Smart Agricultural Technology 14 (2026) 102384

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Zero-shot parcel instance delineation in fragmented agricultural landscapes: Cross-source evaluation and structure-aware validation

Joao Calisto, Massimo Vecchio *, Miguel Pincheira, Fabio Antonelli

Fondazione Bruno Kessler (FBK), Trento, Italy

ARTICLE INFO

Keywords:
Foundation models
Prompt-driven segmentation
Instance segmentation
Vegetation enhancement
Fragmented landscapes
Structural consistency

ABSTRACT

Agricultural parcel and boundary delineation (APBD) remains challenging in fragmented agricultural landscapes, where small, densely adjacent fields increase merge and split errors even when overlap-based scores appear acceptable. This study addresses the problem as a zero-shot, cross-source, and structure-aware segmentation task for operational use without site-specific retraining or annotated training data. We combine vegetation-enhancement (VE) preprocessing with foundation-model-based instance segmentation and evaluate the pipeline across Sentinel-2 (S2, 10 m), PlanetScope (PS, 3 m), and Sentinel-2 Deep Resolution 3.0 (S2DR3, 1 m). Evaluation uses a structure-aware protocol that complements standard detection metrics with explicit merge and split quantification, enabling assessment of geographic information system (GIS)-ready parcel outputs beyond overlap alone. On a manually annotated benchmark of 424 parcels in Trentino (Italy), the best configuration (Vegetation-Enhancement + Segment Anything Model (VE + SAM) with S2DR3) achieves F1 0.79 and recall 0.78, with the lowest Global Under-Segmentation Error (GUSE 0.21). At source level, moving from S2 to S2DR3 nearly triples recall and halves GUSE, confirming that input resolution is a dominant factor in highly fragmented small-parcel settings. External validation on three Dutch tiles (300 km^2), where parcels are larger and fragmentation is lower, shows a weaker resolution effect: S2 is no longer fundamentally inadequate and gains from finer imagery are smaller. In this setting, VE + SAM still maintains a consistent structural advantage over the domain-trained baseline, with systematically lower split and merge error rates, indicating that structural gains from the proposed pipeline persist when resolution is less limiting.

1. Introduction

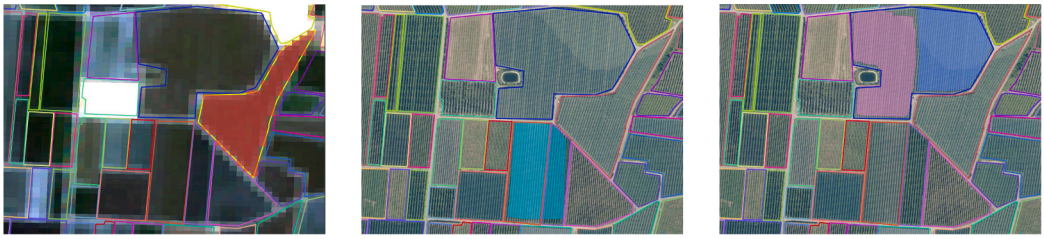
Accurate delineation of agricultural parcels is a practical requirement for operational agricultural workflows because many applications depend on parcel-level spatial units rather than on pixel-level predictions alone [1–3]. These workflows include field-scale crop monitoring and forecasting, machinery path planning, parcel-based change analysis, and Geographic Information System (GIS)-supported farm management [1–3]. In this context, Agricultural Parcel and Boundary Delineation (APBD) aims to identify individual parcels and their boundaries from geospatial imagery so that downstream analyses can operate on coherent management units instead of fragmented image regions [4,5].

Robustness remains limited in fragmented agricultural landscapes, where parcels are often small, irregular, densely adjacent, and weakly separated in the imagery [4,6]. In these settings, weak boundaries often cause adjacent parcels to “stick” together, leading to incorrect merges between neighboring instances [7]. Under these conditions, even small changes in image resolution, seasonal appearance, or parcel shape can affect whether neighboring fields are separated correctly.

For operational parcel mapping, the main failure mode is often structural rather than purely geometric. In fragmented agricultural landscapes, two adjacent parcels may be mistakenly fused into a single unit, or one parcel may be fragmented into multiple instances, even when the predicted contours remain locally plausible [8,9]. A first source of difficulty is coarse spatial resolution itself, which can make parcel boundaries too weak or ambiguous to be localized consistently in fragmented landscapes [10]. Merge and split errors can therefore be more harmful than moderate boundary offsets because they compromise instance integrity, even when overlap-based scores remain acceptable [11,12]. These three delineation problems are illustrated in Fig. 1. As a result, a parcel layer may look acceptable quantitatively while remaining awkward to use for parcel-based analysis or vectorization [13]. The literature also shows a recurring problem: validation is often not reported in enough detail, and evidence on portability across agricultural settings

* Corresponding author.
E-mail addresses: joaopcalisto@fbk.eu (J. Calisto), mvecchio@fbk.eu (M. Vecchio), mpincheiracaro@fbk.eu (M. Pincheira), fantonelli@fbk.eu (F. Antonelli).

https://doi.org/10.1016/j.atech.2026.102384
Received 26 March 2026; Received in revised form 3 July 2026; Accepted 8 July 2026
Available online 9 July 2026



(a) Boundary delineation at coarse spatial resolution.

(b) Merge error: two distinct parcels predicted as a single instance.

(c) Split error: one parcel fragmented into multiple predicted instances.

Fig. 1. Representative delineation difficulties in fragmented agricultural landscapes. Colored outlines indicate ground-truth parcel boundaries, whereas colored patches indicate predicted instances.

is still limited. This makes it hard to tell how well published methods would hold up in fragmented landscapes and under heterogeneous image conditions [4,14].

For these reasons, zero-shot and foundation models are appealing in agricultural settings, where repeated annotation and site-specific retraining can quickly become costly. However, current evidence suggests that these approaches still face important limitations. For example, Segment Anything Model (SAM), in automatic mask generation (AMG) mode, tends to over-segment agricultural scenes, producing high recall at the cost of precision, as documented in independent evaluations of SAM in farming contexts [4,15,16].

Motivated by these limitations, we study APBD as a zero-shot, cross-source, and structure-aware problem oriented toward operational agricultural use. Rather than asking only whether overlap scores improve, we ask whether modern segmentation strategies can return coherent parcel instances that remain usable in downstream GIS workflows when applied across different image sources. In particular, we examine how imagery characteristics and preprocessing interact with segmentation behavior across Sentinel-2, PlanetScope, and Sentinel-2 Deep Resolution 3.0 imagery, with emphasis on fragmented agricultural landscapes where structural errors are especially consequential [17,18].

The study makes three contributions. First, we present a zero-shot APBD pipeline that combines vegetation-enhancement preprocessing with foundation-model-based instance segmentation and is designed to operate across heterogeneous optical sources without source-specific retraining.

Second, we provide a controlled cross-source evaluation of two contrasting segmentation paradigms, namely prompt-driven foundation models and a detector-style baseline, under a consistent zero-shot setting across Sentinel-2, PlanetScope, and Sentinel-2 Deep Resolution 3.0 imagery. This comparison clarifies practical trade-offs between open-access and commercial data, as well as between zero-shot generalist and domain-specific approaches [18].

Third, we adopt an evaluation protocol that extends standard instance-detection metrics with structure-aware error analysis, specifically merge and split quantification, to assess whether the resulting parcel outputs are usable in downstream GIS workflows [11–13], going beyond what global overlap scores alone can reveal.

Taken together, these choices address the three gaps identified above: the lack of controlled cross-source comparisons under consistent evaluation settings, limited evidence on zero-shot transferability across imaging conditions, and the underemphasis on structure-aware validation in fragmented agricultural landscapes.

2. Related work

2.1. Supervised APBD methods

Traditional and handcrafted approaches. Early automated parcel delineation relied on classical computer-vision techniques. Edge-detection operators such as Canny, Sobel, and Laplacian of Gaussian highlighted reflectance discontinuities to approximate parcel borders [19, 20, 21], while unsupervised segmentation algorithms including watershed, graph-based methods, and Simple Linear Iterative Clustering (SLIC) grouped spectrally similar pixels into coherent regions [22]. These methods proved tractable in homogeneous, regularly structured landscapes but degraded substantially in heterogeneous and fragmented settings, where boundary evidence is weak or discontinuous due to crop variability, topography, and management practices. Their sensitivity to radiometric noise and inability to generalise across sensors motivated the shift toward data-driven approaches.

CNN-based supervised methods.. As open satellite archives and computational resources expanded, supervised deep learning became the dominant paradigm. Encoder-decoder Convolutional Neural Networks (CNN) architectures (e.g., U-Net, SegNet) demonstrated end-to-end field boundary extraction from high-resolution imagery [21,22]. A landmark contribution in this direction is the work by Waldner and Diakogiannis [10], who showed that CNNs trained on edge imagery could extract field boundaries from satellite data at scale, establishing a widely used baseline for deep-learning-based boundary delineation. Subsequent task-specific designs addressed the instance separation challenge more directly through multi-task and edge-aware supervision strategies. A semantic edge-aware multi-task neural network was proposed to jointly predict parcel interiors and boundaries, improving delineation accuracy in densely adjacent agricultural fields [23]. Similarly, CT-HiffNet proposed cross-task hierarchical feature fusion, combining boundary and interior prediction branches to resolve the merging artefacts common in adjacent parcels [24], while a Point-Line-Region interactive multi-task model decomposed parcel geometry into complementary geometric primitives to improve boundary precision in high-resolution imagery [25]. A generalized framework for agricultural field delineation further demonstrated that supervised pipelines can be designed for transferability across different satellite sources [26]. Multi-task architectures such as MSCPUnet added an auxiliary boundary-detection head alongside the main mask branch to alleviate the “sticking” problem between adjacent parcels [7]. These designs substantially improved delineation quality, but all remain dependent on annotated training data and exhibit performance degradation when applied outside their training distribution.

Transformer-based supervised methods. Vision Transformers and hybrid CNN–Transformer architectures improved long-range context modelling and cross-domain transfer compared with purely convolutional designs [27]. Applied to cropland field parcel mapping, HBGNet exploited hierarchical boundary-guided attention to capture multi-scale parcel structure [28], and CTHBNet combined CNN local features with Transformer global context and explicit boundary guidance for improved separation of adjacent parcels in high-resolution remote sensing imagery [29]. Multi-Swin Mask Transformer (MSMTransformer), Mask2Former-based, a query-based masked-attention Transformer trained end-to-end

for instance segmentation, has also been applied to agricultural parcel delineation in a supervised fine-tuning setting, demonstrating strong boundary-level accuracy on annotated parcel datasets [30]. Despite these advances, Transformer-based supervised methods retain the same fundamental deployment constraints as CNN-based ones: large annotated datasets, site-specific tuning, and sensitivity to domain shift. Recent studies also highlight that robustness in fragmented landscapes remains sensitive to parcel complexity and scene heterogeneity [4,7].

2.2. Foundation and zero-shot approaches

Foundation models changed how segmentation can be approached by enabling promptable inference and, in some cases, zero-shot use without task-specific retraining. In agricultural evaluations, SAM showed clear potential for rapid mask generation and semi-automatic annotation workflows [15]. Comparative studies confirm that fine-tuned variants such as LoRA-SAM can achieve superior instance separation in complex European agricultural landscapes relative to vanilla SAM, indicating that the zero-shot ceiling is not absolute [4]. In parcel-delineation settings, adaptations such as fabSAM applied these foundation-model concepts to farmland extraction [31]. More recently, hybrid approaches have begun to combine SAM's boundary-detection capability with supervised task-specific networks: TFNet integrates SAM-derived boundary priors as auxiliary spatial guidance within a supervised cropland extraction pipeline, demonstrating that foundation models can function as feature providers rather than end-to-end segmentors in agricultural settings [32]. At the same time, other state-of-the-art approaches have moved toward large-scale, task-specific instance-segmentation architectures. In particular, Delineate Anything (DelAny) and Delineate Anything Flow are built around YOLOv11-based instance segmentation models and target large-scale, zero-shot field-boundary mapping across diverse geographies [33,34]. This distinction matters because these approaches do not simply extend SAM-based prompting to the same problem setting, but instead represent a separate line of development oriented toward scalable, resolution-agnostic field instance delineation.

However, the current evidence base also points to important limitations. In automated settings, foundation-model-based segmentation may over-segment textured ground patterns, miss weak parcel boundaries, or fail to separate tightly adjacent field instances [15]. More recent task-specific instance-segmentation frameworks have addressed some of these limitations by improving scalability and zero-shot transfer across resolutions and geographic settings [33,34]. Nevertheless, comparative evidence still suggests that delineation performance remains sensitive to parcel complexity, and both SAM-based and task-specific large-scale pipelines require careful assessment under realistic agricultural image conditions [4,35].

