parcel system by leveraging existing well-annotated regions as source domains to map vast, unlabeled parts of the world. Additionally, our method can be transferred to other domains with scarce labels, such as building extraction, road extraction, and other object detection and classification tasks.

While the proposed two-stage training paradigm is computationally intensive, it is crucial to distinguish between the training and inference phases. The complex teacher-student architecture and mutual learning are employed only during model development. The final deployed model for practical application is a standard Mask R-CNN network with 63 M parameters and achieves an inference speed of 13.5 FPS. Therefore, its parameter counts and runtime efficiency during inference are comparable to other state-of-the-art instance segmentation models, without the significant overhead of the UDA training framework. This

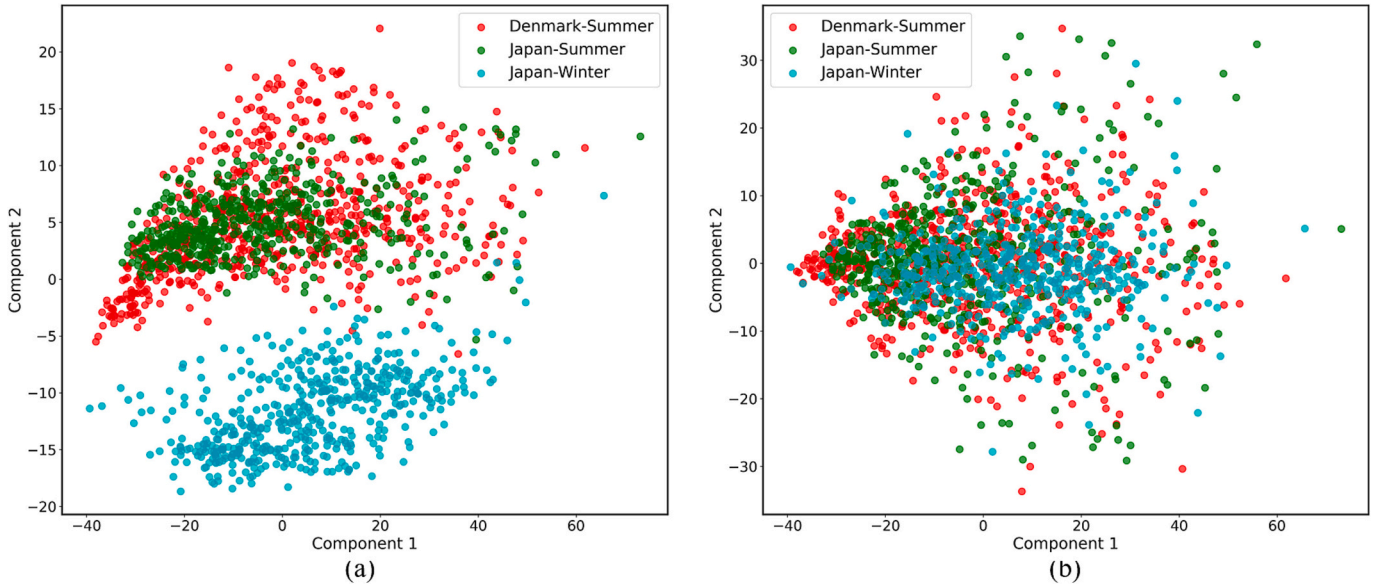

Fig. 24. Cross-scene Distribution Visualization of Seasonal Data Using PCA Components. (a) Distribution of Denmark-Summer, Japan-Summer, and Japan-Winter before domain adaptation. (b) Distribution of Denmark-Summer, Japan-Summer, and Japan-Winter after domain adaptation.

