

Fields of The World: A Machine Learning Benchmark Dataset For Global Agricultural Field Boundary Segmentation

Hannah Kerner^{1*}, Snehal Chaudhari^{1*}, Aninda Ghosh^{1*}, Caleb Robinson^{2*}, Adeel Ahmad^{3,4}, Eddie Choi⁴, Nathan Jacobs⁴, Chris Holmes⁵, Matthias Mohr⁵, Rahul Dodhia², Juan M. Lavista Ferres², Jennifer Marcus⁵

¹Arizona State University, Tempe, AZ 85281 USA, hkerner@asu.edu

²Microsoft AI for Good Research Lab, Redmond, WA 98052 USA

³Taylor Geospatial Institute, St Louis, MO 63108 USA

⁴Washington University in St Louis, St Louis, MO 63130 USA

⁵Taylor Geospatial Engine, St Louis, MO 63130 USA

Abstract

Crop field boundaries are foundational datasets for agricultural monitoring and assessments but are expensive to collect manually. Machine learning (ML) methods for automatically extracting field boundaries from remotely sensed images could help realize the demand for these datasets at a global scale. However, current ML methods for field instance segmentation lack sufficient geographic coverage, accuracy, and generalization capabilities. Further, research on improving ML methods is restricted by the lack of labeled datasets representing the diversity of global agricultural fields. We present Fields of The World (FTW)—a novel ML benchmark dataset for agricultural field instance segmentation spanning 24 countries on four continents (Europe, Africa, Asia, and South America). FTW is an order of magnitude larger than previous datasets with 70 462 samples, each containing instance and semantic segmentation masks paired with multi-date, multi-spectral Sentinel-2 satellite images. We provide results from baseline models for the new FTW benchmark, show that models trained on FTW have better zero-shot and fine-tuning performance in held-out countries than models that aren't pre-trained with diverse datasets, and show positive qualitative zero-shot results of FTW models in a real-world scenario – running on Sentinel-2 scenes over Ethiopia.

Code — <https://github.com/fieldsoftheworld/ftw-baselines>

Datasets — <https://beta.source.coop/repositories/kerner-lab/fields-of-the-world/>

1 Introduction

Crop field boundary datasets are urgently needed in agricultural monitoring, sustainable agriculture, and development applications (Nakalembe and Kerner 2023). However, these datasets do not exist for most of the world. Automatic field delineation in globally available satellite imagery offers a promising solution, but semantic reasoning about globally diverse agricultural landscapes in satellite imagery remains

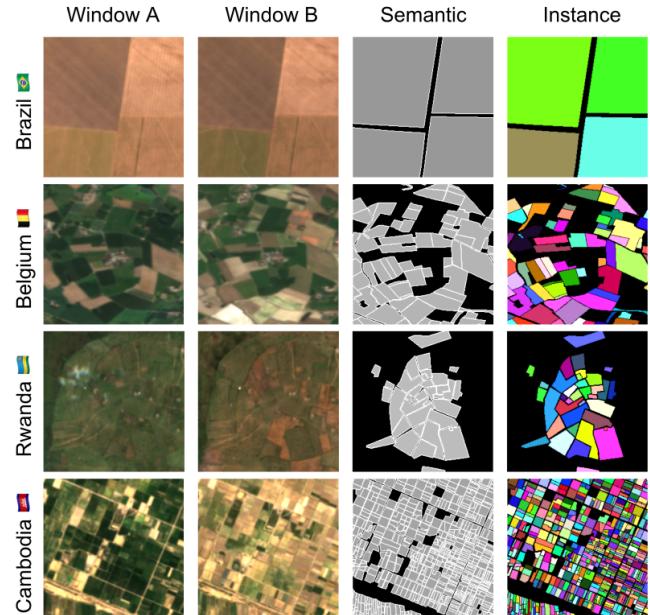


Figure 1: Training samples from four continents, demonstrating the diversity within Fields of The World.

challenging. Field morphologies, agricultural practices, and climate patterns vary greatly across the world. For example, average field sizes range from 1.6 hectares (ha) in Sub-Saharan Africa to 121 ha in North America (Debats et al. 2016). Motivated by this global diversity, we present a novel dataset for agricultural field instance segmentation: Fields of The World (FTW). FTW spans diverse landscapes in 24 countries on 4 continents (Europe, Asia, Africa, and South America). Fields of The World aims to catalyze field instance segmentation research and enable consistent, granular evaluation of different modeling approaches.

Field boundary datasets enable field-scale monitoring of crop conditions, yield, pest/diseases, farming practices, resource utilization, and other agricultural characteris-

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tics (Nakalembe and Kerner 2023). They are also in high demand for conservation and climate change policies and programs that require Measurement, Reporting, and Verification (MRV) of greenhouse gas emissions, carbon sequestration, and sustainable land management practices, such as the European Union Deforestation Regulation (European Parliament and Council of the European Union 2023). Field boundaries also simplify challenging tasks like crop-type classification by enabling classification at the object level rather than over individual pixels (Garnot, Landrieu, and Chehata 2022). Statistics agencies use field boundaries for ground-based survey design (Nakalembe and Kerner 2023). Field boundary maps over multiple years enable analyses of environmental and socioeconomic land change dynamics such as aggregation or fragmentation of farm parcels over time (Estes et al. 2022; Sullivan et al. 2023).

Previous work has demonstrated good performance for field instance segmentation in European countries, enabled by a combination of novel algorithms and labeled datasets (e.g., Wang, Waldner, and Lobell (2022); Sainte Fare Garnot and Landrieu (2021)). Research progress has been driven by benchmark datasets such as PASTIS (Sainte Fare Garnot and Landrieu 2021; Garnot, Landrieu, and Chehata 2022), AI4Boundaries (d’Andrimont et al. 2023), and AI4FoodSecurity (Planet, TUM, DLR and Radiant Earth 2021). While these datasets have been critical research catalysts, they do not fully capture the diversity and complexity of global agricultural landscapes. Existing datasets have limited geographic diversity, with labels concentrated in a handful of (usually European) countries (Table 2).

Fields of The World captures greater geographic diversity, morphological diversity, and agro-climatic diversity than any previous dataset. It includes fields of various sizes (Figure 3), shape, and orientation (Figure 2). FTW is also an order of magnitude larger than previous datasets, with 70 462 samples covering a total geographic area of 166 293 km².

Fields of The World provides harmonized ML-ready inputs from the optical Sentinel-2 satellite. Each example in Fields of The World includes four spectral channels (red, green, blue, and near-infrared) from two contrasting dates. Labels include instance and semantic segmentation masks. We provide the polygon label annotations in a standardized format using the *fiboa* (field boundaries for agriculture) specification (fiboa contributors 2024). This makes previously siloed datasets interoperable and enables users to obtain custom satellite data or other inputs corresponding to geo-referenced labels. The provided metadata also allows users to easily subset the dataset depending on their needs (e.g., commercial license or specific location).

We include training, validation, and test sets for each country, using existing splits when possible to maximize compatibility with existing work. We propose benchmark tasks that mimic real-world scenarios relevant to downstream users of field boundary datasets (e.g., region-specific evaluation, transfer learning, and zero-shot generalization). Finally, we perform experiments to demonstrate the value of the FTW dataset and provide baseline results for benchmark tasks. We release code via Github, data via Source Cooperative, and data loaders and pre-trained models via TorchGeo.

2 Dataset Description

Annotations

Field boundary representations Field boundary annotations are typically in the form of geo-referenced polygons. Since field boundaries at the same location may change across growing seasons, these polygons should also be temporally referenced to specify when the boundary is valid. Field polygons may be farmer-reported, manually drawn on high-resolution satellite images with GIS software, or recorded by walking the field perimeter with a handheld location-recording device. Polygons can then be paired with satellite data from the same location and time.

We conducted a comprehensive search for field polygons from government databases, published literature, and other websites. We looked for datasets with diverse geographic coverage, high-quality and trustworthy polygon annotations, and licenses that permit reuse. We included all datasets meeting these criteria in FTW. We considered author-reported quality assessment in each dataset’s documentation, previous use of the dataset in ML analyses, and visual inspection (e.g., closed polygons, polygons consistent with satellite images from reported dates, etc). Table 1 lists the 24 source datasets selected for Fields of The World.

Presence/absence labels *Presence/absence labels* are binary labels providing information about both the occurrence and non-occurrence of a phenomenon across sampled locations or time periods—for example, the presence of a field boundary or its absence. Most of the datasets in Fields of The World have presence/absence labels. However, some have *presence-only* labels, i.e., they are partially labeled. These indicate the presence of some, but not necessarily all, field boundaries in the sampled locations. Some pixels in presence-only label masks have unknown labels that might be labeled as background. Partial labels are a common challenge in field boundary segmentation (Wang, Waldner, and Lobell 2022). The Rwanda example in Figure 1 (row 3) illustrates presence-only labels while the Cambodia example (row 4) illustrates presence/absence labels.

Semantic filtering We focused Fields of The World on field boundaries for annual crops. Annual crops, also called temporary crops, are planted, grown, and harvested within a single growing season or year. Common examples include wheat, rice, maize, soybeans, and barley. This does not include permanent or perennial crops, which are cultivated for longer than one year and are not replanted annually, such as fruit trees, nut trees, and some grasses. We also excluded parcels used for pasture, fallow land, or other non-crop agricultural activities, such as grazing, orchards, vineyards, and forestry. If a dataset included parcels that were not active annual crops, we filtered them out (details in supplement).

Sample grids Many datasets in Table 1, particularly those sourced from the European Union government websites and EuroCrops (Schneider and Körner 2022), have millions of dense annotations spanning the entire country. Including all of these annotations in Fields of The World would bias the dataset toward large European countries. We sub-sampled these datasets by: 1) creating a bounding box enclosing the

Table 1: Key dataset details for each country in FTW, with green and brown cells indicating Windows A and B respectively.

Country	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Presence/absence labels		# train, val, test	Source
													Sub-sampled?	# train, val, test		
Austria														✓	5303, 637, 745	Schneider and Körner (2022)
Belgium														✓	1554, 189, 198	Schneider and Körner (2022)
Cambodia														✗	245, 27, 25	Persello et al. (2023)
Croatia														✓	2778, 351, 353	ARKOD (2024)
Denmark														✓	2868, 360, 332	Ministry of Food, Agriculture and Fisheries of Denmark (2021)
Estonia														✓	5348, 681, 684	Schneider and Körner (2022)
Finland														✓	4527, 550, 588	Finnish Food Authority (2021)
Corsica														✓	1974, 240, 258	The Service and Payment Agency (ASP) (2024)
France														✓	2773, 339, 396	The Service and Payment Agency (ASP) (2024)
Germany														✗	306, 30, 350	Kondmann et al. (2021)
Latvia														✓	5529, 668, 741	Schneider and Körner (2022)
Lithuania														✓	4208, 522, 528	Schneider and Körner (2022)
Luxembourg														✓	643, 81, 84	Administration of technical agricultural services (2024)
Netherlands														✓	3110, 381, 388	Netherlands Enterprise Agency (Government) (2021)
Portugal														✓	47, 9, 10	Instituto de Financiamento da Agricultura e Pescas (2024)
Slovakia														✓	3275, 390, 408	Slovak Republic Government (2024)
Slovenia														✓	1733, 216, 228	Schneider and Körner (2022)
South Africa														✗	590, 72, 85	Planet et al. (2021)
Spain														✓	2015, 201, 216	Schneider and Körner (2022)
Sweden														✓	3802, 442, 516	The Swedish Agency for Agriculture (2024)
Vietnam														✗	228, 36, 23	Persello et al. (2023)
Country	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Presence-only labels		# train, val, test	Source
													Sub-sampled?	# train, val, test		
Brazil														✗	1289, 130, 188	Oldoni et al. (2020)
India														✗	1261, 300, 399	Wang, Waldner, and Lobell (2023)
Kenya														✗	316, 20, 55	Pula Advisors (2022)
Rwanda														✗	57, 6, 7	NASA Harvest and Radiant Earth Foundation (2024)

entire dataset, 2) splitting the bounding box into a grid where each grid cell covered an area between 3300 to 5000 km², and 3) selecting 2-4 grid cells per country that captured a mixture of high-density and low-density agricultural areas.

For the eight datasets in Table 1 that were not sub-sampled, we used all field boundaries provided by the source dataset. Four of these datasets were published with predefined grids, which we used without modification: Germany (Kondmann et al. 2021), South Africa (Planet et al. 2021), Cambodia (Persello et al. 2023), and Vietnam (Persello et al. 2023). The Kenya (Pula Advisors 2022) and Brazil (Oldoni et al. 2020) label polygons were highly clustered but did not have predefined grids or clusters. We used k -means clustering to cluster the label polygons (using the center latitude/longitude of each polygon as the features). We defined a rectangular grid spanning the bounds of each cluster. We chose k by visually inspecting each dataset and balancing between over-clustering (resulting in high overlap between cluster grids) or under-clustering (re-

sulting in sparse grids with large unlabeled areas). In the India (Wang, Waldner, and Lobell 2023) and Rwanda (NASA Harvest and Radiant Earth Foundation 2024) datasets, polygon labels were in small clusters (e.g., 5 fields per cluster for India). We did not define grids for these datasets because we created sample chips directly from each small cluster.

Sample chip ROIs We tiled each sample grid into 1536m × 1536m sample patch ROIs (regions of interest), which we call ‘chips’. For India and Rwanda, we created a 1536m × 1536m chip around the center of each label cluster.