2.3. Data-centric enhancement

Alongside model development, input quality strongly influences how clearly parcel structures can be observed in the imagery. Super-resolution and related enhancement methods are therefore relevant because they can make narrow or discontinuous boundaries more visible in otherwise coarse open-data inputs [17,18]. In this context, frameworks such as SIDEST and FieldSeg suggest that source characteristics and preprocessing can influence delineation behavior when field boundaries are extracted from heterogeneous imagery [35,36].

2.4. Research gap

Taken together, the literature leaves three practical gaps. First, controlled comparisons across methodological families remain limited under consistent evaluation settings, especially when parcel complexity is expected to influence performance [4,5]. Supervised methods achieve strong results in-domain, but their annotation dependency and sensitivity to domain shift mean that operational practitioners without retraining capacity face a substantially different evidence landscape. Second, evidence on cross-source robustness is still constrained, and validation transparency remains uneven across agricultural remote-sensing studies; heterogeneous split definitions and evaluation settings also make it difficult to judge how well published methods generalize when regions, sources, or image conditions change [14]. Third, many studies still underemphasize structure-aware validation, even though the practical usefulness of parcel delineation depends on more than global overlap scores [12,13]. These gaps motivate our evaluation framework, which jointly examines zero-shot behavior, cross-source transfer, and the structural usefulness of parcel outputs in downstream GIS workflows in fragmented agricultural landscapes.

These gaps motivate our evaluation framework, which jointly examines zero-shot behavior, cross-source transfer, and the structural usefulness of parcel outputs in downstream GIS workflows in fragmented agricultural landscapes.Beyond annotation cost, a less-discussed limitation of supervised fine-tuning concerns the quality of available training targets. Reference datasets such as the Land Parcel Identification System (LPIS) and Geospatial Aid Application (GSAA) map administrative ownership units rather than physical field boundaries [37]. As a result, they systematically carry temporal update lags, digitization errors, and administrative over and under-segmentation, where multiple physically distinct fields share a single cadastral record, or conversely, a single uniformly managed field is split across several administrative parcels. Training on such data risks biasing the model toward replicating administrative conventions rather than actual boundary geometry. Deploying foundation models in a zero-shot setting avoids this label-contamination pathway by relying on pre-trained visual primitives without task-specific supervision.

3. Data sources and models

3.1. Data sources

A key design property of the proposed approach is source-agnostic applicability: the same segmentation pipeline operates across heterogeneous optical inputs without source-specific retraining or adaptation. To evaluate this characteristic under controlled conditions, three input products with distinct spatial properties and access profiles are selected: Sentinel-2 (S2), PlanetScope (PS), and S2 Deep Resolution 3.0 (S2DR3). Together, these sources span the range from open-access medium-resolution imagery to commercial very-high-resolution data, enabling a systematic assessment of how source resolution and access profile interact with zero-shot segmentation behavior.

S2 imagery was retrieved through the Copernicus Data Space Ecosystem (CDSE) and PS data through the Planet API, in both cases using Python-based SDKs that support fully programmatic spatiotemporal queries and automated download. Table 1 reports the main sensor and acquisition characteristics for all three sources.

S2DR3 is a super-resolved product derived from S2 imagery [17]. It preserves the spectral and radiometric characteristics of S2 while increasing spatial detail through CNN-based reconstruction, achieving a target resolution of 1 m. Including S2DR3 allows the analysis to quantify how far resolution enhancement can close the gap between open-access and commercial very-high-resolution imagery for parcel delineation, without requiring access to a separate sensor.

3.2. Model families

Two distinct segmentation paradigms are compared under a common zero-shot evaluation setting. The first comprises promptable foundation models from the SAM family, which generate instance masks through automatic or prompt-based inference without task-specific training. The second is a detector-style instance segmentation model specifically designed for agricultural field delineation. This paradigm distinction is

Table 1
Input imagery sources and key sensor, resolution, and access characteristics used in this study.

Parameter	Sentinel-2	PlanetScope	S2DR3
Satellite / Sensor	Sentinel-2 MSI	PlanetScope SuperDove	S2-derived product
Spatial Resolution (m)	10, 20, 60	3	1 (reconstructed)
Spectral Resolution	13 bands (VNIR–SWIR)	8 bands (VNIR)	13 bands (VNIR–SWIR)
Temporal Resolution (days)	5	~1	5
Data Access	Copernicus Data Space Ecosystem API	Planet API	Proprietary Package[a]

[a] S2DR3 is distributed as an encrypted Python package, freely available for small-scale testing via Google Colab, but requires a commercial agreement for large-scale processing.

central to the study: it enables a direct comparison between general-purpose zero-shot segmentation and a domain-specific approach trained on large-scale agricultural data, under identical inference conditions and across the same imagery sources. It is important to note that this comparison is not symmetric with respect to pre-training data. SAM-family models are trained on SA-1B, SA-V, and SA-Co, which are broad general-purpose segmentation datasets, whereas DelAny is fine-tuned on FBIS-22M, a large-scale agricultural boundary dataset. This difference is intentional: the study compares two operational paradigms rather than simply two architectures. The practical question addressed is whether training-free zero-shot inference with a general-purpose foundation model can approach the performance of a domain-trained specialist under identical inference and evaluation conditions. Any performance gap between these paradigms should therefore be interpreted as reflecting this training-data distinction, not architectural capability alone. Correspondingly, SAM-based results in this study represent a conservative estimate of what general-purpose foundation models can achieve without agricultural or local fine-tuning.

SAM-based models. To map a realistic spectrum of zero-shot capabilities, the selected models trace the technological evolution and diverse prompting interfaces of the Segment Anything paradigm. The first paradigm is represented by three successive generations of the Segment Anything Model. SAM [38] is a Vision Transformer foundation model for prompt-based segmentation using a ViT-H backbone. SAM2 [39] extends the architecture to the video domain by introducing a hierarchical Vision Transformer (Hiera) and a streaming memory mechanism for real-time spatiotemporal object tracking. SAM3 [40] further extends the family with concept-aware segmentation, coupling a DETR-based detector with a SAM2 backbone to enable text and exemplar-based prompting across a large concept vocabulary. In this study, SAM and SAM2 are operated in automatic mask generation (AMG) mode, while SAM3 is used in text-prompted inference mode, as native AMG support is not available. Unlike SAM and SAM2, which generate masks exhaustively in AMG mode, SAM3 requires a text string to define the concept category to be segmented. To identify a prompt that SAM3 could meaningfully associate with individual agricultural parcels rather than broader land cover classes or irrelevant objects, a small set of candidate formulations was qualitatively checked on a randomly selected image region outside the evaluation area. This check was not a performance optimization: no metric was computed and no ground truth was used. Its sole purpose was to verify that the chosen prompt elicited instance-level parcel-shaped outputs rather than coarse semantic regions. Structural descriptors (e.g., shape and geometry cues) achieved this more reliably than semantic or crop-type formulations. Inference configurations for all three models are reported in Table 2, while the full parameter setup of the different models used for this study, including the complete set of evaluated SAM3 prompt candidates, is provided as a yaml file in [41].

Detector-style model. The second paradigm is represented by Delineate Anything (DelAny, [33]), a domain-specific instance segmentation model built on a YOLOv11-L backbone and fine-tuned on FBIS-22M, a large-scale benchmark containing approximately 22 million agricultural field boundaries across multiple countries and imagery sources.

Fig. 2. LangSAM failure case on agricultural parcel delineation. The red rectangle denotes the Grounding DINO bounding box and the green overlay the corresponding SAM mask. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In our experimental design, DelAny serves as the primary supervised control baseline. Because it represents a heavy-duty, domain-specific architecture explicitly pre-trained on tens of millions of actual agricultural shapes, including it allows us to rigorously test whether our zero-shot preprocessing pipeline can genuinely challenge or complement state-of-the-art fully supervised networks without requiring local dataset retraining. Unlike the SAM-based models, DelAny generates parcel instances through a proposal-driven pipeline that predicts bounding boxes and per-instance masks directly from image tiles without requiring prompts. DelAny represents a recent domain-specific baseline specifically designed for agricultural field delineation: it has been explicitly designed and trained for the agricultural field delineation task addressed in this study, making it the most relevant reference point for evaluating the practical value of the proposed zero-shot pipeline. Its inference configuration is reported in Table 2.

Additional models. Two further models are included at the margins of the comparison for specific purposes. FastSAM [42] is a lightweight CNN-based instance segmentation model built on YOLOv8-seg that produces instance masks in a single forward pass, included as a computational efficiency reference representing the lower end of the precision–cost trade-off; its inference configuration is reported in Table 2. LangSAM [43] extends SAM with language-guided prompting by coupling it with Grounding DINO [44] for open-vocabulary segmentation. In the tested configuration, LangSAM grouped the cultivated area into a single broad region rather than separating adjacent fields, as illustrated in Fig. 2, and did not produce valid parcel-level instance outputs. It is therefore retained here for completeness but excluded from the main quantitative comparison in Section 7.

Table 2
Inference configuration summary for all evaluated models. Parameters omitted for a model are not applicable (—). Full configurations are available in the companion repository. The arrow ↓ indicates that lower values are preferable.

Parameter	SAM	SAM2	SAM3	FastSAM	DelAny
Architecture and training					
Backbone	ViT-H	Hiera-L	DETR + SAM2	YOLOv8-x	YOLOv11-L
Inference mode	AMG	AMG	text prompt	everything	trained det.
Training data	SA-1B	SA-1B	SA-Co 4M +	SA-1B (2%)	FBIS-22M
Key inference parameters					
Points per side	64	64	—	—	—
IoU / score thresh.	0.80	0.80	0.35	0.40	0.25
Stability thresh.	0.95	0.95	—	—	—
NMS IoU thresh.	0.70	0.70	—	0.90	0.70
Image size (px)	—	—	—	1024	640
Text prompt	—	—	"Rectangular Agricultural Fields"	—	—
Computational efficiency					
Runtime (ms) ↓	4870	2630	200	30	20 / 23

Inference settings. All models are evaluated in zero-shot mode using publicly released pretrained weights with fixed inference-time parameters. SAM-based models are run using the Segment-Geospatial (SamGeo) package, version 0.16.0[1], and DelAny is run from its official open-source repository[2]. The inference parameters most relevant to the analysis are summarised in Table 2. Processing time values were measured on the same hardware configuration (NVIDIA Ampere A100 SXM 64GB), averaged over three runs, and reported as milliseconds per tile. These values should be interpreted together with each model's input-handling strategy. SAM-family models internally resize inputs to a fixed 1024 px inference grid, making runtime largely independent of the original tile size. DelAny processes native input dimensions through a convolutional architecture, so runtime scales with tile size; for this reason, DelAny is reported for both 256×256 px and 1024×1024 px tiles. To isolate core model performance, DelAny timings reflect the direct forward-pass execution of the repository's open-source model weights[3], excluding all external data-loading loops and downstream vector-cleaning operations.

4. Zero-shot APBD inference pipeline

The proposed pipeline is zero-shot by design: all models operate using publicly released pretrained weights with fixed inference-time parameters, no site-specific fine-tuning is performed, and no annotated parcel data is required at any stage. This property is what enables the cross-source evaluation in Section 3 and the direct comparison with the detector-style baseline in Section 7 to be conducted under identical, reproducible inference conditions across all evaluated sources. The pipeline consists of four sequential stages: input preprocessing, tiling and model inference, post-processing and cross-tile consolidation, and raster-to-vector conversion. The workflow presented in Fig. 3 provides a high-level overview of the complete pipeline, and each stage is described in the subsections that follow. We do not introduce a new segmentation architecture or algorithmic component. The pipeline is used as a reproducible experimental protocol, not as an algorithmic contribution. It standardizes preprocessing, tiling, prompt generation, inference, and post-processing across models, enabling a controlled comparison between zero-shot foundation models and a supervised agricultural baseline under common evaluation conditions.

4.1. Input preprocessing

In the first stage, imagery inputs described in Section 3.1 undergo standardized preprocessing that converts raw spectral bands into red, green and blue (RGB) representations and applies contrast stretching. A vegetation-enhancement step then operates in hue, saturation and value (HSV) space, where adaptive histogram equalization is applied to the saturation and value channels, while the hue channel remains unchanged to preserve crop color identity. This step increases local color and brightness contrast within vegetation and reduces illumination effects, which produces more uniform field interiors and sharper transitions at parcel boundaries. The adjusted images are converted back to RGB and passed to the compared segmentation models, where the stronger local contrast supports mask separation between adjacent agricultural fields. Section 7 demonstrates the effect of this image filtering on segmentation quality.

