PROMPT-SWITCHABLE TOPOLOGICAL VECTORIZATION: STRUCTURED GEOMETRY DECODING FROM A FROZEN SEGMENTATION FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Downstream geospatial applications often require segmentation outputs to be represented as polygons satisfying hard geometric and topological constraints: one polygon per instance, no overlapping interiors, and exact shared boundaries between adjacent polygons. Converting masks to polygons is usually left to post-processing, which makes topology something to repair rather than something to guarantee. We move these constraints into the decoder: a frozen promptable segmentation backbone is adapted with LoRA and a geometry head on the same features, 7.1M trainable parameters versus 841M frozen, and the predicted geometry fields are decoded into a planar graph whose faces satisfy these constraints by construction. Our experiments reveal two broader findings. Instance metrics are close to blind to structural decodability: dropping the geometric supervision costs half a point of micro F1 while raising uncovered instances from 11.3% to 96.7%. And an open vocabulary at the prompt does not reach the output: switching the prompt alone gives 0.0800 micro F1, whereas ten annotated images give 0.8173. On three cropland domains, the topological output outperforms all evaluated baselines, with a 17-point gain on the one genuine zero-shot transfer, while returning ten to sixteen times the shared boundary of a contour pipeline, and still seven to nine times after that pipeline's output has been through a standard GIS repair.

1 INTRODUCTION

Cadastral registries, subsidy audits and planning pipelines consume polygons, not pixels, and a parcel map is usable only when adjacent parcels share their boundary exactly: no slivers, no overlaps, one edge owned by two faces. Segmentation models return masks, the step from mask to polygon is left to post-processing, and topology becomes something to repair after the fact, unreliably, because two independently traced contours have no reason to agree on where their common boundary lies.

Two lines of work each hold one half of what this needs. Prompt-driven segmentation models accept an open vocabulary but stop at raster output (Carion et al., 2025; Li et al., 2025; Song et al., 2026). Specialised vectorization models emit instance polygons (Zorzi et al., 2022; Xu et al., 2023a; Adimoolam et al., 2025; Lavreniuk et al., 2026), each trained for one category fixed before training begins: asked outside it, a cropland model scores 0.0291 micro F1 and two building models between 0.0004 and 0.0191. Neither line gives structured instance polygons with topological guarantees while retaining a prompt-driven category interface.

We freeze the backbone and adapt it in two places: LoRA in the global attention blocks and a geometry head on the same features, so the geometric loss reaches the backbone through them (7.1M trainable, 0.84%). The edge field decodes into a planar graph whose faces the masks label, and two adjacent faces reference one arc, so the constraint holds by construction.

Two findings outlast the system. Detection metrics barely see whether a representation can be decoded into structure: dropping the geometric supervision costs half a point of F1 and raises uncovered instances from 11.3% to 96.7%. And an open vocabulary at the prompt does not reach the output, since buildings give 0.0800 zero-shot and 0.8173 from ten images; we therefore call the model prompt-switchable rather than open-vocabulary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

On the three cropland domains the topological output reaches 0.6023, 0.4945 and 0.6410 micro F1, above every baseline we could run; the middle number is a zero-shot transfer to a different country and sensor, 17 points above the strongest closed-set cropland model there. A contour decoder on the same weights scores 0.3 to 4.6 points higher, and the topological output returns ten to sixteen times its shared boundary; a standard GIS repair of the contour output costs under half a point of F1 and leaves it with a seventh to a ninth of that shared boundary, since it removes errors without creating structure. Buildings are less flattering: a training-free open-vocabulary baseline reaches 0.8665, which our topological output does not reach at any annotation budget.

Our contributions are threefold:

- We formulate mask-to-polygon conversion as constrained structured decoding, satisfying non-overlap and shared boundaries through a planar graph rather than a loss term (Section 3), and measure what that buys in the units a map consumer uses (Section 4.5).
- We show that coupling geometric supervision to parameter-efficient adaptation is what makes that decoding usable at all, and that standard instance metrics do not reveal it (Section 4.6).
- We build one prompt-switchable model that leads all five baselines on three cropland domains and reaches 0.8173 on buildings from ten annotated images, where closed-set baselines fall below 0.17 outside their training category (Sections 4.3 and 4.4).

2 Related Work

2.1 Promptable segmentation and open-vocabulary transfer

Promptable segmentation started with a geometric interface: SAM and SAM2 turn a point or a box into a mask (Kirillov et al., 2023; Ravi et al., 2025), and the prompt says where, not what. Open-vocabulary recognition supplied the other half, aligning region features with CLIP text embeddings (Radford et al., 2021) for detection (Gu et al., 2022; Li et al., 2022; Liu et al., 2024b) and segmentation (Xu et al., 2023b); SAM3 joins the two, a noun phrase in and every matching instance out (Carion et al., 2025). Remote sensing adopted it within months, training-free, with calibrated prompts, or on small parameter budgets (Li et al., 2025; Song et al., 2026; Wang et al., 2026; Blushtein-Livnon et al., 2025; Dabaja & Celik, 2026; Zhang et al., 2026a). Every one of them returns a raster. We keep the concept interface and change what leaves it.

2.2 Structured polygon output and graph decoding

Building footprint models have tried nearly every output parameterisation (Li et al., 2019; Adimoolam et al., 2025; Liang et al., 2020; Zorzi et al., 2022; Girard et al., 2021; Xu et al., 2023a; Wang et al., 2024), and what they share matters more here: each predicts one instance's vertices independently of its neighbours, for a category fixed before training. Independence is nearly free on buildings, which stand apart, and expensive on parcels, which tile the plane, so work there restores the shared boundary after the fact (Lavreniuk et al., 2025; 2026; Long et al., 2022; Liu et al., 2025; Duan et al., 2026; Zheng et al., 2026). Our decoder instead reads a planar graph out of raster features, as road and wireframe parsers do (Bastani et al., 2018; He et al., 2020; Xu et al., 2022; Xue et al., 2020); those stop at nodes and edges, we continue to the faces, where non-overlap and shared boundaries stop costing anything. Concurrent work reaches the same corner (Jiao et al., 2026; Yan et al., 2026; Wang et al., 2025; Jiao et al., 2025); none of it couples the structured output to a prompt-driven backbone.

