GTPBD: A Fine-Grained Global Terraced Parcel and Boundary Dataset

Zhiwei Zhang[1], **Zi Ye**[1], **Yibin Wen**[1], **Shuai Yuan**[2],
Haohuan Fu[3,4], **Jianxi Huang**[5,6], **Juepeng Zheng**[1,4,*]
[1]Sun Yat-Sen University [2]The University of Hong Kong [3]Tsinghua University
[4]National Supercomputing Center in Shenzhen
[5]Southwest Jiaotong University [6]China Agricultural University

Abstract

Agricultural parcels serve as basic units for conducting agricultural practices and applications, which is vital for land ownership registration, food security assessment, soil erosion monitoring, etc. However, existing agriculture parcel extraction studies only focus on mid-resolution mapping or regular plain farmlands while lacking representation of complex terraced terrains due to the demands of precision agriculture. In this paper, we introduce a more fine-grained terraced parcel dataset named **GTPBD** (**G**lobal **T**erraced **P**arcel and **B**oundary **D**ataset), which is the first fine-grained dataset covering major worldwide terraced regions with more than 200,000 complex terraced parcels with manually annotation. GTPBD comprises 47,537 high-resolution images with three-level labels, including pixel-level boundary labels, mask labels, and parcel labels. It covers seven major geographic zones in China and transcontinental climatic regions around the world. Compared to the existing datasets, the GTPBD dataset brings considerable challenges due to the: (1) terrain diversity; (2) complex and irregular parcel objects; and (3) multiple domain styles. Our proposed GTPBD dataset is suitable for four different tasks, including semantic segmentation, edge detection, terraced parcel extraction and unsupervised domain adaptation (UDA) tasks. Accordingly, we benchmark the GTPBD dataset on eight semantic segmentation methods, four edge extraction methods, three parcel extraction methods and five UDA methods, along with a multi-dimensional evaluation framework integrating pixel-level and object-level metrics. GTPBD fills a critical gap in terraced remote sensing research, providing a basic infrastructure for fine-grained agricultural terrain analysis and cross-scenario knowledge transfer. The code and data are available at `https://github.com/Z-ZW-WXQ/GTPBD/`.

1 Introduction

In ancient China, terraced fields were given the reputation of "Dragon Spines" ("龙脊" in Chinese), which is not only a poetic description of their winding and magnificent forms, but also a high praise in agricultural civilization. Nowadays, approximately 120 million acres of terraced fields worldwide still support more than 500 million mountainous populations [36], reducing soil erosion by more than 23.7 billion tons annually, which indicates significant ecological and economic values [12]. Therefore, a better and fine-grained understanding of terraced area is needed for a sustainable agricultural ecosystem, which is vital for land ownership registration, food security assessment, etc [30, 2].

With the development of artificial intelligence techniques, terrace mapping have progressed significantly, from pixel-based and object-oriented methods to machine learning and deep learning-based

[*] Corresponding author. Email: `zhengjp8@mail.sysu.edu.cn`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

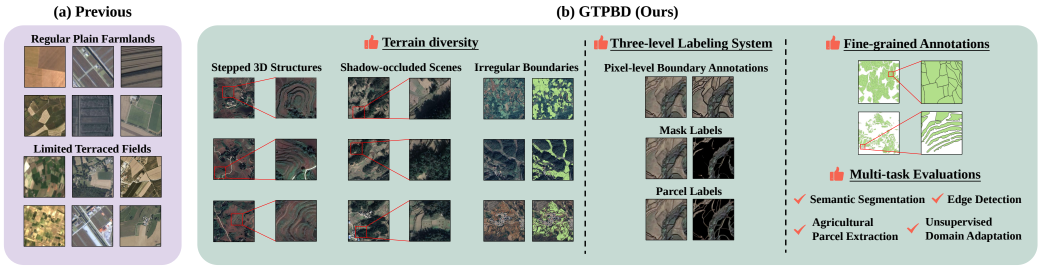


Figure 1: Comparisons between previous dataset that focus on regular plain farmlands with limited terraced fields, and our GTPBD that offers more challenging and comprehensive evaluations.

methods from remote sensing images. Existing terrace mapping only focus on mid-resolution mapping [4, 21, 11] using Sentinel-1/2, Landsat, MODIS, etc. Meanwhile, some researchers have collected public large-scale agricultural parcels and boundaries datasets from high-resolution remote sensing images. To improve the agricultural parcel extraction, advanced deep learning methods have been used in parcel-level extraction and vectorization [44, 26, 22, 19].

Despite the impressive capabilities demonstrated by deep learning for agricultural parcel extraction, fine-grained terraced parcel and boundary benchmarks for agricultural remote sensing remain scarce due to they are **extremely complex and unexpectedly irregular**. Existing agricultural parcel dataset primarily focus on scenarios of regular plain farmlands (See Fig. 1 (a)). Datasets designed for agricultural parcel dataset from high-resolution remote sensing images exhibit notable limitations, primarily in terms of **terrain diversity**, **annotation difficulty** and **multiple tasks**, as shown in Fig. 1.

To begin with, current parcel dataset suffer from regular plain farmlands and limited terraced fields [26, 10], limiting their ability to insufficient terrain diversity (See Fig. 1 (a)). To achieve a more precise agricultural mapping and yield estimation for global terraced field, it is imperative to develop a more fine-grained terraced dataset covering terrain diversity that contains stepped 3D structures, shadow-occluded scenes and irregular boundaries that combine three-dimensional topological structures (See Fig. 1 (b)). In addition, most of existing related datasets only provide binary classification mask labels [51, 10], failing to explicitly distinguish between two topological relationships of adjacent terraced ridges: shared edges (where adjacent parcels share the same boundary) and non-shared edges (where adjacent parcels have separate boundaries). It is highly demanded to collect a refined terraced parcel dataset with fine-grained boundaries. Furthermore, existing agricultural parcel dataset often rely on oversimplified task design such as binary classification or semantic segmentation [28, 39], and have not discussed the transferability of agricultural parcel extraction. In order to improve and benchmark their generalizability, appropriate dataset that includes multiple domain styles is required.

In response to the aforementioned issues, this paper proposes a more challenging terraced parcel dataset named **GTPBD** (**G**lobal **T**erraced **P**arcel and **B**oundary **D**ataset), which is the first fine-grained dataset covering major worldwide terraced regions with more than 200,000 complex terraced parcels with manually annotation.

In summary, the contributions of this work can be concluded as:

- We construct GTPBD which is the first global terraced dataset on parcel level, containing 47,537 images and collectively covering 885 km^2 of terraced farmland across varied terrains, with mountainous terraces accounting for over 80% of the dataset. The dataset covers seven major geographic zones in China and transcontinental climatic regions around the world.

- We propose GTPBD, which contains three-level labels for each pixel, including boundary labels, mask labels, and parcel labels. GTPBD is suitable for four different tasks, including semantic segmentation, edge detection, terraced parcel extraction and unsupervised domain adaptation (UDA) tasks with three different domains and six transfer tasks.

- We conduct a comprehensive evaluations of 8 semantic segmentation methods, 4 edge detection methods, 3 parcel extraction methods and 5 UDA methods, long with a multi-dimensional evaluation framework integrating pixel-level and object-level metrics. Experimental results reveal the limitations of current deep learning methods in terraced parcel extraction and its generalization capacity.

Table 1: Comparison between GTPBD and the main agricultural field datasets.

Dataset	Source	Images	Area (km^2)	Resolution (m)	Classes	Global Coverage	Task			
							SS	APE	ED	UDA
FHAPD [51]	GaoFen-1/2	68,982	$<$1000	1-2	2		✓	✓	✓	
GTM [21]	Sentinel-2	108,300	853,161	10	2	✓	✓			
FTW [18]	Sentinel-2	70,462	166,293	10	2,3	✓	✓	✓	✓	
AI4Boundaries (Sentinel-2) [10]	Sentinel-2	7,831	51,321	10	2		✓	✓	✓	
AI4Boundaries (Ortho) [10]	Aerial	7,598	1,992	1	2		✓	✓	✓	
PASTIS [13]	Sentinel-2	2,433	3,986	10	19		✓	✓	✓	
CropHarvest [35]	Sentinel-1/2	95,186	>5,000,000	10-60	348	✓	✓	✓	✓	
GFSAD30 [32]	Landsat-8	64,800	18,740,000	30	2	✓	✓			
GloCAB [14]	Sentinel-2	190,832	N/A	10	2		✓	✓	✓	
AI4SmallFarms [27]	Sentinel-2	62	1,550	10	2		✓	✓	✓	
India Smallholder Boundaries [39]	Airbus SPOT-6/7 & PlanetScope	10,000	30-50	1.5-4.8	2		✓	✓	✓	
GTPBD (Ours)	**GF-2 & Google Earth**	**47,537**	**885**	**0.5–0.7**	**3**	✓	✓	✓	✓	✓

• **SS**: Semantic Segmentation; **APE**: Agricultural Parcel Extraction; **ED**: Edge Detection; **UDA**: Unsupervised Domain Adaptation.

2 Related works

2.1 Agricultural parcel and boundary datasets

In recent years, considerable progress has been made in developing agricultural parcel and boundary datasets (see Table 1). Early datasets such as GFSAD30 [32] provide global coverage but at a coarse resolution of 30 m, making them unsuitable for fine-grained parcel delineation. More recent efforts like PASTIS [13] and FTW [18], built on Sentinel-2 imagery (10 m resolution), improved spatial granularity but still mainly represent regular, planar agricultural fields. Higher-resolution datasets such as FHAPD [51] and AI4Boundaries(Ortho) [10] offer imagery at 1 m resolution, enhancing spatial detail. However, these remain limited to structured flat farmland, with little focus on irregular terraced landscapes. Importantly, none of the existing datasets support unsupervised domain adaptation (UDA), despite its necessity for cross-region generalization.

To overcome these limitations, we introduce a more challenging terraced parcel dataset named **GTPBD**, specifically designed to address the gaps in existing agricultural datasets. GTPBD is distinguished by its global scope, encompassing diverse regions spanning tropical, subtropical, and temperate climatic zones. Furthermore, GTPBD supports multiple tasks, including semantic segmentation, agricultural parcel extraction, unsupervised domain adaptation, and edge detection, offering a unified benchmark for advancing the analysis of complex terraced agricultural landscapes.

2.2 Remote sensing tasks

Semantic Segmentation (SS) assigns a class label to every pixel in imagery. Early encoder–decoder CNNs such as U-Net [29], DeepLabV3 [6], and PSPNet [52], recover spatial detail via skip connections, atrous convolutions, and pyramid pooling. More recently, transformer-based frameworks including SegFormer [41] and Mask2Former [8], leverage self-attention to capture long-range dependencies and improve robustness to local appearance variations [8, 41, 5, 54]. Despite their advances, these models have been benchmarked almost on coarse-grained land-cover datasets [51, 1, 34], which focus on urban scenes or flat fields. To fill this gap, we evaluate all of the above architectures on GTPBD, which includes fine-grained annotations, to deliver the comprehensive semantic segmentation results for complex terraced agricultural landscapes.

Agricultural Parcel Extraction (APE) delineates farmland units for precision agriculture [48, 37]. Most CNN-based methods treated APE as single-task segmentation, using encoder–decoder backbones to predict parcel masks but often yielding unclosed or jagged boundaries in heterogeneous fields [25]. To address this, edge-aware multi-task architectures have emerged: SEANet [19] jointly predicts masks, distances, and edges to enforce closed parcels, REAUNet [22] integrates Sobel filters and attention for multi-scale edge enhancement, HBGNet [51] extracts low-level boundary features and high-level parcel semantics in parallel with a dual-branch framework. Despite these models perform well on planar benchmarks like FHAPD [51] and AI4Boundaries [10], their robustness to terraced ridge geometries is untested. We therefore benchmark them on GTPBD to provide systematic evaluation of APE methods in complex terraced agricultural landscapes.

Edge Detection (ED) precisely localizes pixel-level boundaries and facilitates high-fidelity terrace morphology analysis and erosion monitoring [3, 42]. We benchmark four state-of-the-art models

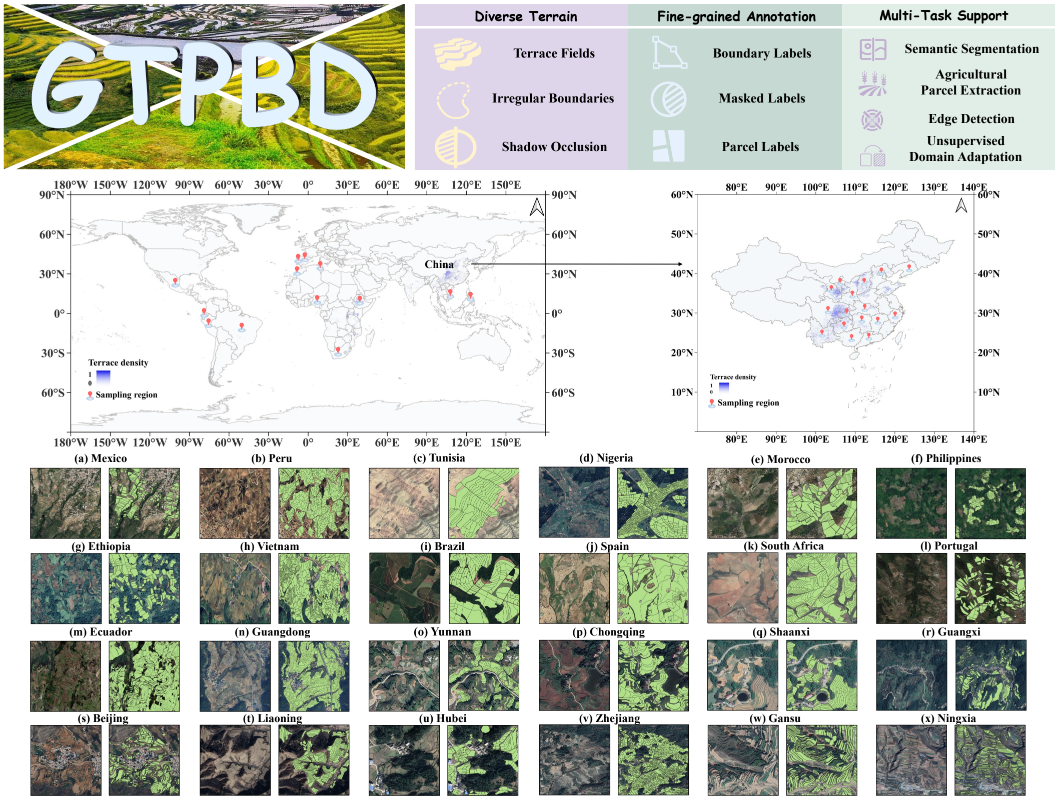


Figure 2: The characteristics of our proposed GTPBD with diverse terrain, fine-grained annotation and multi-task support. GTPBD covers seven major geographic zones in China and transcontinental climatic regions around the world.

on GTPBD: REAUNet-Sober [22], which embeds Sobel-style filters into a U-Net backbone; MuGE [55], employing multi-scale gating units to enhance edge features; PiDiNet [31], a lightweight pixel-difference network optimized for real-time inference; and UEAD [56], leveraging unsupervised edge-aware representation learning. Although these models have demonstrated strong performance on urban or synthetic benchmarks, their efficacy on complex terraced agricultural landscapes has not been evaluated. Our results on GTPBD's fine-grained boundary annotations therefore provide accurate assessment of edge detection performance in terraced farmland environments.

