AI4SmallFarms: A Dataset for Crop Field Delineation in Southeast Asian Smallholder Farms

Claudio Persello, *Senior Member, IEEE*, Jeroen Grift, Xinyan Fan, Claudia Paris, *Senior Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, Mila Koeva, and Andrew Nelson

***Abstract*— Agricultural field polygons within smallholder farming systems are essential to facilitate the collection of geo-spatial data useful for farmers, managers, and policymakers. However, the limited availability of training labels poses a challenge in developing supervised methods to accurately delineate field boundaries using Earth observation (EO) data. This letter introduces an open dataset for training and benchmarking machine learning methods to delineate agricultural field boundaries in polygon format. The large-scale dataset consists of 439 001 field polygons divided into 62 tiles of approximately 5 × 5 km distributed across Vietnam and Cambodia, covering a range of fields and diverse landscape types. The field polygons have been meticulously digitized from satellite images, following a rigorous multistep quality control process and topological consistency checks. Multitemporal composites of Sentinel-2 (S2) images are provided to ensure cloud-free data. We conducted an experimental analysis testing a state-of-the-art deep learning (DL) workflow based on fully convolutional networks (FCNs), contour closing, and polygonization. We anticipate that this large-scale dataset will enable researchers to further enhance the delineation of agricultural fields in smallholder farms and to support the achievement of the Sustainable Development Goals (SDGs). The dataset can be downloaded from https://doi.org/10.17026/dans-xy6-ngg6.**

***Index Terms*— Cambodia, crop field boundaries, deep learning (DL), Sentinel-2 (S2) data, smallholder farms, Vietnam.**

I. Introduction

CROP field polygons enable digital agriculture services and can record specific information such as crop type, soil characteristics, yield, and farming practices in a spatial database [1]. However, field boundaries are not yet available in many countries, especially in Asian and African regions where smallholder farms with fields smaller than two hectares comprise 70% of the cropland [2]. This information gap hampers the achievement of Target 2.3 of the United Nations Sustainable Development Goals (SDGs), which emphasize the need to improve the agricultural productivity of small-scale food producers to achieve food security, improved nutrition, and sustainable agriculture. Moreover, crop field boundaries often correspond to visible cadastral parcels, which are essential parts of the country's land administration systems. Land tenure security offers many benefits to farmers, such as access to credit and investments, government and insurance services, and reduced conflicts. The importance of secure tenure rights to land, with legally recognized documentation, is recognized by Target 1.4 of the SDG agenda as an essential factor in ending poverty.

Recent developments in deep learning (DL) and Earth observation (EO) show that field boundaries can be effectively delineated using very high resolution (VHR) images [1], [3]. However, the main bottleneck that limits the generalization ability and general performance of such methods in smallholder farms is the lack of a large amount of reference data [4]. While several benchmark datasets focus on crop-type mapping in Europe [5], [6], [7] and a few in Africa [8], [9], little has been done so far for crop boundary delineation. d'Andrimont et al. [10] introduced a dataset of images and field boundary labels collected from seven European countries representing 14.8 M parcels and covering 372 K km^2. It is far more challenging to access crop field boundary data in smallholder farms that are common in many low- and middle-income countries. Wang et al. [4] publicly released a dataset containing 10 000 Indian field boundary labels. However, only a fraction of the fields in each image are labeled.

In this letter, we introduce AI4SmallFarms, a benchmark dataset for the automated extraction of field polygons in small-scale farms located in two South Asian countries: Vietnam and Cambodia. These countries are characterized by fragmented agricultural areas with small fields of less than one hectare. Although VHR images can provide accurate results, their high acquisition cost hinders the use of commercial images for extracting and updating crop boundaries. Moreover, their restricted policies prevent the creation of open datasets. Therefore, we explore here the potentials of Sentinel-2 (S2) data for field boundary delineation [11], which are available openly and freely to all the users. Nevertheless, the coarser resolution of 10 m poses significant challenges in accurately delineating the small-size fields present in smallholder farms.