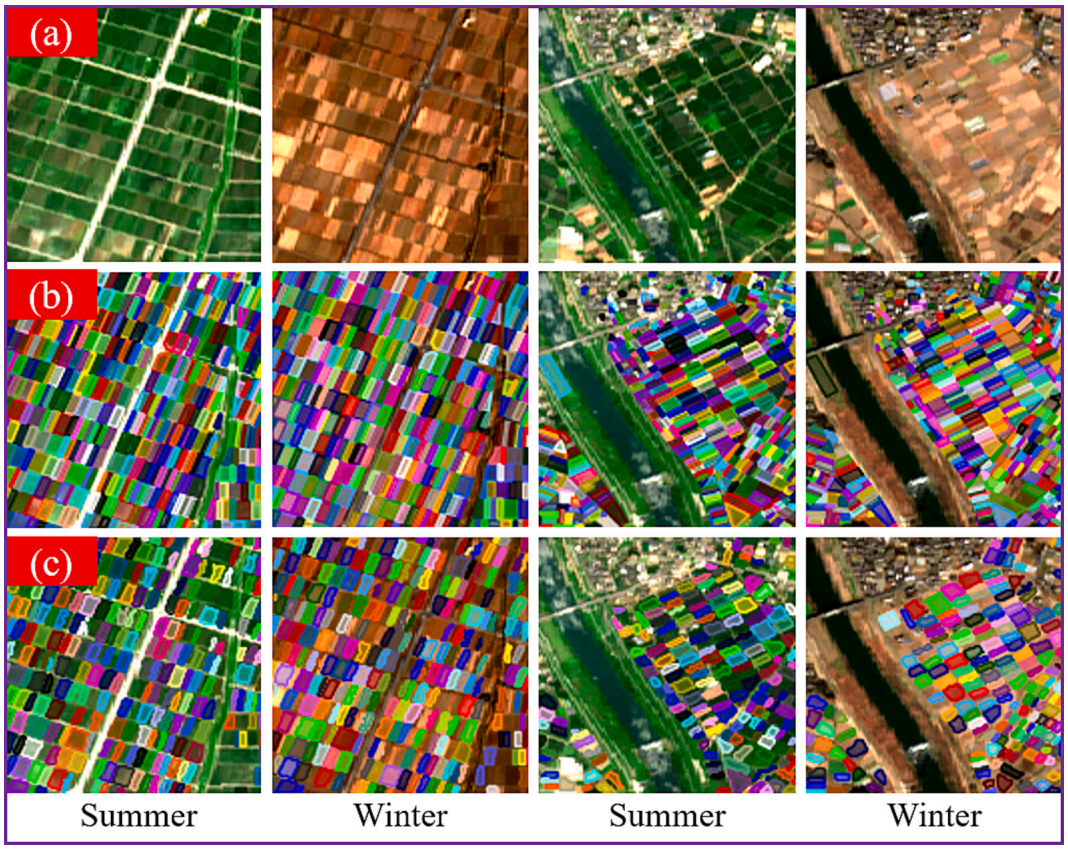

Fig. 25. Visualization results of the different seasons in Japan. (a) Sentinel-2 imagery of two locations in Japan, with the left two columns showing the same location in summer (July, J^S) and winter (December, J^W), and the right two columns showing another location in summer and winter. (b) Corresponding annotations. (c) UDA-FT's results.

makes it feasible for large-scale deployment where inference speed is a key consideration.

Despite its strong performance, our framework has several limitations that offer avenues for future research. Firstly, model performance in the target domain is inherently dependent on the quality and relevance of the source domain. A significant morphological discrepancy between source and target parcels, such as adapting from small, square fields to a domain with elongated, narrow fields, can hinder the model's ability to accurately segment complete shapes. Future research can attempt to address this issue at the algorithmic level or through data