2.3 Topological constraints and parameter-efficient adaptation

Topology can instead be pushed into the loss (Hu et al., 2019; Clough et al., 2022; Hu et al., 2021; Shit et al., 2021; Mosinska et al., 2018). When the consumer is a GIS layer, where non-overlap and shared boundaries belong to the data model rather than to its quality (Egenhofer & Franzosa, 1991), we think that is the wrong place: a loss lowers the probability of a violation, it does not remove the violation from the output space, and nothing here argues against a topological loss as well. Our adaptation follows the parameter-efficient line (Hu et al., 2022; Houlsby et al., 2019; Jia et al., 2022;

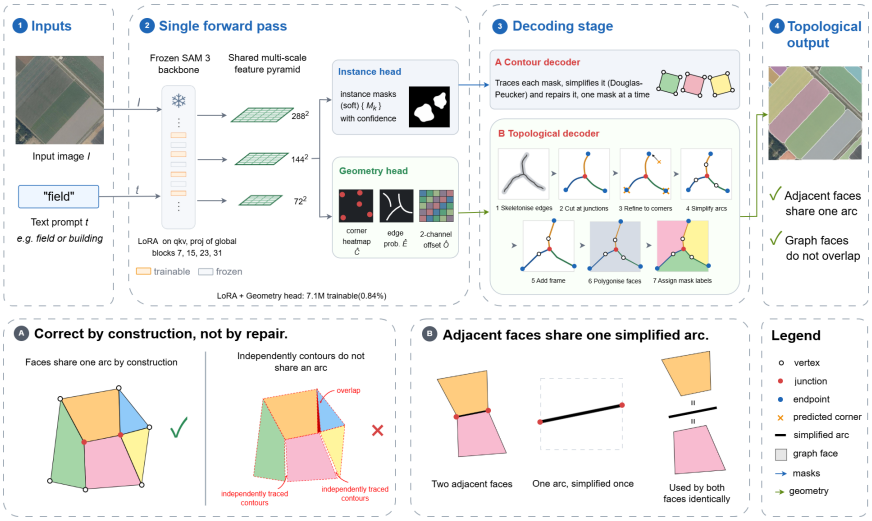


Figure 1: One pass through the model, and the two ways out of it. Instance masks leave through the instance head, the edge, corner and offset fields through the geometry head, both from one feature pyramid. Decoding selects one route at inference: the contour path starts from the masks, the topological path from the geometry fields, returning faces of a planar graph and consulting the masks only to label them (step 7). Panel A contrasts the output with a repaired contour output; panel B shows two adjacent faces sharing one arc.

Chen et al., 2022; Liu et al., 2024a; Chen et al., 2023), differing in the one way Section 4.6 measures: there the added supervision is what the backbone was pre-trained on, ours is geometric and arrives only through shared features.

3 Method

3.1 Overview

Let $\Omega \subset \mathbb{R}^2$ be the image domain. Given an image $I : \Omega \to \mathbb{R}^3$ and a text prompt t naming one category, we return simple polygons with labels and confidences, $\mathcal{P} = \{(P_i, \ell_i, s_i)\}_{i=1}^{N}$ with $P_i \subset \Omega$. Accuracy alone does not make such a set usable; a mapping pipeline also needs the polygons to form a partition-like arrangement, which we write as membership in

$$\mathcal{A} = \Big\{ \{P_i\}_{i=1}^{N} \;:\; \operatorname{int}(P_i) \cap \operatorname{int}(P_j) = \emptyset \\ \text{and } \partial P_i \cap \partial P_j \text{ is a union of shared vertices and arcs}, \quad i \neq j \Big\}. \tag{1}$$

The second condition is stronger than it looks: it is not enough that two neighbouring boundaries coincide geometrically, they must be the same arc, so that an edit to one is an edit to the other and no sliver can open between them.

Where that condition is enforced is the design decision of this paper. A topological penalty gives an objective $\min_\theta \mathbb{E}[\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{topo}}]$ that lowers the probability of a violation without removing violations from the range of the model: whatever λ is chosen, $\mathcal{P}_\theta \in \mathcal{A}$ remains an event rather than a fact. We move the condition into the decoder $\mathcal{D}$, built so that

$$\operatorname{range}(\mathcal{D}) \subseteq \mathcal{A}. \tag{2}$$

Section 3.3 constructs such a $\mathcal{D}$ and is explicit about the one path around it. Predictions are matched to ground truth by greedy assignment at an intersection-over-union threshold of 0.5 (Lin et al., 2014); the structural measures are ours and Section 4.1 defines them.

A single forward pass does the recognition (Figure 1). Image and prompt enter a frozen segmentation backbone carrying injected LoRA matrices, and two things come out: instance masks $\{M_k\}$ with

confidences from the instance head, and three geometry fields from the geometry head, an edge probability $\hat{E}$, a corner heatmap $\hat{C}$ and a two-channel sub-pixel offset field $\hat{O}$. Both heads read the same feature pyramid, which is the point of the design rather than an implementation detail, and Section 4.6 tests that claim. Decoding runs as a separate stage over those outputs, along one of two main routes. The topological route satisfies equation 2: it reads $\hat{E}$ as a planar graph and returns its faces, so two neighbouring faces reference a single arc and cannot overlap. That guarantee has one hole we report rather than hide, since polygons added by the residual fallback lie outside the graph and put roughly 0.004 of overlap back.

3.2 Geometry-coupled parameter-efficient adaptation

Attaching a geometry head to a frozen backbone is not the same as training one alongside it. Because the geometric loss reaches the backbone only through the features the instance head also reads, it changes the direction in which the adaptation moves rather than adding an independent output. When Section 4.6 separates the two, the difference turns out to be almost invisible to F1 and decisive for everything else.

We inject LoRA (Hu et al., 2022) into the query-key-value and output projections of the four global attention blocks of the vision trunk, at rank 16 with scaling 16, which comes to $393{,}216$ parameters. The class-agnostic geometry head aligns the three pyramid levels to 256×256, fuses them with a 1×1 convolution, and splits the result into branches for edges, corners and an attraction field that steers the other two through channel attention. Three predictors emit an edge logit map, a CenterNet-style corner heatmap with output bias initialised to -2.19, and the two-channel sub-pixel offset field. Those $6{,}684{,}894$ parameters bring the trainable total to $7{,}078{,}110$, or 0.84% of the $840{,}902{,}966$ in the backbone.

The three fields are supervised by a CenterNet-style focal loss on the corner heatmap against Gaussian soft targets, an ℓ_1 loss on the offset field restricted to vertex pixels and normalised by their count, and a class-balanced binary cross-entropy on the edge map; the three are combined under learned uncertainty weights and join the detection loss with both weights at 1.0. Appendix C gives the definitions and the training schedule. The joint model sees AI4Boundaries and iFLYTEK together, which is why both are called training domains throughout and neither is reported as a zero-shot result.

3.3 Topological decoding

Shared boundaries hold here because the representation leaves them no way to fail. For a planar graph $G = (V, E)$ embedded in Ω with the frame $\partial\Omega$ among its arcs, write

$$F(G) \;=\; \text{the bounded connected components of } \Omega \setminus \textstyle\bigcup_{e \in E} e. \tag{3}$$

Any two distinct $f, f' \in F(G)$ satisfy $\operatorname{int}(f) \cap \operatorname{int}(f') = \emptyset$ and $\bar{f} \cap \bar{f}'$ is a union of shared vertices and arcs, which is exactly the pair of conditions in equation 1: membership in $\mathcal{A}$ is a property of the representation, inherited by any decoder that returns faces of a graph.