To the best of our knowledge, AI4SmallFarms is the first large-scale open dataset available to the public for smallholder farming in Southeast Asia. The dataset aims to support the development of operational mapping and monitoring systems of crop boundaries in Asia. It also aims to facilitate the development of machine learning methods and support the achievement of SDG 2 "Zero Hunger" and ongoing initiatives of the Food and Agriculture Organization (FAO) fostering the implementation of effective food security measures. This activity has been carried out in collaboration with the Image

Manuscript received 21 July 2023; revised 21 September 2023; accepted 30 September 2023. Date of publication 10 October 2023; date of current version 7 November 2023. This work was supported by GRSS and the Ministry of Agriculture, Forestry and Fisheries (MAFF) of Japan and Implemented by the Statistics Division of the Food and Agriculture Organization (FAO) of the United Nations, within the Framework of the "Developing a spatially-explicit agricultural database in support of agricultural planning" under Project GCP/GLO/1019/JPN. *(Corresponding author: Claudio Persello.)*

Claudio Persello, Jeroen Grift, Claudia Paris, Mila Koeva, and Andrew Nelson are with the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500 AE Enschede, The Netherlands (e-mail: c.persello@utwente.nl).

Xinyan Fan is with the Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, The Netherlands.

Ronny Hänsch is with the German Aerospace Center (DLR), 82234 Weßling, Germany.

Digital Object Identifier 10.1109/LGRS.2023.3323095

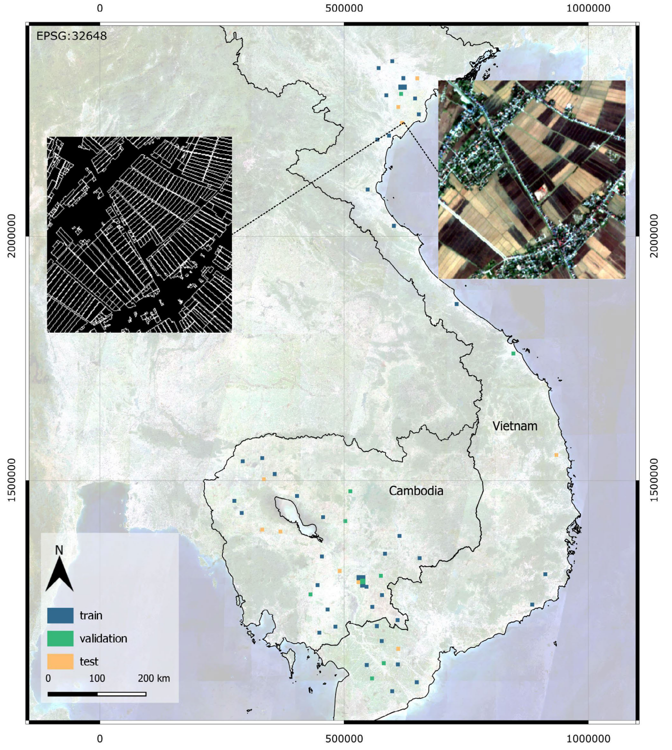

Fig. 1. Spatial distribution of the training, validation, and test tiles shown in blue, yellow, and green, respectively. For one tile, the S2 image and the corresponding reference boundaries are displayed.

Analysis and Data Fusion (IADF) Technical Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

II. Study Areas and Benchmark Dataset

Although agriculture is a key sector in Vietnam and Cambodia, involving a large portion of their population, these countries lack comprehensive, high-quality, and up-to-date field boundary data. Having rice paddy occupying nearly 80% of the harvested area in Cambodia, and Vietnam being one of the largest rice exporters at the global level, both the countries are mainly focused on rice production. Their agricultural system is predominantly characterized by small-scale farming, with subsistence farming systems located in rural areas.

A. Dataset Organization

The proposed AI4SmallFarms dataset is made up of 62 tiles having a size of approximately 5 × 5 km, where professional photograph interpreters manually digitized all the visible agricultural fields. Fig. 1 shows the spatial distribution of the reference tiles across most of the agricultural regions present in Vietnam and Cambodia. The tiles are split into spatially disjoint subsets for training, validating, and testing machine learning algorithms. Fig. 1 also shows an example of an S2 image tile and the corresponding digitized crop boundaries.