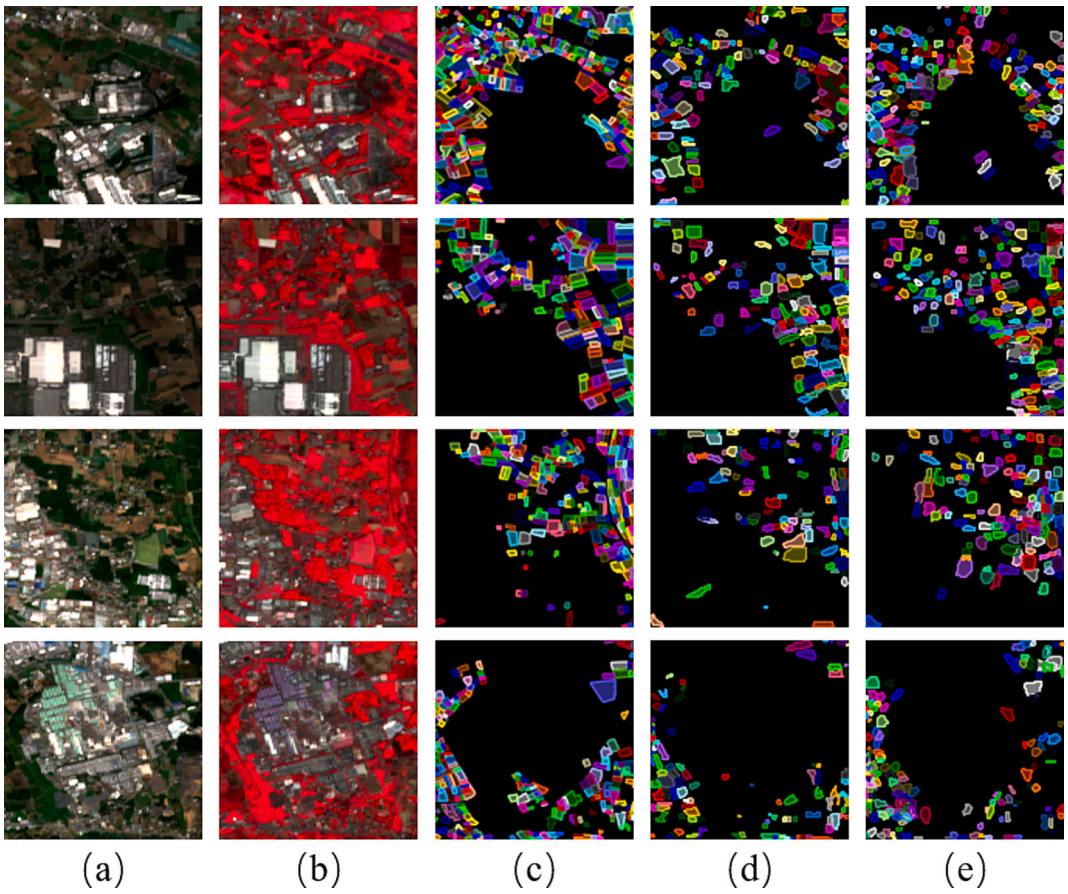

Fig. 26. Visualization results of UDA-FT with different band combinations in Japan. Here uses a black background to better visualize the extracted agricultural field parcel instances. (a) RGB imagery of four locations in Japan, captured in July during the summer growing season. (b) Corresponding false-color composite imagery. (c) Ground truth annotations. (d) UDA-FT's results using RGB bands. (e) UDA-FT's results using RGBN bands.

Table 14
Quantitative evaluation results of RGB and RGBN Bands on the Japan Target Domain.

Methods	RGB UDA-FT	RGBN UDA-FT	RGB Oracle	RGBN Oracle	RGB source	RGBN source
AP	0.105	0.128	0.193	0.227	0.022	0.031
AP_{50}	0.190	0.245	0.272	0.319	0.045	0.059
F1	0.272	0.331	0.361	0.410	0.072	0.084

augmentation. Furthermore, the current framework relies on a single source domain for training. Future work could focus on several key areas. A viable solution to the morphological discrepancy issue lies in broadening the source domain types and enhancing the morphological diversity of the training data. Investigating the effectiveness of incorporating multiple source domains could potentially enhance the model's generalization ability and robustness across various target domains. Additionally, improving computational efficiency for large-scale deployment through techniques like knowledge distillation or model pruning remains a valuable direction.