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge from a labeled source domain to an unlabeled target domain [33]. UDA methods for remote sensing can be broadly grouped into adversarial training and self-training approaches. Adversarial training such as Cycada [15] and CLAN [23] leverage discriminators to align feature distributions, they incur complex optimization overhead and may not directly address style discrepancies. Representative self-training frameworks include DAFormer [16], which integrates rare-class sampling and pseudo-label denoising. In addition, input-level adaptation methods like FDA [43] perform only low-frequency transfer in the Fourier domain, requiring no adversarial network or pseudo label. Although these UDA methods have been widely used in remote sensing community with some improvements, existing public remote sensing datasets have been limited by urban scenes [20, 7] or land cover mapping [38, 53]. To this end, the GTPBD dataset is proposed for a more challenging benchmark, promoting future research of cross-regional or temporal agricultural and terraced parcel extraction algorithms and its applications.

3 Dataset

3.1 Image collection and manual annotation

We collect our images in our dataset based on the Global Terrace Map (GTM) with 10-m resolution [21]. To ensure data quality and spatial coverage, we adopted the following screening strategies:

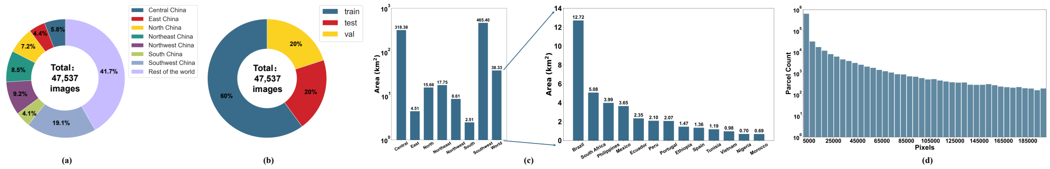


Figure 3: Statistics for the number and area of GTPBD dataset. (a) Distribution of images across different regions. (b) Distribution of images across different dataset splits. (c) Distribution of area across different regions. (d) Distribution of the parcel sizes (logarithmic scale in vertical axis).

Spatial Coverage and Resolution Following GTM [21], we prioritized regions with dense terrace distribution and complex topography, and further included representative terraces across diverse landforms—such as alpine terraces in southwestern China and the Sa Pa terraces in Vietnam—to assess model generalization across varied terrains. The dataset spans fourteen countries worldwide, including China, Vietnam, Tunisia, Ethiopia, Peru, and Mexico, as well as seven major geographic regions of China (Central, South, North, East, Northwest, Northeast, and Southwest). High-resolution images were collected from GF-2 and Google Earth with spatial resolutions ranging from 0.1 m to 1 m, covering different seasons and lighting conditions. All images were selected from cloud-free scenes acquired between 2021 and 2025 to ensure consistency and research suitability. Examples are shown in Fig.2, with additional samples in AppendixB.1.

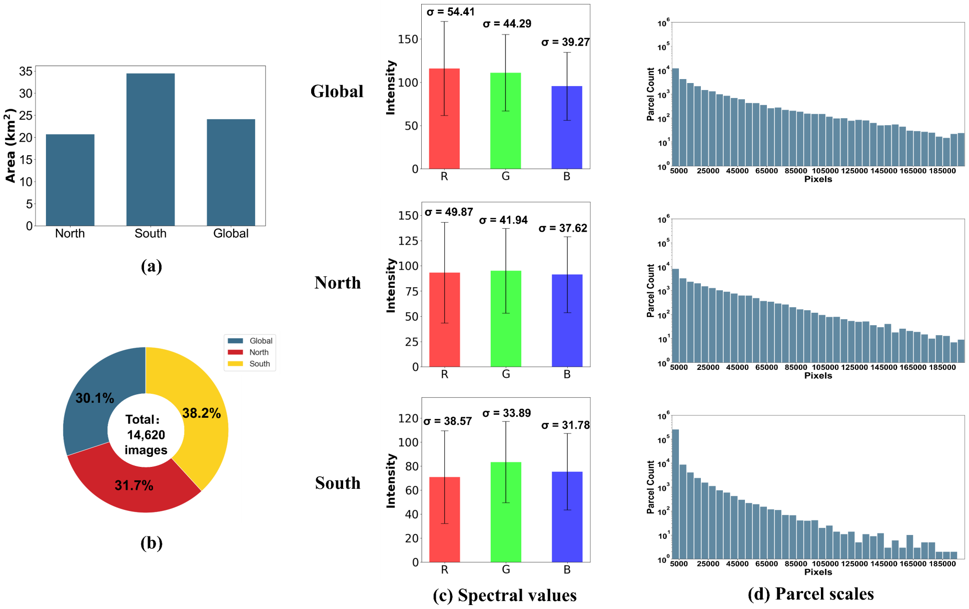
Parcel Annotation We construct the topological vectorization and annotations in terraced fields through QGIS. We recruit over 50 participants, comprising undergraduate and graduate students, to manually annotate the fine-grained terraced parcel and boundary with strict quality inspection. The specific implementation strategy is as follows: for areas with dense hierarchical structures and narrow boundaries (less than 0.5m), the common edge annotation method is adopted - prioritizing vectorization of large continuous parcels, and generating sub-parcel topologies through internal line feature cutting; For areas where the field ridge width is greater than or equal to 0.5 m, bilateral segmentation annotation is applied — independent vector boundaries are generated along both sides of the field ridge, and the ridge itself is removed to form a non-standard edge topology structure.

Three-Level Labels As for mask labels, we use GDAL's rasterize function for fully connected rasterization (all touched strategy), establish a binary semantic segmentation benchmark, and assign 1 to terraced areas and 0 to backgrounds. As for boundary labels, we use a 3×3 rectangular grid mask to perform a single morphological etching and construct a 3-pixel-wide edge detection dedicated label. As for parcel labels, we generate pure parcel regions through mask boundary using the XOR operation to achieve terraced parcel extraction tasks. Appendix B.2 displays the differences between previous dataset and our GTPBD.

Dataset Construction To address potential spatial correlation or leakage between patches, we first resampled all source images and corresponding labels to a unified resolution of 0.5–0.7 m. Images from the same geographic region share the same target resolution and were split into training (60%), validation (20%), and testing (20%) sets *before* cropping, ensuring spatial independence between subsets. Following common practice in high-resolution agricultural benchmarks such as AI4Boundaries[10], Agriculture-Vision [9], and FUSU [45], we adopted a patch size of 512 × 512 pixels. Compared to 256 × 256 patches, this size preserves more complete geometric patterns and semantic context, such as parcel shapes and boundaries.

3.2 Statistics

This section will introduce some statistics of the GTPBD dataset. With the collection of Global Terrace Map (GTM) [21], the number of labeled images from different regions has been count and the images belonging to each region are divided into training, testing and validation sets with the same scale structure. As is shown in the Fig. 3 (a), our proposed GTPBD dataset contains a large number of images from Southwest China which is followed by Rest of the world, which reflects the global distribution of terraces well. About dataset splits, the majority of the images (60%) are allocated to the training set, while the testing and validation sets each contain 20% of the total images (Fig. 3 (b)). Fig. 3 (c) shows a detailed construction of the spatial coverage across various regions, which suggests that GTPBD dataset is collected mainly from Southwest China, Central China and other parts of the world. Fig. 3 (d) illustrates the parcel size distribution, and because logarithmically scaled vertical coordinate is used, it can be seen that the majority of terraced parcels are smaller than



(a)

(b)

(c) Spectral values

(d) Parcel scales

Figure 4: Statistics for three different domains. (a) Distribution of area across different domains. (b) The number of images across different domains. (c) Spectral statistics of mean and standard deviation (σ) for different domains. (d) Distribution of parcel sizes across different domains (logarithmic scale).

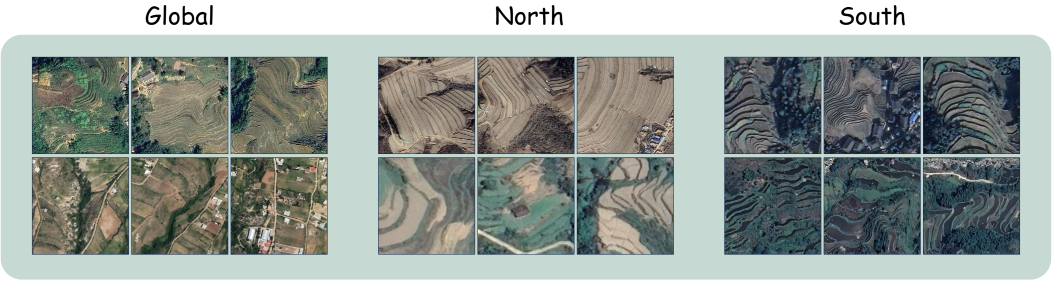


Figure 5: Some cases for three different domains: **Global**, **North** and **South**.

5000 pixels, which presents a challenge to the parcel extraction task. More regional statics of terraced parcel size could be found in Appendix B.3.

3.3 Differences among three domains

We seperate our GTPBD into three domains: South China (**South**), North China (**North**) and other regions except China (**Global**). The detailed statics of these three domains are shown in Fig. 4 (a) and (b). According to Fig. 5, the characteristics from different regions are really different.

For the spectral statistics, the mean values are similar (Fig. 4 (c)) because of the large-scale homogeneous geographical areas and diverse land cover types. We could observe that the **South** domain have lower standard deviations. As is shown in Fig. 4 (d), although all domains exist "long tail" phenomenon (logarithmic scale in vertical axis), the **South** domain has the most of the parcels have relatively small scales. When faced with large-scale terraced parcel and boundary extraction tasks, the differences between different scenes or regions bring new challenges to the model generalization and transferability.

Table 2: Comparisons of different semantic segmentation methods on the GTPBD dataset (%). More visualization results and the performance of regions could be found in Appendix E.1.3 and E.1.2.

Method	Prec.↑	Rec.↑	IoU↑	OA↑	F1-score↑
UNet [29]	74.11	54.93	46.09	75.46	63.09
DeepLabV3 [6]	69.64	73.45	57.04	78.28	71.58
PSPNet [52]	68.33	72.41	54.22	76.65	72.31
Nonlocal [40]	**75.06**	70.27	51.48	**79.52**	72.58
OCRNet [46]	73.05	71.56	53.87	76.95	72.30
K-Net [50]	74.56	61.04	50.52	77.17	67.13
SegFormer [41]	74.45	69.07	55.84	78.14	71.66
Mask2Former [8]	71.22	**74.33**	**57.16**	78.73	**72.74**

Table 3: Comprehensive performance comparison of edge detection models on the GTPBD dataset.

Method	ODS↑	OIS↑	AP↑
MuGE [55]	62.56	61.93	65.12
PiDiNet [31]	53.70	53.12	52.92
UEAD [56]	25.88	26.01	17.94
REAUNet-Sober [22]	**65.06**	**63.73**	**70.09**

4 Experiments

4.1 Experimental setup

The data splits followed Sec. 3.1. During the training, we used the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. For the data augmentation, 512×512 patches were randomly cropped with random mirroring and rotation. We implemented all our experiments on NVIDIA RTX 4090 GPU. We benchmark the GTPBD dataset on eight semantic segmentation methods (U-Net [29], DeepLabV3 [6], PSPNet [52], NonLocal [40], OCRNet [46], K-Net [50], SegFormer [41] and Mask2Former [8]), four edge extraction methods (MuGE [55], PiDiNet [31], UEAD [56] and REAUNet-Sober [22]), three parcel extraction methods (REAUNet [22], SEANet [19], HBGNet [51]) and five UDA methods (Source only, FDA [43], PiPa [24], HRDA [17] and DAFormer [16]). More implementation details are provided in the Appendix C.

4.2 Evaluation metrics

We brief list our comprehensive evaluation metrics for different tasks including pixel-level, object-level and edge detection of GTPBD. More details could be found in Appendix D.

(1) **Pixel-level evaluation metrics:** To evaluate segmentation accuracy on the GTPBD dataset, we adopt five standard pixel-level metrics commonly used in semantic segmentation and unsupervised domain adaptation (UDA) tasks, including Precision **(Prec.)**, Recall **(Rec.)**, Intersection over Union **(IoU)**, Overall Accuracy **(OA)** and **F1-score**; (2) **Object-level geometric metrics:** To evaluate the geometric accuracy of parcel delineation on the GTPBD dataset, we adopt three object-level metrics: Global Over-Classification Error **(GOC)**, Global Under-Classification Error **(GUC)**, and Global Total Classification Error **(GTC)**. These indicators provide a comprehensive assessment of shape fidelity and segmentation precision at the object level; (3) **Edge detection evaluation metrics:** For edge detection tasks on the GTPBD dataset, we evaluate model performance using three standard metrics: Optimal Dataset Scale F1-score **(ODS)**, Optimal Image Scale F1-score **(OIS)**, and Average Precision **(AP)**. These metrics jointly assess the adaptability and stability of boundary predictions.