Table I shows the number of tiles, the number of field polygons, and the average crop size present in the training, validation, and test sets. The reference boundary are manually digitized resulting in 318 088 and 120 913 field polygons in Cambodia and Vietnam, respectively. Differently from [4], all the visible crop boundaries within each tile are manually digitized, resulting in a total number of 439 001 polygons.

TABLE I

Number of Tiles, Field Polygons, and Average Crop Size in Each Split of the Proposed Benchmark Dataset

	N. of Tiles	N. of Polygons	Avg. Crop Size (m^2)
Training	43	304777	2920
Validation	9	83066	2385
Test	10	51158	4042
Total	62	439001	2950

The field boundaries are published in a vector format as polygons and polylines. Using the geo-referenced vector data, any EO satellite data can be associated with the crop field boundaries. Details of the delineation procedure are provided in Section II-B. The benchmark dataset also includes preprocessed S2 imagery (details in Section II-C).

To facilitate the training process on S2 data, a freely available agricultural field boundary dataset [Basisregistratie Gewaspercelen (BRP)] and S2 data from The Netherlands are used to pretrain the DL-based workflow. This dataset consists of 87 tiles of 10 × 10 km^2 and 192 321 field polygons and is released together with AI4SmallFarms to enable the reproducibility of the experiments.

To assess the impact of the spatial resolution in delineating crop boundaries in the considered complex agricultural area, experiments are also carried out considering freely downloadable VHR Google Map (GM) RGB images with a spatial resolution of about 0.5 m (not included in the benchmark dataset because of redistribution policy restrictions). The satellite imagery along with the reference polygons are used to create image/label pairs. This is done for all the plots in Vietnam and Cambodia and for both S2 data and VHR GM data.[1] For the purpose of training and accuracy assessment, the rasterized reference boundaries of the VHR GM data are buffered with a distance of 1 m, while S2 reference data are not buffered.

B. Digitization Method

The digitization of reference polygons has been carried out by visual interpretation of S2 and VHR GM images acquired in August 2021.[2] Each tile is fully digitized, including all the visible agricultural boundaries. The process followed a rigorous multistep approach involving two quality control gates and topological consistency checks. First, agricultural field boundaries are visually detected and digitized in polygon format by the production team. Second, all the polygons are checked by an independent quality control team to verify shape correctness, missing polygons, and software-driven checks to identify topological errors (e.g., to avoid overlap between polygons or unnecessary gaps between adjacent fields). All the minor errors have been corrected by the quality control team. In case of major errors, data have been reverted to the production team for rework. Third, 10% of the vectors have been further checked by a second quality control team to verify the projection system, data completeness, nomenclatures of files, and output format.

[1] Processed in FME Workbench 2023.0.

[2] The digitization process has been carried out with the support of RMSI https://www.rmsi.com/ in collaboration with IADF and with the financial support of GRSS.