7. Conclusions

This study presents a novel unsupervised domain adaptation approach for agricultural crop field instance segmentation, addressing the critical challenge of adapting models across diverse scenes or sensor types without the need for target domain labels. Our proposed method, UDA-FT, successfully bridges the gap between source and target domains, demonstrating significant improvements in segmentation accuracy compared to source-only model. Quantitative evaluations demonstrate substantial improvements across multiple metrics for both the cross-scene and cross-sensor case studies. In the cross-scene case, our method achieves notable increases in AP_{50} and F1-score compared to the source-only model across all four regions: Japan (AP_{50}: +0.113, F1-score: +0.442), China (AP_{50}: +0.141, F1-score: +0.519), Cambodia (AP_{50}: +0.331, F1-score: +0.579) and Vietnam (AP_{50}: +0.322, F1-score: +0.550), highlighting its effectiveness in adapting to different geographical contexts. In the cross-sensor dataset, our method shows significant improvements across all four sensors combinations: JiLin-1 to Esri Imagery (AP_{50}: +0.196, F1-score: +0.553), Esri Imagery to JiLin-1 (AP_{50}: +0.166, F1-score: +0.316), PlanetScope to Esri Imagery (AP_{50}: +0.147, F1-score: +0.342), Esri Imagery to PlanetScope (AP_{50}: +0.141, F1-score: +0.438), underscoring its adaptability to various sensor characteristics. Qualitative analysis through visualizations further demonstrates the superiority of UDA-FT over the source-only method, particularly in preserving fine-grained details and accurately delineating small-scale agricultural field parcels. The effectiveness of UDA-FT can be attributed to its innovative integration of strong data augmentation, adversarial learning techniques and consistency mutual learning strategies. This approach enables the model to learn domain-invariant features while maintaining robustness to the inherent variabilities in satellite imagery, including differences in resolution, field sizes, geographical regions, color, and imaging angles. In conclusion, this research contributes a valuable tool to the field of agricultural remote sensing, offering a practical solution to the pervasive problem of domain shift in satellite imagery analysis. By enabling more accurate and generalizable field instance segmentation across diverse regions and

sensor types, our method has the potential to significantly promote global agricultural monitoring, resource management, and food security efforts.

CRediT authorship contribution statement

Ren Wei: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lin Yang:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Xiang Li:** Writing – review & editing, Visualization. **Chenxu Zhu:** Validation. **Lei Zhang:** Writing – review & editing. **Jie Wang:** Validation. **Jie Liu:** Writing – review & editing. **Liming Zhu:** Writing – review & editing. **Chenghu Zhou:** Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is supported by the National Key Research and Development Program Plan (Grant No. 2022YFC3800802), the National Natural Science Foundation of China (Project No. 42471468) and the Fundamental Research Funds for the Central Universities (0209-14380115).

Data availability

Data will be made available on request.

References

Atzberger, C., 2013. Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs. Remote Sens. 5, 949–981. https://doi.org/10.3390/rs5020949.

Avbelj, J., Muller, R., Bamler, R., 2015. A metric for polygon comparison and building extraction evaluation. IEEE Geosci. Remote Sens. Lett. 12, 170–174. https://doi.org/10.1109/lgrs.2014.2330695.

Biasetton, M., Michieli, U., Agresti, G., Zanuttigh, P., 2019. Unsupervised Domain Adaptation for Semantic Segmentation of Urban Scenes. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 0–0).

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. Hybrid Task Cascade for Instance Segmentation. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974–4983.

Chen, Y., Song, S., Li, S., Wu, C., 2020. A graph embedding framework for maximum mean discrepancy-based domain adaptation algorithms. IEEE Trans. Image Process. 29, 199–213. https://doi.org/10.1109/TIP.2019.2928630.

Chen, X., Pan, S., Chong, Y., 2022a. Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2022.3200246.

Chen, J., Zhu, J., Guo, Y., Sun, G., Zhang, Y., Deng, M., 2022b. Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2021.3140108.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023a. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogramm. Remote Sens. 195, 129–152. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Chen, F., Zhao, H., Roberts, D., Van De Voorde, T., Batelaan, O., Fan, T., Xu, W., 2023b. Mapping center pivot irrigation systems in global arid regions using instance segmentation and analyzing their spatial relationship with freshwater resources. Remote Sens. Environ. 297, 113760. https://doi.org/10.1016/j.rse.2023.113760.

Cheng, T., Ji, X., Yang, G., Zheng, H., Ma, J., Yao, X., Zhu, Y., Cao, W., 2020. DESTIN: a new method for delineating the boundaries of crop fields by fusing spatial and temporal information from WorldView and planet satellite imagery. Comput. Electron. Agric. 178, 105787. https://doi.org/10.1016/j.compag.2020.105787.

Cho, H., Nishimura, K., Watanabe, K., Bise, R., 2022. Effective pseudo-labeling based on heatmap for unsupervised domain adaptation in cell detection. Med. Image Anal. 79, 102436. https://doi.org/10.1016/j.media.2022.102436.