4.2. Tiling and model inference

After preprocessing, the images are divided into spatial tiles, which are then processed independently by the evaluated segmentation models. A 50% overlap is introduced to reduce border artefacts and improve instance coverage: parcels located near tile boundaries in one pass appear closer to the center in a neighboring tile, where they are observed under a more favorable spatial context. This strategy is particularly important for SAM-based models, whose mask proposals are sensitive to local image context and may miss some parcels in a single pass. Overlap, therefore, increases the likelihood that parcels missed in one tile are recovered in an adjacent one.

For SAM-based models, tile size is adapted to the spatial resolution of each source. Tiles of 1024×1024 pixels are used for 1 m imagery, 512×512 for PlanetScope at about 3 m resolution, and 256×256 for Sentinel-2 at 10 m resolution. The number of automatic grid prompts remains fixed at 64 points per side across all sources. This resolution-aware tiling strategy preserves the default encoder input size for 1 m imagery while maintaining comparable parcel detectability across sources with different spatial characteristics. At inference time, each tile is processed according to the corresponding model pipeline. For SAM and SAM2, AMG returns candidate instance masks, which are filtered using the predicted IoU and stability thresholds. SAM3 outputs prompt-based instance masks directly. In LangSAM, Grounding DINO first localizes candidate parcel regions from the text prompt, and SAM then refines them into pixel-level instance masks. FastSAM and DelAny produce instance masks in a single forward pass and are subsequently filtered using the `conf` and `iou` thresholds.

4.3. Post-processing and cross-tile consolidation

The resulting predictions are post-processed to remove spurious instances, consolidate overlapping tile outputs, and generate final parcel geometries. First, predicted instances are filtered against an external

[1] https://samgeo.gishub.org/
[2] https://lavreniuk.github.io/Delineate-Anything/
[3] https://huggingface.co/MykolaL/DelineateAnything

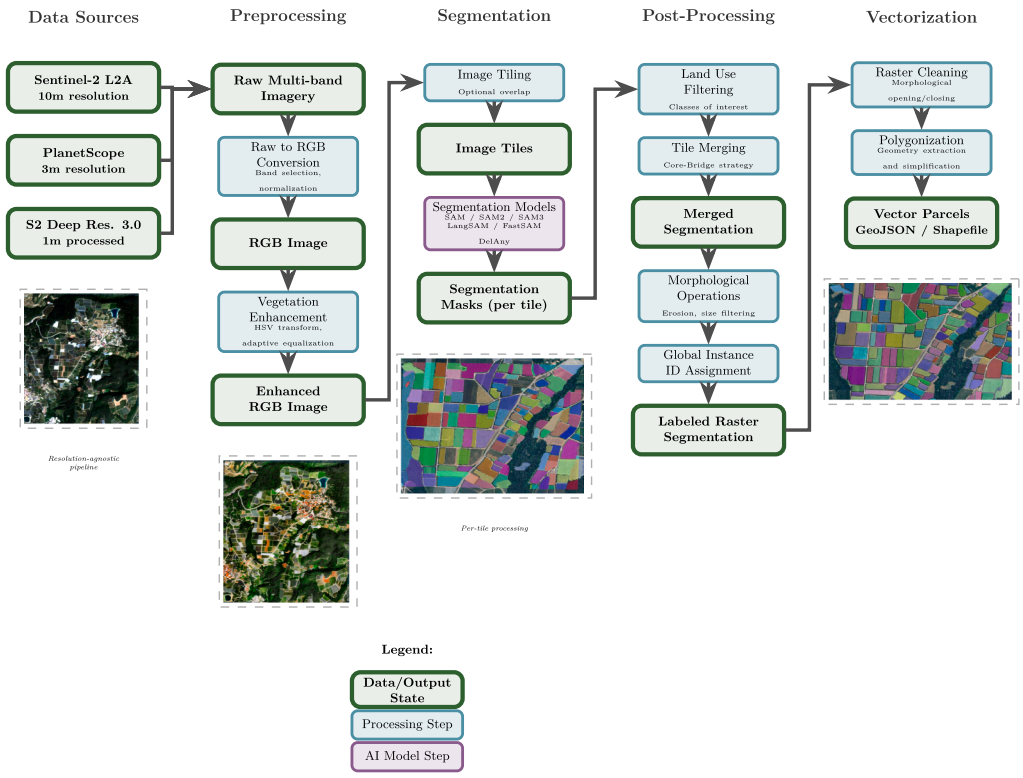


Fig. 3. End-to-end agricultural parcel segmentation pipeline, including preprocessing, optional vegetation enhancement (VE), segmentation, post-processing, and vectorization.

land-use raster, such as a national or regional land-cover map. In this work, the Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) National Land Cover Map is used for the Trentino imagery, and OSMlanduse is used for the Netherlands validation data [45]. Only agriculture-related land-use classes are retained as valid targets. The raster is reprojected to each tile's coordinate reference system and spatial resolution through nearest-neighbor resampling, and instances with an overlap ratio below 0.5 are discarded. This step removes predictions falling on roads, buildings, or water bodies.

Filtered tile outputs are then merged into a scene-wide mosaic using a two-stage interleaved tile-merging procedure. Tiles are partitioned into alternating sets based on their grid positions. The first set populates the mosaic, while the second fills the remaining gaps and improves continuity across tile boundaries. Before merging, instance identifiers are remapped to globally unique values to avoid conflicts across tiles.

The assembled mosaic is subsequently regularized by instance-wise morphological erosion using a metric kernel converted to pixel values at each source resolution. To avoid excessive shape loss, the original mask is retained whenever erosion removes most of the instance or reduces it below a minimum size. Disconnected fragments are then relabeled, enclosed holes are filled, and instances smaller than 500 m^2 are discarded[4] A final relabeling step produces a compact scene-wide instance map.

4.4. Raster-to-vector conversion

The cleaned raster label map is converted into vector parcels by tracing connected regions of equal pixel value into polygon geometries. Before polygonization, each instance mask undergoes one morphological opening followed by one closing, using a square structuring element whose 1 m side length is converted to pixels from the raster affine transform. This step suppresses boundary halos and closes small intra-parcel gaps without distorting the overall shape. At coarse resolutions, where the metric radius falls below 0.75 pixels, the full sequence is replaced by a single lightweight morphological pass. If no valid kernel can be formed, the operation is skipped. Connected components smaller than 500 m^2 are discarded prior to polygonization[5] to suppress fragmented predictions produced by the opening-and-closing sequence. Polygon boundaries are then simplified using the Douglas-Peucker algorithm. By default, the tolerance is set to one pixel width in meters so that boundary smoothing scales consistently with raster resolution. Topological correctness is enforced throughout simplification, preventing self-intersections and ring reversals. A final geometry-cleaning step removes over-segmentation artifacts. Each polygon is compared against spatially coincident candidates, and any polygon whose area overlaps a larger enclosing polygon by at least 99% of its own area is dissolved into the outer one through geometric union. This containment merge removes erosion remnants and spurious sub-parcels fully nested within a larger field boundary. The final geometries are exported as GeoJSON or Shapefile for downstream GIS use.

Implementation details, configuration files, and processing scripts for the full pipeline will be made publicly available through a GitHub repository archived on Zenodo upon publication [41].

5. Evaluation protocol

Evaluation follows a structure-aware protocol designed to assess both detection quality and structural consistency of GIS-ready parcel outputs. Standard overlap-based metrics alone are insufficient in fragmented agricultural landscapes, where merge and split errors can compromise parcel integrity even when global scores remain acceptable. The protocol therefore operates at two complementary levels: instance-level detection and overlap assessment, and structure-level quantification of

[4] The area threshold is converted to pixels as $\lceil A_{\min}/(p_w \cdot p_h) \rceil$, where p_w and p_h denote pixel width and height in metres.

[5] As in Section 4.3, the area threshold is converted to pixels according to raster resolution.

merge and split behavior. This dual assessment is what makes the evaluation structure-aware, and it is reported consistently across all evaluated model-source combinations throughout the following sections.

The first level, instance detection and overlap assessment, is performed at parcel level using overlap-based matching between predicted and reference instances, as described below.

Specifically, we use Intersection over Union (IoU), a standard overlap metric in segmentation that jointly penalizes over-segmentation and under-segmentation and is consistent with common instance-evaluation protocols [12]. For a predicted parcel p_i and a reference parcel g_j, IoU is defined as:

$$\mathrm{IoU}(p_i, g_j) = \frac{|p_i \cap g_j|}{|p_i \cup g_j|} \tag{1}$$

Similarly to [35], we adopt a one-to-one matching strategy between predicted and reference parcels. In our implementation, the assignment is solved with the Hungarian algorithm [46] over the IoU matrix, with predictions on rows and references on columns. Only pairs with $\mathrm{IoU} \geq \tau$ are retained as valid matches. Here, τ denotes the minimum IoU required for a predicted-reference pair to be considered a valid match: lower values are more permissive, whereas higher values enforce stricter spatial agreement. In all experiments, we set $\tau = 0.5$. The choice of matching algorithm is particularly important in fragmented agricultural mosaics, where dense parcel adjacency and frequent merge/split errors can make greedy assignment methods sensitive to matching order and prone to overly optimistic results.

Under this strategy, evaluation starts from the scene-level parcel instances produced by the inference pipeline for each data source and model configuration. Predicted and reference parcels are compared through the IoU matrix to identify matched and unmatched instances. A matched pair (p_i, g_j) with $\mathrm{IoU}(p_i, g_j) \geq \tau$ is counted as a true positive (TP), whereas unmatched predictions and references are counted as false positives (FP) and false negatives (FN), respectively. From these counts, we compute Precision, Recall, F1, and mIoU, to quantify detection and localization performance.

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{4}$$

$$\mathrm{mIoU} = \frac{1}{|\mathcal{M}|} \sum_{(p_i, g_j) \in \mathcal{M}} \mathrm{IoU}(p_i, g_j), \tag{5}$$

The second level, structure-level error analysis, quantifies merge and split behavior explicitly through global over-segmentation (GOSE) and global under-segmentation (GUSE), adapted from the geometric evaluation framework of [12]:

$$\mathrm{GOSE} = \frac{1}{|G_{split}|} \sum_{g \in G_{split}} \left(1 - \frac{\max_{p \in P_g} |p \cap g|}{|g|} \right), \tag{6}$$

$$\mathrm{GUSE} = \frac{1}{|P_{merge}|} \sum_{p \in P_{merge}} \left(1 - \frac{\max_{g \in G_p} |p \cap g|}{|p|} \right). \tag{7}$$

Here we use a parcel-level variant that isolates fragmentation and merging events. In the equations above, G and P denote the sets of reference and predicted parcels, respectively, and $|\cdot|$ is the area operator. For each reference parcel g, P_g denotes the set of overlapping predicted parcels; conversely, for each predicted parcel p, G_p denotes the set of overlapping reference parcels. To focus on topological errors rather than minor boundary noise, $G_{split} \subset G$ includes reference parcels overlapped by more than one predicted parcel, each covering at least 10% of the reference area, whereas $P_{merge} \subset P$ includes predicted parcels overlapping multiple reference parcels. GOSE therefore measures split tendency as the portion of a reference parcel not covered by its largest predicted fragment, while GUSE measures merge tendency as the portion of a predicted parcel not covered by its dominant reference parcel. For both metrics, lower values indicate better structural consistency, with values closer to zero corresponding to fewer split and merge errors.

Finally, results are aggregated by data source and segmentation-model configuration to enable consistent comparison across imagery sources and model families. All benchmark metrics are computed as micro-averaged aggregates: true positives, false positives, and false negatives are summed across all evaluation instances before Precision, Recall, F1, GOSE, and GUSE are derived. This produces a single exact deterministic score over the complete test set, weighting each instance equally regardless of spatial origin. Because the evaluated models operate with fixed pretrained weights and deterministic inference, repeated executions on the same input reproduce the same prediction identically and do not constitute independent statistical replicates. The reported metrics should therefore be interpreted as exact benchmark measurements under the tested conditions, not as sample-based estimates requiring confidence intervals or significance tests. Comparative claims are framed accordingly as benchmark-specific evidence rather than population-level claims of model superiority over arbitrary agricultural landscapes.