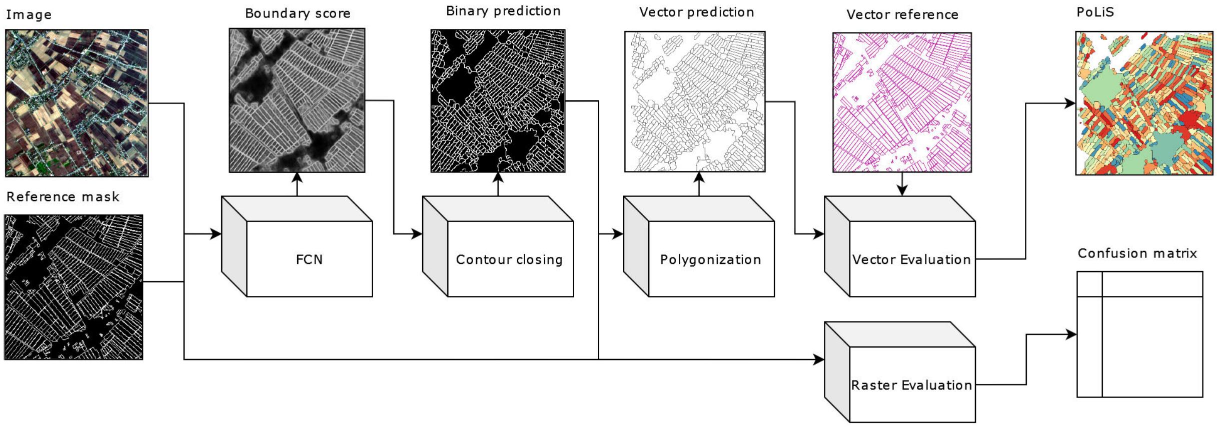

Fig. 2. Overview of the adopted DL-based workflow for field boundary delineation in polygon format.

C. S2 Satellite Data

AI4SmallFarms is designed to advance the capability of using S2 data for crop boundary delineation in smallholder farms. To ensure the reproducibility of our experimental analysis and foster further developments, the dataset includes the preprocessed S2 images used to perform the experiments reported in this letter. Because of the heavy cloud coverage typically present in the considered study areas, the atmospherically corrected S2 L2A images are preprocessed to generate monthly composites almost free of clouds using the standard statistics-based approach described in [6]. This approach aggregates all the S2 images acquired within a month into a composite by calculating the median pixel value, excluding the pixels that are affected by clouds or shadows. In the considered experimental analysis, we focused on the least cloudy months of the dry season, i.e., mainly January and February (and also November, December, and May in some cases), by selecting only the S2 images having cloud cover $< 40\%$. For each tile, the monthly composites obtained were visually inspected to select the least cloudy one. Only the blue (B2–490 nm), green (B3–560 nm), red (B4–665 nm), and near -infrared (B8-842 nm) bands are considered, since these are the most relevant spectral bands for crop boundary delineation and the ones acquired at the highest spatial resolution, i.e., 10 m.

III. Experimental Analysis

We conducted an experimental analysis using both the S2 and GM images with a state-of-the-art DL-based workflow inspired from [1] and depicted in Fig. 2 consisting of three main consecutive steps: 1) semantic segmentation using an fully convolutional network (FCN); 2) contour closing; and 3) polygonization.

The first step generates pixelwise predictions of boundary probability scores. For this purpose, we adopted a UNet-like architecture [12].[3] Since input image/label tiles have varying sizes, the data were split into patches of fixed dimensions. We used patches of 256 × 256 pixels for S2, and 512 × 512 for the GM images. We compared three models: 1) trained from scratch with S2 data; 2) pretrained in The Netherlands with S2 data and fine-tuned on AI4SmallFarms; and 3) trained from scratch with GM data. The hyperparameters of the three networks were tuned using the validation set. All the models were trained with the Adam optimizer (initial learning rate = 0.001, decay step = 10 000, and decay rate = 0.9), a focal cross-entropy loss function, and an early stopping procedure (max 1000 epochs). Table II reports the values of the other hyperparameters.

The output of the FCN does not guarantee obtaining closed contours and often results in fragmented lines that cannot be directly converted into polygons. For this reason, we adopted a contour closing procedure that allows us to derive a closed segment for each field.[4] The procedure consists of the following steps: 1) extended-minima transform; 2) impose minima; 3) connected component labeling; and 4) watershed transform. In the polygonization step, the closed contours were converted into vector format (polylines), simplified with the Douglas-Peucker algorithm, and finally, converted into polygons.[5]

To assess the accuracy of the predicted boundaries, both the raster-based and vector-based metrics are used. The confusion matrix is used to calculate raster metrics, i.e., precision, recall, and F1-score of the class boundary. These metrics are calculated per tile using the binary output of the contour closing step and the corresponding reference boundaries. However, raster metrics do not fully capture the agreement between polygons. Therefore, we propose to adopt the PoLiS metric [13], which analyzes the difference in position and shape between the predicted and reference polygons. Since the computation of PoLiS takes matched pairs of the prediction–reference polygons as input, its value is independent of the spatial resolution and the thickness of the raster boundaries, which critically affects the raster-based metrics. For this reason, we consider PoLiS a more suitable metric for field boundary delineation tasks, especially when comparing results obtained from images of different resolutions.