Metadata standardization We converted the label polygon datasets to the fiboa specification (fiboa contributors 2024). If a dataset was sub-sampled from a larger dataset, we only converted the sub-sampled version. Per the fiboa core specification, each dataset has per-polygon attributes `id` (unique field identifier), `determination_datetime` (last timestamp at which the field boundary existed/was observed), `area` (field area in hectares), and `geometry` (field

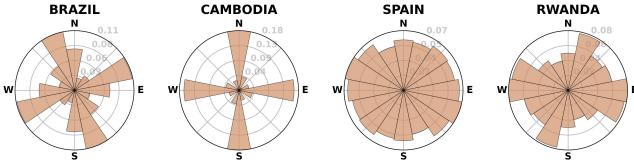


Figure 2: Field orientation histograms for selected countries.

polygon geometry; we use the WGS84/EPSC:4326 coordinate reference system). We included crop type or other attributes when available. Converted GeoParquet files are available on Source Cooperative at <https://beta.source.coop/repositories/kerner-lab/fields-of-the-world/>. The README for each dataset provides details including the source dataset license (extends to the FTW subset using it) and link. We provide GeoParquet files for the sample grids and chip ROIs.

Label masks Previous work explored several approaches to convert (“rasterize”) field label polygons to label masks, which are the ML prediction targets. Approaches include binary field extent masks (field interiors vs background), binary field boundary masks (field boundaries vs background), 3-class masks (field interiors, field boundaries, and background) (Taravat et al. 2021), and distance masks (to field centroids) (d’Andrimont et al. 2023). We provide binary field extent and 3-class semantic masks and instance masks.

Satellite data

We obtained multispectral Sentinel-2 satellite images using Microsoft Planetary Computer (Microsoft Open Source et al. 2022). Images in this catalog are processed to Level 2A (bottom-of-atmosphere) and stored in cloud-optimized GeoTIFF (COG) format. We used the red (B04), green (B03), blue (B02), and near-infrared (B08) spectral bands, all of which have spatial resolution of 10 m per pixel. We used Sentinel-2 because it is the highest-resolution optical satellite dataset that is freely accessible. In the FTW Github repository, we provide a CSV file containing the Sentinel-2 scene ID, cloud percentage, and date ranges for each sample.

Dates Previous work showed that contrasting images from different times during the same year improved crop field segmentation by highlighting the intra-annual variation characteristic of active crop fields (Estes et al. 2022; Debats et al. 2016). This contrast can help models rule out potential false positives such as fallow fields or forest stands.

For each country, we collected images from two date ranges (Table 1). Growing seasons vary greatly globally (and even within countries). To choose the date ranges, we looked at each country’s crop calendar which specifies the planting, mid-season, harvesting, and off-season months for the main crops (USDA Foreign Agricultural Service 2024). We inspected the available satellite images in the sample area(s) in two date ranges spanning the planting/mid-season and harvesting/off-season months. If a country had multiple growing seasons (e.g., winter and summer crops), we visualized both seasons and chose the one appearing most active. We then iteratively adjusted the date ranges to account for

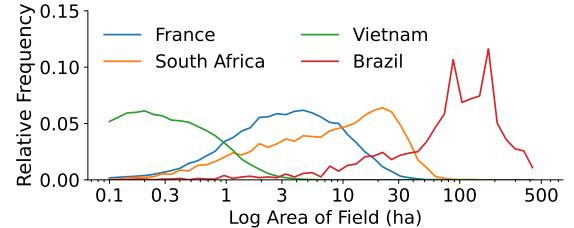


Figure 3: Field area distribution across four countries.

good contrast between images and cloud cover. After adjustment, the date ranges do not necessarily match the growing season stages, so we call them Window A and B.

Cloud filtering For each chip ROI, we searched for Sentinel-2 scenes with <90% scene level cloud cover in the two date ranges. For the resulting scenes, we cropped each scene to the sample chip ROI and computed the cloud percentage in the patch using the Sentinel-2 scene classification layer (SCL) “Cloud medium probability” and “Cloud high probability” classes. We selected the chip with the lowest cloud percentage. If there were no chips with cloud percentage < 10%, we discarded the chip. Table 1 gives the resulting number of Sentinel-2 chips created for each country.

Resizing and normalization We resized each chip to 256×256 pixels. Each chip is stored as a GeoTIFF with EPSG:4326. We normalize images during training by dividing each channel by 3000 (an approximate mean value).

Dataset splits

FTW defines training, validation, and test sets for each country to facilitate evaluation of test metrics at the country scale. Many sample chips are spatially adjacent since they were tiled from large grids. Spatial autocorrelation between adjacent chips may cause leakage between data subsets if chips are randomly split into subsets (Rolf 2023). To reduce the impact of spatial autocorrelation, we implemented a blocked random splitting strategy. We grouped chips into 3×3 blocks and randomly assigned 80% to training, 10% validation, and 10% test. To ensure 3×3 blocks were large enough, we performed an experiment to quantify the sensitivity of test performance to the block size and distance from each test patch to the nearest training patch (see supplement).

3 Dataset Analysis and Related Work

The dramatic differences in field morphology across the globe motivated the construction of FTW (Figure 1). FTW has significant advantages over previous field instance segmentation datasets in terms of (i) geographic representation and extent, (ii) annotation volume, and (iii) annotation/scene complexity. The most relevant datasets for comparison are AI4Boundaries (d’Andrimont et al. 2023), AI4SmallFarms (Persello et al. 2023), PASTIS (Sainte Fare Garnot and Landrieu 2021), and PASTIS-R (Garnot, Landrieu, and Chehata 2022). We only include datasets that

Table 2: Key attributes of Fields of The World and previous field boundary segmentation datasets.

Dataset	# countries	Sensor(s)	Input dim ([H, W, C, T])	# field polygons (million)	# total samples	Sample dim. (m × m)	Total area (km ²)
Fields of The World (FTW)	24	Sentinel-2	[256, 256, 4, 2]	1.63	70,484	1,536	166,293
AI4Boundaries (d'Andrimont et al. 2023)	7	Sentinel-2, Aerial	[256, 256, 4, 12] [512, 512, 3, 1]	1.07 ³	7,831 7,598	2,560 512	51,321 1,992
AI4SmallFarms (Persello et al. 2023)	2	Sentinel-2	[256, 256, 4, 1]	0.44	62	5,000	1,550
PASTIS (Sainte Fare Garnot and Landrieu 2021)	1	Sentinel-2	[128, 128, 10, N] ¹	0.12	2,433	1,280	3,986
PASTIS-R (Garnot, Landrieu, and Chehata 2022)	1	Sentinel-2, Sentinel-1	[128, 128, 10, N] ² [128, 128, 3, 70]	0.12	2,433	1,280	3,986

¹ Varying observations (38–61) taken between September 2018 and November 2019.

² All available observations for the 2019 season.

³ d'Andrimont et al. (2023) reports 2.5M parcels contained in 7,831 4-km samples, however the number of polygons included in dataset sample masks is smaller.

explicitly label individual field instances, excluding semantic segmentation labels since field instances enable a broader range of applications. We also exclude datasets providing field polygons but no imagery, such as the field boundary dataset created by the French Land Parcel Identification System (The Service and Payment Agency (ASP) 2024). The lack of standardized satellite imagery for polygon-only datasets hampers the creation and comparison of automated algorithms for field boundary delineation and related tasks.

Table 2 compares the key attributes for all datasets (see geographic distributions in the supplement). FTW has a significantly broader geographic distribution than all previous datasets, including fields from 24 countries spanning four continents (Europe, Asia, Africa, and South America). Previous datasets include at most 7 countries and are mostly concentrated in Europe, except AI4SmallFarms which only has images from Cambodia and Vietnam. FTW is the largest dataset in terms of number of samples, total area, and total annotations. FTW has an order of magnitude more samples and area covered than the next-largest dataset (AI4Boundaries (d'Andrimont et al. 2023)), with 70,462 sample chips spanning 166,293 km². It also has the highest annotation volume with 1.63M field polygons, compared to the next-largest volume of 1.07M in AI4Boundaries.

The FTW dataset captures greater morphological diversity of agricultural field instances than any other dataset. Figure 1 shows example field boundaries from four countries. The diverse shapes and sizes reflect the unique topographical, environmental, and historical factors that influenced the development of field boundaries in each region. Figure 3 shows the dramatic difference between the field areas in different countries. There is little overlap between the distributions of Vietnam (small fields) and Brazil (large fields). Figure 9 shows the distribution field areas for AI4SmallFarms, AI4Boundaries, and FTW (note that we did not include PASTIS/PASTIS-R since both provide field boundaries in raster, not vector, format so we could not compute morphological statistics). AI4SmallFarms consists mostly of small-area fields, AI4Boundaries consists mostly of medium-area fields, and FTW has a broader distribution of field areas. There is also significant diversity in field shape complexity (see visualizations in supplement). For example, Estonia and Slovakia have complex field shapes, with 111.9 and 95.4 polygon vertices on average, respectively. In con-

trast, Kenya and Rwanda have simpler field shapes, with 4.5 and 5.6 vertices on average, respectively.

FTW provides a more complete representation of the diversity and complexity of agricultural landscapes across the globe than previous datasets. We hope that FTW will lead to more broadly applicable models for field boundary segmentation by providing a large and diverse training dataset and enabling region-specific evaluation and error analysis.

4 Baseline Experiments

Setup and metrics We follow the common approach to field instance segmentation of segmentation then polygonization of predicted raster masks (Persello et al. 2023). Unless specified otherwise, we used a U-net with EfficientNet-b3 backbone with inputs consisting of concatenated 4-channel RGB-NIR images from Window A and B. We found that this architecture performed well compared to other architectures and backbones (see **Architectures** paragraph in this section and Table 7 in supplement). We also found that concatenating both temporal windows and using four spectral bands performed best compared to other configurations (see **Multi-temporal and multispectral channels** paragraph and Table 4). We initialized the RGB channels using ImageNet and NIR channels using random weights. We optimized cross-entropy loss with class weights inversely proportional to each class's frequency in the training set. We trained all models for 100 epochs. We used a fixed random seed for all experiments (randomly chosen). We did not perform hyperparameter tuning. Experiments required 4 A100 and 8 V100 GPUs for approximately one week.

We used semantic (pixel-level) and instance (object-level) segmentation metrics: pixel-level intersection over union (IoU), precision, and recall, and object-level precision and recall (functions in the FTW code repository). We converted segmentation masks to polygons using `rasterio` then computed object-level metrics with IoU threshold of 0.5.

Modeling configuration Fields of The World includes multiple target mask formats and satellite images from two dates and four spectral channels, giving users many modeling choices. In semantic segmentation followed by polygonization, there are other choices such as which architecture to choose. We performed several experiments to assess the impact of these choices on model performance. In each experiment, we evaluate performance using the test set for each

Table 3: Performance metrics for different target mask formats in Slovenia (SVN), France (FRA), and South Africa (ZAF). We compared 2-class field extent and 3-class masks with or without ignoring background (bg) pixels for presence-only samples.

Mask type	Pixel IoU			Pixel precision			Pixel recall			Object precision			Object recall		
	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF
2-class	0.66	0.83	0.83	0.84	0.87	0.87	0.76	0.95	0.94	0.52	0.64	0.54	0.06	0.29	0.24
2-class (ignore presence-only bg)	0.69	0.82	0.81	0.78	0.86	0.84	0.86	0.95	0.96	0.30	0.38	0.44	0.08	0.15	0.19
3-class	0.67	0.83	0.83	0.87	0.88	0.88	0.75	0.94	0.94	0.60	0.71	0.63	0.07	0.45	0.35
3-class (ignore presence-only bg)	0.59	0.79	0.79	0.90	0.89	0.89	0.63	0.88	0.88	0.33	0.55	0.51	0.20	0.58	0.55

Table 4: Ablation results for multispectral (RGB-NIR vs. RGB only) and multi-temporal (Window A and Window B) input channels in Slovenia (SVN), France (FRA), and South Africa (ZAF).

Channels	Pixel IoU			Pixel precision			Pixel recall			Object precision			Object recall		
	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF
Stacked Windows A and B	0.58	0.79	0.80	0.91	0.89	0.90	0.61	0.87	0.87	0.30	0.54	0.55	0.18	0.58	0.54
Stacked Windows A and B (RGB only)	0.58	0.78	0.79	0.90	0.89	0.89	0.62	0.86	0.88	0.27	0.51	0.53	0.18	0.56	0.54
Mean of Windows A and B	0.54	0.77	0.78	0.88	0.89	0.88	0.59	0.86	0.88	0.27	0.50	0.49	0.16	0.55	0.53
Window A only	0.55	0.77	0.78	0.88	0.88	0.88	0.59	0.86	0.87	0.27	0.47	0.47	0.17	0.54	0.52
Window B only	0.52	0.78	0.79	0.87	0.89	0.89	0.57	0.86	0.87	0.24	0.49	0.52	0.15	0.54	0.53

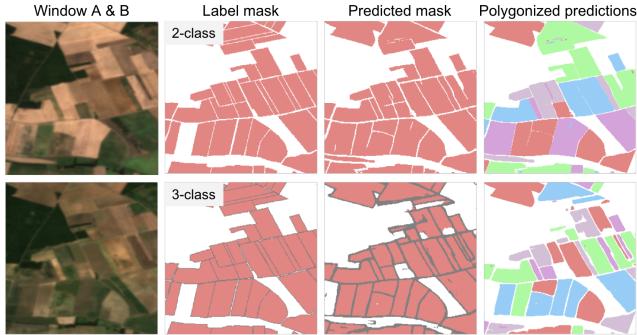


Figure 4: Example France predictions for 2-class and 3-class models in rows 2 and 4 of Table 3.

country. We report results for countries with presence/absence labels chosen to span a range of field sizes: Slovenia - SVN (average field size 0.64 ha), France - FRA (average 5.7 ha), and South Africa - ZAF (average 13.8 ha). We provide results for all countries in the supplementary material. An extensive search of modeling configurations is beyond the scope of this work, but we hope future studies will explore a greater range of options using the FTW dataset.