4.3 Semantic Segmentation Results

Semantic segmentation plays a vital role in agricultural parcel extraction. As summarized in Table 2, these models exhibit notable discrepancies in segmentation performance, reflecting varying capabilities in modeling fine-grained agricultural boundaries. Specifically, the NonLocal model achieved the highest Prec. (75.06%) and OA (79.52%), indicating strong discrimination with minimal false positives. In contrast, Mask2Former demonstrated superior generalization, attaining the best scores in Rec. (74.33%), IoU (57.16%), and F1-score (72.74%). These findings highlight a trade-off between precision and recall among different architectures and emphasize the robustness of transformer-based models under complex terraced scenarios. Fig. 6 visually compare the segmentation results on a sample patch, with red highlighting false positives and blue highlighting false negatives. As can be observed, conventional semantic segmentation can only distinguish farmland regions from non-farmland areas, but fails to delineate precise parcel boundaries — especially for adjacent parcels that share common borders. More visualization results and the performance of regions could be found in Appendix E.1.

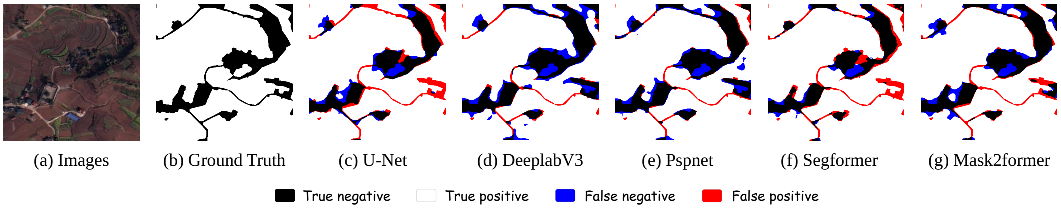

(a) Images (b) Ground Truth (c) U-Net (d) DeeplabV3 (e) Pspnet (f) Segformer (g) Mask2former

Figure 6: Qualitative comparisons of semantic segmentation and error maps among different methods.

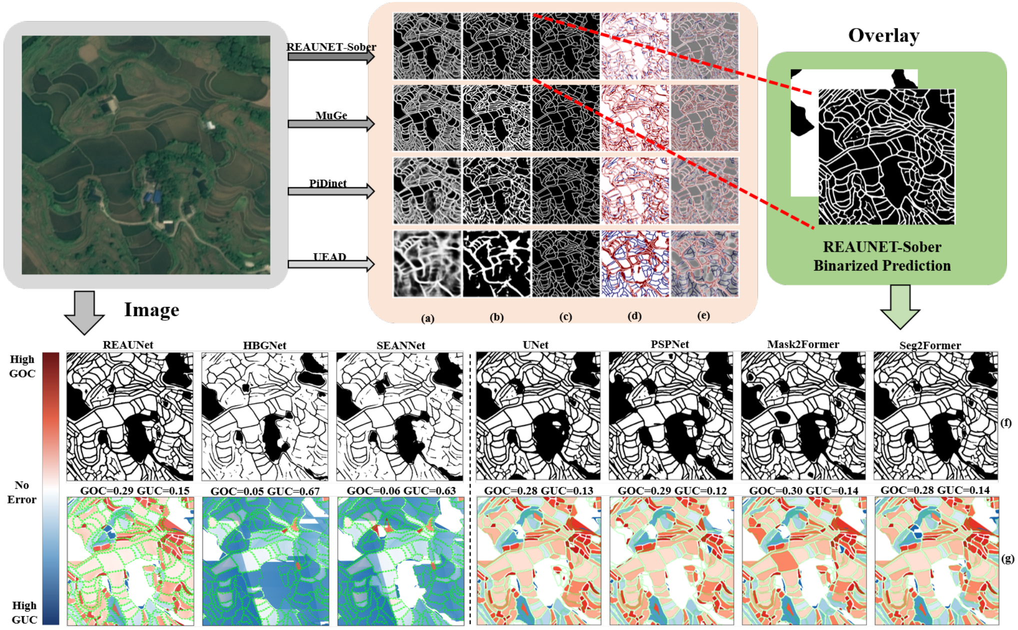


Figure 7: Edge detection and terraced parcel extraction. (a) Probability Map. (b) Binarized Prediction. (c) Ground Truth. (d) Error Regions (Red: FP, Blue: FN). (e) Overlay Visualization (alpha=0.5). (f) The performance of integrating edge results (REAUNET-Sober) and semantic segmentation results. (g) Evaluation of object-level parcel extraction (Red: GOC, Blue: GUC, Green: GT boundary).

4.4 Edge Detection Results

Conventional edge detection datasets typically include multi-scale, manually annotated boundaries. To simulate this for the GTPBD dataset, we apply morphological operations (erosion and dilation) to generate edge labels with widths of 1 to 5 pixels, mimicking human-labeled multi-resolution boundaries. Table 3 presents the results of four edge detection models. Among them, REAUNet-Sober achieves the best performance across all three metrics—ODS (65.06%), OIS (63.73%), and AP (70.09%)—significantly outperforming both general-purpose (e.g., PiDiNet), demonstrating its effectiveness in delineating parcel edges under noisy conditions. Visual examples in Fig. 7 (d) further illustrate model robustness, where false positives are marked in red and missed detections in blue.More visualization results and the performance of regions could be found in AppendixE.2

4.5 Boundary-integration terraced parcel extraction results

Building on the boundary constraint theory proposed by Yuan et al. [47], which highlights the importance of edge orientation and intensity in enhancing object-level delineation, we explore boundary-integration strategies for agricultural parcel segmentation. We compare four general-purpose segmentation models (U-Net, PSPNet, SegFormer, and Mask2Former) with three agriculture-specific ones (HBGNet, REAUNet, and SEANet). To incorporate edge cues, we introduce a boundary detection pipeline (see Fig. 7) that explicitly integrates edge information into the segmentation process. REAUNet-Sober achieves the best performance among the trained edge detectors, and its binarized output is used to guide segmentation refinement. As shown in Table 4, Mask2Former delivers the best object-level performance, achieving the lowest GOC (22.04%) and GTC (35.53%), despite lower

Table 4: Performance Comparison of Parcels Extraction(unit:%)

Method	Prec.↑	Rec.↑	IoU↑	OA↑	F1↑	GOC↓	GUC↓	GTC↓
U-Net [29]	**74.34**	54.29	45.72	75.39	62.75	43.57	**37.04**	40.43
PSPNet [52]	68.55	71.50	53.84	76.59	69.99	22.66	49.90	39.75
SegFormer [41]	74.71	68.23	55.43	79.05	71.32	27.15	41.17	35.84
Mask2Former [8]	71.48	73.42	56.79	78.66	72.44	**22.04**	45.15	**35.53**
SEANet [19]	70.69	79.56	59.89	76.07	74.04	26.80	54.70	44.08
REAUNet [22]	73.03	78.01	60.56	80.60	75.44	27.02	42.25	36.07
HBGNet [51]	72.50	**81.81**	**62.44**	**81.20**	**76.88**	27.40	42.52	35.79

Table 5: Comparisons of UDA and Boundary-integration UDA Performance on GTPBD (%).

		Conventional UDA					Boundary-integration UDA (Integrated by REAUNet-Sober)							
Domain	**Method**	**Prec.↑**	**Rec.↑**	**IoU↑**	**OA↑**	**F1↑**	**Prec.↑**	**Rec.↑**	**IoU↑**	**OA↑**	**F1↑**	**GOC↓**	**GUC↓**	**GTC↓**
S → N	Source Only	61.57	68.75	48.11	71.07	64.96	61.55	68.58	48.01	71.03	64.88	25.49	49.12	39.13
	FDA[43]	75.60	46.72	40.60	73.33	57.75	75.59	46.41	40.36	73.25	57.51	53.95	34.20	45.16
	PiPa [24]	62.74	**84.71**	**56.35**	74.41	**72.09**	62.72	**84.05**	**56.05**	74.29	**71.84**	**13.79**	54.71	40.91
	HRDA [17]	81.48	59.31	52.26	**78.87**	68.65	81.56	58.78	51.88	**78.73**	68.32	42.02	26.27	**35.25**
	DAFormer [16]	**82.12**	58.18	51.64	78.74	68.11	**82.15**	57.79	51.35	78.64	67.85	41.08	**29.49**	35.78
S → G	Source Only	66.26	72.53	52.97	**77.63**	69.25	66.27	72.46	52.94	77.63	69.23	26.62	48.39	39.95
	FDA[43]	**67.03**	53.79	42.54	74.76	59.68	**67.21**	52.71	41.93	74.64	59.08	48.20	44.08	46.80
	PiPa [24]	61.33	**80.99**	**53.61**	75.66	**69.80**	61.54	**79.33**	**53.03**	75.60	**69.31**	**18.06**	54.68	40.71
	HRDA [17]	62.85	73.72	51.35	75.74	67.85	63.15	72.01	50.70	75.69	67.29	22.48	51.45	**39.70**
	DAFormer [16]	66.24	62.43	47.36	75.90	64.28	66.43	61.36	46.84	**75.81**	63.79	38.18	**42.73**	40.51
N → S	Source Only	**82.18**	68.18	59.40	81.21	74.53	**82.36**	68.03	59.38	81.23	74.51	32.40	**33.53**	33.46
	FDA[43]	71.37	62.66	50.07	74.81	66.73	71.62	62.05	49.80	74.78	66.49	40.43	50.03	45.48
	PiPa [24]	79.28	80.71	66.65	83.72	79.99	79.55	79.91	66.29	83.62	79.73	23.88	39.95	32.91
	HRDA [17]	77.19	**84.01**	**67.30**	**83.74**	**80.46**	77.51	**83.13**	**66.98**	**83.77**	**80.22**	**19.23**	41.83	**32.55**
	DAFormer [16]	78.26	82.41	67.05	83.67	80.28	78.59	81.56	66.73	83.60	80.05	19.39	42.21	32.85
N → G	Source Only	73.83	56.90	47.36	78.03	64.27	73.86	56.86	47.33	78.03	64.25	45.92	33.78	40.30
	FDA[43]	66.10	51.38	40.66	73.96	57.82	66.22	50.68	40.27	73.89	57.42	56.64	48.15	52.57
	PiPa [24]	69.60	80.67	59.65	81.05	74.72	69.85	79.30	59.08	80.93	74.28	24.34	42.21	34.45
	HRDA [17]	70.66	**83.45**	**61.98**	**82.22**	**76.52**	70.97	**81.90**	**61.35**	**82.08**	**76.04**	**20.67**	43.05	**32.37**
	DAFormer [16]	**74.96**	67.77	55.26	80.95	71.19	**75.21**	66.54	54.57	80.76	70.61	32.33	**33.14**	32.74
G → S	Source Only	**80.06**	72.26	61.24	81.56	75.96	**80.22**	72.10	61.22	81.58	75.94	28.63	**37.75**	33.50
	FDA[43]	66.15	51.92	41.02	69.90	58.18	66.37	51.42	40.79	69.91	57.95	46.02	49.66	47.88
	PiPa [24]	78.06	**77.84**	**63.86**	82.23	**77.95**	77.96	**76.65**	62.99	81.84	**77.30**	**22.67**	41.10	**33.29**
	HRDA [17]	74.59	76.28	60.54	79.96	75.42	74.90	75.45	60.23	79.91	75.18	25.91	42.75	35.35
	DAFormer [16]	78.72	76.75	63.56	**82.26**	77.72	79.01	76.01	**63.23**	**82.18**	77.48	26.46	40.55	34.04
G → N	Source Only	**77.03**	69.08	57.28	**79.90**	72.84	**77.03**	69.03	57.25	**79.89**	72.81	32.14	**35.97**	**34.10**
	FDA[43]	53.77	75.31	45.72	65.12	62.75	53.72	74.87	45.51	65.04	62.56	25.75	56.99	44.22
	PiPa [24]	62.13	**88.77**	**57.60**	74.51	**73.10**	62.12	**88.10**	**57.31**	74.40	**72.86**	**10.14**	54.53	39.22
	HRDA [17]	58.89	83.01	52.56	70.77	68.90	58.85	82.40	52.28	70.66	68.66	20.83	51.16	39.01
	DAFormer [16]	70.80	72.87	56.03	77.69	71.82	70.81	72.26	55.68	77.56	71.53	28.16	41.19	35.28
Avg	Source Only	73.49	68.12	54.39	77.23	70.51	73.53	68.01	54.29	77.24	70.44	31.70	42.59	36.74
	FDA[43]	70.00	56.13	43.25	71.99	62.22	70.22	55.94	42.92	71.92	61.91	45.17	47.19	47.10
	PiPa [24]	69.02	**82.12**	**59.62**	78.60	**74.38**	69.29	**81.22**	**59.13**	78.46	**74.07**	**17.15**	47.88	36.91
	HRDA [17]	70.78	76.63	58.50	78.45	73.30	71.16	75.61	57.91	78.34	72.95	25.19	42.75	**35.93**
	DAFormer [16]	**75.48**	69.90	56.82	**79.72**	72.40	**75.63**	69.25	56.23	**79.58**	72.22	31.10	**38.22**	36.20

pixel-level scores. In contrast, HBGNet attains the highest pixel-level accuracy (Recall: 81.81%, IoU: 62.4%, OA: 82.20%, F1: 76.88%) but suffers from higher object-level errors, indicating a trade-off between pixel precision and structural integrity.

4.6 Unsupervised domain adaptation results

We evaluate domain robustness of segmentation models through UDA experiments across three distinct domains in GTPBD (see Sec. 3.3). The Source only and other four domain adaptation methods are benchmarked: FDA, DAFormer, HRDA, and PiPa. As shown in Table 5, across all six domain adaptation directions, PiPa achieves the best performance in pixel-level metrics (Rec., IoU, F1), demonstrating stable cross-domain generalization. Notably, HRDA performs best when the source domain is N, achieving superior results across nearly all metrics under this setting. After adding boundary results (integrated by the performance of REAUNet-Sober), further achieves the lowest GOC and GTC, indicating a stronger ability to preserve parcel boundaries and maintain object completeness, while DAFormer attains the lowest GUC, reflecting its advantage in reducing under-segmentation and improving boundary coherence.Visualization comparisons are provided in Appendix E.4.