Corgne, S., Hubert-Moy, L., Betbeder, J., 2016. Monitoring of agricultural landscapes using remote sensing data. In: Land Surface Remote Sensing in Agriculture and Forest. Elsevier, pp. 221–247. https://doi.org/10.1016/B978-1-78548-103-1.50006-6.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.

Duveiller, G., 2010. A conceptual framework to define the spatial resolution requirements for agricultural monitoring using remote sensing. Remote Sens. Environ. 114, 2637–2650.

Ganin, Y., Lempitsky, V., 2025. Unsupervised Domain Adaptation by Backpropagation.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 1–35.

Garnot, V.S.F., Landrieu, L., 2021. Panoptic Segmentation of Satellite Image Time Series With Convolutional Temporal Attention Networks. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4872–4881.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63, 139–144. https://doi.org/10.1145/3422622.

Graesser, J., Ramankutty, N., 2017. Detection of cropland field parcels from Landsat imagery. Remote Sens. Environ. 201, 165–180. https://doi.org/10.1016/j.rse.2017.08.027.

Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y., 2022. Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Trans. Multimed. 24, 2502–2514. https://doi.org/10.1109/TMM.2021.3082687.

Guizilini, V., Li, J., Ambruș, R., Gaidon, A., 2021. Geometric Unsupervised Domain Adaptation for Semantic Segmentation. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8537–8547.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

Hoyer, L., Dai, D., Wang, H., Van Gool, L., 2023. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11721–11732.

Hu, S., Liao, Z., Xia, Y., 2022. Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer Nature, Switzerland, Cham, pp. 650–659. https://doi.org/10.1007/978-3-031-16449-1_62.

Huang, Y., Chen, Z., Yu, T., Huang, X., Gu, X., 2018. Agricultural remote sensing big data: management and applications. J. Integr. Agric. 17, 1915–1931. https://doi.org/10.1016/S2095-3119(17)61859-8.

Huang, Y., Peng, J., Chen, N., Sun, W., Du, Q., Ren, K., Huang, K., 2023. Cross-scene wetland mapping on hyperspectral remote sensing images using adversarial domain adaptation network. ISPRS J. Photogramm. Remote Sens. 203, 37–54. https://doi.org/10.1016/j.isprsjprs.2023.07.009.

Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. Deep Learn. RS Data 145, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014.

Lee, T., Lee, B.-U., Shin, I., Choe, J., Shin, U., Kweon, I.S., Yoon, K.-J., 2022. UDA-COPE: Unsupervised Domain Adaptation for Category-Level Object Pose Estimation. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14891–14900.

Li Dai, X., Ma, C.-Y., Liu, Y.-C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P., 2022. Cross-Domain Adaptive Teacher for Object Detection.

Li, R., Zheng, S., Duan, C., Wang, L., Zhang, C., 2022. Land cover classification from remote sensing images based on multi-scale fully convolutional network. Geo-spat. Inf. Sci. 25, 278–294. https://doi.org/10.1080/10095020.2021.2017237.

Li, M., Long, J., Stein, A., Wang, X., 2023. Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images. ISPRS J. Photogramm. Remote Sens. 200, 24–40. https://doi.org/10.1016/j.isprsjprs.2023.04.019.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.

Litrico, M., Del Bue, A., Morerio, P., 2023. Guiding Pseudo-Labels With Uncertainty Estimation for Source-Free Unsupervised Domain Adaptation. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7640–7650.

Liu, D., Zhang, D., Song, Y., Zhang, F., O'Donnell, L., Huang, H., Chen, M., Cai, W., 2020. Unsupervised Instance Segmentation in Microscopy Images via Panoptic Domain