Edges come first. We threshold the edge map at $\tau_e = 0.6$, skeletonise it, and cut the skeleton into polylines $\{\gamma_m\}$ at every pixel whose eight-neighbourhood degree differs from two, that is at junctions and free endpoints. Corners refine rather than seed: peaks of $\hat{C}$ above $\tau_j = 0.3$, displaced by the offset field to $V = \{\, p + (\sigma(\hat{O}_p) - \tfrac{1}{2}) \,\}$, attract each segment endpoint within $\rho = 3$ pixels. Connectivity comes from the skeleton, not the heatmap, so a missed corner costs precision at one junction instead of erasing an edge.

Simplification is where the shared-boundary property is won. We apply Douglas-Peucker per segment with tolerance $\epsilon = 1$ pixel and both endpoints held fixed, replacing each γ_m by $\mathrm{DP}_\epsilon(\gamma_m)$ before any face exists. Because simplification acts directly on the graph segments, the two faces that later claim a segment inherit one and the same simplified polyline. Simplifying faces instead, which is what a contour pipeline necessarily does, computes $\mathrm{DP}_\epsilon(\partial f)$ and $\mathrm{DP}_\epsilon(\partial f')$ independently; these agree on the shared portion only by accident, and where they disagree a sliver opens.

Faces come last. We add $\partial\Omega$ to the polylines, node the union and polygonise it, clamping endpoints within 3 pixels of the frame so boundary parcels close. Each face f takes the label and confidence

of the mask it overlaps most, $\ell(f) = \arg\max_k |f \cap M_k|$, and is dropped as background when $|f \cap M_{\ell(f)}|/|f| < \tau_c = 0.7$. The assignment need not be injective, which is useful: an under-segmented mask covering two parcels supports two faces and is split along the traced boundary, a repair contour extraction cannot make.

Fallbacks, and what they cost. An image whose graph yields no face falls back to contour decoding, counted in fallback rate at 4.0% to 9.4% across our cropland domains. Instances left with no face are vectorised separately and appended, cutting uncovered ground-truth instances on AI4Boundaries from 11.3% to 2.4%. That second route is the hole in equation 2: appended polygons are not faces of G_ϵ, so overlap rises from 1.0×10^{-5} on the graph output to 3.8×10^{-3} on the delivered set. It stays below the contour path at 4.9×10^{-3} and the closed-set cropland model at 5.2×10^{-3}, but it is not zero: judge the construction on the graph output and the delivered polygons on the combined one. A bridging operator that joins dangling endpoints did not earn its place and is off in the configuration we report.

3.4 Decoding variants

The main paper uses two decoders, named Contour and Topo. The contour path deduplicates the instance masks, cleans them morphologically, extracts contours, simplifies each with Douglas-Peucker and repairs the result, never reading a geometry field. The topological path is the decoder above. It is the output this paper is about, and the contour path is reported beside it as the instance-accuracy reference from the same weights. The two are not ordered: the contour path scores higher under instance matching, the topological path is the only one whose neighbours share an arc, and instance metrics do not see that difference. Section 4.5 puts an exchange rate on it and measures what it is worth downstream.

4 Experiments

4.1 Setup

Every number below comes from one protocol: 351 evaluation images per domain, greedy matching at an intersection-over-union threshold of 0.5, and one post-processing configuration per object category rather than one per domain.

We evaluate on four domains, two of them training domains. AI4Boundaries (d'Andrimont et al., 2023) supplies Dutch parcels at 1 m; we use its full 351-image validation split. iFLYTEK (Zhao et al., 2022) supplies Chinese parcels from Jilin-1 imagery at 0.75 to 1.1 m. Both were in the training set, and neither is ever reported as generalisation. PASTIS (Sainte Fare Garnot & Landrieu, 2021) supplies French parcels from Sentinel-2 and is the only zero-shot domain, since neither its country nor its sensor appears in training. WHU Building (Ji et al., 2019) supplies aerial footprints and is the category-transfer domain; its evaluation set is also 351 images, and the 10, 50 or 100 images quoted for it are the annotated images the model was adapted on, drawn from the training split and disjoint from the 351. The PASTIS, iFLYTEK and WHU subsets are fixed lists reused by every method.

Detection quality is micro precision, recall and F1, pooled from the summed true positives, false positives and false negatives, with mean intersection-over-union and mean PoLiS over matched pairs. Three further measures describe the arrangement itself. For a predicted set $\{P_i\}$ we report the fraction of polygons that are valid, non-empty and of positive area, the overlap rate $\rho_{\text{ovl}} = \sum_{i<j} |P_i \cap P_j| / \sum_i |P_i|$, and the shared-boundary rate, the fraction of total boundary length that runs within δ of some other polygon's boundary,

$$\rho_{\text{shr}} = \sum_i \operatorname{len}\Big(\partial P_i \cap \mathcal{B}\big(\textstyle\bigcup_{j \neq i} \partial P_j,\ \delta\big)\Big) \ \Big/ \ \textstyle\sum_i \operatorname{len}(\partial P_i), \qquad \delta = 1 \text{ pixel}. \tag{4}$$

The tolerance in equation 4 is generous to everyone: a contour method earns shared-boundary length whenever two independently traced outlines pass within one pixel of each other, and the measure never asks whether they are the same arc. Structural rates are computed per image and averaged, each being a ratio internal to one image's polygon set, while coverage is pooled as the fraction of ground-truth instances that no output polygon represents. Section 4.5 adds downstream measures that ask what the arrangement costs a consumer.

Two post-processing configurations appear below, indexed by object category rather than by domain, and they differ only in the minimum-area filter: 400 for cropland, 100 for buildings, whose instances are an order of magnitude smaller. Topological parameters are identical everywhere ($\tau_e = 0.6$, $\rho = 3$, $\epsilon = 1$, $\tau_c = 0.7$, bridging off, residual fallback on).

Both configurations were selected on data we also report on: the cropland grids ran on the same 351 AI4Boundaries images that appear in Table 1, the building configuration on a 50-image subset of the WHU split. No parameter was chosen on a transfer domain, and on iFLYTEK we declined one worth $+1.04$ points of contour F1 because it cost the topological path 0.32. Appendix D gives every grid.

4.2 Baselines and reproducibility

Only one baseline attempts the same task as ours, and the table has to be read with that in mind. SegEarth-OV3 (Li et al., 2025) is the closest comparison, an open-vocabulary model that runs SAM 3 on overhead imagery without training and whose raster output we convert to instances by connected components. Delineate Anything v2 (Lavreniuk et al., 2026) and BsiNet (Long et al., 2022) are closed-set cropland specialists, HiSup (Xu et al., 2023a) and PolyWorld (Zorzi et al., 2022) closed-set building specialists, each run from its authors' released weights without fine-tuning (Appendix A), so every number those four produce is a cross-dataset zero-shot transfer. Cells marked † put a cropland model on buildings or a building model on parcels; they measure how rigidly a closed-set model is bound to its category, and Section 4.4 uses them for exactly that.