5 Conclusion

In this paper, we propose a more challenging terraced parcel dataset named **GTPBD** (**G**lobal **T**erraced **P**arcel and **B**oundary **D**ataset), which is the first fine-grained dataset covering major worldwide terraced regions with more than 200,000 complex terraced parcels with manually annotation. GTPBD comprises 47,537 high-resolution images with three-level labels, including pixel-level boundary labels, mask labels, and parcel labels. It covers seven major geographic zones in China and transcontinental climatic regions around the world. Our proposed GTPBD dataset is suitable for four different tasks, including semantic segmentation, edge detection, terraced parcel extraction and Unsupervised Domain Adaptation (UDA) tasks. Accordingly, we benchmark the GTPBD dataset on eight semantic segmentation methods, four edge extraction methods, three parcel extraction methods and five UDA methods, along with a multi-dimensional evaluation framework integrating pixel-level and object-level metrics. Results highlight the limitations of current models in extracting fine-grained terraced parcels and boundaries, especially under domain shifts. Therefore, our proposed GTPBD fills a critical gap in remote sensing benchmarks and provides a critical infrastructure for complex agricultural terrain analysis and cross-scenario knowledge transfer.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant T2125006 and Grant 42401415; in part by the Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant KJZD20230923115106012; in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, under Project 24xkjc002; and in part by the Jiangsu Innovation Capacity Building Program under Project BM2022028.

References

[1] Hamed Alemohammad and Kevin Booth. Landcovernet: A global benchmark land cover classification training dataset. *arXiv preprint arXiv:2012.03111*, 2020.

[2] José Arnáez, Noemí Lana-Renault, T Lasanta, P Ruiz-Flaño, and J Castroviejo. Effects of farming terraces on hydrological and geomorphological processes. a review. *Catena*, 128:122–134, 2015.

[3] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[4] Bowen Cao, Le Yu, Victoria Naipal, Philippe Ciais, Wei Li, Yuanyuan Zhao, Wei Wei, Die Chen, Zhuang Liu, and Peng Gong. A 30-meter terrace mapping in china using landsat 8 imagery and digital elevation model based on the google earth engine. *Earth System Science Data*, 2020:1–35, 2020.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Wentang Chen, Yibin Wen, Juepeng Zheng, Jianxi Huang, and Haohuan Fu. Ban: A universal paradigm for cross-scene classification under noisy annotations from rgb and hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[9] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G. Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2825–2835, 2020.

[10] Raphaël d'Andrimont, Martin Claverie, Pieter Kempeneers, Davide Muraro, Momchil Yordanov, Devis Peressutti, Matej Batič, and François Waldner. Ai4boundaries: an open ai-ready dataset to map field boundaries with sentinel-2 and aerial photography. *Earth System Science Data*, 15(1):317–329, 2023.

[11] Hu Ding, Jiaming Na, Shangjing Jiang, Jie Zhu, Kai Liu, Yingchun Fu, and Fayuan Li. Evaluation of three different machine learning methods for object-based artificial terrace mapping—a case study of the loess plateau, china. *Remote Sensing*, 13(5):1021, 2021.

[12] Cruz Ferro-Vázquez, Carol Lang, Joeri Kaal, and Daryl Stump. When is a terrace not a terrace? the importance of understanding landscape evolution in studies of terraced agriculture. *Journal of Environmental Management*, 202:500–513, 2017.

[13] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[14] Joanne V Hall, Fernanda Argueta, and Louis Giglio. Glocab cropland field boundary dataset. *Data in Brief*, 55:110739, 2024.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[16] Peter Hofmann, Janis Lenssen, and Cordelia Schmid. Daformer: Improving network generalization for domain-adaptive semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2203.03605.

[17] Sven Hoyer, Frederik Kroeger, and Bernt Schiele. Hrda: Hierarchical representation domain adaptation for cross-domain semantic segmentation. In *European Conf. on Computer Vision (ECCV)*, pages 663–679, 2022.

[18] Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, et al. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28151–28159, 2025.

[19] Mengmeng Li, Jiang Long, Alfred Stein, and Xiaoqin Wang. Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 200:24–40, 2023.

[20] Qingmei Li, Yibin Wen, Juepeng Zheng, Yuxiang Zhang, and Haohuan Fu. Hyunida: Breaking label set constraints for universal domain adaptation in cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[21] Yifan Li, Fuyou Tian, Miao Zhang, Hongwei Zeng, Shukri Ahmed, Xinli Qin, Yanxu Liu, Lizhe Wang, Runyu Fan, and Bingfang Wu. A 10-meter global terrace mapping using sentinel-2 imagery and topographic features with deep learning methods and cloud computing platform support. *International Journal of Applied Earth Observation and Geoinformation*, 139:104528, 2025.

[22] Rui Lu, Yingfan Zhang, Qiting Huang, Penghao Zeng, Zhou Shi, and Su Ye. A refined edge-aware convolutional neural networks for agricultural parcel delineation. *International Journal of Applied Earth Observation and Geoinformation*, 133:104084, 2024.

[23] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3940–3956, 2021.

[24] Rohan T. Mahapatra, Yiming Wang, Qiaoni Liu, and Zhuowen Tu. Pipa: Pixel- and patch-wise self-supervised domain adaptation for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11544–11553, 2021.

[25] Jingshan Pan, Zhiqiang Wei, Yuhan Zhao, Yan Zhou, Xunyu Lin, Wei Zhang, and Chang Tang. Enhanced fcn for farmland extraction from remote sensing image. *Multimedia Tools and Applications*, 81(26):38123–38150, 2022.

[26] Yang Pan, Xinyu Wang, Liangpei Zhang, and Yanfei Zhong. E2evap: End-to-end vectorization of smallholder agricultural parcel boundaries from high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203:246–264, 2023.

[27] Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. Ai4smallfarms: A dataset for crop field delineation in southeast asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

[28] Joana Reuss, Jan Macdonald, Simon Becker, Lorenz Richter, and Marco Körner. The eurocropsml time series benchmark dataset for few-shot crop type classification in europe. *Scientific Data*, 12(1):664, 2025.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[30] Joseph E Spencer and Gary A Hale. The origin, nature, and distribution of agricultural terracing. *Pacific viewpoint*, 2(1):1–40, 1961.

[31] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021.

[32] Prasad S Thenkabail, Pardhasaradhi G Teluguntla, Jun Xiong, Adam Oliphant, Russell G Congalton, Mutlu Ozdogan, Murali Krishna Gumma, James C Tilton, Chandra Giri, Cristina Milesi, et al. Global cropland-extent product at 30-m resolution (gcep30) derived from landsat satellite time-series data for the year 2015 using multiple machine-learning algorithms on google earth engine cloud. Technical report, US Geological Survey, 2021.

[33] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies*, 8(2):35, 2020.

[34] Xin-Yi Tong, Qikai Lu, Gui-Song Xia, and Liangpei Zhang. Large-scale land cover classification in gaofen-2 satellite imagery. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 3599–3602. IEEE, 2018.

[35] Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[36] Unicef et al. The state of food security and nutrition in the world 2024. 2024.

[37] François Waldner and Foivos I Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote sensing of environment*, 245:111741, 2020.

[38] Junjue Wang, Zhuo Zheng, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[39] Sherrie Wang, François Waldner, and David B Lobell. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *Remote Sensing*, 14(22):5738, 2022.

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[41] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[42] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[43] Yinghao Yang, Gongjie Zhang, Jianfeng Yang, and Dimitris Metaxas. Frequency domain adaptation for semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6748–6757, 2020.

[44] Mingge Yu, Xiaoping Rui, Weiyi Xie, Xijie Xu, and Wei Wei. Research on automatic identification method of terraces on the loess plateau based on deep transfer learning. *Remote sensing*, 14(10):2446, 2022.

[45] Shuai Yuan, Guancong Lin, Lixian Zhang, Runmin Dong, Jinxiao Zhang, Shuang Chen, Juepeng Zheng, Jie Wang, and Haohuan Fu. Fusu: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 132417–132439. Curran Associates, Inc., 2024.

[46] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[47] Yuhui Yuan, Jingyi Xie, and Xilin Chen & Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision – ECCV 2020*, 2020.

[48] Florian Zabel, Birgitta Putzenlechner, and Wolfram Mauser. Global agricultural land resources–a high resolution suitability evaluation and its perspectives until 2100 under climate change conditions. *PloS one*, 9(9):e107522, 2014.

[49] Han Zhai and Ruoheng Zhang. A deeply supervised semantic-edge dual u-net with shared structure for cropland parcel extraction from high-resolution remote sensing imagery. *Available at SSRN 4940114*.

[50] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

[51] Hang Zhao, Bingfang Wu, Miao Zhang, Jiang Long, Fuyou Tian, Yan Xie, Hongwei Zeng, Zhaoju Zheng, Zonghan Ma, Mingxing Wang, et al. A large-scale vhr parcel dataset and a novel hierarchical semantic boundary-guided network for agricultural parcel delineation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:1–19, 2025.

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[53] Juepeng Zheng, Yibin Wen, Mengxuan Chen, Shuai Yuan, Weijia Li, Yi Zhao, Wenzhao Wu, Lixian Zhang, Runmin Dong, and Haohuan Fu. Open-set domain adaptation for scene classification using multi-adversarial learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:245–260, 2024.

[54] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[55] Caixia Zhou, Yaping Huang, Mengyang Pu, Qingji Guan, Ruoxi Deng, and Haibin Ling. Muge: Multiple granularity edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25952–25962, 2024.

[56] Caixia Zhou, Yaping Huang, Mengyang Pu, Qingji Guan, Li Huang, and Haibin Ling. The treasure beneath multiple annotations: An uncertainty-aware edge detector. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 15507–15517, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims are summarized in Fig. 1, with detailed explanations provided in Sec. 3 (Dataset) and Sec. 4 (Experiment).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limitations of this work have been discussed in Appendix F of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explained our implementation details in Sec. 4.1 and hyperparameters in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and code, and GitHub link is offered.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details are summarized in Sec. 4.1 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Since running many repeated experiments would require significant computational resources and time, the paper reports only the direct results on our dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources in Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We affirm that our research fully complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work only investigates generalization on existing models, and poses no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper have been properly credited and the license and terms of use explicitly have been mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code and accompanying documentation in our GitHub link.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

GTPBD: A Fine-Grained Global Terraced Parcel and Boundary Dataset (Supplementary material)

Table of Contents in Appendix

A Broader impact

The proposed GTPBD dataset provides the first large-scale, fine-grained benchmark for terraced parcel analysis across diverse global terrains. Its release is intended to facilitate progress in computer vision methods for agricultural mapping, with potential societal benefits including improved food security analysis, precision agriculture, and climate-adaptive land management. By focusing on underrepresented mountainous regions, GTPBD promotes algorithmic development that extends beyond conventional flat farmland, contributing to more inclusive and geographically equitable AI research.This dataset may assist governmental and environmental agencies in monitoring land use, identifying erosion risks, and planning sustainable agricultural practices. It may also support digital infrastructure for land ownership registration in developing regions, where documentation remains limited or inaccessible.However, the deployment of models trained on GTPBD in real-world scenarios must be approached with care. Potential negative impacts include misuse of parcel delineation algorithms for land commodification, surveillance, or disempowerment of local communities, especially in regions with contested or informal land tenure. Furthermore, any bias introduced during data annotation or model training could disproportionately affect underrepresented regions or farming systems.

To mitigate these risks, we strongly encourage practitioners to collaborate with local stakeholders and domain experts, ensure transparency in model usage, and align applications with ethical land governance principles. The dataset should serve as a scientific tool for advancing equitable and sustainable land analysis, not for enabling exploitative practices.

B More image cases

B.1 More image cases in GTPBD

GTPBD covers seven major geographic zones in China and transcontinental climatic regions around the world. There are different terraces in different regions, and Fig. 8 supplements the images of terraces of the 12 regions mentioned in Fig. 2 . The dataset's comprehensive labeling system features three distinct levels of annotation, with detailed examples illustrated in Fig. 9.

B.2 Some cases in previous datasets

Fig. 10 displays typical annotation–imagery pairs from six widely used parcel datasets: (a) FHAPD [51], (b) FTW [18], (c) AI4Boundaries [10], (d) PASTIS [13], (e) CP-Set [49], and (f) GFSAD30 [32]. Although each benchmark has advanced parcel delineation research, several common limitations are evident. FHAPD and CP-Set rely on binary masks with coarse edges, leaving thin ridges either over-smoothed or missing. FTW and PASTIS are built on Sentinel-2 imagery (10 m GSD), so images appear blurry and smallholder fields are merged. AI4Boundaries depicts large, orthogonal parcels from mechanised farms, providing little shape diversity. GFSAD30, at 30 m resolution, captures only blocky field outlines and omits fine-scale geometry altogether. These issues—low spatial resolution, overly regular parcel shapes, and imprecise or incomplete boundaries limit the evaluation of models on fragmented, irregular terrains.

B.3 More statistics of GTPBD

In Sec. 3.2, detailed statistical analyses of the data have been carried out. To clearly illustrate the distribution of parcel sizes across the eight regions, Fig. 11 provides individual bar charts for each area. These charts highlight that small-sized terraced blocks dominate in every region.