6. Primary benchmark setup

6.1. Site description and landscape characteristics

The primary benchmark for this study is a manually annotated parcel dataset in Trentino (northeastern Italy), a representative heterogeneous smallholder-like agricultural landscape. The region is characterized by complex topography, with elevations ranging from below 200 m to about 3300 m above sea level, and by a fragmented valley-hillside spatial pattern shaped by the Eastern Alps. Agricultural activity is mainly concentrated in valleys, plateaus, and sun-exposed hillsides, whereas mountainous areas are largely forested. According to the 7th General Census of Agriculture (conducted in 2021, reference year 2020) by the Italian National Institute of Statistics (ISTAT), the Province of Trento includes 14,236 farms and 121,790 ha of utilized agricultural area. Compared with the national context, farm structure in Trento is more fragmented: average farm size is about 8.6 ha versus 11.1 ha in Italy. This pattern is consistent with the predominance of individual/family-run holdings (13,288 of 14,236 farms, about 93.3%) and with intensive fruit-tree and vineyard systems, where average specialized surfaces are about 2.3 ha for apple orchards and 1.8 ha for vineyards [47]. This structural fragmentation reflects the Alpine land constraints of the region, where narrow valley floors and terraced hillsides limit farm consolidation and have historically fostered a system of family-owned smallholdings supported by region-wide cooperative structures [48]. From a delineation perspective, this setting is particularly challenging because adjacent parcels within the same crop system share similar spectral and phenological signatures, stressing geometric instance separation rather than crop-class discrimination. Parcel boundaries are therefore often short, irregular, and locally ambiguous.

To ensure the highest possible reference quality, the annotation effort was focused on a representative region managed by a single agricultural consortium. This targeted collaboration facilitated a rigorous manual delineation process and enabled a field-informed verification step for ambiguous boundaries. By prioritizing this expert-led verification over a larger unverified sample, we ensured high internal consistency in the reference data across the sampled terrain and management conditions. Table 3 summarizes the attributes of the main evaluation area.

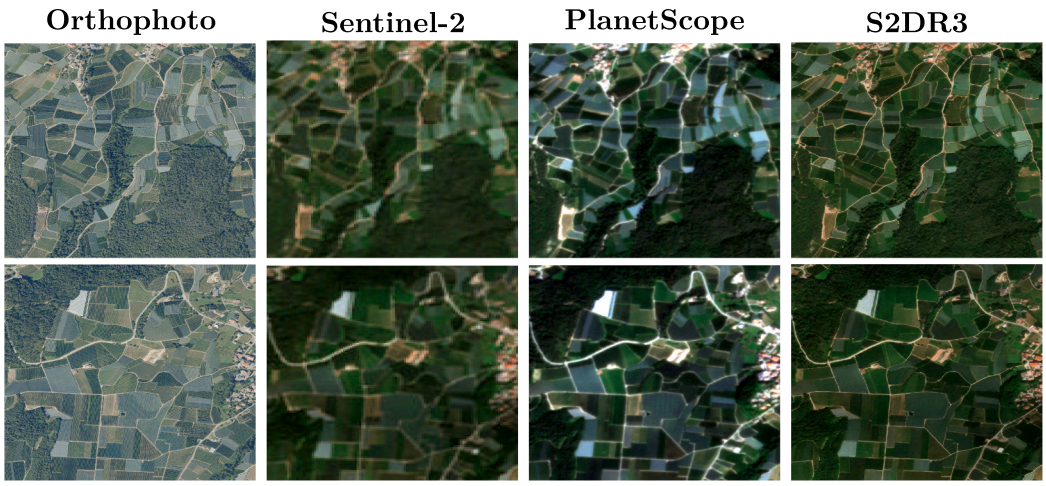


Fig. 4. RGB composites for two representative areas in Trentino. Orthophoto imagery is shown as a visual reference, alongside Sentinel-2 (S2), PlanetScope (PS), and the derived S2DR3 product. The first and second rows correspond to Region 1 and Region 2, respectively. The S2 and S2DR3 images were acquired on 25/06/2023, while the PS imagery was acquired on 24/06/2023.

Table 3
Characteristics of the main evaluation site. The site was selected and annotated in collaboration with the local agricultural consortium to represent the structural complexity of intensive apple orchard management.

Attribute	Value
Consortium	Tres
Dominant Crop System	Apple Orchards
Evaluation Area (ha)	294.5
Number of Orthophoto Tiles	3
Number of Reference Parcels	424
Median Parcel Size (ha)	0.373
Elevation Range (m)	640–840
Boundary Complexity	High / Fragmented

6.2. Data acquisition and temporal alignment

Trentino lies between approximately 45.8 and 46.5 deg N, a latitude range favorable to regular optical satellite coverage. With a two-satellite constellation (S2A and S2B), Sentinel-2 offers a 5-day revisit cycle, enhanced in the Trentino region by overlap between adjacent swaths, while PS is approximately daily. Tres, the main evaluation area summarized in Table 3, is used as a representative example in Fig. 4.

In addition to the model input sources, an orthophoto of Trentino from 2023 is used to support evaluation-set annotation (Ortofoto 2023, AGEA – Agenzia per le Erogazioni in Agricoltura, Roma[6]). The orthophoto has 20 cm spatial resolution and includes RGB and near-infrared bands. Source imagery was acquired between July and September 2023; satellite acquisitions used in this study were selected in the same seasonal window. The orthophoto is referenced in ETRS89/ETRF2000 and distributed in UTM projection as slightly overlapping GeoTIFF tiles covering the full region. Fig. 5 shows tile layout and evaluation-area location.

6.3. Reference parcel annotation and quality control

Reference parcel instances are generated from the orthophoto described in Section 6.2 within the selected evaluation areas. Polygon geometries are digitized per parcel and stored in GeoJSON format. Each parcel record includes geometry and supporting attributes (*field_id*, *confidence*, *verified*, *observations*). These references were manually digitized in QGIS using the 2023 Trentino orthophoto as the primary visual source. The annotation process followed visible parcel boundaries, including roads, tracks, field margins, crop-row discontinuities, and other stable separators observable at 20 cm resolution. The annotation process distinguishes between hard and soft boundaries. Hard boundaries are visually clear and usually correspond to roads or non-vegetated separators. Soft boundaries separate adjacent parcels with weak or gradual vegetation transitions and are therefore more uncertain in single-date imagery. Ambiguous cases, especially soft boundaries between adjacent parcels with similar vegetation, were flagged during digitization and reviewed with local agronomic support before inclusion in the final reference layer; the *confidence* and *observations* attributes were used to document uncertainty and corrections prior to validation use. We did not compute formal inter-annotator agreement statistics, because the benchmark was produced through an expert-informed quality-control workflow rather than through independent replicated annotations. We therefore treat the absence of inter-annotator overlap metrics as a limitation of the dataset, while noting that the expert review step was used to improve boundary consistency in locally ambiguous cases.

Expert-curated reference construction from high-resolution imagery is a common practical approach in agricultural parcel and boundary delineation [8,19], particularly when the evaluation target is the physical field boundary rather than an administrative parcel record. In this study, the resulting reference layer should therefore be interpreted as an expert-curated reference benchmark rather than as an uncertainty-free ground truth. This qualification is particularly relevant for soft boundaries between adjacent parcels with similar vegetation patterns, where alternative but plausible delineations may exist in single-date imagery. Consequently, the reported Precision, Recall, F1, mIoU, GOSE, and GUSE values quantify model agreement with this expert-curated reference layer, not absolute agreement with an independently replicated annotation consensus. The absence of formal inter-annotator overlap metrics remains a limitation of the dataset; however, the combination of high-resolution orthophoto imagery, explicit hard/soft boundary distinction, uncertainty attributes, and local agronomic review was adopted to constrain annotation subjectivity and improve the internal consistency of the benchmark.

7. Results on the trentino primary benchmark

We first isolate the effect of vegetation enhancement, then compare model families and imagery sources, and finally examine how the main performance differences relate to source resolution and model behavior.

[6] https://www.agea.gov.it

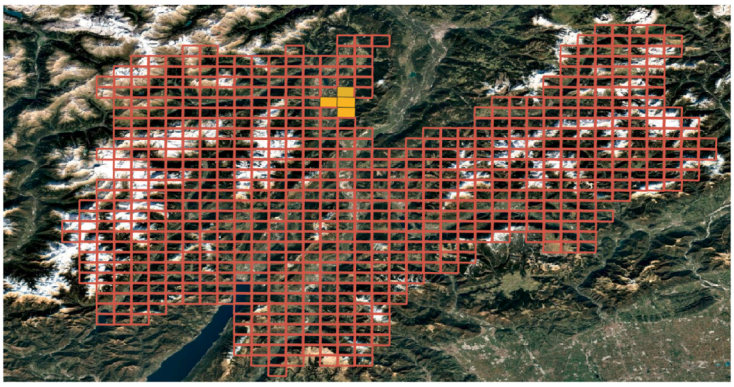


Fig. 5. Spatial layout of the GeoTIFF tiling scheme for Trentino over OpenStreetMap. Yellow tiles denote the main evaluation areas.

Table 4
Source-stratified ablation of vegetation enhancement (VE): zero-shot SAM on original RGB versus VE-enhanced RGB across the three evaluated imagery sources (S2, S2DR3, and PS). The arrows ↑ and ↓ indicate that higher and lower values are preferable, respectively.

Source	Model	Seg. Quality ↑	Detection Completeness ↑			Structural Error ↓		Δ (VE-SAM)		
		mIoU	Precision	Recall	F1	GOSE	GUSE	ΔPrecision	ΔRecall	ΔF1
S2	SAM	**0.6743**	**0.6784**	0.4627	**0.5501**	**0.2916**	0.5159	-0.0313	+0.0144	-0.0009
	VE + SAM	0.6579	0.6471	**0.4771**	0.5492	0.3157	**0.4837**			
S2DR3	SAM	**0.7949**	**0.8270**	0.7373	0.7796	**0.1470**	**0.2054**	-0.0205	+0.0458	+0.0150
	VE + SAM	0.7783	0.8065	**0.7831**	**0.7946**	0.1928	0.2084			
PS	SAM	**0.7946**	**0.7928**	0.4241	0.5526	**0.0843**	**0.3874**	-0.0435	+0.2169	+0.1383
	VE + SAM	0.7464	0.7493	**0.6410**	**0.6909**	0.3108	0.4423			
Avg. Δ (all sources)								-0.0318	+0.0924	+0.0508

In all result tables, bold values denote the best result within each metric column.

7.1. Effect of vegetation enhancement across imagery sources

The first analysis is a source-stratified ablation experiment designed to isolate the contribution of vegetation-enhancement (VE) preprocessing across each of the three imagery sources independently. Using zero-shot SAM as a reference configuration, we evaluate Sentinel-2, S2DR3, and PS imagery under two input conditions: original RGB and vegetation-enhanced RGB. Table 4 reports the resulting detection, overlap, and structural error metrics.

The results show a highly consistent trade-off across all three sources: vegetation enhancement always increases recall and always reduces precision. On one hand, it recovers parcels that were previously missed, thereby reducing false negatives; on the other hand, it also introduces additional false positives. Overall, the recall gain is sufficient to offset the precision loss, yielding average increases of 9.2% in recall and 5.0% in F1 across the evaluated sources.

In Table 4, the Δ columns report the difference between vegetation-enhanced and original SAM outputs (*i.e.*, Δ = VE − SAM). This pattern is most clearly visible in the Δ Recall and Δ Precision columns: Δ Recall is positive across all three sources, whereas Δ Precision is consistently negative. Positive values are highlighted in green to indicate a beneficial effect of vegetation enhancement, while negative values are highlighted in red.

This trade-off is favorable for PS and S2DR3, where the gain in recall is substantial enough to improve overall detection quality. On PS, in particular, the increase in recall is much larger than the associated drop in precision, leading to the strongest F1 improvement. On Sentinel-2, by contrast, the recall gain is too small to offset the precision loss, and F1 remains essentially unchanged. This suggests that vegetation enhancement is most useful when the source already contains enough spatial detail for the additional vegetation contrast to translate into recoverable parcel instances.

These results support adopting vegetation-enhancement as a fixed preprocessing component in the subsequent SAM-based experiments. This combination is referred to as the VE + SAM pipeline throughout the remainder of the paper.

7.2. Main segmentation performance

Table 5 reports the results for all evaluated model-source combinations on the Trentino benchmark. As anticipated in Section 3.2, LangSAM is excluded from this comparison because its language-guided detection setup does not produce valid parcel-level instance outputs for the evaluated delineation task. Further implementation-specific details are outside the scope of the present results section.