IV. Experimental Results

A. Quantitative Results

Table III shows the quantitative results of our experiments using S2 (with and without pretraining) and GM images as input to our workflow. For the S2 experiments, we observe

[3] https://deepsense.ai/deep-learning-for-satellite-imagery-via-image-segmentation

[4] https://imagej.net/plugins/morphological-segmentation

[5] Processed in ArcGIS Pro 3.1.1.

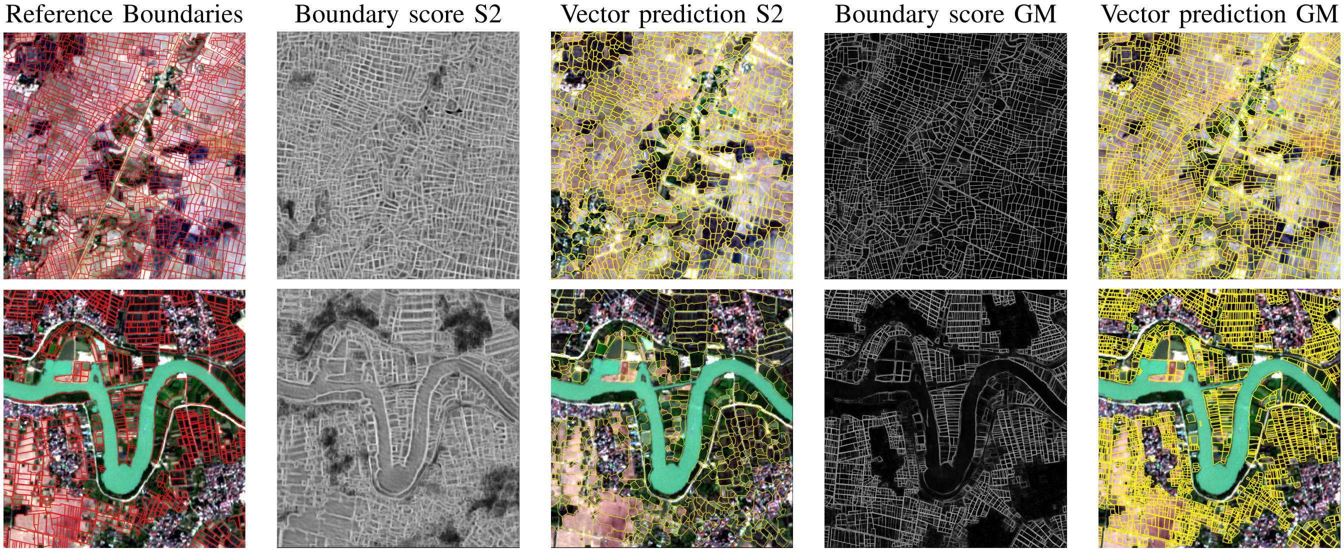

Fig. 3. Qualitative results of tiles Cambodia 12 (top) and Vietnam 27 (bottom).

TABLE II

SELECTED HYPERPARAMETERS OF THE THREE MODELS

	S2 (scratch)	S2 (pretrained)	GM
Batch size	8	2	8
Segmentation tolerance	0.05	0.1	0.1
Simplify tolerance	10	10	2

TABLE III

QUANTITATIVE AVERAGE RESULTS ACROSS ALL TEST TILES

	Precision	Recall	F1-score	PoLiS (m)
S2 (from scratch)	0.42	0.36	0.38	27.0
S2 (pre-trained)	0.48	0.33	0.39	26.7
GM Images	0.47	0.52	0.49	20.3

that the model pretrained on the Dutch dataset and fine-tuned in Cambodia and Vietnam resulted in an average F1-score of 0.39 and an average PoLiS distance of 26.7 m, and it is only slightly superior to the model trained from scratch, which results in an average F1-score of 0.38 and an average PoLiS distance of 27.0 m. This result shows that our benchmark dataset is large enough to obtain accurate results without the need for pretraining.