C Model details

C.1 Semantic segmentation

MMSegmentation is an open-source semantic segmentation toolbox built on the OpenMMLab ecosystem and PyTorch [2]. It implements a wide range of state-of-the-art architectures, such as U-Net,

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

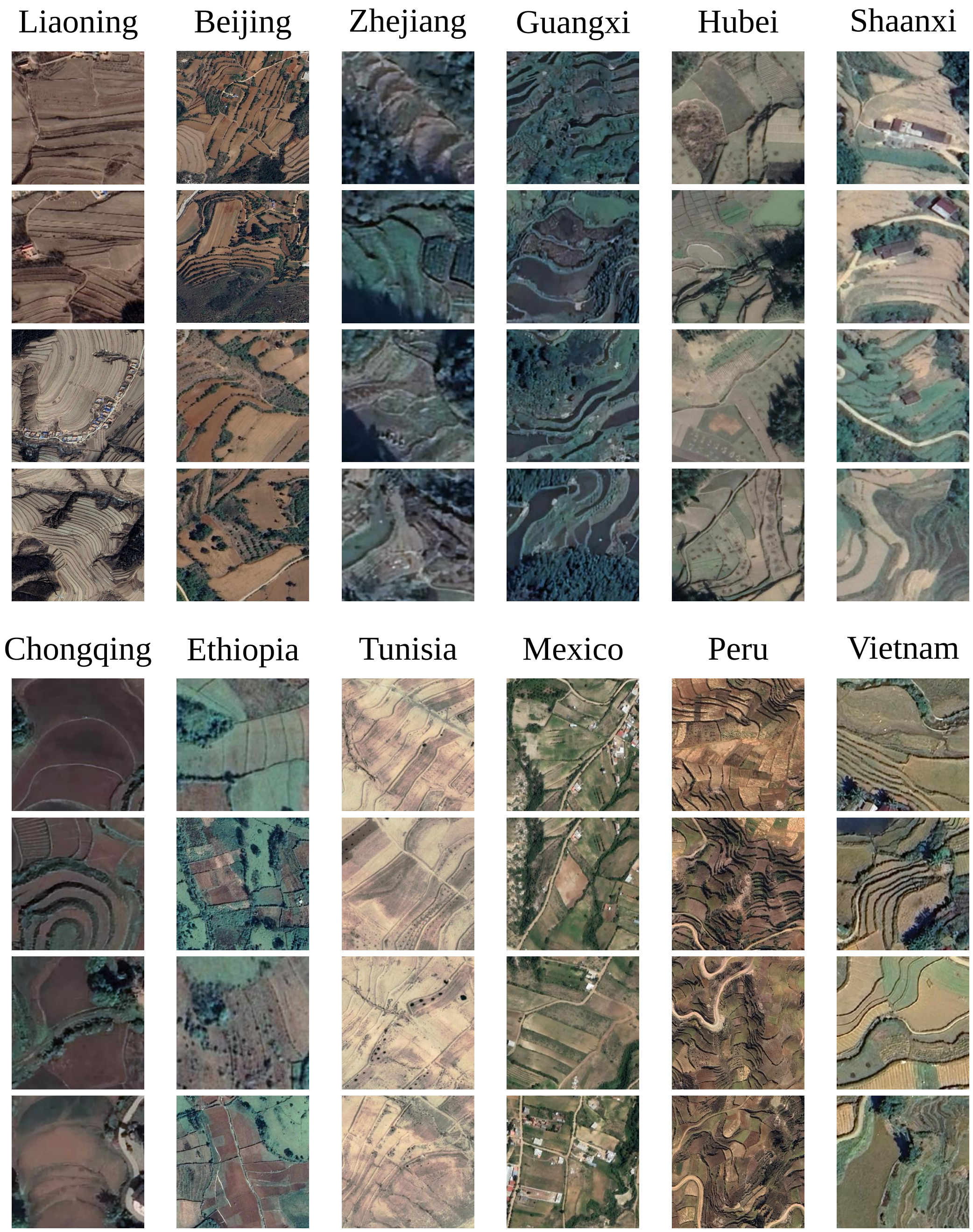


Figure 8: More images cases of different regions in GTPBD.

DeepLabV3, PSPNet, and Segformer, thereby facilitating reproducible comparisons across models. Our semantic segmentation experiments are all completed under the MMsegmention framework.

C.1.1 U-Net

U-Net, proposed by Ronneberger et al. [29], adopts a symmetric encoder–decoder design that captures contextual information via a contracting path and enables precise localization through an expansive path. In our experiments using MMSegmentation's default U-Net implementation, the key hyperparameters are configured as Table 6.

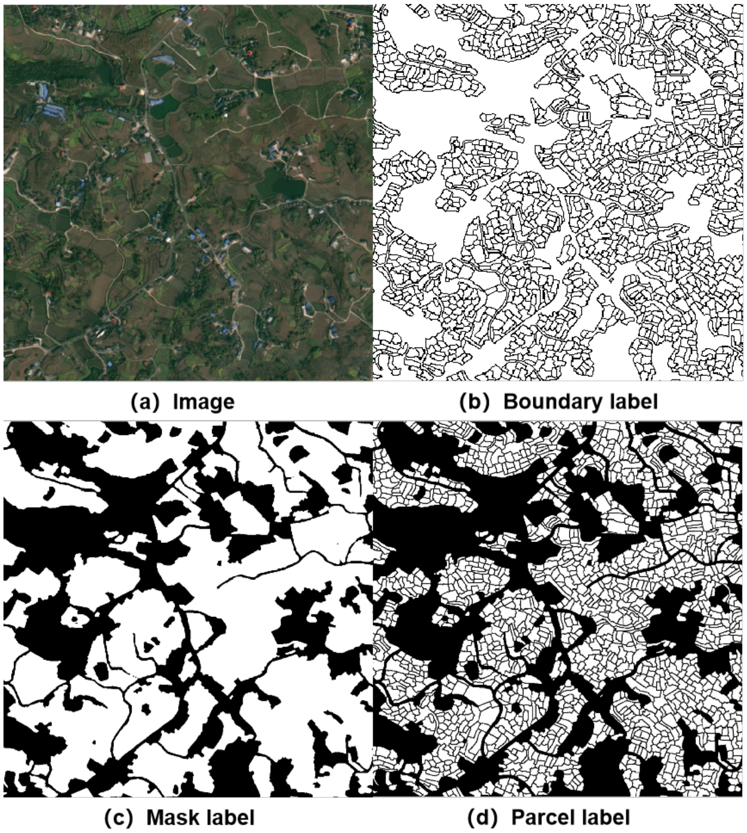


Figure 9: More images cases of three-level labels in GTPBD.

Table 6: The hyperparameters of U-Net

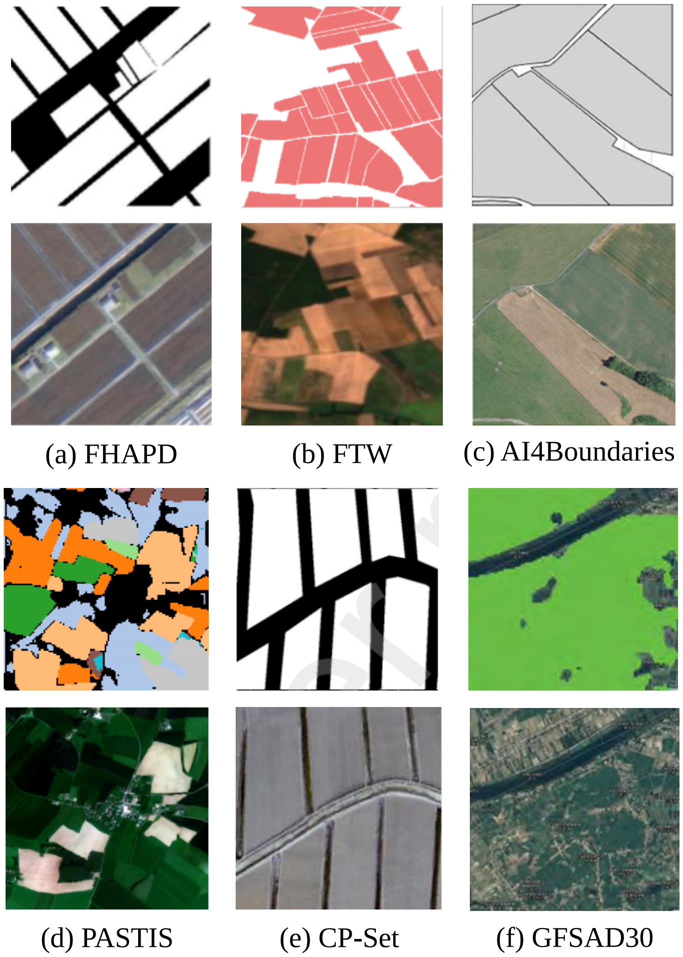
Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone base channels	Number of filters in the first convolution layer	64
Number of stages	Levels of encoder–decoder (depth of U-Net)	5
Convolutions per stage	Number of conv layers in each block	2
Optimizer	Optimization algorithm and initial learning rate	Adam, lr=$1e^{-4}$
Sliding-window crop size	Sliding-window inference patch size	256 × 256
Sliding-window stride	Stride between adjacent sliding-window patches	170
Training iterations	Total number of training iterations	20,000

C.1.2 DeepLabV3

DeepLabV3 [6] enhances semantic segmentation by employing atrous (dilated) convolutions to enlarge the receptive field without losing resolution, and by integrating an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale context. In our experiment, All hyper-parameters remain the same as in the mmsegmentation framework and are summarized in Table 7.

C.1.3 PSPNet

PSPNet [52] introduces a Pyramid Pooling Module that aggregates context information at multiple spatial scales by pooling features into different bin sizes, and then upsampling and concatenating them with the original feature map. This design enables the network to capture both global context



(a) FHAPD (b) FTW (c) AI4Boundaries

(d) PASTIS (e) CP-Set (f) GFSAD30

Figure 10: Representative Samples from Existing Agricultural Parcel Datasets

and fine details simultaneously. In our experiments, we rely on the out-of-the-box configuration provided by MMSegmentation without further modification, and summarize the key hyperparameters in Table 8.

C.1.4 NonLocal

Non-local Neural Networks [40], a generic non-local operation that computes pairwise interactions between any two positions on the feature map, enabling the model to capture long-range dependencies and context beyond local receptive fields. We adopt the standard MMSegmentation implementation without modifying its default settings and report the core hyperparameters in Table 9.

C.1.5 OCRNet

OCRNet [46] enhances pixel-wise segmentation by introducing an object-contextual representation module that aggregates contextual information from object regions to refine per-pixel predictions. Specifically, it first generates a coarse object region map via soft object-region pooling, then computes a pixel-wise contextual representation by weighting features according to their belongingness to these regions. This two-stage design allows the network to capture both local details and global object semantics effectively. Table 10 summarize the main hyperparameters in our experiment.

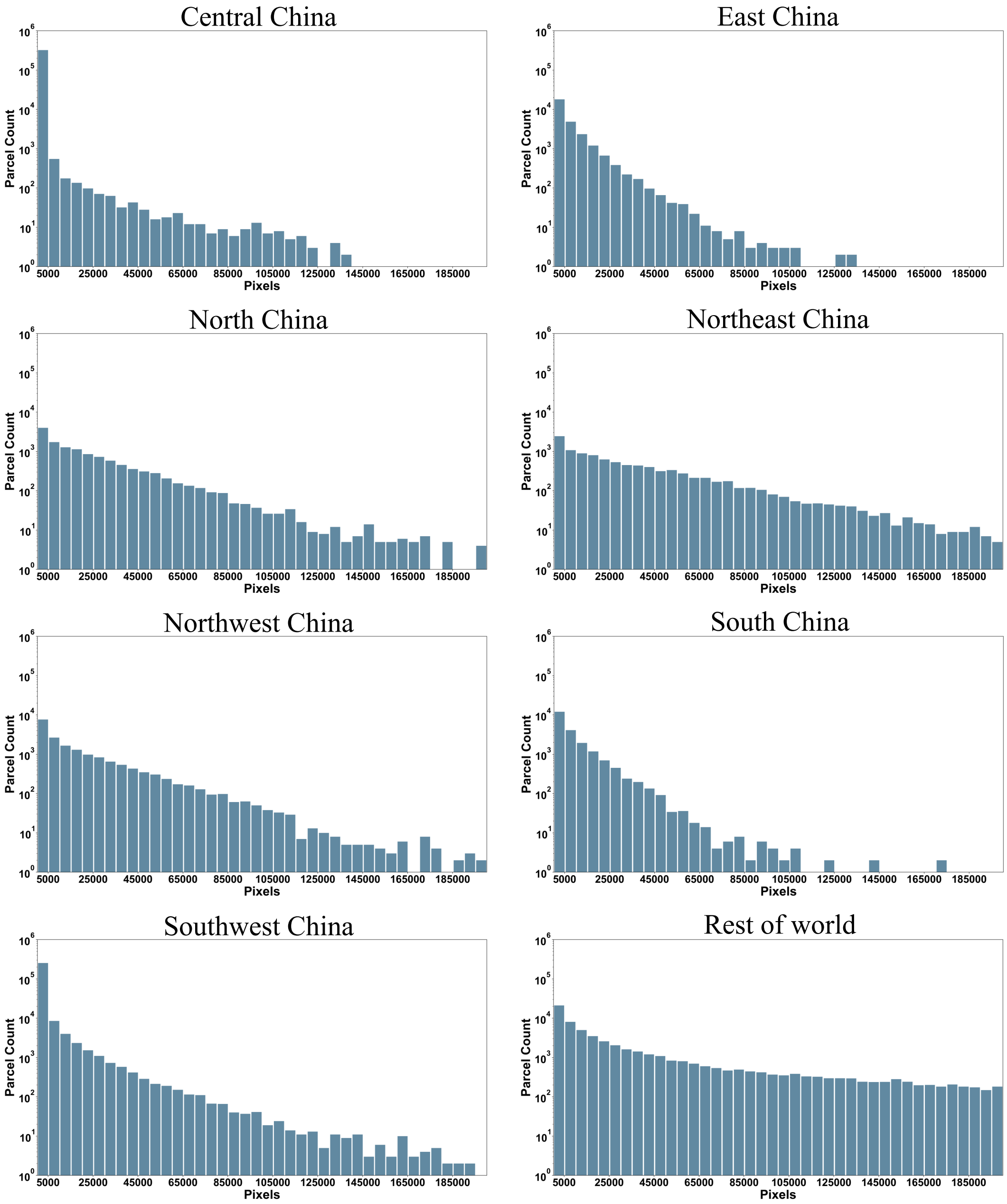


Figure 11: Distribution of parcel sizes across different regions. (logarithmic scale)

C.1.6 K-Net

K-Net [50] formulates semantic segmentation as a dynamic kernel-based grouping problem, enabling adaptive context aggregation and object-level reasoning. Each kernel generates an attention map that segments the feature map into distinct semantic regions, and these kernels are updated via a lightweight transformer-style interaction. In our experiments, we employ MMSegmentation's K-Net configuration without further modifications and summarize the core hyperparameters in Table 11.