No single configuration is best for every metric. Two quantitative patterns stand out in Table 5. First, S2DR3 yields the strongest results for every model family, whereas S2 is consistently associated with the weakest overall results across models, suggesting that source resolution dominates the comparison more strongly than model identity alone. Second, the best overall configuration is SAM with S2DR3, which achieves the highest recall (0.7831) and F1 (0.7946), while also recording the lowest GUSE (0.2084). However, it is not best for every metric: DelAny with S2DR3 attains the highest mIoU (0.7974), whereas SAM2 with S2DR3 attains the highest precision (0.8639).

This comparison shows that the top configurations remain relatively close in mIoU, with differences on the order of only a few hundredths, whereas the largest numerical differences appear in recall, F1, and GUSE. Relative to SAM on S2DR3, DelAny reaches slightly higher mIoU (+0.0191) and precision (+0.0249), but with substantially lower recall (-0.2602), lower F1 (-0.1526), and a much higher GUSE (+0.1709). SAM2 on S2DR3 shows a different profile: its precision exceeds SAM by 0.0574 and its mIoU by 0.0109, but recall is lower by 0.0795 and

Table 5
Performance comparison of segmentation models across satellite data sources in the Trentino study area.

Model	Source	Seg. Quality ↑	Detection Completeness ↑			Structural Error ↓	
		mIoU	Precision	Recall	F1	GOSE	GUSE
SAM	S2	0.6579	0.6471	0.4771	0.5492	0.3157	0.4837
	S2DR3	0.7783	0.8065	**0.7831**	**0.7946**	0.1928	**0.2084**
	PS	0.7464	0.7493	0.6410	0.6909	0.3108	0.4423
SAM2	S2	0.7035	0.8169	0.2795	0.4165	0.1108	0.5000
	S2DR3	0.7892	**0.8639**	0.7036	0.7756	0.1133	0.2396
	PS	0.7842	0.8571	0.1446	0.2474	0.0169	0.3857
SAM3	S2	0.7005	0.7368	0.0337	0.0645	**0.0096**	0.4737
	S2DR3	0.7849	0.8392	0.6916	0.7583	0.1518	0.2719
	PS	0.7514	0.8072	0.4843	0.6054	0.1687	0.4378
DelAny	S2	0.6905	0.3564	0.0867	0.1395	0.2241	0.8812
	S2DR3	**0.7974**	0.8314	0.5229	0.6420	0.1494	0.3793
	PS	0.7333	0.6920	0.3952	0.5031	0.2892	0.5823
FastSAM	S2	0.6699	0.7377	0.1084	0.1891	0.0482	0.6721
	S2DR3	0.7679	0.7747	0.5470	0.6412	0.0843	0.2628
	PS	0.7265	0.7176	0.2265	0.3443	0.0964	0.4962

Table 6
Impact of imagery source on segmentation performance, averaged across the evaluated models.

Source	Seg. Quality ↑		Detection Completeness ↑						Structural Error ↓	
	mIoU ↑ mean ± std	Δ%	Precision ↑ mean ± std	Δ%	Recall ↑ mean ± std	Δ%	F1 ↑ mean ± std	Δ%	GOSE ↓ mean ± std	GUSE ↓ mean ± std
S2	0.685 ± 0.018	—	0.659 ± 0.161	—	0.197 ± 0.162	—	0.272 ± 0.182	—	0.142 ± 0.113	0.602 ± 0.157
PS	0.748 ± 0.020	+9.3%	0.765 ± 0.060	+16.0%	0.378 ± 0.178	+92.0%	0.478 ± 0.163	+75.7%	0.176 ± 0.112	0.469 ± 0.067
S2DR3	**0.784 ± 0.010**	**+14.5%**	**0.823 ± 0.030**	**+24.9%**	**0.650 ± 0.099**	**+229.6%**	**0.722 ± 0.067**	**+165.8%**	**0.138 ± 0.037**	**0.272 ± 0.058**

Table 7
Model-normalized resolution impact expressed as the median relative change with respect to S2, summarized across the evaluated models.

Source	Seg. Quality ↑	Detection Completeness ↑			Structural Error ↓	
	mIoU ↑	Precision ↑	Recall ↑	F1 ↑	GOSE ↓	GUSE ↓
PS	+8.4 [+7.3, +11.5]	+9.6 [+4.9, +15.8]	+108.9 [+34.4, +355.8]	+82.1 [+25.8, +260.6]	+29.0 [−1.6, +100.0]	−22.9 [−26.2, −8.6]
S2DR3	**+14.6** [+12.2, +15.5]	**+13.9** [+5.8, +24.6]	**+404.6** [+151.7, +503.1]	**+239.1** [+86.2, +360.2]	+2.3 [−33.3, +74.9]	**−56.9** [−57.0, −52.1]

GUSE is higher by 0.0312. A similar pattern is observed on PS and S2. On PS, SAM remains the strongest configuration in F1 (0.6909), followed by SAM3 (0.6054), DelAny (0.5031), FastSAM (0.3443), and SAM2 (0.2474). On S2, the gap between configurations widens sharply and most models lose a substantial fraction of the reference parcels: only SAM and SAM2 remain above 0.40 in F1, whereas DelAny and FastSAM drop below 0.20 and SAM3 collapses in recall and F1. Taken together, these comparisons suggest that SAM remains the most robust configuration across sources, while the gap between configurations increases as source quality decreases.

7.3. Resolution effect across sources

To isolate the role of imagery source, Table 6 reports source-level means and standard deviations across all evaluated models, together with the corresponding relative changes with respect to S2. For each higher-resolution source $S \in \{\mathrm{PS}, \mathrm{S2DR3}\}$, these changes are expressed as percentages through $\Delta_S = (S - \mathrm{S2})/\mathrm{S2} \times 100$. For mIoU, Precision, Recall, and F1, positive Δ_S values denote improvement, whereas for GOSE and GUSE lower values indicate better performance. Additionally, Table 7 provides a model-normalized view of the same source effect displaying median and IQR of per-model Δ, which is discussed below together with Table 8.

The source comparison shows a clear ordering across the Trentino study area: S2 is consistently the weakest source, PS is intermediate, and S2DR3 yields the strongest average results. For this reason, S2 is used as the baseline for the relative improvements reported as Δ percentages in Tables 6 and 7. Specifically, relative to S2, moving to S2DR3 increases mIoU from 0.685 to 0.784, precision from 0.659 to 0.823, recall from 0.197 to 0.650, and F1 from 0.272 to 0.722 (+14.5%, +24.9%, +229.6%, and +165.8%, respectively). PS improves on the same baseline as well, but less strongly, reaching 0.748 mIoU, 0.378 recall, and 0.478 F1 (+9.3%, +92.0%, and +75.7%, respectively).

Table 8
Source ranking per model (1 = best). Avg. Rank denotes the mean rank across the six reported metrics.

Model	Source	mIoU	Prec.	Recall	F1	GOSE	GUSE	Avg. Rank
SAM	S2DR3	1	1	1	1	1	1	**1.00**
	PS	2	2	2	2	2	2	2.00
	S2	3	3	3	3	3	3	3.00
SAM2	S2DR3	1	1	1	1	3	1	**1.33**
	PS	2	2	3	3	1	2	2.17
	S2	3	3	2	2	2	3	2.50
SAM3	S2DR3	1	1	1	1	2	1	**1.17**
	PS	2	2	2	2	3	2	2.17
	S2	3	3	3	3	1	3	2.67
DelAny	S2DR3	1	1	1	1	1	1	**1.00**
	PS	2	2	2	2	3	2	2.17
	S2	3	3	3	3	2	3	2.83
FastSAM	S2DR3	1	1	1	1	2	1	**1.17**
	PS	2	3	2	2	3	2	2.33
	S2	3	2	3	3	1	3	2.50

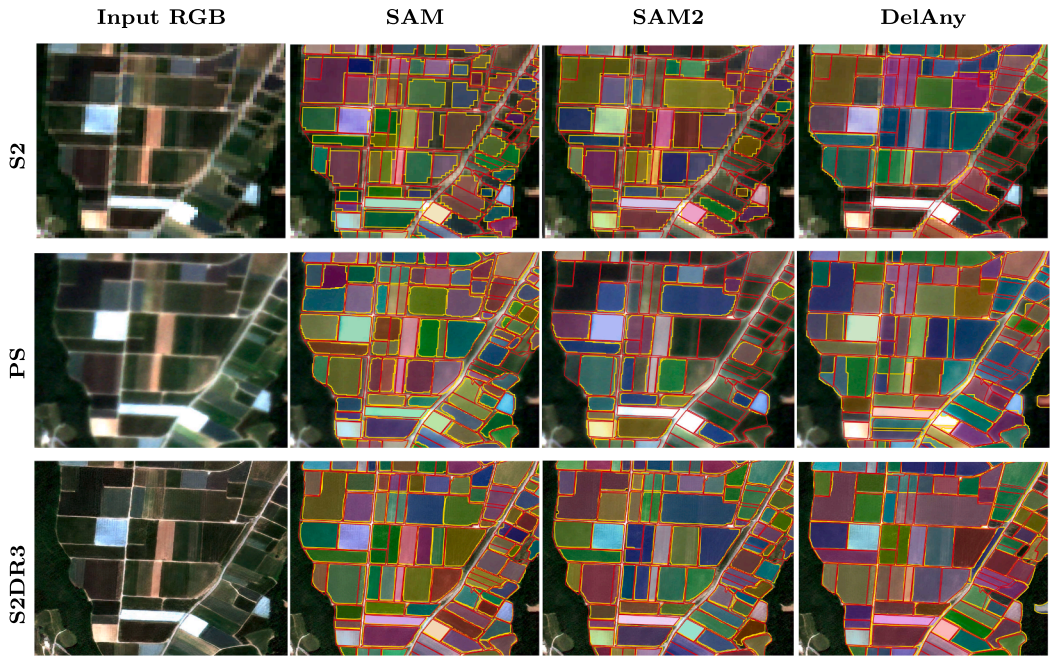


Fig. 6. Qualitative comparison of crop-field segmentation across imagery sources and the three most informative model families. Reference parcel instances are indicated by red borders, while model predictions are delineated by yellow borders containing semi-transparent, instance-specific color fills to highlight individual parcel boundaries. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Structural error metrics follow the same ranking. Average GUSE drops from 0.602 for S2 to 0.469 for PS and to 0.272 for S2DR3, while average GOSE is lowest for S2DR3 (0.138). The only exception is PS, whose GOSE is slightly higher than that of S2 (0.176 vs. 0.142). This local deviation does not alter the overall ordering, but it indicates that the effect of additional spatial detail is not fully uniform across structural error metrics, especially when detection remains weak.

Because Table 6 averages across models, it is useful to verify whether the same source ordering remains visible at the level of individual model responses. For this reason, Table 7 reports the model-normalized source effect. Each entry is obtained by first computing, for each model separately, the relative change with respect to S2 and then summarizing those model-specific changes by their median; the bracketed interval reports the interquartile range from Q1 to Q3. Relative to S2, the median gain remains positive for both higher-resolution sources across all accuracy metrics, and it is especially large for recall and F1. The strongest model-normalized improvement again occurs for S2DR3, where the median recall gain reaches +404.6% and the median F1 gain reaches +239.1%. Table 8 confirms the same ordering from a third perspective: S2DR3 ranks first for every evaluated model, PS ranks second, and S2 ranks last. The only local departures from the best rank for S2DR3 occur on GOSE, where it ranks third with SAM2 and second with SAM3. Taken together, Tables 6–8 confirm that S2DR3 consistently yields the strongest source-level performance. On the basis of the per-source results in Table 5, SAM, SAM2, and DelAny emerge as the three most informative model configurations for the qualitative comparison that follows, as they represent the strongest performers across complementary metrics and sources.

7.4. Qualitative comparison across models and sources

Fig. 6 complements the quantitative results by showing how the three most informative model families behave on the same scene across S2, PS, and S2DR3. Rather than introducing new evidence, the visual comparison helps relate the reported differences in recall, F1, and GUSE to visible changes in parcel continuity, boundary separation, and instance fragmentation. Rows correspond to S2, PS, and S2DR3, whereas columns report the RGB input, the reference parcels, and the segmentation outputs of SAM, SAM2, and DelAny.

The visual comparison is consistent with the quantitative patterns reported in Tables 5 and 6. On S2, all models lose structural detail, in agreement with the marked reduction in Recall and F1 observed in the quantitative comparison, which reflects fewer true-positive parcel matches and a larger number of missed reference parcels. On PS and S2DR3, SAM and SAM2 preserve parcel separation more clearly than DelAny, which more often merges adjacent fields; this is coherent with the lower GUSE values reported for the SAM-based configurations.