Adaptation and Task Re-Weighting. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4243–4252.
Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.-W., Woo, J., 2022. Deep unsupervised domain adaptation: a review of recent advances and perspectives. SIP 11. https://doi.org/10.1561/116.00000192.
Long, J., Li, M., Wang, X., Stein, A., 2022. Delineation of agricultural fields using multi-task BsiNet from high-resolution satellite images. Int. J. Appl. Earth Obs. Geoinf. 112, 102871. https://doi.org/10.1016/j.jag.2022.102871.
Long, J., Shelhamer, E., Darrell, T., 2025. Fully Convolutional Networks for Semantic Segmentation.
Luo, S., Ma, L., Yang, X., Luo, D., Du, Q., 2024. Self-training-based unsupervised domain adaptation for object detection in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 62, 1–21. https://doi.org/10.1109/TGRS.2024.3457789.
Mao, L., Zheng, Z., Meng, X., Zhou, Y., Zhao, P., Yang, Z., Long, Y., 2022. Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images. Landsc. Urban Plan. 222, 104384. https://doi.org/10.1016/j.landurbplan.2022.104384.
Mei, W., Wang, H., Fouhey, D., Zhou, W., Hinks, I., Gray, J.M., Van Berkel, D., Jain, M., 2022. Using deep learning and very-high-resolution imagery to map smallholder field boundaries. Remote Sens. 14, 3046. https://doi.org/10.3390/rs14133046.
Oza, P., Sindagi, V.A., VS, V., Patel, V.M., 2024. Unsupervised domain adaptation of object detectors: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 46, 4018–4040. https://doi.org/10.1109/TPAMI.2022.3217046.
Pande, C.B., Moharir, K.N., 2023. Application of hyperspectral remote sensing role in precision farming and sustainable agriculture under climate change: A review. In: Pande, C.B., Moharir, K.N., Singh, S.K., Pham, Q.B., Elbeltagi, A. (Eds.), Climate Change Impacts on Natural Resources, Ecosystems and Agricultural Systems. Springer International Publishing, Cham, pp. 503–520. https://doi.org/10.1007/978-3-031-19059-9_21.
Papadomanolaki, M., Vakalopoulou, M., Karantzalos, K., 2021. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. IEEE Trans. Geosci. Remote Sens. 59, 7651–7668. https://doi.org/10.1109/TGRS.2021.3055584.
Peng, D., Guan, H., Zang, Y., Bruzzone, L., 2022. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2021.3093004.
Persello, C., Bruzzone, L., 2010. A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48, 1232–1244. https://doi.org/10.1109/TGRS.2009.2029570.
Persello, C., Grift, J., Fan, X., Paris, C., Hänsch, R., Koeva, M., Nelson, A., 2023. AI4SmallFarms: a dataset for crop field delineation in southeast Asian smallholder farms. IEEE Geosci. Remote Sens. Lett. 20, 1–5. https://doi.org/10.1109/LGRS.2023.3323095.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.
Rydberg, A., Borgefors, G., 2001. Integrated method for boundary delineation of agricultural fields in multispectral satellite images. IEEE Trans. Geosci. Remote Sens. 39, 2514–2520. https://doi.org/10.1109/36.964989.
Suzuki, S., 1985. Topological Structural Analysis of Digitized Binary Images by Border Following.
Triantafyllou, A., Sarigiannidis, P., Bibi, S., 2019. Precision agriculture: a remote sensing monitoring system architecture. Information 10, 348. https://doi.org/10.3390/info10110348.
Wachowiak, M.P., Walters, D.F., Kovacs, J.M., Wachowiak-Smolíková, R., James, A.L., 2017. Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. Comput. Electron. Agric. 143, 149–164. https://doi.org/10.1016/j.compag.2017.09.035.
Wei, L.I.U., Zhifeng, W.U., Jiancheng, L.U.O., Yingwei, S.U.N., Tianjun, W.U., Nan, Z., Xiaodong, H.U., Lingyu, W., Zhongfa, Z., 2021. A divided and stratified extraction method of high-resolution remote sensing information for cropland in hilly and mountainous areas based on deep learning. Acta Geodaetica et Cartographica Sinica 50, 105. https://doi.org/10.11947/j.AGCS.2021.20190448.
Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: a meta-review. Remote Sens. Environ. 236, 111402. https://doi.org/10.1016/j.rse.2019.111402.