C.1.7 Segformer

SegFormer [41] leverages a hierarchy of lightweight MLP-based Mix-FFN blocks and multi-level feature fusion to achieve efficient and effective semantic segmentation without positional encodings. Its simple encoder–decoder design employs a series of Transformer-like layers that progressively

Table 7: The hyperparameters of DeeplabV3

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Type and depth of backbone network	ResNetV1c-50
Backbone dilations	Dilation rates for each backbone stage	(1,1,2,4)
Backbone strides	Stride sizes for each backbone stage	(1,2,1,1)
ASPP dilations	Dilation rates in Atrous Spatial Pyramid Pool	(1,12,24,36)
ASPP channels	Number of intermediate channels in ASPP head	512
Auxiliary head channels	Number of channels in auxiliary FCN head	256
Number of classes	Number of segmentation target classes	2
Training iterations	Total number of training iterations	20,000

Table 8: The hyperparameters of PSPNet

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Type and depth of backbone network	ResNetV1c-50
Backbone dilations	Dilation rates for each backbone stage	(1,1,2,4)
Backbone stages	Number of residual stages used	4
PSP pooling scales	Sizes of the pyramid pooling bins	(1,2,3,6)
PSP intermediate channels	Number of feature channels after pooling	512
Auxiliary head channels	Number of channels in auxiliary FCN head	256
Loss weights	Main and auxiliary segmentation loss weights	1.0 (main), 0.4 (auxiliary)
Training iterations	Total number of training iterations	20,000

downsample the input, and fuse multi-scale representations via a lightweight MLP decoder. Key parameters are shown in Table 12 in our experiment.

C.1.8 Mask2Former

Mask2Former [8] presents a universal architecture for segmentation tasks by modeling pixel-wise, instance-wise, and semantic-level grouping using a mask-based Transformer decoder. It decouples mask prediction from task-specific heads, employing dynamic queries and multi-scale features to produce segmentation masks across different granularities. We employ Mask2Former configuration directly without any additional tuning and the key hyperparameters can be seen in Table 13.

C.2 Edge Detection Methods

C.2.1 UEAD

UEAD [56] addresses the annotation ambiguity and subjectivity inherent in edge detection by modeling label uncertainty. Instead of relying on deterministic pixel-wise labels, it learns a Gaussian distribution over annotations to capture labeling variance. The estimated variance serves as a measure of pixel-level uncertainty, and an adaptive weighting loss emphasizes learning from uncertain (i.e., hard) samples. UEAD can be integrated with various encoder–decoder backbones and consistently improves performance across multiple edge detection benchmarks.

C.2.2 MuGE

MuGE[55] introduces a novel edge detection framework that captures edge ambiguity by generating edge maps at multiple controllable granularities. It first predicts edge granularity from annotations,

Table 9: The hyperparameters of NonLocal

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Type and depth of backbone network	ResNetV1c-50
Backbone dilations	Dilation rates for each backbone stage	(1,1,2,4)
NLhead intermediate channels	Number of feature channels in the NLHead	512
NL reduction ratio	Channel reduction factor inside the non-local block	2
NL operation mode	Similarity function used in non-local computation	embedded_gaussian
Dropout ratio	Dropout probability applied in the NLHead	0.1
Training iterations	Total number of training iterations	20,000

Table 10: The hyperparameters of OCRNet

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Type and depth of backbone network	HRNetV2-W18
Number of cascade stages	Number of encoder–decoder stages	2
FCNhead intermediate channels	Total channels after concatenating multi-scale features	270
OCRHead channels	Number of feature channels in OCR contextual module	512
OCR channels	Reduced channel dimension inside OCR module	256
Training iterations	Total number of training iterations	20,000

then injects this granularity into multi-scale feature maps to produce edge maps ranging from coarse contours to fine textures. By decomposing feature maps into frequency components, MuGE enables fine control over edge detail, resulting in both interpretability and high accuracy.

C.2.3 PiDiNet

PiDiNet[31] offers a lightweight and efficient solution for edge detection by combining traditional gradient operators with modern convolutional design. It introduces pixel difference convolutions that emulate classical detectors like Canny and Sobel, enabling real-time inference with high accuracy. With less than 1M parameters, PiDiNet achieves human-level performance on BSDS500 and maintains strong efficiency–accuracy trade-offs across diverse benchmarks.

C.3 Agricultural Parcel Extraction Methods

C.3.1 REAUNet

REAUNet[22] is a tailored edge-aware network designed for accurate agricultural parcel delineation from both medium- and high-resolution remote sensing imagery (Sentinel-2 and GF-2). It addresses the common issue of unclosed and fragmented parcel edges by integrating four key components: an edge detection block, a dual attention block, a deep supervision mechanism, and a refine module. These modules collectively enhance the model's sensitivity to boundary information and improve multi-scale consistency. Experiments show that all four components are critical, with full REAUNet outperforming partial variants. Compared with SEANet and MPSPNet, REAUNet achieves improvements in both thematic (F1, IoU) and geometric (GTC) accuracy, and demonstrates promising transferability to unseen regions with limited fine-tuning data.

C.3.2 SEANet

SEANet[19] is a multi-task neural network that simultaneously predicts semantic masks, edges, and distance maps for precise and closed parcel delineation. By explicitly modeling boundary extraction as an edge detection task, SEANet enhances geometric accuracy and is particularly effective for

Table 11: The hyperparameters of K-Net

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Type and depth of backbone network	ResNetV1c-50
Decode head stages	Number of iterative decoding stages	3
Kernel update FFN channels	Hidden dimension in feed-forward networks of update heads	2048
Kernel update attention heads	Number of attention heads in each KernelUpdateHead	8
Convolution kernel size	Kernel size for pointwise conv in update heads	1
Training iterations	Total number of training iterations	20,000

Table 12: The hyperparameters of Segformer

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone	Type of hierarchical Transformer encoder	MixVisionTransformer
Embed dims	Dimension of the token embeddings at stage 1	32
Number of stages	Levels of hierarchical feature extraction	4
Layers per stage	Number of attention heads in each KernelUpdateHead	8
Layers per stage	Transformer blocks in each stage	[2,2,2,2]
Attention heads	Number of self-attention heads per stage	[1,2,5,8]
Patch sizes	Size of convolutional patches at each stage	[7, 3, 3, 3]
SR ratios	Spatial reduction ratios before attention	[8,4,2,1]
Training iterations	Total number of training iterations	20,000

small, irregular parcels. It incorporates a multi-level edge feature extraction mechanism and a task uncertainty-aware loss to improve generalization. Extensive experiments on both high-resolution (GF-2) and medium-resolution (Sentinel-2) images across China and Europe show that SEANet produces more accurate parcel layouts and demonstrates robust cross-region transferability.

C.3.3 HBGNet

HBGNet[51] is a hierarchical dual-branch framework designed to fully exploit boundary semantics for robust agricultural parcel extraction. It consists of a core AP extraction branch and an auxiliary boundary branch enhanced by Laplacian-based convolution. To improve adaptability across varying parcel sizes and morphologies, HBGNet integrates global–local context aggregation and boundary-guided fusion modules. It also introduces FHAPD, the first large-scale VHR agricultural parcel dataset in China, to support comprehensive evaluation. HBGNet outperforms eight existing methods across multiple datasets (FHAPD, AI4Boundaries, Sentinel-2) in both attribute and geometric metrics, achieving up to 7.5.

C.4 Unsupervised Domain Adaptation Methods

C.4.1 FDA

FDA[43] is a simple yet effective UDA method that reduces domain gaps by aligning the low-frequency components of source and target images in the Fourier domain. The core idea is that low-frequency information in images (e.g., color and style) mainly accounts for the domain shift, while high-frequency components capture structural details relevant to semantic content. By replacing the low-frequency spectrum of source images with that of target images, FDA transfers global appearance cues without altering the semantic layout. This process improves feature-level alignment and enhances segmentation performance in the target domain without requiring architectural changes or additional supervision.

Table 13: The hyperparameters of mask2former

Hyperparameter	Description	Typical Value
Input image size	Dimensions of input images	512×512
Backbone architecture	Feature extractor type and depth	ResNet-50
Number of queries	Number of learnable mask queries	100
Transformer encoder layers	Number of layers in the Deformable DETR transformer encoder	6
Transformer decoder layers	Number of layers in the Mask2Former transformer decoder	9
Classification loss weight	Weight for the classification loss in mask prediction	2.0
Mask & Dice loss weights	Weights for mask and Dice losses	mask=5.0, dice=5.0
Training iterations	Total number of training iterations	20,000

C.4.2 DAFormer

DAFormer[16] introduces a Transformer-based architecture for UDA in semantic segmentation, leveraging the superior representation capability of vision Transformers. The model integrates a Transformer encoder with a multi-level, context-aware decoder and employs three essential training strategies: rare class sampling to improve pseudo-labels, ImageNet feature distance regularization to stabilize transfer, and learning rate warmup to reduce overfitting. These components collectively enable DAFormer to effectively adapt to the target domain and learn rare or hard classes. DAFormer achieves significant improvements over previous UDA methods, setting new benchmarks on synthetic-to-real segmentation tasks.

C.4.3 HRDA

HRDA[17] proposes a multi-resolution UDA framework to balance fine-detail preservation and long-range context modeling in semantic segmentation. It combines small high-resolution image crops—effective for boundary and small object delineation—with large low-resolution crops that retain broader contextual information. A learned scale attention module adaptively fuses these dual-resolution representations. HRDA significantly improves segmentation quality on domain shift benchmarks while maintaining computational efficiency, and is particularly effective for detailed structures under limited GPU memory.

C.4.4 PiPa

PiPa[24] introduces a unified pixel- and patch-level self-supervised learning framework for UDA in semantic segmentation. Unlike traditional UDA methods that only minimize inter-domain discrepancies, PiPa focuses on enhancing intra-domain consistency by modeling pixel-level intra-class compactness and patch-level context-invariant semantics. This dual-level supervision enables the model to learn more robust and domain-invariant representations. PiPa achieves competitive results on standard UDA benchmarks and can be flexibly combined with other UDA methods to further improve performance without increasing model complexity.

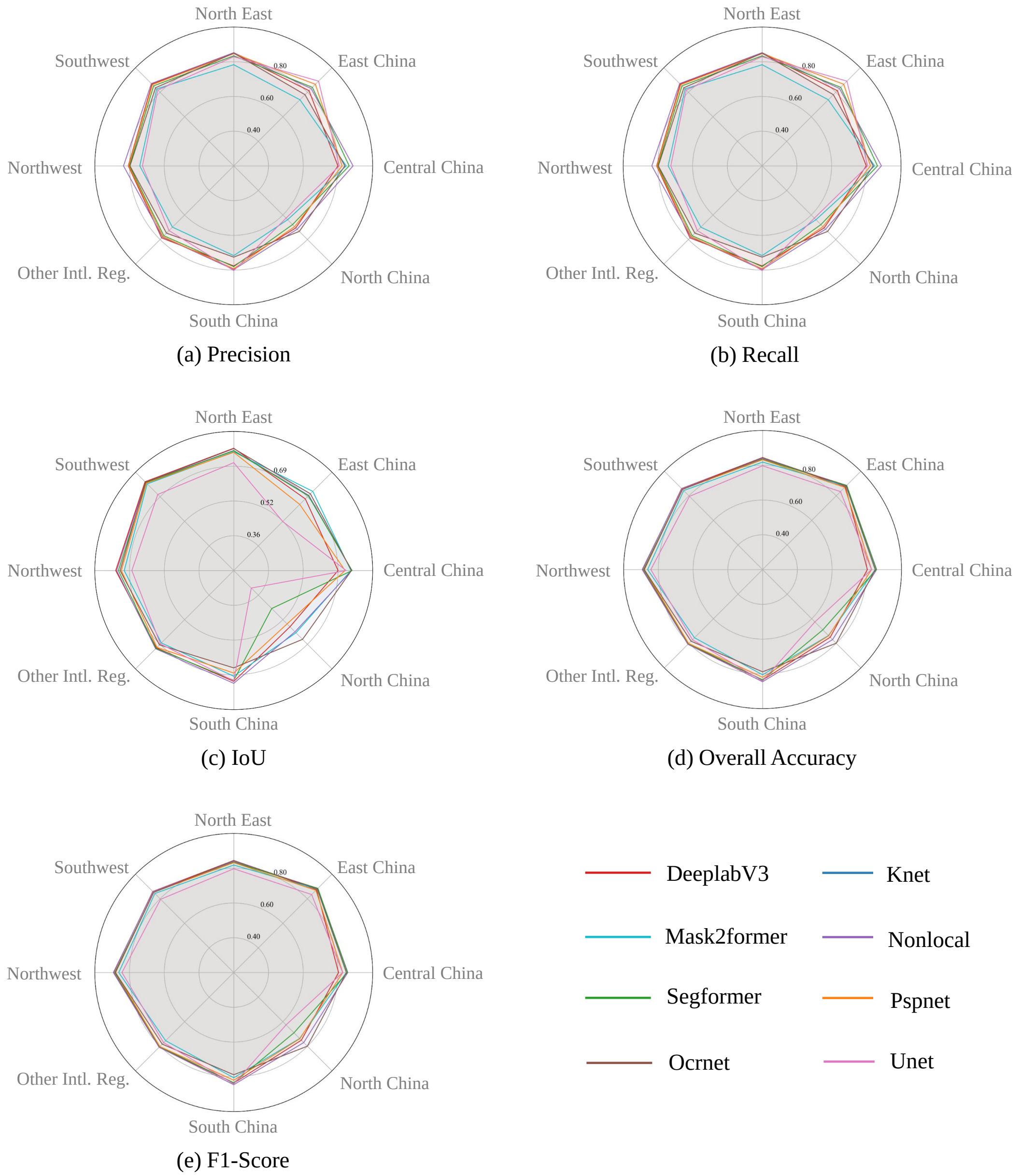
D Evaluation metrics

D.1 Pixel-level evaluation metrics

To evaluate segmentation accuracy on the GTPBD dataset, we adopt five standard pixel-level metrics commonly used in semantic segmentation and unsupervised domain adaptation (UDA) tasks, including Precision **(Prec.)**, Recall **(Rec.)**, Intersection over Union **(IoU)**, Overall Accuracy **(OA)** and **F1-score**.