The reported results are specific to the Trentino study area. The next section therefore extends the analysis through external validation under a different structural and geographic setting.

8. External evaluation: Netherlands agricultural landscape

The primary benchmark in Section 6 establishes pipeline performance in a fragmented smallholder landscape, but the generalizability of these findings to structurally different agricultural settings remains open. This section addresses that question through external evaluation on Dutch agricultural tiles, where parcels are substantially larger and landscape fragmentation is lower. The comparison is restricted to VE + SAM, VE + SAM3, and DelAny across S2 and PS, as the resolution constraint that motivated S2DR3 inclusion in the primary benchmark does not apply in this setting. The external evaluation specifically tests whether the structural advantage of VE + SAM over DelAny persists beyond the small-parcel regime of Trentino, and whether the resolution-segmentation interaction observed there generalizes across parcel-size distributions. Including SAM3 tests whether the pipeline's prompting strategy scales effectively to a landscape with a different parcel distribution.

8.1. Netherlands evaluation tiles

This section describes the construction of the Netherlands evaluation tiles, a set of three 10×10 km tiles (300 km^2 total) used to assess model generalization under controlled conditions. The central methodological

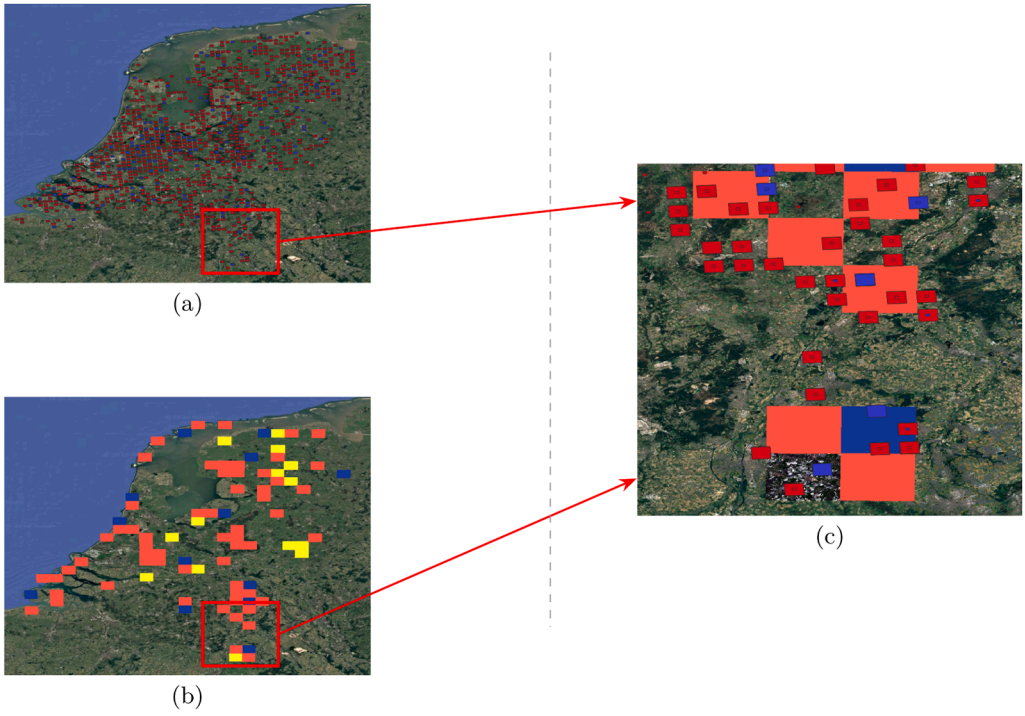


Fig. 7. FBIS-22M coverage, AI4SmallFarms tile distribution, and the selected Dutch evaluation tile used for external validation.

concern is the potential for training data overlap bias affecting DelAny. DelAny was trained on FBIS-22M [33], a large-scale benchmark containing approximately 22 million field boundaries across multiple countries, sensors, and spatial resolutions. FBIS-22M covers the Netherlands extensively and incorporates earlier resources such as AI4Boundaries [37]. As illustrated in Fig. 7(a), its train and test partitions are spatially interleaved at fine granularity, with red and blue patches denoting tiles from different split groups. Any evaluation tile drawn from within this coverage therefore risks introducing optimistic bias in favor of DelAny. The tile selection strategy below is designed explicitly to quantify and control for this risk.

All three evaluation tiles derive their reference boundaries from the Dutch official cadastral data (Basisregistratie Gewaspercelen, BRP), released annually by the Rijksdienst voor Ondernemend Nederland, Den Haag[7]. For this analysis, apart from the AI4SmallFarms original tile, the most recent available version (2024) is used.

Tile 1 is drawn from the AI4SmallFarms test split. AI4SmallFarms also covers the Netherlands and is organized into 10×10 km tiles derived from BRP data. In Fig. 7(b), the red, blue, and yellow patches denote the available AI4SmallFarms tiles, while the red rectangle indicates the region shown in Fig. 7(c), where the selected tile is displayed in RGB. Because AI4SmallFarms covers the same geographic area as FBIS-22M, a residual spatial overlap with DelAny's training data cannot be fully excluded for this tile.

To address this limitation, Tiles 2–3 are placed in areas entirely outside FBIS-22M coverage, where complete 10×10 km tiles can be constructed directly from BRP 2024 boundaries. These tiles provide an evaluation setting free from any potential overlap-related bias.

All three tiles are evaluated using S2 and PS. The S2DR3 source, included in the Trentino analysis, is not considered here, as Dutch agricultural parcels are considerably larger than those in Trentino, reducing the resolution constraint that motivated its inclusion. The parcel-size distributions across datasets are reported in Table 10 and discussed in Section 8.3.

[7] https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp-

8.2. Results on Netherlands evaluation tiles

Table 9 reports the results for the entire area and for each source separately, with the table divided into four blocks: Tile 1 corresponds to the AI4SmallFarms tile located in an FBIS-22M overlap region, Tiles 2–3 correspond to the BRP24 tiles outside FBIS-22M coverage, and the final block aggregates all evaluated tiles. The overall block provides the clearest basis for comparing the two models across all tiles and sources.

Across all tiles, the overall block of Table 9 indicates a close comparison, but with a consistent structural advantage for VE + SAM and VE + SAM3 over DelAny. On PS, both VE + SAM and VE + SAM3 attain higher F1 (0.5249 and 0.5546 respectively) and clearly lower GOSE and GUSE, whereas on S2 both attain higher F1 (0.5221 and 0.5111 respectively) and lower structural-error rates. DelAny remains competitive in recall on both sources, but these local advantages do not translate into a better overall balance.

The split between Tile 1 and Tiles 2–3 is also informative. In the AI4SmallFarms tile, which lies within the FBIS-22M overlap region, the comparison is closer and DelAny retains a small advantage on S2 in Recall and F1. Outside that region, however, VE + SAM and VE + SAM3 become clearly more competitive: on BRP24 they lead on mIoU for both sources, on F1 for S2, and on both structural metrics across all reported configurations. This shift is consistent with the possibility that overlap with FBIS-22M favors DelAny in Tile 1, while the non-overlap tiles provide a cleaner estimate of comparative behavior.

The source effect is weaker than in Trentino, but it is still informative. Moving from S2 to PS produces only modest aggregate gains, and the three models respond differently to that increase in spatial detail. VE + SAM and VE + SAM3 gain on recall (with VE + SAM3 also securing a substantial F1 boost on PS), but give up some precision, whereas DelAny gains on precision but not on parcel recovery. Numerically, this indicates that both SAM-based models use the added detail mainly to recover more parcel instances, while DelAny becomes more selective in the instances it commits to.

The most stable pattern across the entire validation concerns structural consistency. VE + SAM and SAM3 systematically record lower GOSE and GUSE across all sources and both tile blocks compared to

Table 9
Validation results on AI4SmallFarms and BRP24 tiles. Tile 1 overlaps with DelAny training data and may introduce optimistic bias toward that model. Tiles 2–3 lie outside FBIS-22M coverage. Bold indicates the best model per metric, source, and Tile.

Source	Model	mIoU ↑	Precision ↑	Recall ↑	F1 ↑	GOSE ↓	GUSE ↓
Tile 1 (AI4SmallFarms)							
PS	VE + SAM	0.7624	0.4948	**0.4465**	0.4694	0.2803	**0.3776**
	VE + SAM3	**0.7625**	**0.5502**	0.4246	**0.4793**	**0.2316**	0.4165
	DelAny	0.7416	0.4919	0.3166	0.3853	0.2843	0.4949
S2	VE + SAM	0.7169	0.4699	0.3512	0.4020	0.3647	**0.5061**
	VE + SAM3	**0.7354**	**0.6265**	0.2649	0.3724	**0.1606**	0.5600
	DelAny	0.7271	0.4876	**0.3785**	**0.4262**	0.4637	0.5730
Tile 2 (BRP24)							
PS	VE + SAM	**0.8424**	0.5064	**0.5605**	0.5321	0.1449	**0.1742**
	VE + SAM3	0.8410	**0.6908**	0.4788	0.5656	**0.0783**	0.2158
	DelAny	0.8063	0.6447	0.5177	**0.5743**	0.1608	0.2876
S2	VE + SAM	0.7966	0.5536	**0.5717**	0.5625	0.2167	**0.2486**
	VE + SAM3	**0.8032**	**0.6925**	0.4923	**0.5755**	**0.0979**	0.2782
	DelAny	0.7509	0.4659	0.5682	0.5120	0.4584	0.3277
Tile 3 (BRP24)							
PS	VE + SAM	0.8265	0.5916	0.5789	0.5852	0.1472	**0.1588**
	VE + SAM3	**0.8270**	**0.7254**	0.5732	**0.6404**	**0.1184**	0.1792
	DelAny	0.8025	0.6524	**0.6033**	0.6269	0.2542	0.2664
S2	VE + SAM	0.7965	0.6052	0.6034	0.6043	0.2056	**0.2209**
	VE + SAM3	**0.8050**	**0.7860**	0.4977	**0.6094**	**0.0981**	0.2357
	DelAny	0.7668	0.5404	**0.6206**	0.5778	0.4220	0.3189
Overall (All Tiles)							
PS	VE + SAM	**0.8081**	0.5304	**0.5196**	0.5249	0.2016	**0.2464**
	VE + SAM3	0.8072	**0.6432**	0.4874	**0.5546**	**0.1552**	0.2880
	DelAny	0.7864	0.5956	0.4632	0.5211	0.2430	0.3500
S2	VE + SAM	0.7752	0.5474	0.4990	**0.5221**	0.2691	**0.3182**
	VE + SAM3	**0.7854**	**0.7070**	0.4002	0.5111	**0.1239**	0.3493
	DelAny	0.7510	0.5020	**0.5145**	0.5080	0.4476	0.3949

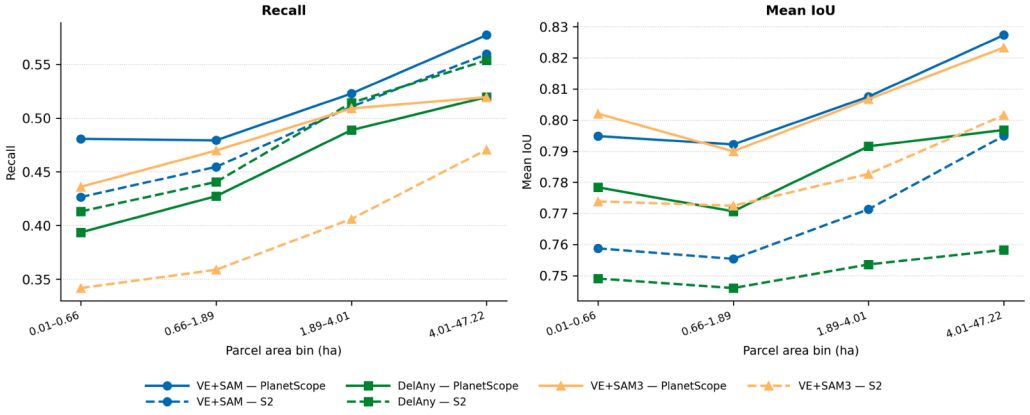


Fig. 8. Per-area-bin evaluation of recall and mean IoU across parcel-size quartiles. Each bin represents an interval of parcel area, ranging from smallest (Q1) to largest (Q4). Line color distinguishes models and line style distinguishes image source.

DelAny, including the spatially biased AI4SmallFarms tile, with SAM3 consistently driving GOSE and VE + SAM driving GUSE to their respective lowest rates. This means that both frameworks break the typical structural trade-off, lowering both split and merge errors simultaneously relative to the baseline rather than improving one at the expense of the other.