Xiao, R., Liu, Z., Wu, B., 2021. Teacher-student competition for unsupervised domain adaptation, in: 2020 25th international conference on pattern recognition (ICPR). In: Presented at the 2020 25th International Conference on Pattern Recognition (ICPR), pp. 8291–8298. https://doi.org/10.1109/ICPR48806.2021.9413219.
Xie, D., Xu, H., Xiong, X., Liu, M., Hu, H., Xiong, M., Liu, L., 2023. Cropland extraction in southern China from very high-resolution images based on deep learning. Remote Sens. 15, 2231. https://doi.org/10.3390/rs15092231.
Xu, L., Ming, D., Zhou, W., Bao, H., Chen, Y., Ling, X., 2019. Farmland extraction from high spatial resolution remote sensing images based on stratified scale pre-estimation. Remote Sens. 11, 108. https://doi.org/10.3390/rs11020108.
Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z., 2021. End-to-end semi-supervised object detection with soft teacher. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada. https://doi.org/10.1109/iccv48922.2021.00305.
Xu, L., Yang, P., Yu, J., Peng, F., Xu, J., Song, S., Wu, Y., 2023. Extraction of cropland field parcels with high resolution remote sensing using multi-task learning. Eur. J. Remote Sens. 56, 2181874. https://doi.org/10.1080/22797254.2023.2181874.
Xu, F., Yao, X., Zhang, K., Yang, H., Feng, Q., Li, Y., Yan, S., Gao, B., Li, S., Yang, J., Zhang, C., Lv, Y., Zhu, D., Ye, S., 2024. Deep learning in cropland field identification: a review. Comput. Electron. Agric. 222, 109042. https://doi.org/10.1016/j.compag.2024.109042.
Yan, L., Roy, D.P., 2014. Automated crop field extraction from multi-temporal Web Enabled Landsat Data. Remote Sens. Environ. 144, 42–64. https://doi.org/10.1016/j.rse.2014.01.006.
Yang, R., Tian, T., Tian, J., 2025. Versatile teacher: a class-aware teacher–student framework for cross-domain adaptation. Pattern Recogn. 158, 111024. https://doi.org/10.1016/j.patcog.2024.111024.
Yu, F., Wang, D., Chen, Y., Karianakis, N., Shen, T., Yu, P., Lymberopoulos, D., Lu, S., Shi, W., Chen, X., 2022. SC-UDA: Style and Content Gaps Aware Unsupervised Domain Adaptation for Object Detection. Presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 382–391.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241, 111716. https://doi.org/10.1016/j.rse.2020.111716.
Yuan, R., Li, Q., Hou, J., Xu, J., Zhang, Y., Feng, R., Chen, H., 2024. Domain Adaptation Using Pseudo Labels for COVID-19 Detection. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5141–5148.
Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H., 2018. Deep mutual learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Salt Lake City, UT. https://doi.org/10.1109/cvpr.2018.00454.
Zhang, D., Pan, Y., Zhang, J., Hu, T., Zhao, J., Li, N., Chen, Q., 2020. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. Remote Sens. Environ. 247, 111912. https://doi.org/10.1016/j.rse.2020.111912.
Zhang, L., Lan, M., Zhang, J., Tao, D., 2022a. Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2021.3104032.
Zhang, J., Xu, S., Sun, J., Ou, D., Wu, X., Wang, M., 2022b. Unsupervised adversarial domain adaptation for agricultural land extraction of remote sensing images. Remote Sens. 14, 6298. https://doi.org/10.3390/rs14246298.
Zhao, Z., Liu, Y., Zhang, G., Tang, L., Hu, X., 2022. The winning solution to the iFLYTEK challenge 2021 cultivated land extraction from high-resolution remote sensing image, in: 2022 14th International Conference on Advanced Computational Intelligence (ICACI). pp. 376–380. Doi: https://doi.org/10.1109/ICACI55529.2022.9837765.
Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. Proceed. AAAI Conf. Artif. Intell. 34, 13001–13008. https://doi.org/10.1609/aaai.v34i07.7000.
Zhou, K., Ming, D., Lv, X., Fang, J., Wang, M., 2019. CNN-based land cover classification combining stratified segmentation and fusion of point cloud and very high-spatial resolution remote sensing image data. Remote Sens. 11, 2065. https://doi.org/10.3390/rs11172065.
Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J., 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. Presented at the Proceedings of the European Conference on Computer Vision (ECCV), pp. 289–305.