The average F1-score of the GM experiment is 0.49. The high spatial resolution of the GM results in an average PoLiS value of 20.3 m, which is significantly better than the average PoLiS value of 26.7 m obtained using S2 data.

B. Qualitative Results

We selected two tiles (tile 12 and 27) to illustrate and assess the quality of the predicted polygons. Fig. 3 shows the reference data, the FCN output (boundary score), and the final vector predictions for both the tiles. As expected, vector predictions from the GM experiment are more regular compared with the predictions from the S2 experiment. Next to that, the boundary scores from the S2 experiment are sometimes too coarse, resulting in missing boundaries. These differences are mainly due to the higher spatial resolution of the GM data. Both the differences can also be observed in the calculated PoLiS metric. The average PoLiS distance for tile 12 is 11.1 and 17.1 m for the GM and S2 predictions, respectively. For tile 27, the average PoLiS metric is 19.4 and 35.0 m for the GM and S2 predictions, respectively. Fig. 4 shows a closeup of the results on tile 12. The GM predictions almost fully cover the reference data and are relatively regular, whereas the S2 predictions have a lot of missing boundaries and are not as regular as the GM predictions.

V. CHALLENGES

This section summarizes the main challenges posed by the proposed dataset.

A. Crop Field Size

Due to the small average crop sizes (far below one hectare), it is difficult to accurately delineate crop boundaries in smallholder farms using S2 images. However, these free and open-access data enable the regular update of crop boundary databases. For this reason, we see the need for developing super-resolution semantic boundary detection strategies to further increase the spatial resolution of results, e.g., by exploiting the spatial-contextual information in label space [11] or adopting generative models.

B. Satellite Data Preparation

Because of the tropical climate, these areas are typically affected by heavy cloud cover. This may hamper the possibility of using specific temporal acquisitions that can emphasize the phenological state of the crops. The least cloudy period is the dry season, when crops may exhibit low contrast to the background, leading to poor crop delineation results. Possible future work may investigate the use of synthetic aperture radar (SAR) data.

C. Imbalanced Classification Problem

The presented crop boundary delineation task poses an extremely imbalanced classification problem due to the prevalence of nonboundary pixels. For this reason, we adopted

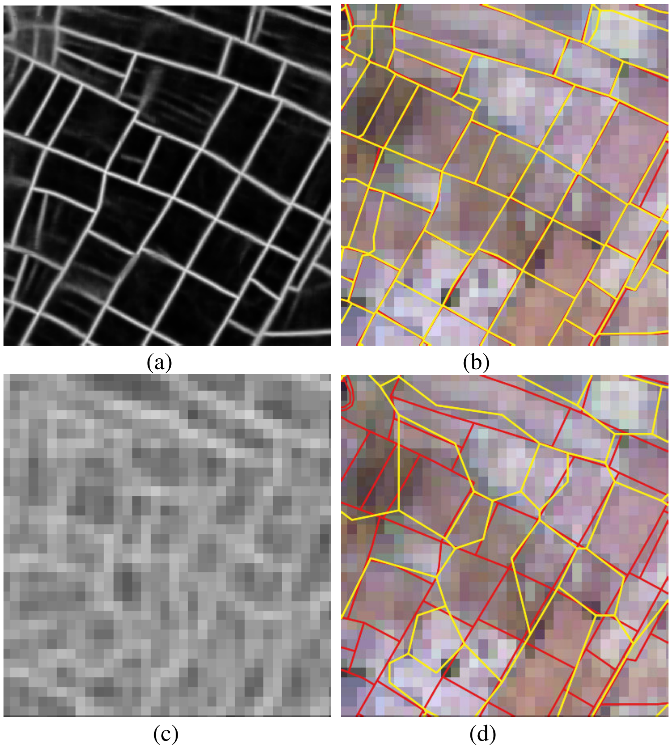

Fig. 4. Closeup of tile Cambodia 12 showing the limitation of S2 with respect to GM in capturing small-sized fields. Reference boundaries are in red and predictions in yellow. (a) Boundary score GM. (b) Prediction/reference GM. (c) Boundary score S2. (d) Prediction/reference S2.

the focal cross-entropy loss in our experiments. Nevertheless, developing a loss function tailored to the specific properties of the crop boundary delineation task is expected to improve the result and mitigate the extreme class imbalance.