8.3. Field-size and fragmentation effects

Field size and landscape fragmentation interact with both spatial resolution and model behavior in ways that aggregate metrics alone cannot reveal. The Trentino results suggest that SAM outperforms DelAny and that higher resolution consistently improves segmentation, but it remains unclear whether these effects hold uniformly across parcel sizes or are amplified in the small-parcel regime that dominates that dataset. Table 10 reports the parcel-size distributions across Trentino, AI4SmallFarms, and BRP24, contextualizing the structural difference between the two study areas. Fig. 8 then examines both effects jointly, reporting per-bin recall and mean IoU on the Dutch validation data for three models (VE + SAM, VE + SAM3, and DelAny) and two sources (S2 and PS), isolating how parcel size modulates model and resolution effects across a wider size range than that represented in Trentino.

The difference across the two sources stands in stark contrast to the Trentino findings. S2 is no longer insufficient for accurate segmenta-

Table 10
Parcel size distribution in hectares across the three datasets.

Dataset	*n*	Q1	Median	Mean	Q3
Trentino	415	0.211	0.373	0.457	0.578
AI4SmallFarms	3298	0.485	1.183	1.949	2.650
BRP24	4366	0.965	2.569	3.422	4.906

tion, and moving to finer resolution yields much smaller gains than in Trentino. This behavior helps isolate the main limiting factor of automatic image-based parcel segmentation: parcel size and structural fragmentation. The Trentino dataset contains fields averaging 0.46 ha, whereas the Netherlands tiles average 1.95 ha in AI4SmallFarms and 3.42 ha in BRP24. This confirms that parcels in Trentino are substantially smaller, making accurate segmentation considerably more challenging at limited spatial resolution.

Several patterns emerge from the per-bin analysis. First, recall increases monotonically with field size for all model and source combinations, confirming that larger parcels are inherently easier to detect regardless of the approach used. Second, the added value of PS over S2 is most pronounced for VE + SAM3, which shows the largest source-driven recall gains across all bins, followed by VE + SAM; DelAny benefits least from the resolution increase. Third, VE + SAM surpasses DelAny in recall across all field-size bins and for both imagery sources; VE + SAM3 also exceeds DelAny on PS but falls below it on S2, where it records the lowest recall of all three models. This result supports the observation that the proposed pipeline is particularly effective in the parcel-size regime most relevant to fragmented smallholder landscapes, though VE + SAM3 is more dependent on high-resolution input to realize that advantage.

9. Discussion and concluding remarks

9.1. Implications for zero-shot parcel delineation

The zero-shot pipeline described in Section 4 operates without site-specific retraining across three imagery sources, making spatial resolution the main variable governing segmentation behavior in fragmented agricultural settings. As analyzed in Section 7.3, narrow boundaries and soft transitions between adjacent parcels become difficult to resolve at coarse spatial resolution. This helps explain why source resolution emerges as a dominant factor in the cross-source comparison for Trentino and why the largest gains are concentrated in detection-oriented metrics rather than only in boundary-overlap scores. Under the tested conditions, S2DR3 consistently provides the strongest overall source-level performance, whereas PS remains intermediate and S2 the weakest option. Because S2DR3 is a CNN-based reconstruction rather than a native 1m observation, its outputs may contain artificial high-frequency structures or locally sharpened gradients [49]. In a highly fragmented landscape like Trentino, where under-segmentation (merging adjacent fields) is the primary structural failure mode, these sharpened edges are highly beneficial. S2DR3 prompts the evaluated models to predict more boundaries, effectively reducing under-segmentation errors (GUSE). This structural recovery is explicitly quantified across the benchmarked models in Table 7, where upgrading the imagery source from native S2 to S2DR3 yields a dramatic model-normalized median drop in under-segmentation errors (ΔGUSE = −56.9%). However, this introduces a distinct data-centric trade-off: downstream segmentation models can occasionally misinterpret reconstruction artifacts or edge ringing as actual physical borders. This can create incorrect extra boundaries that increase over-segmentation errors (GOSE). Our dual structure-aware validation protocol mathematically captures this operational penalty, revealing that the substantial reduction in merged fields triggers only a minor, localized median over-segmentation penalty (ΔGOSE = +2.3%). For this specific smallholder system, the net structural benefit of resolving merged fields generally outweighs these localized over-segmentation penalties. Because these findings are bound to the fragmented characteristics of the Trentino benchmark, we interpret S2DR3 as a highly effective, landscape-specific trade-off rather than a universally flawless resolution upgrade.

Within the SAM-based pipeline, vegetation-enhancement, as evaluated in Section 7.1, does not act as a universal accuracy booster. Instead, it increases recall, meaning that fewer real parcels are missed and a larger share of the reference fields is successfully detected, at the cost of a systematic precision reduction, shifting predictions toward more inclusive parcel masks. In agricultural parcel mapping, this trade-off is useful because missing field interiors and incomplete parcel coverage are a more limiting failure mode than moderate reductions in geometric conservativeness.

The comparison between SAM and SAM2 described in Section 7.4 is best read as a trade-off between coverage and conservativeness. SAM produces denser parcel-instance masks and therefore reaches higher recall, especially for small and closely adjacent fields. SAM2 behaves more conservatively and produces fewer candidate regions, leaving part of the parcel instances undetected. This difference also appears in the error profile: SAM favors more exhaustive instance discovery, whereas SAM2 records the highest precision and cleaner delineations on the parcels it detects.

SAM3 completes the SAM-based family, and the NL validation results refine its characterization. As discussed in Section 7.4, its text-prompted inference setting yields acceptable results on PS and S2DR3, yet this behavior does not extend to S2 at 10 m spatial resolution, where several fields remain undetected. This pattern, consistent across both Trentino and the Dutch validation, suggests that text-prompted grounding is more sensitive to spatial resolution than embedding-based prompting: semantic disambiguation of field boundaries may require sufficient visual detail to align with the language prior. In overall terms, it remains less convincing than VE + SAM for the present delineation setting, particularly when high-resolution imagery is unavailable. LangSAM provides a useful contrast: in the tested configuration, its language-guided detection setup did not produce valid parcel-level instance outputs, which further highlights the difficulty of transferring broad semantic prompting directly to parcel delineation.

Conversely, DelAny exhibits a different behavior, as discussed in Section 7.4. It tends to cover cultivated areas broadly, but its main limitation lies in instance separation, since adjacent or visually similar parcels are frequently merged into a single mask. This pattern persists across sources, even when precision, recall, and mIoU improve with finer spatial detail. FastSAM remains quantitatively close to DelAny, but does not materially alter the overall ranking of the evaluated model families. In practical terms, the SAM-based pipeline appears to be the more effective option when parcel separation is the main objective.

The results show that, in highly fragmented agricultural landscapes, input resolution can dominate apparent model performance: the transition from Sentinel-2 to finer-resolution sources substantially improves parcel detectability because many boundaries are too narrow or weakly expressed at 10 m. Model comparisons should therefore be interpreted within each imagery source rather than as resolution-independent rankings of segmentation architectures.

The external validation introduced in Section 8.2 and further examined in Section 8.3 clarifies that the role of spatial resolution is strongly conditioned by parcel size and landscape structure. Indeed, in the Dutch tiles, where parcels are substantially larger than in Trentino, S2 is no longer fundamentally inadequate and the gains from finer-resolution imagery are much smaller. At the same time, the proposed VE + SAM pipeline continues to record lower structural-error rates across sources and tile blocks, indicating fewer split and merge errors in parcel delineation, including in the spatially biased AI4SmallFarms tile. Because the same structural tendency was already observed in Trentino, its persistence under external validation supports the interpretation that the proposed pipeline provides a real advantage in boundary localization rather than a dataset-specific artifact. Because the pipeline is tile-based and does not require site-specific retraining, it is technically com-

patible with regional or national-scale deployment where suitable imagery and LPIS-like reference data are available. The per-tile processing times reported in Table 2 provide a first operational basis for estimating computational load under a defined tiling scheme. These values should not be interpreted as a direct one-to-one estimate of national production time, because large-scale deployment also depends on image availability, cloud filtering, tile overlap, I/O, post-processing, and compute parallelization. Nevertheless, they make the computational implications of scaling the pipeline more transparent. The resulting vector outputs would also support downstream analyses such as parcel-size distributions, fragmentation metrics, and area-based summaries. For comparative benchmarking, national-scale evaluation further requires careful control of training-data overlap, especially for supervised baselines trained on large datasets that may already include the target country.

9.2. Conclusions, limitations, and future work

This paper addressed APBD as a zero-shot, structure-aware problem in fragmented smallholder-like landscapes. The proposed pipeline combines vegetation-enhancement preprocessing, foundation-model-based instance segmentation, and a structure-aware evaluation protocol, designed for zero-shot, cross-source applicability without site-specific retraining.

Across the Trentino benchmark, the proposed VE + SAM pipeline applied to S2DR3 provided the strongest overall configuration, while SAM2 combined with S2DR3 offered the highest precision under a more conservative delineation profile. Overall, these results show that performance in highly fragmented landscapes depends not only on model family, but also on the interaction between model behavior and source resolution.

The external validation over tiles in the Netherlands confirmed the main structural finding of the study. Even in a setting with larger parcels and a weaker dependence on spatial resolution, the proposed pipeline remained competitive with DelAny and more consistently reduced structural errors. This suggests encouraging transfer beyond the specific landscape conditions represented in the Trentino benchmark, while still requiring broader validation across additional agricultural settings.

The Trentino benchmark was designed as a controlled case-study evaluation, not as a multi-site replicated experiment from which formal statistical significance tests could be derived. Model differences should therefore be interpreted as comparative benchmark results under the tested imagery sources, landscape structure, and evaluation protocol, rather than as population-level estimates of model superiority. The Netherlands evaluation provides an additional cross-setting check over independent tiles, but a full statistical assessment of model ability would require a larger multi-region sampling design with replicated landscapes, acquisition conditions, imagery sources, and parcel-size distributions.

Several limitations bound the present findings. The analysis remains centered on fragmented agricultural settings, and additional validation is needed before extending the conclusions to landscapes with substantially different parcel structure. The reported behavior also depends on the specific zero-shot pipeline configurations and preprocessing choices adopted in this study. Existing benchmark datasets in this domain may inherit annotation noise or management-unit ambiguities from LPIS- or GSAA-derived sources. In addition, SAM-based pipelines require auxiliary land-use information for scene-level post-filtering. This dependency introduces a potential mode of error propagation: if agricultural areas are omitted from the external layer, valid parcel predictions are incorrectly discarded, whereas unmapped non-agricultural areas can cause spurious masks to be retained. Temporal mismatch can introduce similar effects when reference maps lag behind recent land-use changes. In this study, this risk is mitigated by the broad semantic level of the filtering step and the established reliability of the reference layers, which typically exhibit precision and recall rates exceeding 90% and 85% for agricultural macro-classes [45]. Furthermore, wide-scale continental inventories show that annual cross-category land transitions are highly stable, generally remaining below 0.5% [50], meaning a minor 1-to-2-year version lag introduces negligible masking noise. Conversely, while this post-filtering structure leaves the zero-shot pipeline dependent on external dataset currency, traditional detector-style approaches instead inherit structural and semantic biases directly from the training data on which they are built.

Future work should investigate multitemporal input conditioning, which may help resolve weak or ambiguous boundaries that remain difficult to identify in single-date imagery. More broadly, progress in this domain will also depend on more accurate and geographically diverse parcel-level ground truth, especially for fragmented smallholder environments where even 1 m imagery does not yet deliver fully reliable zero-shot delineation.

CRediT authorship contribution statement

Joao Calisto: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation; **Massimo Vecchio:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Miguel Pincheira:** Writing – review & editing, Writing – original draft, Validation, Methodology; **Fabio Antonelli:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

This research was conducted within the IT4LIA AI FACTORY project, funded by the Italian Ministry of University and Research (MUR) and the National Cybersecurity Agency (ACN), and co-funded by the European Commission through the European High Performance Computing Joint Undertaking (EuroHPC JU) under Grant Agreement No. 101234224.