Target mask We evaluated the two types of target masks provided in FTW: 2-class (field interior vs. background) and 3-class (field interior, boundary, and background). In presence-only countries, pixels with an unknown class are labeled as background. We evaluated two scenarios: 1) when computing the loss, ignore all pixels labeled background for presence-only countries, and 2) compute the loss for all pixels treating unknown labels as background. Before computing test metrics, we converted outputs from all models to binary field extent masks to ensure a common evaluation basis.

Rows 1 and 3 of Table 3 show that almost all metrics are higher with 3-class masks than binary masks. Rows 2 and 4

compare when unknown-label pixels are counted as background or ignored for presence-only samples when training with 3-class masks. Although object precision is lower when ignoring unknown pixels, object recall is significantly higher, especially for Slovenia. Figure 4 shows that the segmentation masks predicted by both models are good, but 3-class masks improve the delineation of contiguous fields and thus object recall. From these results, we concluded that training with 3-class masks and ignoring presence-only background is most likely to give good performance across all regions and used this setup for subsequent experiments.

Multi-temporal and multispectral channels We did an ablation experiment to evaluate the benefit of the two contrasting image dates (Window A and B) in FTW. We also evaluated a mean of both windows. Finally, we evaluated with/without the NIR channel. Table 4 shows the best performance comes from both time windows and all spectral channels. Overall, removing one of the time windows causes a greater drop in metrics than removing the NIR channel. This is consistent with previous work that showed performance improvements were greater when adding more timesteps compared to more spectral channels (Debats et al. 2016).

Architecture We evaluated U-net (Ronneberger, Fischer, and Brox 2015) and DeepLabv3+ (Chen et al. 2018) models with 5 different backbones: ResNet-18, ResNet-50, ResNeXt-50, EfficientNet-b3, and EfficientNet-b4. Overall performance is similar across different architectures and backbones, though U-net models tend to outperform DeepLabv3+ models (results in supplement).

Transfer learning Some countries have many labels while others have few (Table 1). Prior work showed that performance on a data-scarce region could be improved by pre-training models on a country with a large labeled dataset and then fine-tuning on the target region. Wang, Waldner, and Lobell (2022) fine-tuned a model for India after pre-training

Table 5: Transfer learning results for models pre-trained on France (FRA) or Netherlands (NLD), AI4Boundaries countries (Austria/AUT, Spain/ESP, FRA, Luxembourg/LUX, NLD, Slovenia/SVN, and Sweden/SWE), or FTW minus the target region. Models are fine-tuned and tested on the target region. We report recall metrics only for India since it has presence-only labels. Each cell gives two results: no fine-tuning / after fine-tuning using the target training set.

Analogous work	Fine-tune/test	Pre-train	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Wang, Waldner, and Lobell (2022)	India	FRA	0.51 / 0.50	-	0.60 / 0.55	-	0.03 / 0.13
		AUT, ESP, FRA, LUX, NLD, SVN, SWE	0.16 / 0.54	-	0.16 / 0.59	-	0.05 / 0.16
		FTW – {India}	0.57 / 0.55	-	0.63 / 0.60	-	0.14 / 0.19
Persello et al. (2023)	Cambodia Vietnam	NLD	0.04 / 0.48	0.77 / 0.90	0.04 / 0.51	0.02 / 0.22	0.00 / 0.18
		AUT, ESP, FRA, LUX, NLD, SVN, SWE	0.16 / 0.51	0.90 / 0.90	0.16 / 0.54	0.12 / 0.27	0.04 / 0.22
		FTW – {Cambodia, Vietnam}	0.43 / 0.55	0.92 / 0.91	0.44 / 0.58	0.21 / 0.29	0.16 / 0.24

on France. Persello et al. (2023) fine-tuned for Cambodia and Vietnam after pre-training on the Netherlands.

Direct comparisons to these works are not possible due to differences in data formats. Instead, we performed three analogous experiments to evaluate the improvement of pre-training on FTW compared to smaller, more geographically limited datasets as in prior work: 1) pre-training on one data-rich country (France or Netherlands), 2) pre-training on the countries included in AI4Boundaries (to emulate pre-training on AI4Boundaries), and 3) pre-training on FTW with the target country held-out. We then fine-tuned each model for 200 epochs using the target country (India or Cambodia+Vietnam) training set and evaluated on its test set.

Table 5 shows that models pre-trained with FTW outperform models trained on more geographically limited subsets in both target regions. The performance of FTW models without any fine-tuning is especially impressive. FTW models with no fine-tuning perform similarly or better than fully fine-tuned versions of compared models.

Deployment readiness Motivated by the impressive performance of FTW pre-trained models without fine-tuning in Table 5, we used a FTW pre-trained model to predict field boundaries in Ethiopia, a challenging region not in FTW (Figure 5). The results show good qualitative performance that could be improved with local fine-tuning and post-processing. This shows the high potential of FTW to be immediately used in practice with little adaptation effort.

5 Discussion and Conclusion

ML research on automatic extraction of agricultural field boundaries from remotely sensed imagery is limited by a lack of ML-ready datasets to train and evaluate models on the global diversity of crop fields. These datasets are urgently needed in many applications for agriculture, climate change, and development. We designed Fields of The World to improve ML model performance for field boundary segmentation in diverse global agricultural landscapes and enable granular country-scale evaluation for more countries than any prior dataset. Our experiments established a performance baseline for the new FTW benchmark and showed that FTW-trained models perform better than more geographically limited datasets analogous to existing benchmarks.

Future work could build on these baselines by testing more model architectures, including instance segmentation architectures (e.g., Mask-RCNN (He et al. 2017) or

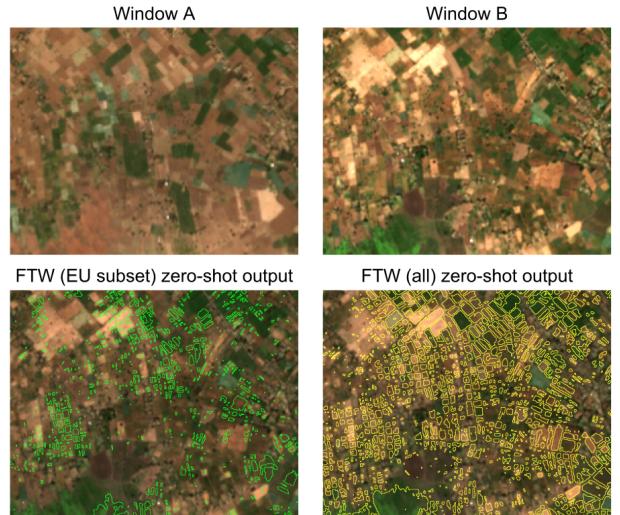


Figure 5: Zero-shot predictions with no post-processing for a 20 sq km region in Ethiopia (8°05'N, 38°51'E). A FTW pre-trained model achieves (qualitatively) good performance, even though Ethiopia is not in the FTW training set and is a challenging region to delineate fields.

SAM (Kirillov et al. 2023)) and geospatial foundation models (e.g., SatMAE (Cong et al. 2022) or Presto (Tseng et al. 2023)). Future work could also explore other methods of constructing target masks, motivated by our result that training with 3-class masks performed better than 2-class masks.

We provide complete metadata for sample grids, sample chips, and field boundary polygons to enable future extensions of Fields of The World. For example, future work could add more spectral channels or sensors, timesteps, or sample locations. FTW can be extended as field polygons become available for more countries. We hope the community will build on FTW as research on this important task grows.

Benchmarking on FTW We hope this study will inspire researchers to develop new methods for field boundary segmentation and measure improvement using the FTW benchmark. We suggest benchmarking performance on the per-country test sets and reporting individual country results, the mean across all countries, or the minimum across countries (worst-case performance). Supplement Table 13 reports these metrics for the best model evaluated in this paper.

Ethics statement

Researchers, practitioners, and other users of Fields of The World must be aware of important ethical considerations raised by the digitization of field boundaries and other information from publicly accessible satellite data. Digitized field boundary data could inadvertently expose the practices and characteristics of individual land parcels, which could infringe on the privacy of local landowners who may be unaware of this digitization or its implications. There are also risks that private or public entities may use digitized field boundary data in a way that marginalizes vulnerable individuals such as smallholder farmers. Rolf et al. (2024) summarized distinct ethical concerns of machine learning applied to satellite data. In line with the recommendations of Rolf et al. (2024), we suggest that users of FTW work with local organizations and communities to build and release responsible field boundary datasets and ensure their project goals and practices align with local needs and regulations.

Acknowledgments

This project was supported by funding from the Taylor Geospatial Engine and a NASA Supplemental Open Source Software Award.

References

- Administration of technical agricultural services. 2024. Flik Plot Repository - Open Data. <https://data.public.lu/en/datasets/referentiel-des-parcelles-flik/#resources>.
- ARKOD. 2024. Agency for Payments in Agriculture, Fisheries and Rural Development. <https://www.aprarr.hr/prostorni-podaci-servisi/>.
- Beck, H. E.; Zimmermann, N. E.; McVicar, T. R.; Vergopolan, N.; Berg, A.; and Wood, E. F. 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1): 1–12.
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.
- Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 197–211.
- d'Andrimont, R.; Claverie, M.; Kempeneers, P.; Muraro, D.; Yordanov, M.; Peressutti, D.; Batić, M.; and Waldner, F. 2023. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. *Earth System Science Data*, 15(1): 317–329.
- Debats, S. R.; Luo, D.; Estes, L. D.; Fuchs, T. J.; and Caylør, K. K. 2016. A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes. *Remote Sensing of Environment*, 179: 210–221.
- Estes, L. D.; Ye, S.; Song, L.; Luo, B.; Eastman, J. R.; Meng, Z.; Zhang, Q.; McRitchie, D.; Debats, S. R.; Muhando, J.; et al. 2022. High resolution, annual maps of field boundaries for smallholder-dominated croplands at national scales. *Frontiers in Artificial Intelligence*, 4: 744863.
- European Parliament; and Council of the European Union. 2023. Regulation (EU) 2023/1115 on Deforestation-free Products. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023R1115>. Accessed: 2024-08-08.
- fiboa contributors. 2024. Field Boundaries for Agriculture (fiboa) specification.
- Finnish Food Authority. 2021. Agricultural parcel containing spatial data. <https://www.ruokavirasto.fi/en/about-us/open-information/inspire/>.
- Garnot, V. S. F.; Landrieu, L.; and Chehata, N. 2022. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187: 294–305.
- Hall, J. V.; Argueta, F.; and Giglio, L. 2024. GloCAB crop-land field boundary dataset. *Data in Brief*, 55: 110739.
- He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.
- Instituto de Financiamento da Agricultura e Pescas. 2024. Sistema de Informação de Parcelas.
- Jung, S.; Rasmussen, L. V.; Watkins, C.; Newton, P.; and Agrawal, A. 2017. Brazil's national environmental registry of rural properties: implications for livelihoods. *Ecological Economics*, 136: 53–61.
- Kehs, A.; McCloskey, P.; Chelal, J.; Morr, D.; Amakove, S.; Plimo, B.; Mayieka, J.; Ntango, G.; Nyongesa, K.; Pamba, L.; et al. 2021. From village to globe: A dynamic real-time map of African fields through PlantVillage. *Frontiers in Sustainable Food Systems*, 5: 514785.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*.
- Kondmann, L.; Toker, A.; Rußwurm, M.; Camero, A.; Peressuti, D.; Milcinski, G.; Mathieu, P.-P.; Longépé, N.; Davis, T.; Marchisio, G.; et al. 2021. DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.
- Microsoft Open Source; McFarland, M.; Emanuele, R.; Morris, D.; and Augspurger, T. 2022. microsoft/PlanetaryComputer: October 2022. <https://doi.org/10.5281/zenodo.7261897>.
- Ministry of Food, Agriculture and Fisheries of Denmark. 2021. LandbrugsGIS. <https://landbrugsgeodata.fvm.dk/>.
- Nakalembe, C.; and Kerner, H. 2023. Considerations for AI-EO for agriculture in Sub-Saharan Africa. *Environmental Research Letters*, 18(4): 041002.
- NASA Harvest and Radiant Earth Foundation. 2024. Rwanda Field Boundary Competition. Accessed: 2024-08-07.