D. Crop Variability

AI4SmallFarms consists of tiles showing substantial variability among fields, different landscape conditions, soil characteristics, crop management practices, and crop arrangements. This dataset provides an opportunity to assess the generalization capabilities of crop boundary delineation approaches. Indeed, it is hard for a single model to achieve accurate delineation results across all the considered tiles. This dataset aims to support the development of robust models that show improved generalization ability when encountering unseen landscapes. Future studies may explore advanced DL solutions based on semisupervised learning or self-training. Moreover, we expect that recent semantic segmentation models based, for example, on vision transformers (e.g., Swin Transformer) may further improve the feature extraction performance.

VI. Conclusion

This letter presented AI4SmallFarms, a large-scale dataset for field boundary delineation in fragmented agricultural areas characterized by small fields. The dataset comprises 439 001 field polygons divided into 62 nonoverlapping tiles of approximately 5 × 5 km distributed across Vietnam and Cambodia. The field polygons were meticulously digitized using S2 and GM images, following a rigorous multistep quality control procedure and topological consistency checks. For each tile, all the visible crops were manually digitized. To ensure the reproducibility of the experimental results, the dataset is provided into three spatially disjoint sets, i.e., training, validation, and test tiles. The experimental results obtained with a state-of-the-art DL-based workflow highlight the main challenges of the proposed dataset and its properties. AI4SmallFarms is publicly available to support the community in advancing the development of supervised machine learning methods for crop boundary delineation in agricultural areas with smallholder farms using S2 data.

Acknowledgment

The authors would like to thank FAO and GRSS for financing this research activity. They also would like to thank RMSI, New Delhi, India, for the meticulous digitization work.

References

[1] C. Persello, V. A. Tolpekin, J. R. Bergado, and R. A. de By, “Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping,” *Remote Sens. Environ.*, vol. 231, Sep. 2019, Art. no. 111253.

[2] M. Lesiv et al., “Estimating the global distribution of field size using crowdsourcing,” *Global Change Biol.*, vol. 25, no. 1, pp. 174–186, Jan. 2019.

[3] C. Persello et al., “Deep learning and Earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities,” *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 172–200, Jun. 2022.

[4] S. Wang, F. Waldner, and D. B. Lobell, “Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision,” *Remote Sens.*, vol. 14, no. 22, p. 5738, Nov. 2022.

[5] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, “BreizhCrops: A time series dataset for crop type mapping,” *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2-2020, pp. 1545–1551, Aug. 2020.

[6] G. Weikmann, C. Paris, and L. Bruzzone, “TimeSen2Crop: A million labeled samples dataset of sentinel 2 image time series for crop-type classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4699–4708, 2021.

[7] D. Sykas, M. Sdraka, D. Zografakis, and I. Papoutsis, “A Sentinel-2 multiyear, multicountry benchmark dataset for crop classification and segmentation with deep learning,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3323–3339, 2022.

[8] C. Bocquet, *Dalberg Data Insights Uganda Crop Classification*. Washington, DC, USA: Version 1.0, Radiant MLHub, Nov. 2022, doi: 10.34911/RDNT.EII04X.

[9] J. Rineer, *Drone Imagery Classification Training Dataset for Crop Types in Rwanda*. Washington, DC, USA: Version 1.0, Radiant MLHub, 2021, doi: 10.34911/rdnt.r4p1fr.

[10] R. d’Andrimont et al., “AI4Boundaries: An open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography,” *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 317–329, Jan. 2023.

[11] K. M. Masoud, C. Persello, and V. A. Tolpekin, “Delineation of agricultural field boundaries from Sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks,” *Remote Sens.*, vol. 12, no. 1, p. 59, Dec. 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.

[13] J. Avbelj, R. Müller, and R. Bamler, “A metric for polygon comparison and building extraction evaluation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 170–174, Jan. 2015.