References

[1] Y. Sadeh, X. Zhu, D. Dunkerley, J.P. Walker, Y. Chen, K. Chenu, Versatile crop yield estimator, Agron. Sustain. Dev. 44 (4) (2024) 42. https://doi.org/10.1007/s13593-024-00974-4

[2] H. Wang, Z. Ma, Y. Ren, S. Du, H. Lu, Y. Shang, S. Hu, G. Zhang, Z. Meng, C. Wen, W. Fu, Interactive image segmentation based field boundary perception method and software for autonomous agricultural machinery path planning, Comput. Electron. Agric. 217 (2024) 108568. https://doi.org/10.1016/j.compag.2023.108568

[3] Z. Huang, X. Yang, Y. Liu, Z. Wang, Y. Ma, H. Jing, X. Liu, Multi-type change detection and distinction of cultivated land parcels in high-resolution remote sensing images based on segment anything model, Remote Sens. 17 (5) (2025) 787. https://doi.org/10.3390/rs17050787

[4] A. Hadir, M. Adjou, O. Assainova, G. Palka, M. Elbouz, Comparative study of agricultural parcel delineation deep learning methods using satellite images: validation through parcels complexity, Smart Agric. Technol. 10 (2025) 100833. https://doi.org/10.1016/j.atech.2025.100833

[5] J. Zheng, Z. Ye, Y. Wen, J. Huang, Z. Zhang, Q. Li, Q. Hu, B. Xu, L. Zhao, H. Fu, A comprehensive review of agricultural parcel and boundary delineation from remote sensing images: recent progress and future perspectives (2025). arXiv:2508.14558

[6] L.P. Osco, Q. Wu, E.L. de Lemos, W.N. Gonçalves, A.P.M. Ramos, J. Li, J. Marcato, The segment anything model (SAM) for remote sensing applications: from zero to one shot, Int. J. Appl. Earth Obs. Geoinform. 124 (2023) 103540. https://doi.org/10.1016/j.jag.2023.103540

[7] K. Fang, S. Zhang, Y. Han, L. Yang, M. Luo, L. Liu, Q. Zhang, B. Wang, Mscpunet: a multi-task neural network for plot-level crop classification in complex agricultural areas, Smart Agric. Technol. 9 (2024) 100660. https://doi.org/10.1016/j.atech.2024.100660

[8] B. Watkins, A. van Niekerk, A comparison of object-based image analysis approaches for field boundary delineation using multi-temporal sentinel-2 imagery, Comput. Electron. Agric. 158 (2019) 294–302. https://doi.org/10.1016/j.compag.2019.02.009
[9] M. Torre, B. Remeseiro, P. Radeva, F. Martinez, Deepnem: deep network energy-minimization for agricultural field segmentation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 726–737. https://doi.org/10.1109/JSTARS.2020.2971061
[10] F. Waldner, F.I. Diakogiannis, Deep learning on edge: extracting field boundaries from satellite images with a convolutional neural network, Remote Sens. Environ. 245 (2020) 111741. https://doi.org/10.1016/j.rse.2020.111741
[11] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: common objects in context, arXiv preprint arXiv:1405.0312 (2015). https://arxiv.org/abs/1405.0312.
[12] T. Su, S. Zhang, Local and global evaluation for remote sensing image segmentation, ISPRS J. Photogramm. Remote Sens. 130 (2017) 256–276. https://doi.org/10.1016/j.isprsjprs.2017.06.003
[13] Y. Pan, X. Wang, L. Zhang, Y. Zhong, E2evap: end-to-end vectorization of smallholder agricultural parcel boundaries from high-resolution remote sensing imagery, ISPRS J. Photogramm. Remote Sens. 203 (2023) 246–264. https://doi.org/10.1016/j.isprsjprs.2023.08.001
[14] D.S. Costa, A. Barriguinha, I. Areosa, M. Caetano, I. Teixeira, L. Vanneschi, Remote sensing-based approaches for automatic vineyard area identification: a systematic review, Smart Agric. Technol. 13 (2026) 101812. https://doi.org/10.1016/j.atech.2026.101812
[15] A. Carraro, M. Sozzi, F. Marinello, The segment anything model (SAM) for accelerating the smart farming revolution, Smart Agric. Technol. 6 (2023) 100367. https://doi.org/10.1016/j.atech.2023.100367
[16] B. Awad, I. Erer, Boundary sam: improved parcel boundary delineation using sam's image embeddings and detail enhancement filters, IEEE Geosci. Remote Sens. Lett. (2025) 1. PP (99).
[17] M. Chanev, I. Kamenova, P. Dimitrov, L.H. Filchev, Evaluation of sentinel-2 deep resolution 3.0 data for winter crop identification and organic barley yield prediction, Remote Sens. 17 (6) (2025) 957. https://doi.org/10.3390/rs17060957
[18] Y. Akhtman, Sentinel-2 Deep Resolution, 2024, https://medium.com/@ya_71389/sentinel-2-deep-resolution-3-0-c71a601a2253. Accessed: 23 January 2024.
[19] H.C. North, D. Pairman, S.E. Belliss, Boundary delineation of agricultural fields in multitemporal satellite imagery, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (1) (2019) 237–251. https://doi.org/10.1109/JSTARS.2018.2884513
[20] R. Hong, J. Park, S. Jang, H. Shin, H. Kim, I. Song, Development of a parcel-level land boundary extraction algorithm for aerial imagery of regularly arranged agricultural areas, Remote Sens. 13 (6) (2021) 1167. https://doi.org/10.3390/rs13061167
[21] M. Wang, J. Wang, Y. Cui, J. Liu, L. Chen, Agricultural field boundary delineation with satellite image segmentation for high-resolution crop mapping: a case study of rice paddy, Agronomy 12 (10) (2022) 2342. https://doi.org/10.3390/agronomy12102342
[22] X. Zhang, T. Zhang, G. Wang, P. Zhu, X. Tang, X. Jia, L. Jiao, Remote sensing object detection meets deep learning: a meta-review of challenges and advances, 2023, arXiv:2309.06751 [cs.CV]
[23] M. Li, J. Long, A. Stein, X. Wang, Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images, ISPRS J. Photogramm. Remote Sens. 200 (2023) 24–40. https://doi.org/10.1016/j.isprsjprs.2023.04.019
[24] H. Wu, J. Xie, W. Deng, A. Lin, A.R.M. Shariff, S. Akmalov, W. Wu, Z. Li, Q. Yu, Q. Wang, J. Zhang, X. Mei, Q. Hu, ct-hiffnet: a contour-texture hierarchical feature fusion network for cropland field parcel extraction from high-resolution remote sensing images, Comput. Electron. Agric. 239 (2025) 111010. https://doi.org/10.1016/j.compag.2025.111010
[25] M. Li, C. Lu, M. Lin, X. Xiu, J. Long, X. Wang, Extracting vectorized agricultural parcels from high-resolution satellite images using a point-line-region interactive multitask model, Comput. Electron. Agric. 231 (2025) 109953. https://doi.org/10.1016/j.compag.2025.109953
[26] Y. Zhu, Y. Pan, T. Hu, D. Zhang, C. Zhao, Y. Gao, A generalized framework for agricultural field delineation from high-resolution satellite imageries, Int. J. Digit. Earth 17 (1) (2024) 2297947. https://doi.org/10.1080/17538947.2023.2297947
[27] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: a survey, 2024, arXiv:2304.00685.
[28] H. Zhao, B. Wu, M. Zhang, J. Long, F. Tian, Y. Xie, H. Zeng, Z. Zheng, Z. Ma, M. Wang, J. Li, A large-scale vhr parcel dataset and a novel hierarchical semantic boundary-guided network for agricultural parcel delineation, ISPRS J. Photogramm. Remote Sens. 221 (2025) 1–19. https://doi.org/10.1016/j.isprsjprs.2025.01.034
[29] J. Xie, H. Wu, W. Wu, L. Hong, L. He, Q. Yu, L. Liu, A. Lin, J. Som-ard, A cnn-transformer hybrid network with boundary guidance for mapping cropland field parcels from high-resolution remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–22. https://doi.org/10.1109/TGRS.2026.3651284
[30] B. Zhong, T. Wei, X. Luo, B. Du, L. Hu, K. Ao, A. Yang, J. Wu, Multi-swin mask transformer for instance segmentation of agricultural field extraction, Remote Sens. 15 (3) (2023) 549. https://doi.org/10.3390/rs15030549
[31] Y. Xie, H. Wu, H. Tong, L. Xiao, W. Zhou, L. Li, T.C. Wanger, fabSAM: a farmland boundary delineation method based on the segment anything model, 2025, https://arxiv.org/abs/2501.12487.
[32] J. Long, H. Zhao, M. Li, X. Wang, C. Lu, Integrating segment anything model derived boundary prior and high-level semantics for cropland extraction from high-resolution remote sensing images, IEEE Geosci. Remote Sens. Lett. 21 (2024) 1–5. https://doi.org/10.1109/LGRS.2024.3454263
[33] M. Lavreniuk, N. Kussul, A. Shelestov, B. Yailymov, Y. Salii, V. Kuzin, Z. Szantoi, Delineate anything: resolution-agnostic field boundary delineation on satellite imagery, (2025). https://arxiv.org/abs/2504.02534.
[34] M. Lavreniuk, N. Kussul, A. Shelestov, Y. Salii, V. Kuzin, S. Skakun, Z. Szantoi, Delineate anything flow: fast, country-level field boundary detection from any source (2025). http://arxiv.org/abs/2511.13417.
[35] L.B. Ferreira, V.S. Martins, U.R. Aires, N. Wijewardane, X. Zhang, S. Samiappan, FieldSeg: a scalable agricultural field extraction framework based on the segment anything model and 10-m sentinel-2 imagery, Comput. Electron. Agric. 232 (2025) 110086. https://doi.org/10.1016/j.compag.2025.110086
[36] H. Sun, Z. Wei, W. Yu, G. Yang, J. She, H. Zheng, C. Jiang, X. Yao, Y. Zhu, W. Cao, T. Cheng, I. Ali, SIDEST: a sample-free framework for crop field boundary delineation by integrating super-resolution image reconstruction and dual edge-corrected segment anything model, Comput. Electron. Agric. 230 (2025) 109897. https://doi.org/10.1016/j.compag.2025.109897
[37] R. d'Andrimont, M. Claverie, P. Kempeneers, D. Muraro, M. Yordanov, D. Peressutti, M. Batič, F. Waldner, AI4Boundaries: an open AI-ready dataset to map field boundaries with sentinel-2 and aerial photography, Earth Syst. Sci. Data 15 (1) (2023) 317–329. https://doi.org/10.5194/essd-15-317-2023
[38] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, 2023, https://arxiv.org/abs/2304.02643.
[39] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K.V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, C. Feichtenhofer, Sam 2: segment anything in images and videos, (2024). https://arxiv.org/abs/2408.00714.
[40] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K.V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H.K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, C. Feichtenhofer, Sam 3: segment anything with concepts, (2025). https://arxiv.org/abs/2511.16719.
[41] C. Joao, M. Vecchio, M. Pincheira Caro, F. Antonelli, Zero-Shot Parcel Instance Delineation in Fragmented Agricultural Landscapes: Cross-Source Evaluation and Structure-Aware Validation, 2026, https://doi.org/10.5281/zenodo.20667660.
[42] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, J. Wang, Fast segment anything (2023). https://arxiv.org/abs/2306.12156.
[43] L. Medeiros, Langsam : language segment-anything, 2023, (https://github.com/luca-medeiros/lang-segment-anything).
[44] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang, Grounding dino: marrying dino with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499 (2024). https://arxiv.org/abs/2303.05499.
[45] M. Schultz, H. Li, Z. Wu, D. Wiell, M. Auer, Z. Alexander, Osmland use a dataset of European union land use at 10 m resolution derived from openstreetmap and sentinel-2, Sci. Data 12 (1) (2025). https://doi.org/10.1038/s41597-025-04703-8
[46] H.W. Kuhn, The hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1955) 83–97. https://doi.org/10.1002/nav.3800020109
[47] Istituto Nazionale di Statistica (ISTAT), Censimento generale dell'agricoltura: primi risultati, 2022, https://www.istat.it/it/files//2022/06/REPORT-CENSIAGRI_2021-def.pdf.
[48] M. Carrieri de Souza, O.J. Rover, F. Forno, Food networks and agroecology in the province of Trento – Italy, Front. Sustain. Food Syst. 7 (2023). https://doi.org/10.3389/fsufs.2023.1130082
[49] J. Yang, H. Ren, Z. He, M. Zeng, Deep learning-based remote sensing image super-resolution: recent advances and challenges, Neurocomputing 674 (2026) 132939. https://doi.org/10.1016/j.neucom.2026.132939
[50] K. Winkler, R. Fuchs, M. Rounsevell, M. Herold, Global land use changes are four times greater than previously estimated, Nat. Commun. 12 (1) (2021) 2501. https://doi.org/10.1038/s41467-021-22702-2