4.3 Main results on cropland

The topological path is above every baseline on all three cropland domains, at 0.6023, 0.4945 and 0.6410 micro F1 against 0.5910, 0.3247 and 0.2575 for the strongest closed-set cropland model, and it does so while returning between five and twelve times the shared boundary of that model and between ten and sixteen times that of the contour path from our own weights (Table 1); no other closed-set baseline except DAv2 exceeds a shared-boundary rate of 0.005 on any domain. The grey reference row is the contour decoder run on the same weights; it scores higher still under instance matching, 0.6320, 0.4978 and 0.6711, and the two halves of the table say why we report it: the upper block is what a detection metric sees, the lower block is what a map consumer sees, and no single output is best in both.

PASTIS provides the cleanest evidence of cross-domain generalization because neither its country nor its sensor was seen during training. The topological path reaches 0.4945 there against 0.3247 for the strongest closed-set cropland model, a gap of 17.0 points, and gives up only 0.33 points to the contour path from the same weights. iFLYTEK's numbers are larger and a weaker claim, since that domain was in training.

On AI4Boundaries the lead is won on recall and geometry rather than precision, which sits below DAv2's (Appendix A, Table 6). Section 4.5 asks what the lower block of the table is worth to a consumer of the map; buildings are the subject of the next section.

4.4 Category transfer to buildings

Buildings test the prompt-switching claim, and Table 9 is where its limit shows (Appendix E). Switching the prompt alone gets the cropland-adapted model to 0.0800 micro F1 for the topological output; ten annotated images take it to 0.8173, and going from ten to one hundred adds only 0.02, with a sample standard deviation of 0.011 or less across three seeds. Nearly all of the usable transfer is bought by the first ten images rather than by prompt switching alone. The decoder transfers with them: at ten shots the topological output leaves 1.3% of ground-truth buildings uncovered after the residual fallback, against 75% zero-shot, so the geometry fields, not only the masks, have moved to the new category. Against the baselines the result is mixed: ten images are enough to pass HiSup at 0.7884, which was trained on the full Inria set, while SegEarth-OV3 reaches 0.8665 with no annotation at all, a number the topological output does not reach at any budget and only the contour reference passes, from 50 images. Section 5 explains why buildings suit a connected-component baseline and why the same procedure collapses to 0.23–0.30 on cropland.

Table 1: Upper block: micro F1 at IoU 0.5 on 351 evaluation images per domain. Lower block: shared-boundary rate ρ_{shr} of the same output. † marks cross-category cells, where a model is asked for a category it was never trained on; these are not method comparisons. WHU columns give results after adapting on $10/50/100$ annotated images; the lower block reports the 10-shot model. Bold is the best value per column among methods; the grey row is the contour decoder on our own weights, an instance-accuracy reference for the topological path (Section 3.4), not ranked.

Method	Specialism	AI4B (NL)	PASTIS (FR)	iFLYTEK (CN)	WHU (buildings)
Micro F1 ↑					
SegEarth-OV3	open vocab.	0.2992	0.2589	0.2257	**0.8665**
DAv2	cropland	0.5910	0.3247	0.2568	0.0291†
BsiNet	cropland	0.2265	0.1387	0.2575	0.1642†
HiSup	buildings	0.0004†	0.0008†	0.0014†	0.7884
PolyWorld	buildings	0.0104†	0.0191†	0.0181†	0.3728
Ours (Topo)	prompt-switch.	**0.6023**	**0.4945**	**0.6410**	0.8173 / 0.8362 / 0.8370
Ours (Contour)	*reference*	*0.6320*	*0.4978*	*0.6711*	*0.8635 / 0.8765 / 0.8773*
Shared-boundary rate ρ_{shr} ↑					
SegEarth-OV3	open vocab.	0.0007	0.0009	0.0009	0.0002
DAv2	cropland	0.0511	0.0367	0.0233	0.0150†
BsiNet	cropland	0.0044	0.0047	0.0040	0.0035†
HiSup	buildings	0.0002†	0.0000†	0.0004†	0.0004
PolyWorld	buildings	0.0000†	0.0000†	0.0005†	0.0004
Ours (Topo)	prompt-switch.	**0.3175**	**0.1929**	**0.2869**	**0.0708**
Ours (Contour)	*reference*	*0.0304*	*0.0122*	*0.0282*	*0.0048*

Two readings of that curve are available and only one is defensible. Asked outside their training category, all four closed-set baselines remain below 0.17 micro F1; we reach 0.8173 from ten images. That contrast is real and does not license the stronger claim, because our own zero-shot number is 0.0800, in the same region as the baselines. The prompt buys a model adaptable to a new category from ten images rather than retrained for it, not a category-agnostic polygon extractor, which is why we call the model prompt-switchable and not open-vocabulary. Passing HiSup from ten images is a statement about data efficiency under a frozen backbone, not about a better building model.

4.5 Structural quality, and what it buys downstream

The topological path buys shared boundary and pays in F1. Across four domains it returns between 10.2 and 15.8 times the shared boundary length of the contour path on the same images, for between 0.3 and 4.6 points of micro F1 (Table 1; Appendix B adds overlap, fallback and coverage per domain). PASTIS is the clearest cell, 0.33 points for 15.8 times; the same exchange costs 2.97 points on AI4Boundaries and 4.62 on buildings. The overlap figure needs the qualification promised in Section 3.3: faces of the planar graph overlap at 1.0×10^{-5} on AI4Boundaries, numerical noise around the zero the construction guarantees, while the delivered set, faces plus appended fallback polygons, overlaps at 0.0038, below the contour path at 0.0049 and DAv2 at 0.0052 but not zero. Buildings have a low shared-boundary rate in absolute terms, 0.0708 against 0.19 to 0.32 on cropland, because detached buildings do not share walls; the ratio against the contour path, 14.9×, is what transfers. Figure 2 shows what the ratio looks like: in the contour output neighbouring parcels are separated by two nearly parallel lines with a thin wedge between them, in the topological output there is one line, and in the building panel the polygons are correct with almost no shared boundary because the buildings do not touch.

Table 2 reports measures that need no matching and correspond to the checks a GIS runs before a layer is accepted. Topology errors, counted per image above a one-pixel tolerance for every method, are overlaps, pairs of polygons whose interiors intersect, and gaps, holes in the union of the output smaller than the minimum-area filter that are not a hole of any single polygon. Count error is $|N_{\text{pred}} - N_{\text{gt}}|/N_{\text{gt}}$ and area error is $|\sum_i |P_i| - \sum_j |G_j||/\sum_j |G_j|$, summed over the 351 images of a domain before the ratio is taken, as a census or an audit would compute them. The last two columns apply the repair a GIS would apply, cutting each overlap in favour of the higher-confidence polygon