Precision (Prec.) measures the proportion of correctly predicted agricultural pixels among all pixels predicted as agricultural. A higher precision indicates fewer false positives in land parcel classification.

$$Prec. = \frac{TP}{TP + FP} \tag{1}$$



(a) Precision

(b) Recall

(c) IoU

(d) Overall Accuracy

(e) F1-Score

Figure 12: Regional Segmentation Performance Across Models

where TP and FP indicate true positive and false positive, respectively. TP indicates the number of pixels correctly identified as agricultural parcels, while FP indicate the number of pixels mis-identified as agricultural parcels (i.e., mistakes).

Recall (Rec.) captures the proportion of true agricultural pixels that are correctly identified. High recall reflects a model's ability to minimize missed detections.

$$Rec. = \frac{TP}{TP + FN} \tag{2}$$

where FN indicates false negative and the number of pixels mis-identified as non-agricultural parcels (i.e., omissions).

Intersection over Union (IoU) quantifies the spatial agreement between the predicted and ground-truth regions. IoU is widely adopted in segmentation benchmarks due to its robustness to class

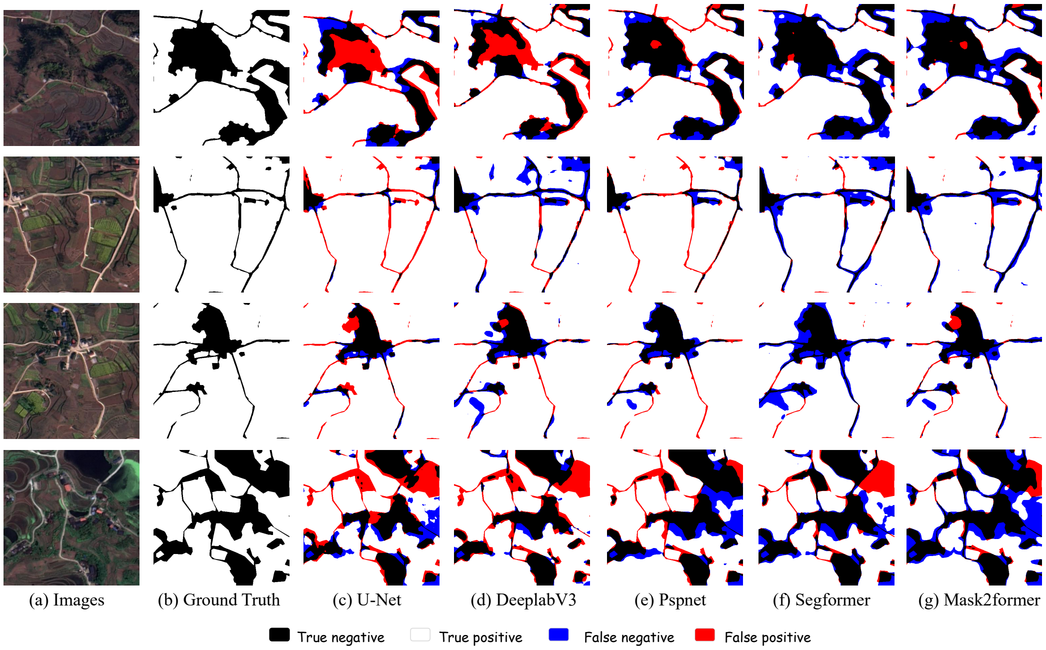

(a) Images (b) Ground Truth (c) U-Net (d) DeeplabV3 (e) Pspnet (f) Segformer (g) Mask2former

Figure 13: Error Map Comparison of different Semantic Segmentation Models

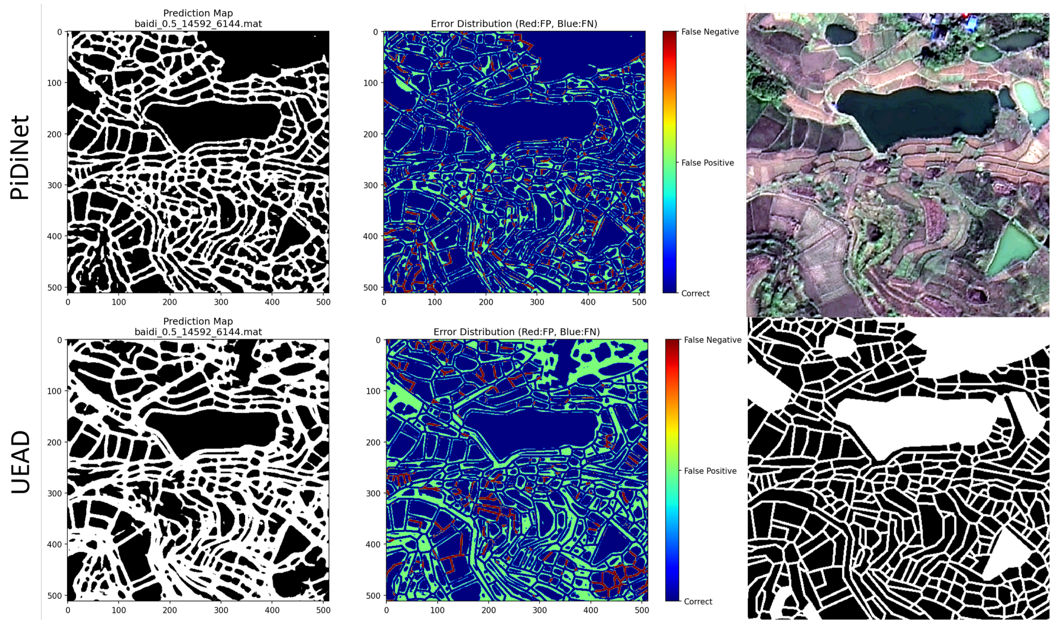


Figure 14: Qualitative edge detection comparisons between PiDiNet [31] and UEAD [56].

imbalance.

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

Overall Accuracy (OA) calculates the proportion of correctly classified pixels across the entire image. OA provides a global view of model performance and is especially informative for datasets with class imbalance.

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

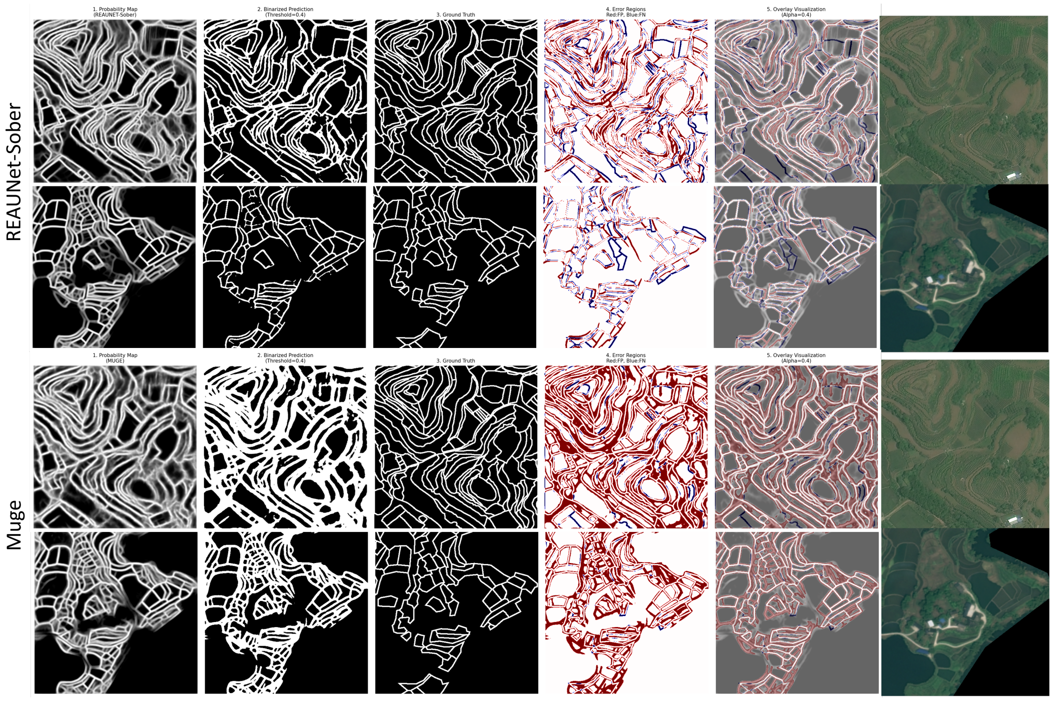

REAUNet-Sober
1. Probability Map (REAUNET-Sober)
2. Binarized Prediction (Threshold=0.4)
3. Ground Truth
4. Error Regions Red:FP, Blue:FN
5. Overlay Visualization (Alpha=0.4)
Muge
1. Probability Map (MUGE)
2. Binarized Prediction (Threshold=0.4)
3. Ground Truth
4. Error Regions Red:FP, Blue:FN
5. Overlay Visualization (Alpha=0.4)

Figure 15: Qualitative edge detection comparisons between REAUNet-Sober [22] and MuGE [55].

where TN indicates true negative and the number of pixels correctly identified as non-agricultural parcels.

F1-score is the harmonic mean of precision and recall, offering a balanced evaluation metric particularly useful under domain shift conditions or in presence of noise and uncertainty.

$$F1 = \frac{2 \cdot Prec. \times Rec.}{Prec. + Rec.} \tag{5}$$

D.2 Object-level geometric metrics

To comprehensively evaluate the geometric quality of delineated agricultural parcels, we adopt three object-level geometric metrics: Global Over-Classification Error (GOC), Global Under-Classification Error (GUC), and Global Total Classification Error (GTC). These indicators quantify segmentation accuracy in terms of spatial overreach, omission, and overall geometric consistency.

Let S_i denote the i-th predicted parcel (segmentation), and let O_i represent the ground truth parcel that has the largest intersection area with S_i. Denote m as the number of predicted parcels. The object-wise evaluation is defined as follows:

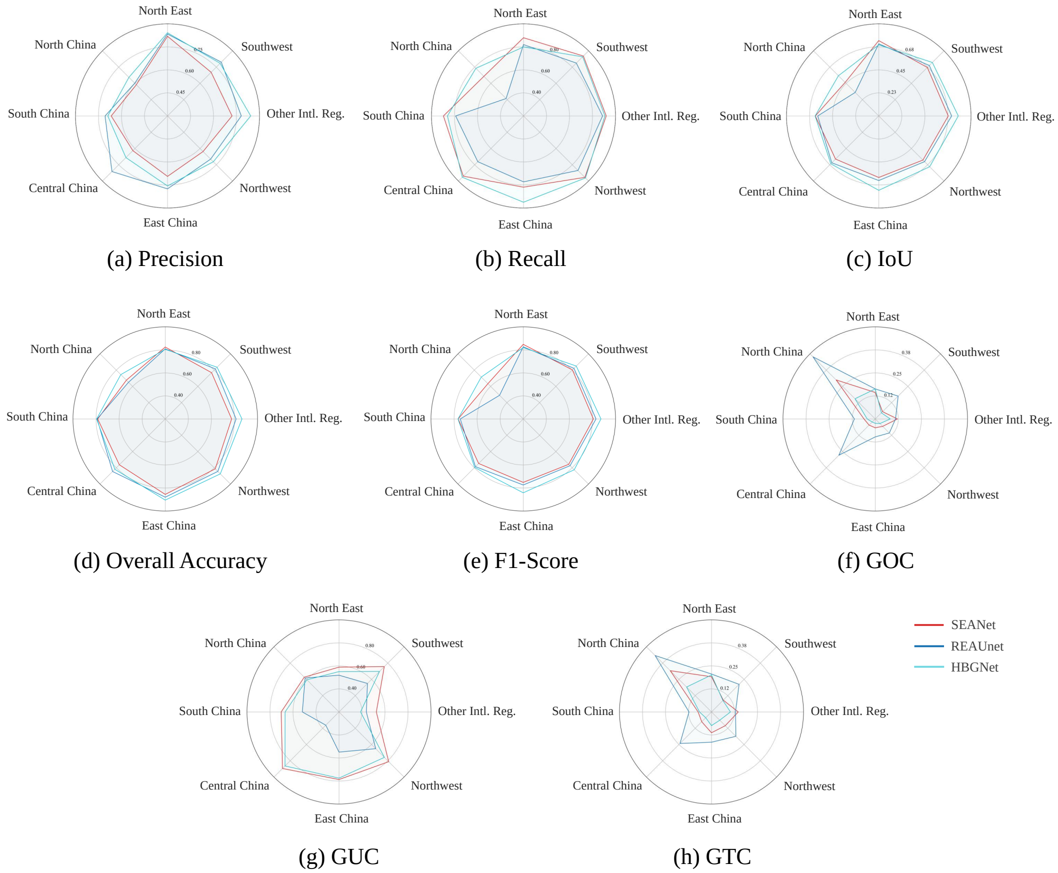
Global Over-Classification Error (GOC) measures the average extent to which predicted parcels exceed the spatial extent of their matched ground truth objects:

$$\mathrm{OC}(S_i) = 1 - \frac{\mathrm{area}(S_i \cap O_i)}{\mathrm{area}(O_i)}, \tag{6}$$

$$\mathrm{GOC} = \sum_{i=1}^{m} \left(\mathrm{OC}(S_i) \cdot \frac{\mathrm{area}(S_i)}{\sum_{k=1}^{m} \mathrm{area}(S_k)} \right), \tag{7}$$

where area$(\cdot)$ denotes the number of pixels in the respective region.