- Netherlands Enterprise Agency (Government). 2021. Dataset: Basic registration Crop plots (BRP). <https://www.pdok.nl/atom-downloadservices-/article/basisregistratie-gewaspercelen-brp->.
- Oldoni, L. V.; Sanches, I. D.; Picoli, M. C. A.; Covre, R. M.; and Fronza, J. G. 2020. LEM+ dataset: for agricultural remote sensing applications. Mendeley Data, V1.
- Persello, C.; Grift, J.; Fan, X.; Paris, C.; Hänsch, R.; Koeva, M.; and Nelson, A. 2023. AI4SmallFarms: A Data Set for Crop Field Delineation in Southeast Asian Smallholder Farms. *IEEE Geoscience and Remote Sensing Letters*.
- Planet; Foundation, R. E.; of Agriculture, W. C. D.; and (DLR), G. A. C. 2021. A Fusion Dataset for Crop Type Classification in Western Cape, South Africa (Version 1.0). <https://doi.org/10.34911/rdnt.gqy868>. Radiant MLHub. [Date Accessed].
- Planet, TUM, DLR and Radiant Earth. 2021. AI4FoodSecurity Challenge. <https://platform.ai4eo.eu/ai4food-security-south-africa/data>.
- Pula Advisors. 2022. Bird's Eye. https://ecass-project-documentation.readthedocs.io/en/latest/modules/about_Ecass.html.
- Rolf, E. 2023. Evaluation challenges for geospatial ML. *arXiv preprint arXiv:2303.18087*.
- Rolf, E.; Klemmer, K.; Robinson, C.; and Kerner, H. 2024. Position: Mission Critical–Satellite Data is a Distinct Modality in Machine Learning. In *Forty-first International Conference on Machine Learning*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.
- Sainte Fare Garnot, V.; and Landrieu, L. 2021. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. *ICCV*.
- Schneider, M.; and Körner, M. 2022. EuroCrops.
- Slovak Republic Government. 2024. Spatial Data and Services. Accessed: 2024-08-07.
- Sullivan, J. A.; Samii, C.; Brown, D. G.; Moyo, F.; and Agrawal, A. 2023. Large-scale land acquisitions exacerbate local farmland inequalities in Tanzania. *Proceedings of the National Academy of Sciences*, 120(32): e2207398120.
- Taravat, A.; Wagner, M. P.; Bonifacio, R.; and Petit, D. 2021. Advanced fully convolutional networks for agricultural field boundary detection. *Remote Sensing*, 13(4): 722.
- The Service and Payment Agency (ASP). 2024. The Graphical Parcel Register (Registre parcellaire graphique (RPG)). <https://geoservices.ign.fr/rpg#telechargementrpg2021>.
- The Swedish Agency for Agriculture. 2024. Agricultural block. <https://www.geodata.se/geodataportalen/srv/swe/catalog.search;jsessionid=6C2D281619D69AC2356E1BD4C1923A3A#/metadata/df439ba5-014e-44ec-86cb-ddb9e5ba306c>.
- Tseng, G.; Cartuyvels, R.; Zvonkov, I.; Purohit, M.; Rolnick, D.; and Kerner, H. 2023. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*.
- USDA Foreign Agricultural Service. 2024. Crop Calendar. <https://ipad.fas.usda.gov/ogamaps/cropcalendar.aspx>.
- Wang, S.; Waldner, F.; and Lobell, D. B. 2022. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *Remote Sensing*, 14(22): 5738.
- Wang, S.; Waldner, F.; and Lobell, D. B. 2023. 10,000 crop field boundaries across India.

A Supplementary Information

Annotation filtering

Additional details on semantic filtering As described in the *Semantic filtering* section, we excluded all classes that were not annual (temporary) crops if they were included in a field boundary dataset. In `ftw-semantic-filters.csv` (found at <https://github.com/fieldsoftheworld/ftw-datasets-list>), we list and justify the classes included and excluded in each dataset. We also give the exact dates used to filter each country's satellite images for Window A and Window B. In Figure 6, we visualize the selected and discarded (filtered-out) fields in Luxembourg.

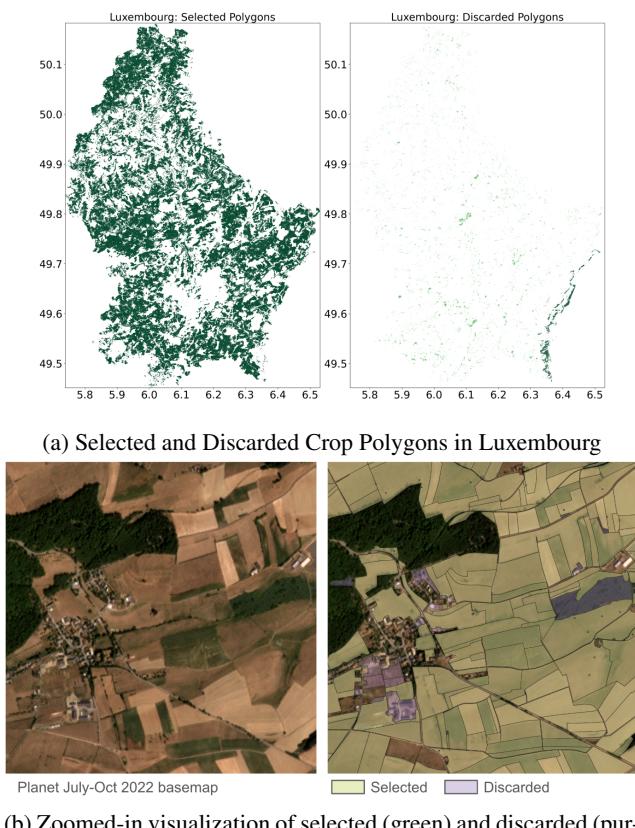


Figure 6: Visualization of polygons that were included (selected) and filtered out (discarded) in the Luxembourg dataset.

Field boundary datasets not included in FTW We conducted a comprehensive search for field polygons from government databases, published literature, and other websites to use as annotations in FTW. We looked for datasets with diverse geographic coverage, high-quality and trustworthy polygon annotations, and licenses that permit reuse. We included all datasets meeting these criteria in FTW. We considered author-reported quality assessment in each dataset's



Figure 7: Example field boundaries in the GloCAB dataset (Hall, Argueta, and Giglio 2024). Some boundaries did not align with the boundary apparent in satellite images, especially for center-pivot fields in Ukraine.

documentation, previous use of the dataset in ML analyses, and visual inspection (e.g., closed polygons, polygons consistent with satellite images from reported dates, etc).

There were a few datasets that did not meet our criteria and thus we decided not to include in FTW:

- Zambia: The same source data provider of our Kenya dataset, ECAAS, also published a dataset in Zambia. The polygons in this dataset were extremely sparse and did not appear to align with satellite imagery from the same year. The dataset can be obtained from <https://drive.google.com/drive/folders/1nEhHxWzsZxqozO2LZa-uUl6DoKNVYZVZ> and metadata from https://ecass-project-documentation.readthedocs.io/en/latest/modules/data_access.html.
- Romania: This documentation of this dataset did not specify the year the field boundaries were valid for. We decided not to use the dataset because we did not know what year of satellite imagery it should be paired with. The dataset can be found at <https://github.com/maja601/EuroCrops/wiki/Romania>.
- Kenya: Kehs et al. (2021) published a crop type dataset with field boundary polygons in Kenya. However, the paper describes limited quality assessment and our visual inspection showed some fields did not align well with contemporaneous satellite imagery.
- Brazil: The Cadastro Ambiental Rural (CAR) (<https://dados.agricultura.gov.br/it/dataset/cadastro-ambiental-rural>) provides geo-referenced data for land parcels including agriculture (Jung et al. 2017). However, we were not able to determine the appropriate attributes and attribute values to determine how to filter the parcels for temporary crops. We will try to obtain this

information in future work to include in a later version of FTW.

- GloCAB (Brazil, Ukraine, USA, Canada, and Russia): Hall, Argueta, and Giglio (2024) published the GloCAB of 190,832 manually-digitized field boundaries in 22 regions of various sizes spanning 5 countries: Brazil, Ukraine, United States of America, Canada, and Russia. While this dataset seems promising for inclusion in FTW, visual inspection revealed many boundaries that did not align with the apparent field extent from the satellite imagery, particularly around center-pivot irrigated fields in Ukraine (see Figure 7). We will continue to investigate using this dataset in future work and hope to include a filtered version of it in future FTW versions.
- USA (California): The Kern County Department Of Agriculture And Measurement publishes crop field boundaries annually since 1997. We were not able to include this dataset in FTW because their website does not specify a license for the data that allows for reuse.

Effect of random spatial splits

As described in the *Dataset splits* section, we perform a blocked random splitting strategy to partition 3×3 groups of patches into training, validation, and test splits for each country in the dataset. Figure 8 shows an example of this splitting strategy for a section of the France dataset. As such, patches in the test sets are adjacent to patches in the train sets, which may allow for leakage between train and test due to spatial autocorrelation in imagery in labels.

We tested for this effect by grouping test patches by the number of training patches they are adjacent to, then computing model performance for each group (using the entire FTW dataset). If autocorrelation was causing data leakage, then we would expect to observe higher model performance among test patches that are adjacent to training patches compared to test patches that are isolated (e.g., in the middle of the 3×3 blocks). We compared the distribution of Pixel IoU per patch using the 2-class (ignore presence-only background) model between the group with no adjacent training patches to those with some adjacent training patches per country with an independent sample t-test. We did not find a statistically significant difference in performance for any country and concluded that spatial autocorrelation is not influencing test set results.

Dataset characteristics

In this section, we provide additional dataset visualizations to show the diversity in field morphology between countries and within the Fields of The World dataset.

Figure 9 shows the distribution of (log) field area across FTW and two previous benchmark datasets.

Figure 10 shows the distribution of field polygon elongation across four countries. To compute elongation, we first compute a minimum bounding rectangle for each field polygon. The elongation is then computed as the ratio of height (i.e., short-side length) to width (i.e., long-side length) of the minimum bounding rectangle, resulting in a value between 0

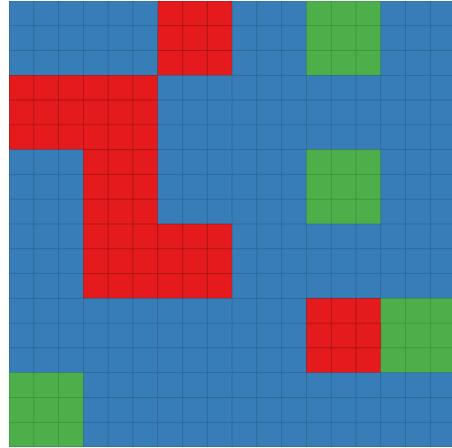


Figure 8: Example of block splits in France where red patches are in the test set, blue patches are in the train set, and green patches are in the validation set.

and 1. This shows, for example, that Austria has many long-narrow fields while the fields in Vietnam and South Africa are typically less elongated.

Figure 11 shows distributions of the Convex Hull Deviation Ratio for different countries within the FTW dataset. Let f be the area of the field polygon and c be the area of its convex hull. The Convex Hull Deviation Ratio is defined as $\frac{c-f}{c}$. This ratio is zero when the field is convex and increasingly close to one for highly non-convex field polygons. This shows that South Africa has heavy tails for the distribution, reflecting the relatively high number of highly non-convex field polygons, especially when compared to Brazil and Vietnam.

Figure 12 presents a comparative analysis of the geographical coverage of various field boundary datasets. AI4Boundaries and PASTIS/PASTIS-R datasets primarily cover European countries, while AI4SmallFarms focuses on two Asian countries. In contrast, the FTW dataset spans multiple continents, including South America, Europe, Africa, and Asia.

Table 6 compares the FTW dataset with other field boundary datasets across current Köppen climate zones of the world (Beck et al. 2018). It shows that the AI4SmallFarms dataset exists only in a single climate zone, the equatorial savannah with dry winter, while the AI4Boundaries dataset spans 9 different climate zones, including two unique zones: Polar tundra and Warm temperate fully humid with a cool summer. The FTW dataset is the most diverse among these, covering 17 different climate zones, including 7 unique zones where no other dataset is present.

Figure 13 shows the distribution of field orientations across all FTW countries. Most countries exhibit diverse field orientations, while a few, such as Austria, Denmark, and Slovenia, have predominantly north-south orientations, and others, like Luxembourg, Portugal, and South Africa, have predominantly east-west orientations.

Figure 14 shows the Convex Hull Index distributions for selected countries within the FTW dataset. Let p be the

perimeter of the original field polygon, and p_c be the perimeter of its convex hull; the Convex Hull Index is defined as $\frac{p_c}{p}$. This ratio provides insight into the complexity of a field's boundary, where values close to 1 indicate that field polygons are nearly convex, and values significantly less than 1 suggest that the polygons are non-convex with more complex boundaries. The figure shows that the field boundaries in South Africa and Estonia are more non-convex/ complex than those in Rwanda and Cambodia.

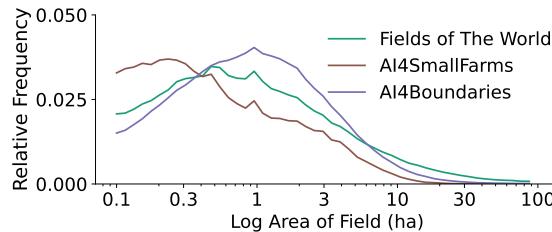


Figure 9: The distribution of (log) field area in FTW and two previous benchmark datasets.

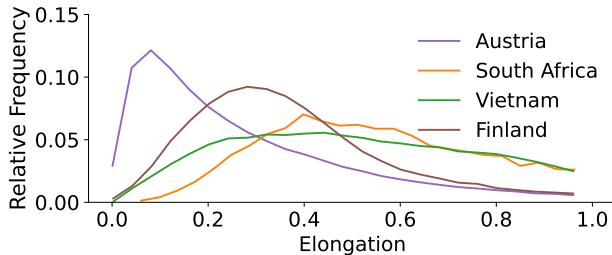


Figure 10: Elongation of field boundaries among different countries within the FTW dataset.

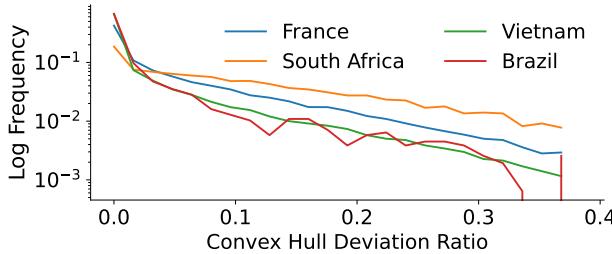


Figure 11: The Convex Hull Deviation Ratio among different countries within the FTW dataset.

Prediction heatmaps

In Figure 15 and Figure 16, we visualize the class prediction heatmaps for the U-Net with EfficientNet-b3 backbone trained on the full FTW dataset (FTW-Full) with 3-class masks (ignore presence-only background). We also visualize the heatmaps for the same model trained on the subset of European countries (FTW-EU) that are in AI4Boundaries

(Austria, Spain, France, Luxembourg, Netherlands, Slovenia, and Sweden).

Figures 15 and 16 illustrate these predictions for a sample of 8 countries, representing both Presence/Absence and Presence-only regions (respectively). The heatmaps use three classes for prediction: Red for the Background class, Green for the Field Extent (Interior) class, and Blue for the Boundary class.