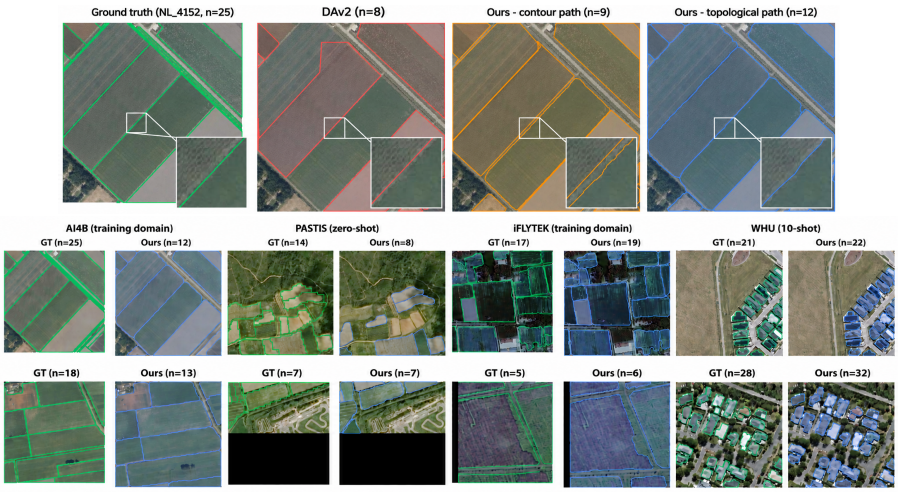


Figure 2: Top: AI4Boundaries ground truth, DAv2, and our two decoders on one tile, each panel zoomed on the same parcel boundary; the contour path leaves two nearly parallel lines there, the topological path the one arc both faces share. Bottom: ground truth (green) and our topological output (blue) on two tiles per domain; buildings, at the right, do not share boundaries.

Table 2: Downstream measures on cropland. Overlaps and gaps are topology errors per image above a one-pixel tolerance; count and area error are over the whole domain; the last two columns are after a standard GIS repair. Lower is better in the first four numeric columns.

		Errors / image				After repair	
Domain	Method	Overlaps	Gaps	Count err.	Area err.	Micro F1	ρ_{shr}
AI4B	DAv2	3.15	**1.17**	0.384	0.403	0.5874	0.0857
	Ours (Contour)	1.11	1.71	0.296	**0.078**	**0.6311**	0.0471
	Ours (Topo)	**0.46**	1.22	**0.208**	0.116	0.6041	**0.3253**
PASTIS	DAv2	1.21	0.30	0.493	0.113	0.3231	0.0572
	Ours (Contour)	0.52	**0.19**	0.070	**0.100**	0.4937	0.0215
	Ours (Topo)	**0.26**	0.46	**0.004**	0.138	**0.4939**	**0.1973**
iFLYTEK	DAv2	0.84	**0.36**	0.697	0.056	0.2523	0.0364
	Ours (Contour)	1.02	1.06	0.196	0.065	**0.6671**	0.0422
	Ours (Topo)	**0.44**	0.92	**0.015**	**0.032**	0.6409	**0.2918**

and dissolving each gap into the neighbour sharing the longest boundary, and re-score the result, so that the paths are compared in a deliverable state.

Three things follow, and the third is the one that matters. First, the topological path leaves fewer errors to repair than the contour path on AI4Boundaries and iFLYTEK, 1.68 against 2.82 and 1.36 against 2.07 per image, and nearly the same on PASTIS, 0.72 against 0.71; DAv2 is at 4.32 on AI4Boundaries. The errors it does leave are the two the construction predicts: overlaps come from the appended fallback polygons, 0.26 to 0.46 per image, and gaps are small faces that no mask claimed, half of them under ten pixels, left unlabelled rather than assigned by guess. Its count error is a fraction of the contour path's on every domain, because a partition returns one face where independent masks return a parcel and its fragments; area error is mixed, the faces slightly over-covering where the annotation leaves margins between parcels.

Second, the repair is cheap. It costs the contour path 0.09, 0.41 and 0.40 points of micro F1, moves DAv2 by less than half a point; anyone who needs a layer that loads can have one from the contour path.

Third, the repair does not make the polygons share boundaries. The contour path's shared-boundary rate rises from 0.0304, 0.0122 and 0.0282 to 0.0471, 0.0215 and 0.0422; the topological path is at

Table 3: Training ablation on AI4Boundaries, topological path. Each row is a separately trained checkpoint evaluated with identical decoding parameters; fallback and uncovered are before the residual fallback.

Trained components	Micro F1	Fallback	Uncovered
LoRA + geometry head (ours)	**0.5926**	**0.0399**	**0.1129**
LoRA only	0.5872	0.7208	0.9669
Geometry head only	0.4761	0.3248	0.4887

0.3253, 0.1973 and 0.2918, still seven to nine times higher. Repair closes enclosed slivers and cuts overlaps, but the two nearly parallel lines between neighbouring parcels in Figure 2 are neither, and they remain two lines. If the product is a layer in which an edit to one parcel's boundary is an edit to its neighbour's, only the topological path delivers it, and the table gives its price in F1.

4.6 Ablation

Removing the geometric supervision costs half a point of micro F1. It costs almost everything else. With LoRA trained alone and the geometry head left out of the loop, 72.1% of images produce no usable graph and 96.7% of ground-truth instances cannot be covered by the decoded planar graph, against 4.0% and 11.3% for the joint model (Table 3). Micro F1 barely reflects this failure.

We think this generalises past our setting. Instance metrics score a region against a region, and a representation can be almost as good at that while being useless for anything that has to walk its boundary; anyone choosing an adaptation recipe by micro F1 would rank LoRA-only within a point of the joint model and never learn that its geometry cannot be decoded. A geometry head on a frozen backbone gives 0.4761: the head cannot compensate for a backbone that was not moved.

5 Discussion and conclusion

The paper leaves two things behind, and they are worth separating. One is a system: a text prompt and an image go in, instance polygons come out, and adjacency among the graph faces is enforced by their representation rather than encouraged by a loss. On three cropland domains its topological output leads every baseline we could run and returns ten to sixteen times the shared boundary of a contour decoder on the same weights, an advantage a standard GIS repair does not close. The other is what building it exposed: the metric used to choose adaptation recipes cannot see whether the result can be decoded into structure at all, since removing the geometric supervision costs half a point of F1 and nearly all of the decodable graph. The first is specific to mapping; we do not think the second is.

The building result is about the targets rather than the models. A training-free baseline that turns a semantic raster into connected components loses almost nothing when the objects it separates do not touch, which helps explain why SegEarth-OV3 reaches 0.8665 on WHU; on cropland, where parcels tile the plane, the same procedure merges neighbours and falls to 0.23–0.30. Results should be read against the topological complexity of the target: a method that wins on detached objects has said little about adjacent ones, and the choice of decoder follows the same logic, the contour path where objects are separated, the topological path where they tile and adjacency has to survive contact with a GIS (Section 4.5). Our claim about the prompt is correspondingly narrow, which is why the paper says prompt-switchable and not open-vocabulary: two categories are not a vocabulary, and we have not shown that ten images suffice for a category geometrically less like cropland than buildings are.