Global Under-Classification Error (GUC) quantifies the proportion of each predicted parcel not covered by the corresponding ground truth:



(a) Precision (b) Recall (c) IoU

(d) Overall Accuracy (e) F1-Score (f) GOC

(g) GUC (h) GTC

Figure 16: Regional Benchmark of Agricultural Parcel Extraction Models across Eight Metrics

$$\text{UC}(S_i) = 1 - \frac{\text{area}(S_i \cap O_i)}{\text{area}(S_i)}, \tag{8}$$

$$\text{GUC} = \sum_{i=1}^{m} \left(\text{UC}(S_i) \cdot \frac{\text{area}(S_i)}{\sum_{k=1}^{m} \text{area}(S_k)} \right). \tag{9}$$

Global Total Classification Error (GTC) synthesizes both over- and under-classification errors into one holistic metric using a root-mean-square formulation:

$$\text{TC}(S_i) = \sqrt{\frac{\text{OC}(S_i)^2 + \text{UC}(S_i)^2}{2}}, \tag{10}$$

$$\text{GTC} = \sum_{i=1}^{m} \left(\text{TC}(S_i) \cdot \frac{\text{area}(S_i)}{\sum_{k=1}^{m} \text{area}(S_k)} \right). \tag{11}$$

D.3 Edge detection evaluation metrics

For edge detection tasks on the GTPBD dataset, we evaluate model performance using three widely adopted metrics: Optimal Dataset Scale F1-score (ODS), Optimal Image Scale F1-score (OIS), and Average Precision (AP). These metrics jointly assess the accuracy, adaptability, and robustness of predicted boundaries.

Let P_t and R_t denote precision and recall computed at threshold t, and let F_t be the corresponding F1-score:

$$F_t = \frac{2 \cdot P_t \cdot R_t}{P_t + R_t}. \tag{12}$$

Optimal Dataset Scale F1-score (ODS) evaluates the global performance of an edge detector across the entire dataset using a single optimal threshold t^*:

$$\text{ODS} = \max_{t \in \mathcal{T}} \left(\frac{2 \cdot P_t^{\text{dataset}} \cdot R_t^{\text{dataset}}}{P_t^{\text{dataset}} + R_t^{\text{dataset}}} \right), \tag{13}$$

where P_t^{dataset} and R_t^{dataset} are aggregated precision and recall over the full dataset under threshold t.

Optimal Image Scale F1-score (OIS) computes the mean of the per-image best F1-scores, reflecting local threshold adaptiveness:

$$\text{OIS} = \frac{1}{N} \sum_{i=1}^{N} \max_{t \in \mathcal{T}} \left(\frac{2 \cdot P_t^{(i)} \cdot R_t^{(i)}}{P_t^{(i)} + R_t^{(i)}} \right), \tag{14}$$

where $P_t^{(i)}$ and $R_t^{(i)}$ denote the precision and recall on the i-th image under threshold t, and N is the total number of images.

Average Precision (AP) is computed as the area under the precision–recall curve:

$$\text{AP} = \int_0^1 P(R), dR, \tag{15}$$

where $P(R)$ is the precision as a function of recall, evaluated across all thresholds.

Dataset	Model	Pixel-level					Edge-level		Object-level		
		Prec.↑	Rec.↑	IoU↑	OA↑	F1↑	OIS↑	ODS↑	GOC↓	GUC↓	GTC↓
FHAPD[51]	UNet	**91.03**	95.03	86.89	90.55	92.99	**81.40**	**70.39**	5.90	**41.97**	**29.97**
	DeepLabV3	90.83	**97.06**	**88.40**	91.60	**93.84**	66.45	56.50	**2.79**	51.24	36.29
	PSPNet	88.67	95.63	85.22	89.06	92.02	55.48	44.06	3.76	55.01	38.99
	SegFormer	90.40	94.61	85.97	89.82	92.46	68.97	57.37	4.80	50.09	35.58
	Mask2Former	90.42	94.51	85.90	89.77	92.42	74.92	64.05	5.66	44.87	31.98
AI4Boundaries (Ortho)[10]	UNet	81.50	78.19	66.40	84.64	79.81	28.75	20.86	31.34	38.92	35.33
	DeepLabV3	81.08	82.31	69.05	85.67	81.69	32.51	22.05	19.46	29.43	24.95
	PSPNet	83.60	82.54	71.04	86.93	83.07	36.43	24.11	21.63	34.25	28.64
	SegFormer	**87.37**	78.45	70.46	87.23	82.67	29.88	20.14	20.57	26.32	23.62
	Mask2Former	84.58	**83.49**	**72.46**	**87.68**	**84.03**	**38.44**	**27.04**	**12.76**	**25.80**	**20.35**
FTW[18]	UNet	83.90	79.24	68.79	90.78	81.51	**73.33**	**62.39**	25.81	34.36	30.39
	DeepLabV3	85.92	**84.02**	**73.86**	**92.37**	**84.96**	70.38	57.28	**15.63**	38.30	29.18
	PSPNet	82.95	81.92	70.11	91.05	82.43	65.92	52.63	18.17	42.23	32.51
	SegFormer	83.47	80.05	69.09	90.82	81.72	68.56	56.70	18.64	38.66	30.35
	Mask2Former	**86.06**	77.88	69.13	91.08	81.75	72.15	61.04	20.44	**30.88**	**26.19**
GTPBD(ours)	UNet	**74.11**	54.93	46.09	75.46	63.09	22.47	15.17	42.69	**37.27**	37.84
	DeepLabV3	69.64	73.45	57.04	71.58	78.28	20.08	13.53	21.38	45.59	35.61
	PSPNet	68.33	72.41	54.22	76.65	70.31	19.66	13.08	21.58	50.06	39.21
	SegFormer	74.45	69.07	55.84	78.14	71.66	22.97	15.70	26.10	41.44	**34.63**
	Mask2Former	71.22	**74.33**	**57.16**	**78.73**	**72.74**	**25.56**	**17.59**	**21.01**	45.26	35.28

Table 14: Semantic segmentation results across datasets and models. Pixel-level metrics include Precision, Recall, IoU, Overall Accuracy (OA) and F1-Score. Edge-level metrics include OIS and ODS. Object-level metrics are GOC, GUC, and GTC.

E More results

E.1 More results on Semantic Segmentation

E.1.1 Results on Multiple Datasets

We provide detailed semantic segmentation results of five representative models (UNet [29], DeepLabV3 [6], PSPNet [52], SegFormer [41], and Mask2Former [8]) on four benchmark datasets: FHAPD [51], AI4Boundaries (Ortho) [10], FTW [18], and the proposed GTPBD dataset. The evaluation considers three categories of metrics, namely pixel-level, edge-level, and object-level, as summarized in Table 14.

Several key observations can be drawn from the results. On GTPBD and AI4Boundaries, Mask2Former achieves the best performance across almost all metrics. On FTW and FHAPD, DeepLabV3 consistently outperforms other models, while UNet also demonstrates competitive boundary detection performance. Notably, GTPBD yields the lowest scores across all models, confirming its higher difficulty and greater diversity compared to existing datasets. These findings highlight that no single model is universally superior across all benchmarks, and that our proposed dataset presents greater challenges for semantic segmentation.

E.1.2 Discussions the results for different regions

The radar plots in Fig. 12 compare five key metrics—Precision, Recall, IoU, Overall Accuracy, and F1-Score—across seven geographic regions for each segmentation architecture. Overall, all models achieve similar performance profiles, though U-Net demonstrates relatively average performance in terms of IoU and F1-Score, trailing behind advanced transformer-based architectures such as Mask2Former across most regions.

E.1.3 More visualization cases

Fig. 13 shows error maps for each model on the same test region, where black denotes true negatives, white true positives, blue false negatives, and red false positives. This visualization highlights that transformer-based methods such as Mask2Former [8] produce fewer false negatives in complex boundary areas. However, overall, the models struggle to accurately segment very fine-grained background details.

E.2 More visualization results on edge detection

Fig. 14 and Fig. 15 present qualitative comparisons of edge detection results on representative regions using four models: UEAD [56], PiDiNet [31], MuGE [55], and REAUNet-Sober [22]. From the visualizations, we observe that UEAD and MuGE tend to produce wider and blurred boundaries, indicating over-smoothed predictions that reduce localization precision. This issue is especially pronounced in densely terraced landscapes, where precise boundary localization is critical. In contrast, PiDiNet and REAUNet-Sober generate sharper and more compact edges, with REAUNet-Sober showing the best alignment with ground truth boundaries across complex spatial structures.

E.3 Discussions the results for different regions for terraced parcel extraction

Fig. 16 compares three parcel-extraction networks—SEANet, REAUNet, and HBGNet—over eight performance metrics (Precision, Recall, IoU, Overall Accuracy, F1-Score, GOC, GUC, and GTC) across eight geographic zones. Overall, performance variance is most pronounced in Southwest and Other International Regions, underscoring the challenge posed by highly heterogeneous terrace morphology.

E.4 Visualization cases on domain adaptation

Fig. 17 illustrate the segmentation results of four UDA methods (FDA, DAFormer, HRDA, and PiPa) on six representative domain transfer tasks, with transfers targeting the **North** (N), **South** (S), and

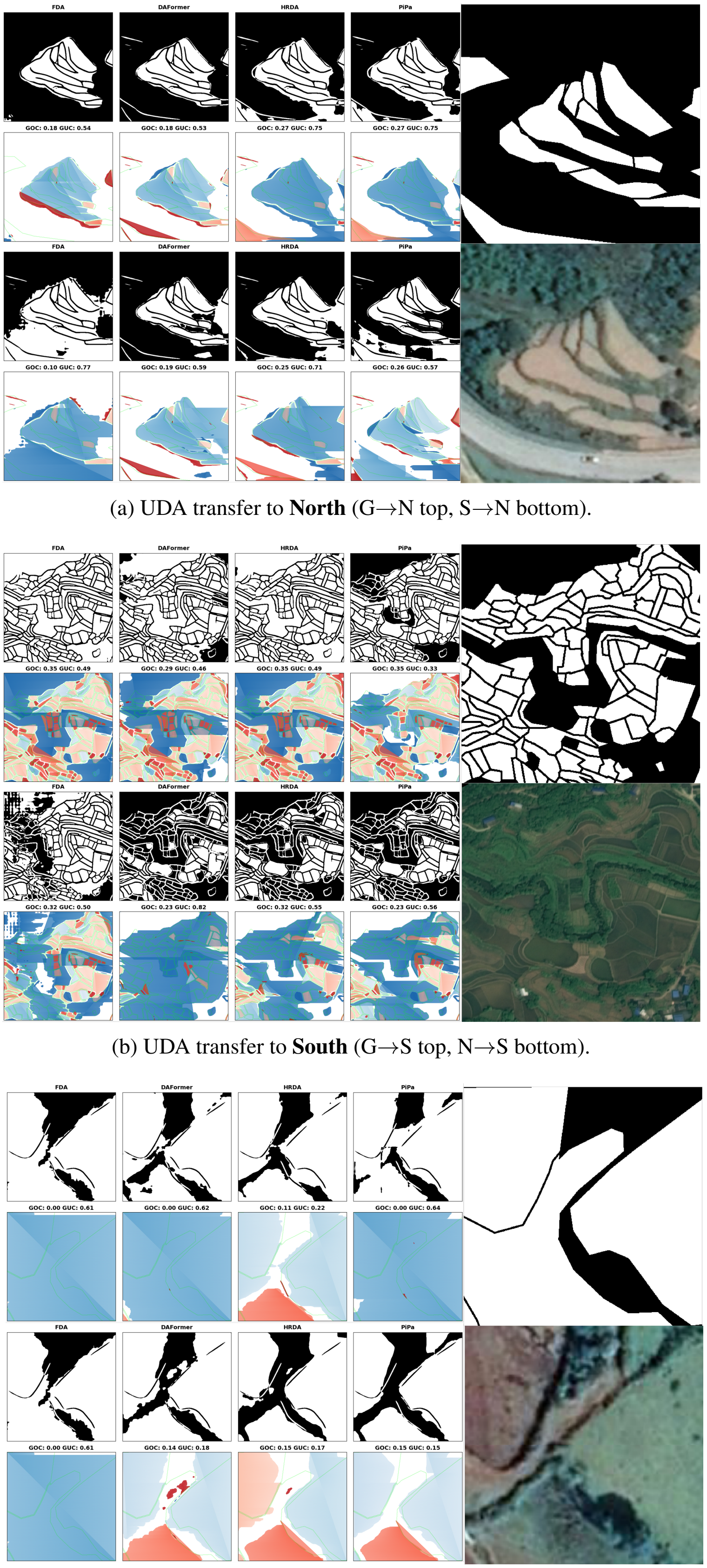
Global (G) domains, respectively. The results show that HRDA and PiPa generally produce finer boundaries and better preserve parcel structures under domain shift.

F Limitations and future work

Although GTPBD encompasses major terraced regions worldwide, the current collection is limited to fourteen well-known terraced areas in countries where terracing is concentrated. Some atypical terraced regions in other countries are not included due to their sparse distribution. The dataset currently consists only of optical images and does not include other remote sensing modalities such as multispectral or infrared imagery.

From the perspective of model training, GTPBD also presents several limitations. First, terraced parcels are highly fragmented and vary significantly in scale, which leads to class imbalance and ambiguous boundaries that pose challenges to current segmentation models. Second, although annotations were carefully generated, large-scale manual delineation inevitably introduces labeling noise, particularly in occluded or irregular regions. Third, the dataset mainly consists of single-season optical imagery, which limits its ability to capture temporal diversity; models trained under such conditions may underperform when applied to multi-season or multi-regional scenarios. Finally, while the resolution of 0.5–0.7 m is generally sufficient, extremely narrow ridges or complex land-cover mosaics remain difficult to capture and segment accurately.

In future work, we will expand the dataset by incorporating additional atypical terrace samples and broadening global coverage. Furthermore, GTPBD will be supplemented with multimodal geographical data, including Digital Elevation Models (DEMs), slope gradients, and spatiotemporal sequences. We also plan to construct question–answer pairs to facilitate integration with multimodal large language models. Finally, since the dataset already covers the primary terrace types, future tasks may focus on domain generalization and adaptation, enabling models not only to delineate known terraces but also to discover and adapt to new or atypical terrace types.



(a) UDA transfer to **North** (G→N top, S→N bottom).

(b) UDA transfer to **South** (G→S top, N→S bottom).

(c) UDA transfer to **Global** (S→G top, N→G bottom).

Figure 17: Qualitative results of UDA transfer to three domains: North, South, and Global. Models compared: FDA, DAFormer, HRDA, and PiPa.