Our results show that the FTW-Full predictions are more aligned with the ground truth for both European and non-European countries. Notably, the FTW-Full model provides more accurate boundary predictions (blue channel) in countries with smaller fields, such as Cambodia, Vietnam, India, and Kenya.

In contrast, the FTW-EU model struggles with accurate predictions in Presence-only regions, particularly for the Field Extent and Boundary classes. However, in some cases, such as France, the FTW-EU confidently predicts the Field Extent class, sometimes more accurately aligning with the ground truth than FTW-Full.

These visualizations help illustrate how using different datasets for training affects the model's predictions in different regions. By comparing the FTW-Full with the FTW-EU model heatmaps, we can see that the FTW-Full heatmaps align better with the ground truth masks across diverse field patterns globally.

Experiments

Model architecture experiment results We evaluated two semantic segmentation model architectures, U-Net (Ronneberger, Fischer, and Brox 2015) and DeepLabv3+ (Chen et al. 2018), with five different backbones: ResNet-18, ResNet-50, ResNeXt-50, EfficientNet-b3, and EfficientNet-b4. We report performance in Table 7 for Slovenia, France, and South Africa. U-Nets performed slightly better than DeepLabv3+ models, and U-Nets with EfficientNet backbones performed best.

Per-country experiment results In the experiment results in Tables 3 and 4 of the main paper, we reported results for a subset of test countries in FTW (Slovenia, France, and South Africa). We provide the full results of those experiments for all test countries in Tables 8-9 and Tables 10-12 (respectively) of the supplement.

Multiple random seeds The results in Tables 3, 4, and 5 of the main paper were run with one arbitrarily-chosen random seed. In supplement Tables 14-15, we report the average results across three random seeds for all test countries for the mask type experiment (Table 3 in the main paper). The standard deviation across random seeds for each experiment and test country are very small (0 or close to 0 for most metrics).

Benchmarking example We suggest benchmarking performance on the per-country test sets and reporting individual country results, the mean across all countries, or the minimum across countries (worst-case performance). Supplement Table 13 reports these metrics for the best model

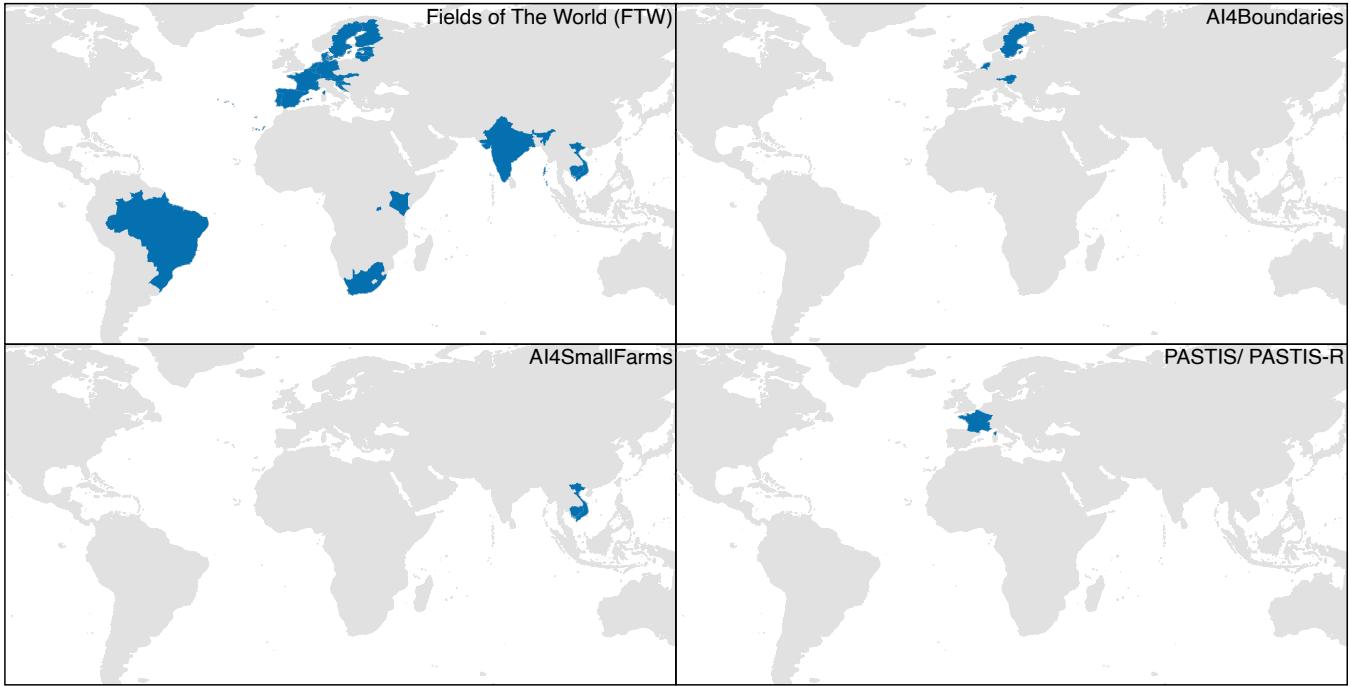


Figure 12: Geographical distribution and comparison of FTW with other field boundary datasets.

evaluated in this paper (U-net with EfficientNet-b3 backbone).

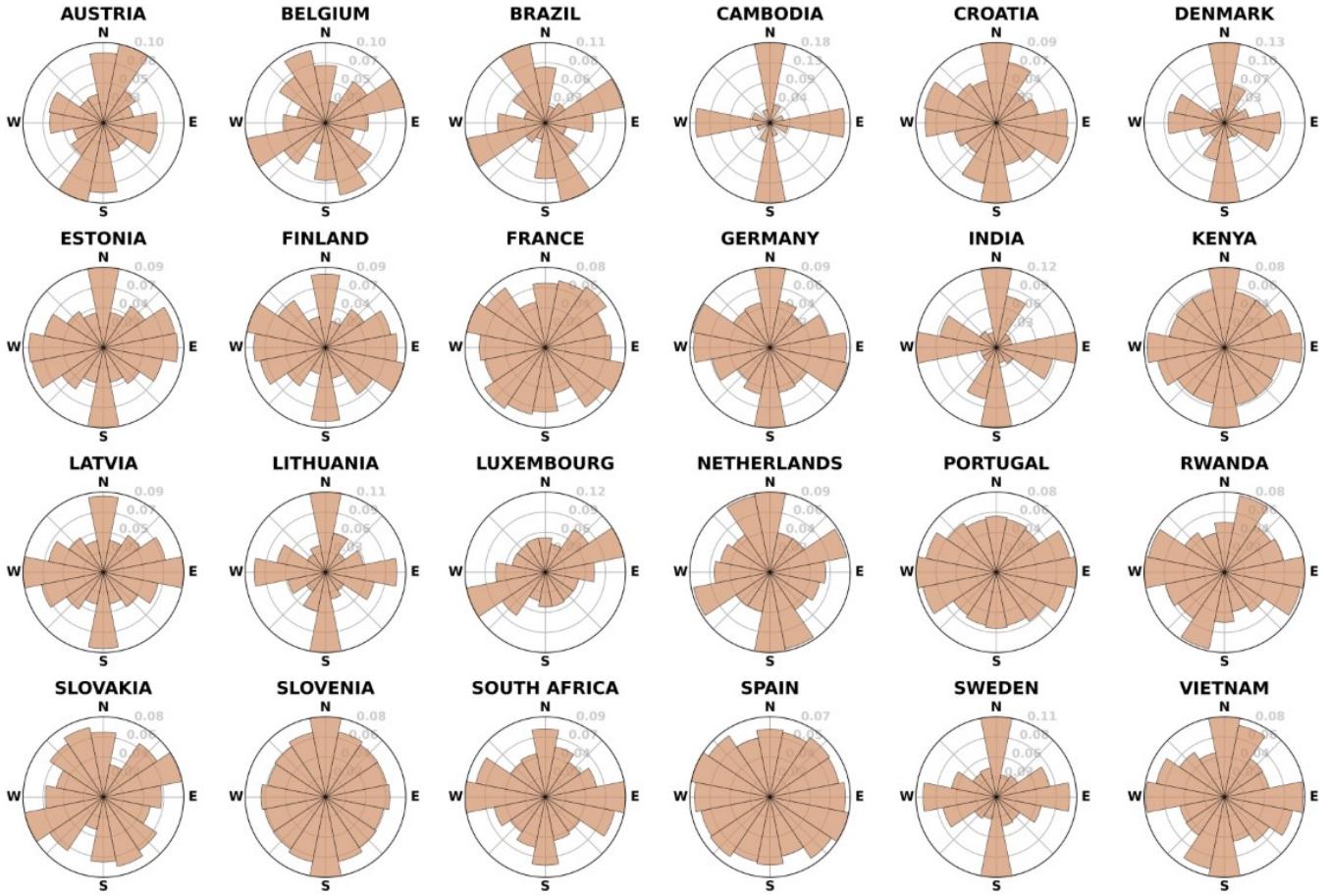


Figure 13: Field orientation histograms of all countries in the Fields of The World (FTW) dataset.

Table 6: Climatological diversity and comparison of Fields of The World (FTW) dataset with previous field boundary datasets, showing the total number of field polygons in each climatic zone.

Köppen Climate zone	Fields of The World (FTW)	AI4Boundaries	AI4SmallFarms
Polar tundra	0	8,273	0
Warm temperate fully humid with cool summer	0	2,293	0
Warm temperate with dry winter and warm summer	62	0	0
Arid Steppe cold	91	11,709	0
Equatorial rainforest, fully humid	100	0	0
Arid desert hot	596	0	0
Equatorial savannah with dry summer	725	0	0
Warm temperate with dry, warm summer	2,686	3,980	0
Arid Steppe hot	3,344	0	0
Warm temperate with dry, hot summer	8,681	129,618	0
Equatorial monsoon	25,587	0	0
Snow fully humid cool summer	38,870	35,109	0
Warm temperate with dry winter and hot summer	51,140	0	0
Warm temperate fully humid with hot summer	114,745	12,508	0
Snow fully humid warm summer	168,334	65,721	0
Equatorial savannah with dry winter	370,259	0	126,672
Warm temperate fully humid with warm summer	842,158	800,012	0
Total	1,627,378	1,069,223	126,672

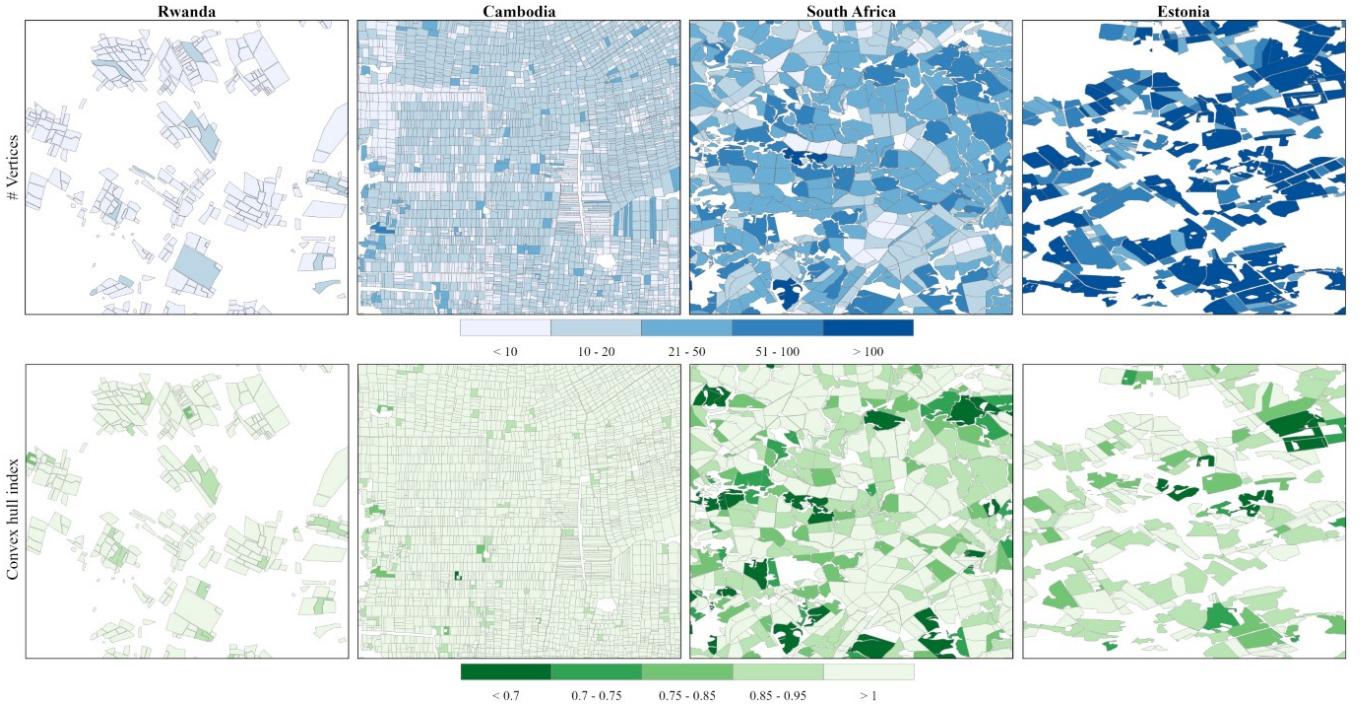


Figure 14: Visualization of the number of Field polygon vertices (above) and convex hull index for selected countries of the FTW dataset.

Table 7: Performance metrics for various model architectures in Slovenia (SVN), France (FRA), and South Africa (ZAF).