The decoder's weakest cell is coverage at coarse resolution. On PASTIS, at 10 m, 21.3% of ground-truth parcels have no face before the residual fallback and 3.2% after it, against 11.3% and 2.4% on AI4Boundaries (Appendix B): where a parcel is a few pixels wide the edge field closes fewer of its boundaries, and the fallback that rescues those instances is also the source of the residual overlap. Predicting the planar graph end to end, rather than decoding it from an edge field, would remove the fallbacks along with the overlap they reintroduce, and is where we would go next.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to provide feedback on research methodology, experimental design, and the interpretation of experimental results. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or assist in the writing of mathematical proofs; the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for language editing and readability improvement, manuscript organization, literature-related assistance, code editing, and suggestions for the presentation and modification of scientific figures. We have reviewed all AI-assisted work. In particular, AI-assisted code was manually reviewed and tested for correctness, literature-related suggestions were checked against the original sources, and AI-assisted text, figures, and interpretations were verified against the actual methodology and experimental results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yeshwanth Kumar Adimoolam, Charalambos Poullis, and Melinos Averkiou. Pix2Poly: A sequence prediction method for end-to-end polygonal building footprint extraction from remote sensing imagery. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 8473–8482, 2025.

AIcrowd. crowdAI mapping challenge. https://www.aicrowd.com/challenges/mapping-challenge, 2018.

Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. RoadTracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4720–4728, 2018. doi: 10.1109/CVPR.2018.00496.

Roni Blushtein-Livnon, Osher Rafaeli, David Ioffe, Amir Boger, Karen Sandberg Esquenazi, and Tal Svoray. On the effectiveness of textual prompting with lightweight fine-tuning for SAM3 remote sensing segmentation. *arXiv preprint arXiv:2512.15564*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, 2022.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. SAM-Adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3359–3367, 2023. doi: 10.1109/ICCVW60793.2023.00361.

James R. Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A. Zimmer, Julia A. Schnabel, and Andrew P. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8766–8778, 2022. doi: 10.1109/TPAMI.2020.3013679.

Mohammad Dabaja and Turgay Celik. Promptable concept segmentation from above: Evaluating SAM 3's zero-shot and one-shot capabilities in remote sensing. *arXiv preprint arXiv:2607.09583*, 2026.

Raphaël d'Andrimont, Martin Claverie, Pieter Kempeneers, Davide Muraro, Momchil Yordanov, Devis Peressutti, Matej Batič, and François Waldner. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. *Earth System Science Data*, 15 (1):317–329, 2023. doi: 10.5194/essd-15-317-2023.

Yaming Duan, Runze Zhao, Xiangde Xu, and Jinshui Zhang. Topology-aware field parcel delineation: Bridging deep semantic features and geometric constraints. *Remote Sensing*, 18(11): 1783, 2026. doi: 10.3390/rs18111783.

Max J. Egenhofer and Robert D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991. doi: 10.1080/02693799108927841.

Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5891–5900, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M. Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. Sat2Graph: Road graph extraction through graph-tensor encoding. In *European Conference on Computer Vision (ECCV)*, pp. 51–67, 2020. doi: 10.1007/978-3-030-58586-0_4.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xiaoling Hu, Yusu Wang, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-aware segmentation using discrete morse theory. In *International Conference on Learning Representations (ICLR)*, 2021.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. doi: 10.1109/TGRS.2018.2858817.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pp. 709–727, 2022. doi: 10.1007/978-3-031-19827-4_41.

Weiqin Jiao, Hao Cheng, George Vosselman, and Claudio Persello. LDPoly: Latent diffusion for polygonal road outline extraction in large-scale topographic mapping. *arXiv preprint arXiv:2504.20645*, 2025.

Weiqin Jiao, Hao Cheng, George Vosselman, and Claudio Persello. ACPV-Net: All-class polygonal vectorization for seamless vector map generation from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.16616.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Mykola Lavreniuk, Nataliia Kussul, Andrii Shelestov, Bohdan Yailymov, Yevhenii Salii, Volodymyr Kuzin, and Zoltan Szantoi. Delineate anything: Resolution-agnostic field boundary delineation on satellite imagery. In *European Conference on Artificial Intelligence (ECAI)*, 2025.

Mykola Lavreniuk, Nataliia Kussul, Andrii Shelestov, Yevhenii Salii, Volodymyr Kuzin, Charlotte Julia Li-Xing Wang, and Zoltan Szantoi. Delineate anything v2: A global foundation model for field delineation. In *European Conference on Computer Vision Workshops (ECCVW)*, 2026.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. SegEarth-OV3: Exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2022. doi: 10.1109/CVPR52688.2022.01069.

Zuoyue Li, Jan Dirk Wegner, and Aurelien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1715–1724, 2019. doi: 10.1109/ICCV.2019.00180.

Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. PolyTransform: Deep polygon transformer for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9128–9137, 2020. doi: 10.1109/CVPR42600.2020.00915.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32100–32121. PMLR, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024b. doi: 10.1007/978-3-031-72970-6_3.

Xiuyu Liu, Jinshui Zhang, Yaming Duan, and Jiale Zhou. CPVF: vectorization of agricultural cultivation field parcels via a boundary–parcel multi-task learning network in ultra-high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226:267–299, 2025. doi: 10.1016/j.isprsjprs.2025.05.019.

Jiang Long, Mengmeng Li, Xiaoqin Wang, and Alfred Stein. Delineation of agricultural fields using multi-task BsiNet from high-resolution satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 112:102871, 2022. doi: 10.1016/j.jag.2022.102871.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the Inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3226–3229, 2017. doi: 10.1109/IGARSS.2017.8127684.

Agata Mosinska, Pablo Marquez-Neila, Mateusz Kozinski, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3136–3145, 2018. doi: 10.1109/CVPR.2018.00331.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4852–4861, 2021. doi: 10.1109/ICCV48922.2021.00483.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice – a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16560–16569, 2021.

Yanghui Song, Nanqing Liu, Haonan Yin, Yingjie Gao, Chengfu Yang, and Qi Ming. Prompt-calibrated SAM 3 for open-vocabulary remote sensing semantic segmentation. *arXiv preprint arXiv:2606.21863*, 2026.

Chenhao Wang, Jingbo Chen, Yu Meng, Yupeng Deng, Kai Li, and Yunlong Kong. SAMPoly-Build: Adapting the segment anything model for polygonal building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:707–720, 2024. doi: 10.1016/j.isprsjprs.2024.09.018.

Kaizheng Wang, Xiaorun Li, and Shuhan Chen. Parameter-efficient and prompt-free adaptation of SAM3 for remote sensing semantic segmentation. In *Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pp. 1–5, 2026. doi: 10.1109/CACML68972.2026.11506752.

Yu Wang, Bo Dang, Wanchun Li, Wei Chen, and Yansheng Li. HoliTracer: Holistic vectorization of geographic objects from large-size remote sensing imagery. *arXiv preprint arXiv:2507.16251*, 2025.

Bowen Xu, Jiakun Xu, Nan Xue, and Gui-Song Xia. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:284–296, 2023a. doi: 10.1016/j.isprsjprs.2023.03.006.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2945–2954, 2023b. doi: 10.1109/CVPR52729.2023.00288.

Zhenhua Xu, Yuxuan Liu, Lu Gan, Yuxiang Sun, Xinyu Wu, Ming Liu, and Lujia Wang. RNGDet: Road network graph detection by transformer in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. doi: 10.1109/TGRS.2022.3186993.

Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip H. S. Torr. Holistically-attracted wireframe parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2788–2797, 2020.

Yinglong Yan, Yunkai Yang, Haoyi Wang, Wei Fu, Linshan Wu, Honghu Pan, Shaobo Xia, Shanghang Zhang, Hao Chen, and Leyuan Fang. Vector map as language: Toward unified remote sensing vector mapping. *arXiv preprint arXiv:2606.10701*, 2026.

Tao Zhang, Shiqing Wei, Shihao Chen, Wenling Yu, Muying Luo, and Shunping Ji. VectorLLM: Human-like extraction of structured building contours via multimodal LLMs. *arXiv preprint arXiv:2507.04664*, 2025.

Xu Zhang, Danyang Li, Yingjie Xia, Xiaohang Dong, Hualong Yu, Jianye Wang, and Qicheng Li. OmniOVCD: Streamlining open-vocabulary change detection with SAM 3. *arXiv preprint arXiv:2601.13895*, 2026a.

Yaoteng Zhang, Julin Zhang, Guangshuai Wang, Jiwei Deng, Hui Sheng, Yasir Muhammad, and Shiqing Wei. PolyBuild: An end-to-end method for polygonal building contour extraction from high-resolution remote sensing images. *arXiv preprint arXiv:2606.08920*, 2026b.

Zhen Zhao, Yuqiu Liu, Gang Zhang, Liang Tang, and Xiaolin Hu. The winning solution to the iFLYTEK challenge 2021 cultivated land extraction from high-resolution remote sensing images. In *International Conference on Advanced Computational Intelligence (ICACI)*, pp. 376–380, 2022. doi: 10.1109/ICACI55529.2022.9837765.

Juepeng Zheng, Zi Ye, Yibin Wen, Jianxi Huang, Zhiwei Zhang, Qingmei Li, Qiong Hu, Baodong Xu, Lingyuan Zhao, and Haohuan Fu. A comprehensive review of agricultural parcel and boundary delineation from remote sensing images: Recent progress and future perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 14:206–237, 2026. doi: 10.1109/MGRS.2026.3658493.

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1848–1857, 2022.

A Baselines and reproducibility

Weights and provenance. No baseline was fine-tuned. Each was run from its authors' released weights, so every cross-domain number is a zero-shot transfer.

Table 4: Baseline provenance. All evaluated on the same 351-image splits with the same matching protocol as our own model.

Baseline	Output	Pre-training data	Adaptation
SegEarth-OV3 (Li et al., 2025)	raster → components	SAM 3, no training	none
DAv2 (Lavreniuk et al., 2026)	instance masks	official checkpoint	none
BsiNet (Long et al., 2022)	instance masks	Xinjiang GF-2	none
HiSup (Xu et al., 2023a)	polygons	Inria (Maggiori et al., 2017)	none
PolyWorld (Zorzi et al., 2022)	polygons	CrowdAI (AIcrowd, 2018)	none

SegEarth-OV3 has no native instance output. We threshold its semantic raster and take connected components as pseudo-instances, which is the most favourable reading available to it and the reason it performs well on detached objects.

Second pass for the structural measures. The shared-boundary rates in the lower block of Table 1 and everything in Table 2 come from one further pass over every method with the parameters of the first pass unchanged, writing each method's polygons to disk so that the structural measures could be computed offline by one implementation. Micro F1 on that pass agreed with the upper block to four decimals for every method but DAv2, whose GPU inference is not deterministic and moved by at most 0.65 points (0.5845 against 0.5910 on AI4Boundaries); the DAv2 values in Table 2 are from the second pass. Our own decoders reproduced to four decimals.

Three baselines we could not run. Our experiment plan named three further methods. Each turned out to be unavailable in a form we could execute, so each was replaced by a reproducible method filling the same role. We report what we found rather than citing their published numbers, which use a different matching protocol and different splits.

Table 5: Replaced baselines. Checked 2026-08-03 and 2026-08-04.

Planned	Code	Weights	Replacement
TopoFP (Duan et al., 2026)	not released	not released	BsiNet
PolyBuild (Zhang et al., 2026b)	no official repository	none	PolyWorld
VectorLLM (Zhang et al., 2025)	repository empty (size 0)	placeholder file	HiSup

Why BsiNet appears on buildings. BsiNet is a cropland specialist and its WHU cell is a cross-category test, not a comparison; we ran it so that the building column has a second cropland model beside DAv2 and no empty cell to explain. It is run exactly as on the cropland domains, from the released Xinjiang GF-2 weights with connected-component instances.

AI4Boundaries breakdown. Table 6 expands the AI4Boundaries column of Table 1. The pattern worth noting is that our precision is below DAv2's while our recall and geometry are well above it. The vertex-guided row is a diagnostic variant on the same weights: it starts from the contour polygons and moves each vertex onto a nearby high-confidence predicted corner where one exists, leaving it in place where none does. It improves mean IoU and PoLiS slightly at a small cost in F1, and, like the contour path, its neighbours do not share an arc; we do not use it elsewhere in the paper.

Table 6: AI4Boundaries validation split, 351 images. Micro scores; mean IoU and PoLiS over matched pairs.

Method	P	R	F1	mean IoU	mean PoLiS	tp / fp / fn
SegEarth-OV3	0.3821	0.2458	0.2992	0.5229	7.740	1087 / 1758 / 3335
DAv2	**0.7811**	0.4754	0.5910	0.5920	4.470	2102 / **589** / 2320
Ours (Contour)	0.7649	0.5384	**0.6320**	0.7963	4.131	2381 / 732 / 2041
Ours (Vertex-guided)	0.7584	0.5339	0.6267	**0.7998**	**4.041**	2361 / 752 / 2061
Ours (Topo, graph only)	0.7037	0.5118	0.5926	0.7642	4.497	2263 / 953 / 2159
Ours (Topo, delivered)	0.6815	**0.5396**	0.6023	0.7776	4.496	**2386** / 1115 / **2036**

B STRUCTURAL MEASURES PER DOMAIN

Table 7 breaks the exchange of Section 4.5 down by domain, with the overlap, fallback and coverage figures of the delivered output.

Table 7: Structural measures for the topological path against the contour path on the same images; overlap and coverage are for the delivered output with the residual fallback on. while Uncov. and After res. report uncovered instances before and after the residual fallback. WHU is the mean over three seeds at 10 shots.

Domain	Topo shared	Contour shared	Ratio	Overlap	Fallback	Uncov.	After res.
AI4Boundaries	0.3175	0.0304	10.4×	0.0038	0.0399	0.1129	**0.0237**
PASTIS	0.1929	0.0122	**15.8**×	0.0035	0.0940	0.2129	**0.0317**
iFLYTEK	0.2869	0.0282	10.2×	0.0079	0.0741	0.1167	**0.0218**
WHU (10-shot)	0.0708	0.0048	14.9×	0.0012	0.0665	0.0781	**0.0133**

C GEOMETRIC LOSSES AND TRAINING SCHEDULE

Write $\hat{c}_p = \sigma(\hat{C}_p)$ for the predicted corner probability at pixel p, c_p for its Gaussian soft target, N_+ for the number of peak pixels, $m_p \in \{0,1\}$ for the vertex-centre mask with mean $\bar{m}$, and n_+, n_-, n for positive, negative and total pixel counts. The three geometric losses, and their combination under learned uncertainty weights s_k, are

$$\mathcal{L}_{\text{cor}} = \frac{1}{N_+}\sum_p \begin{cases} -(1-\hat{c}_p)^{\alpha}\log\hat{c}_p, & c_p = 1 \\ -(1-c_p)^{\beta}\hat{c}_p^{\alpha}\log(1-\hat{c}_p), & \text{otherwise} \end{cases} \qquad (\alpha = 2,\ \beta = 4),$$

$$\mathcal{L}_{\text{off}} = \frac{1}{|\Omega_s|}\sum_p \frac{m_p}{\bar{m}}\left\|\sigma(\hat{O}_p) - \tfrac{1}{2} - O_p\right\|_1, \qquad w_p = \tfrac{n_-}{n} \text{ if } E_p > \tfrac{1}{2},\ \tfrac{n_+}{n} \text{ otherwise},$$

$$\mathcal{L}_{\text{edge}} = -\frac{1}{|\Omega_s|}\sum_p w_p\big[E_p\log\hat{e}_p + (1-E_p)\log(1-\hat{e}_p)\big], \qquad \mathcal{L}_{\text{geo}} = \sum_k \left(e^{-s_k}\mathcal{L}_k + s_k\right). \quad (5)$$

Predictions on the Gaussian skirt of a true corner are penalised less than predictions far from any corner; dividing by $\bar{m}$ keeps the offset gradient independent of how many vertices an image contains; and the edge weights correct for boundary pixels being far rarer than background. Training runs at learning rate 8×10^{-5}, weight decay 10^{-4}, batch size 1, gradient clipping 1.0, full precision, seed 123, early stopping after two epochs without improvement; it stopped at epoch five and kept epoch three.

D SEARCH GRIDS

Two grids selected the cropland configuration used in the main text, and both ran on AI4Boundaries images that also appear in our results. A third grid selected the building configuration, and two diagnostic grids checked whether the cropland domains would have preferred a configuration of their own.

Cropland post-processing (AI4Boundaries, 351 images). A grid over confidence $\in \{0.20, \ldots, 0.60\}$, NMS IoU and min area $\in \{40, 80, 120, 200, 300, 400\}$, 192 configurations in total. The winner was confidence $= 0.6$, NMS IoU $= 0.2$, min area $= 400$ at contour micro F1 0.6320, and PASTIS and iFLYTEK reuse it unchanged.

Topological parameters (AI4Boundaries, 351 images). A grid of 26 configurations over the bridging radius and the residual fallback, Table 8. Bridging alone never cleared our delivery threshold of 5% uncovered instances; the residual fallback did, at the cost of the overlap discussed in Section 3.3. The configuration we report has bridging off.

Table 8: Bridging and residual fallback on AI4Boundaries. *bridges* is the mean number of straight-line bridges inserted per image.

Configuration	F1	shared	overlap	uncovered	after residual	bridges
bridge 0, residual off	0.5926	0.3260	1e-5	0.1129	0.1129	0
bridge 0, residual on	**0.6023**	0.3175	0.0038	0.1129	**0.0237**	0
bridge 8, residual on	0.5997	0.3208	0.0039	0.1042	0.0237	1.64
bridge 24, residual on	0.5776	0.3403	0.0030	0.0758	0.0214	8.86
bridge 24, residual off	0.5721	0.3448	1e-5	0.0758	0.0758	8.86

Building post-processing (WHU). Selected on a 50-image subset of the WHU split we report, in two rounds. The first varied confidence $\times$ min area over $\{0.15, 0.25, 0.35, 0.45, 0.60\} \times \{40, 100, 200, 400\}$ and gave 0.698–0.864 contour F1, peaking at confidence $= 0.6$, min area $= 100$. The second refined min area over $\{60, 80, 100, 120, 150\}$ and confidence up to 0.80, confirming the same optimum; F1 fell monotonically above confidence $= 0.6$, reaching 0.762 at 0.80. Confidence agrees with the cropland optimum; only the minimum area differs, by the order of magnitude that separates a building footprint from a parcel.

Cropland check (PASTIS and iFLYTEK, 40-image diagnostic subsets). We checked whether either transfer domain would have preferred its own configuration. On PASTIS, F1 peaked at confidence $= 0.6$ and was flat in min area from 300 to 800 (0.511–0.513), so the configuration in use is already at that domain's own optimum; as the zero-shot domain, PASTIS was off limits for tuning regardless. On iFLYTEK, confidence $= 0.5$, min area $= 300$ scored 0.687 against 0.681 for the configuration in use. Confirming it on all 351 images gave contour F1 0.6815 against 0.6711, a gain of 1.04 points, while the topological path fell from 0.6410 to 0.6378 and its uncovered rate rose from 0.1167 to 0.1706. We declined it: the two decoders disagree on its direction, and adopting it would replace one cropland configuration with three.

E FEW-SHOT DETAIL

Table 9: WHU building adaptation. Mean $\pm$ sample standard deviation over three seeds; the zero-shot row is a single run of the unchanged cropland model with the prompt set to `"building"`.

Annotated images	Contour F1	Topo F1	Topo shared	Uncov. after res.
0	0.0812	0.0800	0.0009	0.7543
10	0.8635 ± 0.0041	0.8173 ± 0.0045	0.0708 ± 0.0195	0.0133 ± 0.0020
50	0.8765 ± 0.0028	0.8362 ± 0.0085	0.0590 ± 0.0105	0.0215 ± 0.0042
100	0.8773 ± 0.0037	0.8370 ± 0.0111	0.0639 ± 0.0163	0.0259 ± 0.0039

Table 9 summarises the building adaptation curve and Table 10 gives the nine checkpoints behind it, all evaluated with the building post-processing configuration of Appendix D.

Table 10: Per-seed WHU results. Seeds are the annotation-subset draws.

Budget	Seed	Contour F1	Topo F1
10	20260802	0.8678	0.8165
10	20260803	0.8633	0.8133
10	20260804	0.8595	0.8222
50	20260802	0.8774	0.8355
50	20260803	0.8734	0.8280
50	20260804	0.8789	0.8451
100	20260802	0.8771	0.8298
100	20260803	0.8736	0.8314
100	20260804	0.8811	0.8498

Training configuration. The joint model trains at learning rate 8×10^{-5} with weight decay 10^{-4}, batch size 1, gradient clipping at 1.0, full precision, and seed 123, with early stopping after two epochs without validation improvement. It stopped at epoch five and kept epoch three. Detection and geometric losses are weighted equally. Few-shot building checkpoints reuse the same schedule, with the annotation subset drawn by the seed in Table 10.