Architecture + backbone	Pixel IoU			Pixel precision			Pixel recall			Object precision			Object recall		
	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF	SVN	FRA	ZAF
U-net + ResNet-18	0.53	0.77	0.79	0.89	0.88	0.90	0.57	0.86	0.87	0.24	0.49	0.48	0.15	0.54	0.47
U-net + ResNet-50	0.55	0.78	0.79	0.89	0.89	0.90	0.59	0.87	0.87	0.26	0.52	0.47	0.16	0.54	0.50
U-net + ResNeXt-50	0.56	0.79	0.80	0.90	0.89	0.89	0.60	0.88	0.88	0.29	0.54	0.49	0.17	0.56	0.49
U-net + EfficientNet-b3	0.59	0.79	0.80	0.91	0.89	0.89	0.63	0.87	0.88	0.31	0.54	0.53	0.19	0.58	0.56
U-net + EfficientNet-b4	0.59	0.79	0.79	0.90	0.89	0.88	0.63	0.88	0.89	0.31	0.55	0.52	0.19	0.59	0.55
DeepLabv3+ + ResNet-18	0.43	0.75	0.77	0.87	0.88	0.89	0.46	0.84	0.85	0.20	0.47	0.45	0.08	0.47	0.43
DeepLabv3+ + ResNet-50	0.47	0.76	0.79	0.89	0.89	0.90	0.50	0.84	0.87	0.22	0.49	0.48	0.10	0.49	0.44
DeepLabv3+ + ResNeXt-50	0.48	0.77	0.79	0.88	0.89	0.89	0.51	0.85	0.87	0.23	0.50	0.48	0.10	0.49	0.44
DeepLabv3+ + EfficientNet-b3	0.50	0.77	0.78	0.90	0.89	0.89	0.53	0.84	0.87	0.21	0.49	0.47	0.11	0.50	0.47
DeepLabv3+ + EfficientNet-b4	0.50	0.77	0.79	0.91	0.90	0.89	0.53	0.85	0.88	0.24	0.51	0.47	0.12	0.50	0.47



Figure 15: Prediction Heatmaps from Presence/Absence countries (R: Background, G: Fields, B: Boundaries)

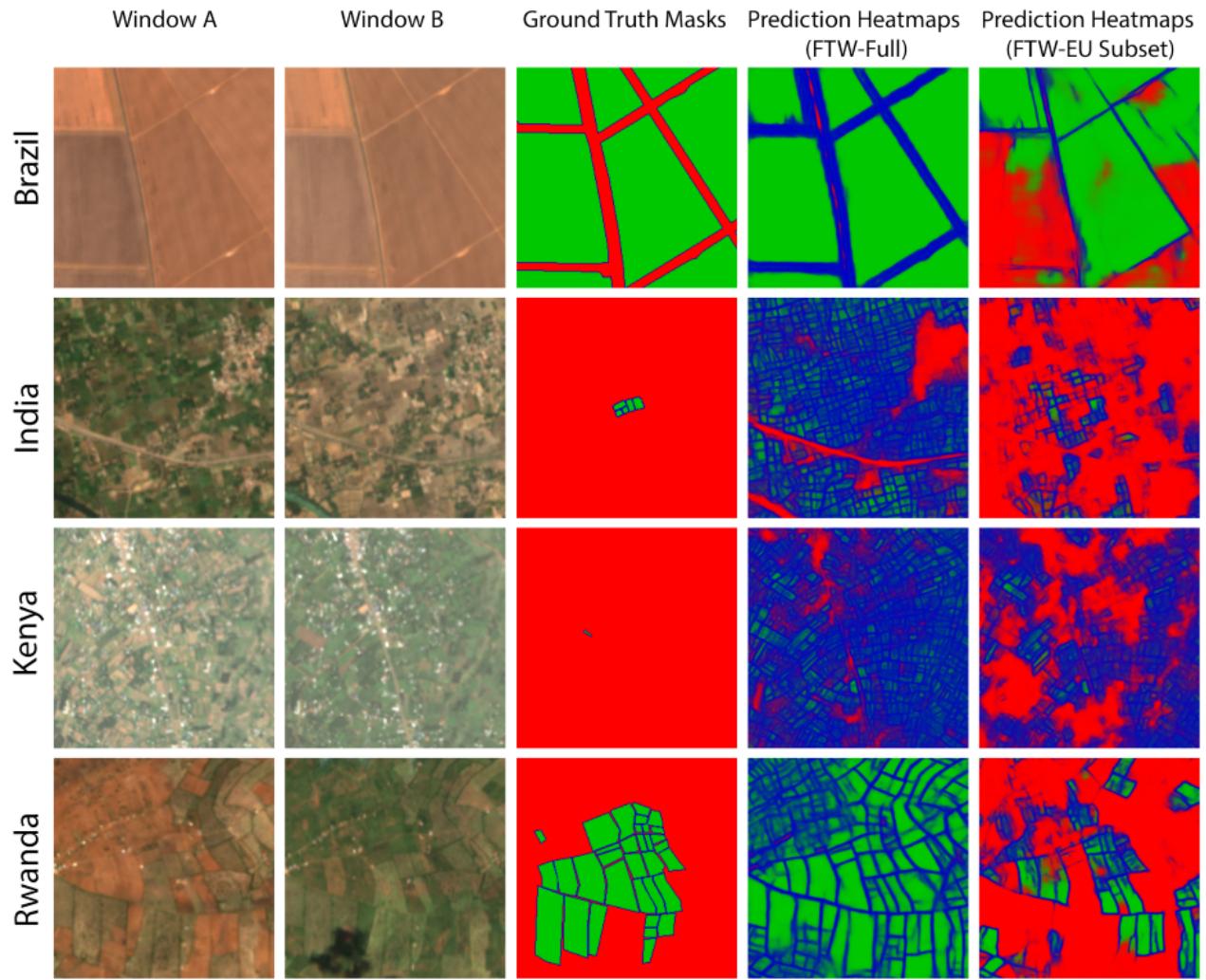


Figure 16: Prediction samples from Presence-Only countries (R: Background, G: Fields, B: Boundaries)

Table 8: Performance metrics for different target mask formats in all test countries (names starting with A-K). We compared 2-class field extent and 3-class masks with or without ignoring background (bg) pixels for presence-only samples. We only report recall metrics for presence-only countries.

Test country	Mask type	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Austria	2-class	0.77	0.84	0.91	0.37	0.11
	2-class (ignore presence-only bg)	0.77	0.83	0.92	0.36	0.11
	3-class	0.71	0.91	0.76	0.44	0.38
	3-class (ignore presence-only bg)	0.70	0.90	0.76	0.44	0.39
Belgium	2-class	0.80	0.86	0.92	0.48	0.24
	2-class (ignore presence-only bg)	0.80	0.86	0.93	0.46	0.23
	3-class	0.75	0.93	0.79	0.56	0.58
	3-class (ignore presence-only bg)	0.75	0.92	0.80	0.57	0.58
Brazil	2-class	-	-	1.00	0.16	0.11
	2-class (ignore presence-only bg)	-	-	1.00	-	0.12
	3-class	-	-	0.96	-	0.60
	3-class (ignore presence-only bg)	-	-	0.96	-	0.58
Cambodia	2-class	0.76	0.78	0.97	0.06	0.00
	2-class (ignore presence-only bg)	0.76	0.78	0.97	0.06	0.00
	3-class	0.40	0.95	0.40	0.22	0.17
	3-class (ignore presence-only bg)	0.43	0.95	0.44	0.26	0.20
Corsica	2-class	0.49	0.69	0.63	0.16	0.07
	2-class (ignore presence-only bg)	0.49	0.68	0.64	0.17	0.09
	3-class	0.45	0.76	0.53	0.21	0.16
	3-class (ignore presence-only bg)	0.48	0.79	0.55	0.21	0.17
Croatia	2-class	0.76	0.80	0.94	0.26	0.08
	2-class (ignore presence-only bg)	0.75	0.79	0.93	0.24	0.08
	3-class	0.67	0.89	0.73	0.25	0.33
	3-class (ignore presence-only bg)	0.68	0.89	0.74	0.25	0.34
Denmark	2-class	0.84	0.89	0.94	0.41	0.28
	2-class (ignore presence-only bg)	0.84	0.89	0.94	0.38	0.25
	3-class	0.83	0.93	0.88	0.45	0.61
	3-class (ignore presence-only bg)	0.83	0.93	0.88	0.45	0.60
Estonia	2-class	0.81	0.88	0.91	0.47	0.29
	2-class (ignore presence-only bg)	0.81	0.88	0.91	0.48	0.29
	3-class	0.80	0.92	0.86	0.46	0.42
	3-class (ignore presence-only bg)	0.79	0.91	0.86	0.47	0.43
Finland	2-class	0.87	0.90	0.96	0.42	0.18
	2-class (ignore presence-only bg)	0.87	0.90	0.96	0.42	0.18
	3-class	0.83	0.96	0.86	0.54	0.56
	3-class (ignore presence-only bg)	0.83	0.96	0.87	0.55	0.57
France	2-class	0.82	0.85	0.95	0.39	0.16
	2-class (ignore presence-only bg)	0.82	0.86	0.95	0.38	0.15
	3-class	0.79	0.89	0.87	0.53	0.57
	3-class (ignore presence-only bg)	0.79	0.89	0.88	0.55	0.58
Germany	2-class	0.80	0.84	0.94	0.40	0.19
	2-class (ignore presence-only bg)	0.79	0.84	0.93	0.37	0.18
	3-class	0.79	0.87	0.89	0.42	0.40
	3-class (ignore presence-only bg)	0.79	0.87	0.90	0.43	0.42
India	2-class	-	-	0.99	-	0.00
	2-class (ignore presence-only bg)	-	-	0.99	-	0.00
	3-class	-	-	0.23	-	0.05
	3-class (ignore presence-only bg)	-	-	0.22	-	0.06
Kenya	2-class	-	-	0.95	-	0.00
	2-class (ignore presence-only bg)	-	-	0.97	-	0.00
	3-class	-	-	0.47	-	0.08
	3-class (ignore presence-only bg)	-	-	0.49	-	0.10

Table 9: Performance metrics for different target mask formats in test countries (names starting with L-Z). We compared 2-class field extent and 3-class masks with or without ignoring background (bg) pixels for presence-only samples. We only report recall metrics for presence-only countries.

Test country	Mask type	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Latvia	2-class	0.84	0.90	0.92	0.43	0.26
	2-class (ignore presence-only bg)	0.84	0.90	0.92	0.44	0.27
	3-class	0.81	0.94	0.85	0.43	0.45
	3-class (ignore presence-only bg)	0.81	0.94	0.86	0.44	0.45
Lithuania	2-class	0.78	0.83	0.93	0.38	0.19
	2-class (ignore presence-only bg)	0.77	0.82	0.93	0.39	0.18
	3-class	0.74	0.88	0.82	0.37	0.41
	3-class (ignore presence-only bg)	0.74	0.88	0.83	0.37	0.41
Luxembourg	2-class	0.85	0.88	0.97	0.25	0.05
	2-class (ignore presence-only bg)	0.86	0.88	0.97	0.24	0.05
	3-class	0.79	0.97	0.81	0.47	0.52
	3-class (ignore presence-only bg)	0.79	0.96	0.82	0.47	0.51
Netherlands	2-class	0.79	0.86	0.91	0.49	0.25
	2-class (ignore presence-only bg)	0.79	0.86	0.91	0.48	0.24
	3-class	0.75	0.92	0.80	0.51	0.45
	3-class (ignore presence-only bg)	0.75	0.92	0.80	0.53	0.45
Portugal	2-class	0.29	0.41	0.51	0.03	0.01
	2-class (ignore presence-only bg)	0.31	0.65	0.37	0.04	0.01
	3-class	0.12	0.79	0.12	0.05	0.02
	3-class (ignore presence-only bg)	0.12	0.67	0.12	0.07	0.03
Rwanda	2-class	-	-	0.99	-	0.00
	2-class (ignore presence-only bg)	-	-	0.98	-	0.00
	3-class	-	-	0.55	-	0.26
	3-class (ignore presence-only bg)	-	-	0.57	-	0.30
Slovakia	2-class	0.93	0.95	0.98	0.59	0.40
	2-class (ignore presence-only bg)	0.93	0.95	0.98	0.59	0.40
	3-class	0.92	0.98	0.94	0.51	0.55
	3-class (ignore presence-only bg)	0.92	0.98	0.95	0.50	0.55
Slovenia	2-class	0.69	0.79	0.85	0.30	0.08
	2-class (ignore presence-only bg)	0.69	0.78	0.86	0.30	0.08
	3-class	0.58	0.90	0.62	0.31	0.19
	3-class (ignore presence-only bg)	0.59	0.90	0.63	0.33	0.20
South Africa	2-class	0.82	0.85	0.96	0.44	0.20
	2-class (ignore presence-only bg)	0.81	0.84	0.96	0.44	0.19
	3-class	0.80	0.90	0.88	0.51	0.55
	3-class (ignore presence-only bg)	0.79	0.89	0.88	0.51	0.55
Spain	2-class	0.84	0.87	0.95	0.33	0.06
	2-class (ignore presence-only bg)	0.84	0.87	0.95	0.32	0.05
	3-class	0.73	0.96	0.75	0.33	0.19
	3-class (ignore presence-only bg)	0.74	0.96	0.76	0.36	0.20
Sweden	2-class	0.83	0.88	0.93	0.37	0.20
	2-class (ignore presence-only bg)	0.83	0.88	0.93	0.34	0.19
	3-class	0.81	0.94	0.85	0.41	0.51
	3-class (ignore presence-only bg)	0.81	0.94	0.85	0.40	0.51
Vietnam	2-class	0.67	0.70	0.94	0.09	0.01
	2-class (ignore presence-only bg)	0.67	0.70	0.95	0.10	0.01
	3-class	0.46	0.89	0.49	0.18	0.13
	3-class (ignore presence-only bg)	0.47	0.89	0.49	0.18	0.13

Table 10: Performance for input channel ablations in all test countries (names starting with A-G): multispectral (RGB-NIR vs. RGB only) and multi-temporal (Window A and Window B) input channels.

Test country	Channels	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Austria	Stacked Windows A and B (RGB only)	0.70	0.90	0.76	0.42	0.37
	Stacked Windows A and B	0.71	0.90	0.77	0.44	0.39
	Mean of Windows A and B	0.66	0.87	0.73	0.39	0.34
	Window A only	0.67	0.90	0.73	0.37	0.34
	Window B only	0.68	0.88	0.74	0.40	0.35
Belgium	Stacked Windows A and B (RGB only)	0.72	0.91	0.77	0.51	0.55
	Stacked Windows A and B	0.75	0.93	0.80	0.58	0.58
	Mean of Windows A and B	0.69	0.89	0.75	0.48	0.53
	Window A only	0.73	0.93	0.77	0.51	0.53
	Window B only	0.65	0.84	0.74	0.42	0.51
Brazil	Stacked Windows A and B (RGB only)	-	-	0.96	-	0.59
	Stacked Windows A and B	-	-	0.96	-	0.58
	Mean of Windows A and B	-	-	0.96	-	0.60
	Window A only	-	-	0.96	-	0.56
	Window B only	-	-	0.96	-	0.60
Cambodia	Stacked Windows A and B (RGB only)	0.35	0.94	0.36	0.20	0.15
	Stacked Windows A and B	0.39	0.95	0.39	0.22	0.17
	Mean of Windows A and B	0.33	0.95	0.33	0.17	0.13
	Window A only	0.35	0.94	0.36	0.15	0.13
	Window B only	0.38	0.94	0.39	0.20	0.15
Corsica	Stacked Windows A and B (RGB only)	0.45	0.79	0.51	0.24	0.16
	Stacked Windows A and B	0.47	0.80	0.54	0.26	0.17
	Mean of Windows A and B	0.42	0.73	0.50	0.18	0.14
	Window A only	0.45	0.81	0.50	0.23	0.17
	Window B only	0.42	0.75	0.49	0.19	0.13
Croatia	Stacked Windows A and B (RGB only)	0.66	0.88	0.72	0.23	0.32
	Stacked Windows A and B	0.67	0.89	0.73	0.25	0.33
	Mean of Windows A and B	0.64	0.88	0.71	0.24	0.29
	Window A only	0.65	0.87	0.72	0.22	0.31
	Window B only	0.66	0.89	0.72	0.22	0.29
Denmark	Stacked Windows A and B (RGB only)	0.82	0.93	0.87	0.43	0.59
	Stacked Windows A and B	0.83	0.93	0.88	0.46	0.60
	Mean of Windows A and B	0.82	0.93	0.88	0.43	0.58
	Window A only	0.82	0.93	0.88	0.44	0.60
	Window B only	0.81	0.93	0.87	0.43	0.58
Estonia	Stacked Windows A and B (RGB only)	0.78	0.91	0.85	0.45	0.41
	Stacked Windows A and B	0.80	0.92	0.86	0.49	0.42
	Mean of Windows A and B	0.78	0.90	0.85	0.45	0.41
	Window A only	0.79	0.93	0.85	0.48	0.41
	Window B only	0.77	0.90	0.84	0.42	0.40
Finland	Stacked Windows A and B (RGB only)	0.82	0.95	0.85	0.52	0.54
	Stacked Windows A and B	0.83	0.96	0.86	0.56	0.56
	Mean of Windows A and B	0.81	0.95	0.84	0.52	0.52
	Window A only	0.82	0.95	0.85	0.51	0.54
	Window B only	0.80	0.95	0.84	0.51	0.52
France	Stacked Windows A and B (RGB only)	0.78	0.89	0.86	0.51	0.56
	Stacked Windows A and B	0.79	0.89	0.87	0.54	0.58
	Mean of Windows A and B	0.77	0.89	0.86	0.50	0.55
	Window A only	0.77	0.88	0.86	0.47	0.54
	Window B only	0.78	0.89	0.86	0.49	0.54
Germany	Stacked Windows A and B (RGB only)	0.78	0.87	0.88	0.41	0.41
	Stacked Windows A and B	0.79	0.87	0.89	0.43	0.41
	Mean of Windows A and B	0.77	0.87	0.87	0.41	0.38
	Window A only	0.77	0.87	0.88	0.37	0.37
	Window B only	0.78	0.87	0.89	0.40	0.41

Table 11: Performance for input channel ablations in all test countries (names starting with H-R): multispectral (RGB-NIR vs. RGB only) and multi-temporal (Window A and Window B) input channels.

Test country	Channels	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
India	Stacked Windows A and B (RGB only)	-	-	0.22	-	0.04
	Stacked Windows A and B	-	-	0.20	-	0.05
	Mean of Windows A and B	-	-	0.20	-	0.04
	Window A only	-	-	0.29	-	0.06
	Window B only	-	-	0.17	-	0.02
Kenya	Stacked Windows A and B (RGB only)	-	-	0.48	-	0.10
	Stacked Windows A and B	-	-	0.47	-	0.10
	Mean of Windows A and B	-	-	0.46	-	0.09
	Window A only	-	-	0.40	-	0.06
	Window B only	-	-	0.47	-	0.08
Latvia	Stacked Windows A and B (RGB only)	0.80	0.94	0.85	0.43	0.44
	Stacked Windows A and B	0.81	0.94	0.86	0.45	0.45
	Mean of Windows A and B	0.79	0.92	0.84	0.41	0.43
	Window A only	0.80	0.93	0.85	0.40	0.43
	Window B only	0.78	0.92	0.83	0.39	0.41
Lithuania	Stacked Windows A and B (RGB only)	0.71	0.86	0.80	0.35	0.39
	Stacked Windows A and B	0.73	0.88	0.82	0.37	0.41
	Mean of Windows A and B	0.69	0.84	0.79	0.33	0.37
	Window A only	0.72	0.87	0.81	0.34	0.37
	Window B only	0.59	0.74	0.74	0.26	0.34
Luxembourg	Stacked Windows A and B (RGB only)	0.78	0.96	0.81	0.46	0.51
	Stacked Windows A and B	0.79	0.97	0.81	0.47	0.51
	Mean of Windows A and B	0.77	0.97	0.80	0.44	0.50
	Window A only	0.78	0.96	0.80	0.42	0.46
	Window B only	0.77	0.96	0.80	0.43	0.49
Netherlands	Stacked Windows A and B (RGB only)	0.71	0.91	0.77	0.47	0.43
	Stacked Windows A and B	0.76	0.92	0.81	0.52	0.45
	Mean of Windows A and B	0.70	0.90	0.76	0.44	0.42
	Window A only	0.65	0.86	0.72	0.38	0.38
	Window B only	0.72	0.91	0.78	0.46	0.42
Portugal	Stacked Windows A and B (RGB only)	0.14	0.82	0.14	0.09	0.03
	Stacked Windows A and B	0.22	0.38	0.34	0.04	0.04
	Mean of Windows A and B	0.15	0.58	0.17	0.05	0.02
	Window A only	0.10	0.80	0.10	0.06	0.02
	Window B only	0.29	0.66	0.35	0.10	0.08
Rwanda	Stacked Windows A and B (RGB only)	-	-	0.61	-	0.26
	Stacked Windows A and B	-	-	0.58	-	0.27
	Mean of Windows A and B	-	-	0.57	-	0.26
	Window A only	-	-	0.50	-	0.23
	Window B only	-	-	0.64	-	0.34

Table 12: Performance for input channel ablations in all test countries (names starting with S-Z): multispectral (RGB-NIR vs. RGB only) and multi-temporal (Window A and Window B) input channels.

Test country	Channels	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Slovakia	Stacked Windows A and B (RGB only)	0.91	0.97	0.94	0.47	0.53
	Stacked Windows A and B	0.92	0.98	0.95	0.52	0.55
	Mean of Windows A and B	0.92	0.97	0.94	0.49	0.54
	Window A only	0.91	0.97	0.94	0.46	0.52
	Window B only	0.91	0.97	0.94	0.48	0.54
Slovenia	Stacked Windows A and B (RGB only)	0.58	0.90	0.62	0.27	0.18
	Stacked Windows A and B	0.58	0.91	0.61	0.30	0.18
	Mean of Windows A and B	0.54	0.88	0.59	0.27	0.16
	Window A only	0.55	0.88	0.59	0.27	0.17
	Window B only	0.52	0.87	0.57	0.24	0.15
South Africa	Stacked Windows A and B (RGB only)	0.79	0.89	0.88	0.53	0.54
	Stacked Windows A and B	0.80	0.90	0.87	0.55	0.54
	Mean of Windows A and B	0.78	0.88	0.88	0.49	0.53
	Window A only	0.78	0.88	0.87	0.47	0.52
	Window B only	0.79	0.89	0.87	0.52	0.53
Spain	Stacked Windows A and B (RGB only)	0.73	0.96	0.75	0.34	0.19
	Stacked Windows A and B	0.73	0.96	0.75	0.34	0.19
	Mean of Windows A and B	0.72	0.96	0.74	0.32	0.18
	Window A only	0.71	0.96	0.74	0.31	0.17
	Window B only	0.71	0.96	0.73	0.32	0.18
Sweden	Stacked Windows A and B (RGB only)	0.80	0.94	0.85	0.40	0.50
	Stacked Windows A and B	0.81	0.94	0.85	0.42	0.51
	Mean of Windows A and B	0.80	0.93	0.84	0.40	0.48
	Window A only	0.81	0.94	0.85	0.39	0.49
	Window B only	0.79	0.92	0.84	0.38	0.47
Vietnam	Stacked Windows A and B (RGB only)	0.36	0.90	0.37	0.11	0.08
	Stacked Windows A and B	0.45	0.89	0.47	0.17	0.12
	Mean of Windows A and B	0.33	0.88	0.34	0.09	0.06
	Window A only	0.35	0.87	0.37	0.09	0.06
	Window B only	0.39	0.90	0.41	0.14	0.10

Table 13: Performance metrics for U-net with EfficientNet-b3 backbone with 3-class masks, ignoring background pixels for presence-only samples. We only report recall metrics for presence-only countries.

Test country	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Austria	0.70	0.90	0.76	0.44	0.39
Belgium	0.75	0.92	0.80	0.57	0.58
Brazil	-	-	0.96	-	0.58
Cambodia	0.43	0.95	0.44	0.26	0.20
Corsica	0.48	0.79	0.55	0.21	0.17
Croatia	0.68	0.89	0.74	0.25	0.34
Denmark	0.83	0.93	0.88	0.45	0.60
Estonia	0.79	0.91	0.86	0.47	0.43
Finland	0.83	0.96	0.87	0.55	0.57
France	0.79	0.89	0.88	0.55	0.58
Germany	0.79	0.87	0.90	0.43	0.42
India	-	-	0.22	-	0.06
Kenya	-	-	0.49	-	0.10
Latvia	0.81	0.94	0.86	0.44	0.45
Lithuania	0.74	0.88	0.83	0.37	0.41
Luxembourg	0.79	0.96	0.82	0.47	0.51
Netherlands	0.75	0.92	0.80	0.53	0.45
Portugal	0.12	0.67	0.12	0.07	0.03
Rwanda	-	-	0.57	-	0.30
Slovakia	0.92	0.98	0.95	0.50	0.55
Slovenia	0.59	0.90	0.63	0.33	0.20
South Africa	0.79	0.89	0.88	0.51	0.55
Spain	0.74	0.96	0.76	0.36	0.20
Sweden	0.81	0.94	0.85	0.40	0.51
Vietnam	0.47	0.89	0.49	0.18	0.13
Mean	0.70	0.90	0.72	0.40	0.37
Minimum	0.12	0.67	0.12	0.07	0.03

Table 14: Performance metrics for different target mask formats in all test countries averaged over 3 random seeds (names starting with A-K). We only report recall metrics for presence-only countries.

Test country	Mask type	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Austria	2-class	0.77 ± 0.00	0.83 ± 0.00	0.91 ± 0.00	0.38 ± 0.01	0.11 ± 0.00
	2-class (ignore presence-only bg)	0.77 ± 0.00	0.83 ± 0.00	0.91 ± 0.00	0.37 ± 0.02	0.12 ± 0.00
	3-class	0.71 ± 0.00	0.90 ± 0.00	0.77 ± 0.00	0.45 ± 0.00	0.39 ± 0.00
	3-class (ignore presence-only bg)	0.71 ± 0.00	0.90 ± 0.01	0.76 ± 0.00	0.45 ± 0.00	0.39 ± 0.00
Belgium	2-class	0.80 ± 0.00	0.85 ± 0.00	0.92 ± 0.00	0.48 ± 0.02	0.23 ± 0.00
	2-class (ignore presence-only bg)	0.80 ± 0.00	0.86 ± 0.00	0.92 ± 0.01	0.48 ± 0.01	0.23 ± 0.00
	3-class	0.75 ± 0.00	0.93 ± 0.00	0.80 ± 0.00	0.57 ± 0.01	0.58 ± 0.00
	3-class (ignore presence-only bg)	0.75 ± 0.00	0.93 ± 0.00	0.80 ± 0.00	0.57 ± 0.01	0.58 ± 0.00
Brazil	2-class	-	-	1.00 ± 0.00	-	0.12 ± 0.02
	2-class (ignore presence-only bg)	-	-	1.00 ± 0.00	-	0.13 ± 0.01
	3-class	-	-	0.96 ± 0.00	-	0.58 ± 0.01
	3-class (ignore presence-only bg)	-	-	0.97 ± 0.00	-	0.58 ± 0.01
Cambodia	2-class	0.76 ± 0.00	0.78 ± 0.00	0.97 ± 0.00	0.04 ± 0.01	0.00 ± 0.00
	2-class (ignore presence-only bg)	0.76 ± 0.00	0.78 ± 0.00	0.97 ± 0.00	0.04 ± 0.02	0.00 ± 0.00
	3-class	0.41 ± 0.02	0.95 ± 0.00	0.42 ± 0.02	0.23 ± 0.01	0.18 ± 0.01
	3-class (ignore presence-only bg)	0.40 ± 0.01	0.95 ± 0.00	0.41 ± 0.01	0.22 ± 0.01	0.17 ± 0.01
Corsica	2-class	0.49 ± 0.01	0.68 ± 0.03	0.64 ± 0.01	0.18 ± 0.01	0.08 ± 0.00
	2-class (ignore presence-only bg)	0.48 ± 0.01	0.67 ± 0.01	0.62 ± 0.01	0.19 ± 0.01	0.09 ± 0.00
	3-class	0.48 ± 0.01	0.79 ± 0.01	0.54 ± 0.02	0.24 ± 0.01	0.17 ± 0.01
	3-class (ignore presence-only bg)	0.46 ± 0.01	0.79 ± 0.01	0.53 ± 0.01	0.24 ± 0.01	0.18 ± 0.00
Croatia	2-class	0.76 ± 0.00	0.80 ± 0.00	0.94 ± 0.00	0.26 ± 0.01	0.08 ± 0.00
	2-class (ignore presence-only bg)	0.76 ± 0.00	0.80 ± 0.01	0.93 ± 0.00	0.26 ± 0.01	0.09 ± 0.00
	3-class	0.68 ± 0.00	0.89 ± 0.00	0.74 ± 0.00	0.25 ± 0.01	0.34 ± 0.01
	3-class (ignore presence-only bg)	0.67 ± 0.00	0.89 ± 0.00	0.74 ± 0.00	0.25 ± 0.00	0.33 ± 0.01
Denmark	2-class	0.84 ± 0.00	0.89 ± 0.00	0.94 ± 0.00	0.40 ± 0.02	0.25 ± 0.00
	2-class (ignore presence-only bg)	0.84 ± 0.00	0.89 ± 0.00	0.94 ± 0.00	0.41 ± 0.03	0.27 ± 0.01
	3-class	0.83 ± 0.00	0.93 ± 0.00	0.88 ± 0.00	0.46 ± 0.01	0.60 ± 0.00
	3-class (ignore presence-only bg)	0.83 ± 0.00	0.93 ± 0.00	0.88 ± 0.00	0.46 ± 0.01	0.60 ± 0.01
Estonia	2-class	0.81 ± 0.00	0.88 ± 0.00	0.92 ± 0.00	0.49 ± 0.01	0.29 ± 0.00
	2-class (ignore presence-only bg)	0.81 ± 0.01	0.88 ± 0.00	0.92 ± 0.00	0.50 ± 0.02	0.30 ± 0.00
	3-class	0.79 ± 0.00	0.92 ± 0.00	0.85 ± 0.01	0.47 ± 0.02	0.42 ± 0.00
	3-class (ignore presence-only bg)	0.79 ± 0.00	0.91 ± 0.00	0.86 ± 0.00	0.47 ± 0.01	0.42 ± 0.00
Finland	2-class	0.87 ± 0.00	0.90 ± 0.00	0.96 ± 0.00	0.44 ± 0.01	0.18 ± 0.00
	2-class (ignore presence-only bg)	0.87 ± 0.00	0.90 ± 0.00	0.96 ± 0.00	0.44 ± 0.02	0.18 ± 0.01
	3-class	0.83 ± 0.00	0.96 ± 0.00	0.86 ± 0.00	0.55 ± 0.00	0.56 ± 0.01
	3-class (ignore presence-only bg)	0.83 ± 0.00	0.96 ± 0.00	0.86 ± 0.00	0.54 ± 0.01	0.56 ± 0.00
France	2-class	0.81 ± 0.00	0.85 ± 0.00	0.95 ± 0.00	0.38 ± 0.01	0.15 ± 0.01
	2-class (ignore presence-only bg)	0.82 ± 0.00	0.86 ± 0.00	0.95 ± 0.00	0.38 ± 0.01	0.15 ± 0.00
	3-class	0.79 ± 0.00	0.89 ± 0.00	0.87 ± 0.00	0.54 ± 0.01	0.58 ± 0.01
	3-class (ignore presence-only bg)	0.79 ± 0.00	0.89 ± 0.00	0.87 ± 0.00	0.54 ± 0.01	0.57 ± 0.00
Germany	2-class	0.80 ± 0.00	0.84 ± 0.00	0.94 ± 0.01	0.40 ± 0.03	0.18 ± 0.01
	2-class (ignore presence-only bg)	0.80 ± 0.00	0.84 ± 0.00	0.94 ± 0.00	0.40 ± 0.03	0.19 ± 0.00
	3-class	0.79 ± 0.00	0.87 ± 0.00	0.89 ± 0.01	0.40 ± 0.02	0.40 ± 0.01
	3-class (ignore presence-only bg)	0.79 ± 0.00	0.87 ± 0.00	0.89 ± 0.00	0.41 ± 0.01	0.40 ± 0.01
India	2-class	-	-	0.99 ± 0.00	-	0.00 ± 0.00
	2-class (ignore presence-only bg)	-	-	0.99 ± 0.00	-	0.00 ± 0.00
	3-class	-	-	0.22 ± 0.03	-	0.05 ± 0.01
	3-class (ignore presence-only bg)	-	-	0.22 ± 0.02	-	0.05 ± 0.01
Kenya	2-class	-	-	0.96 ± 0.01	-	0.00 ± 0.00
	2-class (ignore presence-only bg)	-	-	0.96 ± 0.01	-	0.00 ± 0.01
	3-class	-	-	0.50 ± 0.02	-	0.09 ± 0.02
	3-class (ignore presence-only bg)	-	-	0.47 ± 0.02	-	0.08 ± 0.01

Table 15: Performance metrics for different target mask formats in all test countries averaged over 3 random seeds (names starting with L-Z). We only report recall metrics for presence-only countries.

Test country	Mask type	Pixel IoU	Pixel precision	Pixel recall	Object precision	Object recall
Latvia	2-class	0.83 ± 0.00	0.90 ± 0.00	0.92 ± 0.00	0.44 ± 0.00	0.27 ± 0.00
	2-class (ignore presence-only bg)	0.84 ± 0.00	0.90 ± 0.00	0.92 ± 0.00	0.45 ± 0.01	0.27 ± 0.01
	3-class	0.81 ± 0.00	0.94 ± 0.00	0.85 ± 0.00	0.45 ± 0.01	0.45 ± 0.00
	3-class (ignore presence-only bg)	0.81 ± 0.00	0.94 ± 0.00	0.86 ± 0.00	0.45 ± 0.01	0.45 ± 0.00
Lithuania	2-class	0.77 ± 0.00	0.82 ± 0.00	0.93 ± 0.00	0.39 ± 0.02	0.19 ± 0.00
	2-class (ignore presence-only bg)	0.77 ± 0.00	0.82 ± 0.01	0.93 ± 0.00	0.39 ± 0.01	0.19 ± 0.00
	3-class	0.74 ± 0.01	0.88 ± 0.01	0.83 ± 0.01	0.38 ± 0.01	0.42 ± 0.00
	3-class (ignore presence-only bg)	0.74 ± 0.00	0.88 ± 0.00	0.82 ± 0.00	0.38 ± 0.01	0.41 ± 0.00
Luxembourg	2-class	0.86 ± 0.00	0.88 ± 0.00	0.97 ± 0.00	0.24 ± 0.00	0.05 ± 0.00
	2-class (ignore presence-only bg)	0.86 ± 0.00	0.88 ± 0.00	0.97 ± 0.00	0.25 ± 0.03	0.05 ± 0.01
	3-class	0.79 ± 0.00	0.96 ± 0.00	0.81 ± 0.00	0.45 ± 0.01	0.51 ± 0.00
	3-class (ignore presence-only bg)	0.79 ± 0.00	0.97 ± 0.00	0.81 ± 0.00	0.46 ± 0.01	0.51 ± 0.00
Netherlands	2-class	0.79 ± 0.01	0.86 ± 0.01	0.91 ± 0.00	0.51 ± 0.01	0.24 ± 0.00
	2-class (ignore presence-only bg)	0.79 ± 0.01	0.85 ± 0.00	0.91 ± 0.01	0.50 ± 0.01	0.24 ± 0.00
	3-class	0.74 ± 0.00	0.92 ± 0.00	0.80 ± 0.00	0.52 ± 0.01	0.45 ± 0.00
	3-class (ignore presence-only bg)	0.75 ± 0.00	0.92 ± 0.00	0.80 ± 0.00	0.52 ± 0.01	0.45 ± 0.00
Portugal	2-class	0.31 ± 0.03	0.62 ± 0.03	0.39 ± 0.06	0.06 ± 0.01	0.01 ± 0.00
	2-class (ignore presence-only bg)	0.33 ± 0.04	0.54 ± 0.09	0.49 ± 0.12	0.05 ± 0.01	0.01 ± 0.00
	3-class	0.17 ± 0.06	0.66 ± 0.27	0.24 ± 0.16	0.07 ± 0.02	0.04 ± 0.01
	3-class (ignore presence-only bg)	0.14 ± 0.02	0.76 ± 0.11	0.15 ± 0.03	0.08 ± 0.03	0.03 ± 0.01
Rwanda	2-class	-	-	0.99 ± 0.00	-	0.00 ± 0.00
	2-class (ignore presence-only bg)	-	-	0.99 ± 0.00	-	0.00 ± 0.00
	3-class	-	-	0.60 ± 0.02	-	0.27 ± 0.04
	3-class (ignore presence-only bg)	-	-	0.59 ± 0.02	-	0.27 ± 0.02
Slovakia	2-class	0.93 ± 0.00	0.95 ± 0.00	0.98 ± 0.00	0.60 ± 0.01	0.40 ± 0.01
	2-class (ignore presence-only bg)	0.93 ± 0.00	0.95 ± 0.00	0.98 ± 0.00	0.60 ± 0.03	0.41 ± 0.01
	3-class	0.92 ± 0.00	0.98 ± 0.00	0.95 ± 0.00	0.51 ± 0.01	0.55 ± 0.00
	3-class (ignore presence-only bg)	0.92 ± 0.00	0.98 ± 0.00	0.94 ± 0.00	0.52 ± 0.01	0.55 ± 0.00
Slovenia	2-class	0.69 ± 0.01	0.79 ± 0.00	0.84 ± 0.01	0.30 ± 0.02	0.08 ± 0.00
	2-class (ignore presence-only bg)	0.69 ± 0.00	0.79 ± 0.01	0.85 ± 0.01	0.31 ± 0.02	0.09 ± 0.00
	3-class	0.59 ± 0.01	0.90 ± 0.00	0.63 ± 0.01	0.30 ± 0.00	0.19 ± 0.00
	3-class (ignore presence-only bg)	0.59 ± 0.01	0.91 ± 0.00	0.63 ± 0.01	0.31 ± 0.01	0.19 ± 0.01
South Africa	2-class	0.82 ± 0.00	0.85 ± 0.01	0.96 ± 0.00	0.44 ± 0.01	0.2 ± 0.01
	2-class (ignore presence-only bg)	0.82 ± 0.00	0.85 ± 0.00	0.96 ± 0.01	0.43 ± 0.01	0.2 ± 0.01
	3-class	0.80 ± 0.00	0.89 ± 0.00	0.88 ± 0.00	0.53 ± 0.02	0.54 ± 0.00
	3-class (ignore presence-only bg)	0.80 ± 0.00	0.89 ± 0.00	0.88 ± 0.00	0.52 ± 0.02	0.55 ± 0.01
Spain	2-class	0.84 ± 0.00	0.87 ± 0.00	0.95 ± 0.00	0.34 ± 0.01	0.05 ± 0.00
	2-class (ignore presence-only bg)	0.84 ± 0.00	0.88 ± 0.01	0.95 ± 0.00	0.34 ± 0.03	0.05 ± 0.00
	3-class	0.74 ± 0.00	0.96 ± 0.00	0.76 ± 0.00	0.34 ± 0.01	0.2 ± 0.00
	3-class (ignore presence-only bg)	0.73 ± 0.00	0.96 ± 0.00	0.75 ± 0.00	0.34 ± 0.01	0.19 ± 0.00
Sweden	2-class	0.83 ± 0.00	0.88 ± 0.00	0.93 ± 0.00	0.35 ± 0.01	0.19 ± 0.01
	2-class (ignore presence-only bg)	0.83 ± 0.00	0.89 ± 0.01	0.93 ± 0.00	0.36 ± 0.04	0.2 ± 0.01
	3-class	0.81 ± 0.00	0.94 ± 0.00	0.86 ± 0.00	0.41 ± 0.01	0.51 ± 0.00
	3-class (ignore presence-only bg)	0.81 ± 0.00	0.94 ± 0.00	0.85 ± 0.00	0.41 ± 0.01	0.51 ± 0.01
Vietnam	2-class	0.67 ± 0.00	0.7 ± 0.00	0.94 ± 0.00	0.1 ± 0.01	0.01 ± 0.00
	2-class (ignore presence-only bg)	0.67 ± 0.00	0.7 ± 0.00	0.94 ± 0.00	0.11 ± 0.01	0.01 ± 0.00
	3-class	0.46 ± 0.02	0.89 ± 0.00	0.49 ± 0.02	0.17 ± 0.01	0.12 ± 0.01
	3-class (ignore presence-only bg)	0.47 ± 0.01	0.89 ± 0.01	0.50 ± 0.01	0.19 ± 0.01	0.12 ± 0.00