UniPoly: Unified End-to-End pipeline for Raster2Vector Tasks

Anonymous authors
Paper under double-blind review

Abstract

Spatial vectorization provides structured representations for urban modeling, digital twins, indoor navigation, and Geographic Information Systems (GIS). Existing methods often rely on task-specific architectures or multi-stage pipelines, while evaluation protocols remain fragmented across spatial domains. We propose UniPoly, a unified end-to-end framework for raster-to-vector (R2V) spatial reconstruction that represents diverse spatial objects as polygons. UniPoly adopts a shared polygon-based reconstruction paradigm with configurable components and introduces Poly Relational Attention (PRA) to model interactions within and across polygons. We further establish a unified evaluation protocol for both multi-stage and end-to-end methods, assessing region accuracy, geometric complexity, boundary fidelity, and vector fidelity. Extensive experiments on four indoor and outdoor benchmarks for floorplan, building, farmland, and road reconstruction show that UniPoly achieves the best or second-best performance on 31 of 32 evaluated metrics, with eight metrics per benchmark, while maintaining competitive inference efficiency. These results demonstrate the effectiveness of UniPoly across diverse spatial domains. Code will be released upon acceptance.

1 Introduction

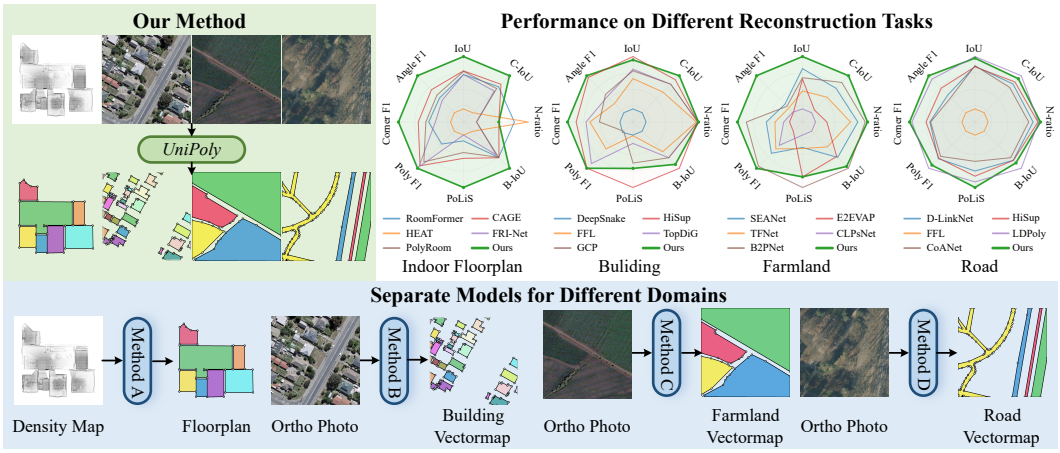


Figure 1: Existing approaches typically develop task-specific models for different spatial reconstruction domains, resulting in fragmented solutions. In contrast, UniPoly provides a unified end-to-end framework that directly reconstructs vector representations from heterogeneous raster inputs, achieving competitive performance across four representative domains.

Structured vector representations have become increasingly important for spatial understanding, supporting applications such as digital twins, urban modeling, geographic information systems, and map updating (Xiong et al., 2023; 2026; 2025; Meng et al., 2025; Jiao

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.

et al., 2026). Compared with raster masks, vector representations explicitly preserve object boundaries and geometric topology, making them more suitable for accurate geometric measurement and subsequent editing. This has driven spatial understanding research beyond pixel-level segmentation toward structured vector reconstruction.

Despite substantial progress, existing vectorization methods are largely application-specific. Indoor floorplan reconstruction (Yue et al., 2023; Sun et al., 2025b; Chen et al., 2023), building contour vectorization (Hu et al., 2024; Zhu et al., 2024b), farmland boundary vectorization (Cheng et al., 2020; Wu et al., 2025b), and road boundary vectorization (Hu et al., 2023a; Mei et al., 2021) are usually treated as independent research problems, each with its own datasets, network architectures, optimization objectives, and evaluation protocols. This naturally raises a fundamental question: Are these truly different reconstruction problems, or simply different manifestations of a common vector contour reconstruction problem?

We argue for the latter. Although boundaries of indoor rooms, buildings, farmland, and roads differ substantially in geometric characteristics, their reconstruction shares the same high-level objective: recovering structured representations from raster observations. Rather than predicting dense raster masks (Xu et al., 2024b; He et al., 2024; Lu et al., 2021), these vectorization tasks aim to extract geometric information from observations and reconstruct structured polygonal contours that faithfully describe spatial objects. This motivates us to investigate a unified reconstruction framework for diverse spatial vectorization tasks.

However, such unification is challenging due to substantial variations in spatial structures. Buildings and rooms tend to be compact and regular, while farmland may have highly irregular boundaries and roads often exhibit elongated and connected structures. A general framework, therefore, needs to capture both geometric consistency within individual polygons and contextual relationships among multiple polygons. Moreover, evaluation practices vary across applications, making results difficult to compare across tasks. More importantly, existing evaluations often focus on raster-level accuracy, providing limited assessment of the final vector representations.

To address these challenges, we propose UniPoly, a unified and end-to-end framework for raster-to-vector(R2V) spatial reconstruction. UniPoly establishes a shared polygon-based reconstruction paradigm while allowing the framework to adapt to different structural characteristics. Within this framework, we introduce Poly Relational Attention(PRA), which explicitly captures geometric dependencies within individual polygons and structural relationships across polygon instances, providing more effective relational modeling for complex outdoor multi-polygon scenes. We further establish a unified evaluation protocol that jointly considers raster fidelity and vector geometric accuracy. The protocol organizes the evaluation metrics into four complementary dimensions, namely region accuracy, geometric complexity, boundary fidelity, and vector fidelity, providing a more comprehensive assessment of vector reconstruction quality. Extensive experiments on representative benchmarks covering indoor floorplans, buildings, farmland, and roads show that UniPoly attains the best or second-best performance on the vast majority of evaluation metrics in every domain. In addition, UniPoly achieves competitive inference speed across different domains, demonstrating its favorable balance between reconstruction accuracy and computational efficiency. These results confirm its effectiveness across heterogeneous vector reconstruction tasks.

Our contributions are summarized as follows:

- We propose **UniPoly**, a unified and adaptable R2V reconstruction framework that employs a shared polygon-based paradigm with PRA for modeling intra- and inter-polygon relationships across diverse spatial domains.
- We establish a **unified evaluation protocol** that jointly considers raster fidelity and vector geometric accuracy, enabling comprehensive and consistent evaluation across heterogeneous spatial vectorization tasks.
- We conduct extensive experiments on representative benchmarks covering buildings, roads, farmlands, and indoor floorplans, demonstrating state-of-the-art or competitive performance across multiple spatial vectorization tasks with favorable computational efficiency.

2 Related Work

2.1 Spatial Vector Reconstruction

Spatial vector reconstruction aims to recover structured geometric representations from raster observations. Existing methods follow three representative reconstruction paradigms, namely segmentation-based vectorization, sequential vectorization, and direct vector reconstruction.

Segmentation-based vectorization. Early spatial vectorization methods rely on raster-based prediction, where semantic masks, boundary maps, or centerlines are first estimated and subsequently converted into vector representations through handcrafted or learning-based polygonization. Representative examples include building contour vectorization methods such as HiSup (Xu et al., 2023), PolygonCNN (Chen et al., 2020), SAMPoly-Build (Wang et al., 2024), PCP (Wang et al., 2025), road vectorization methods such as D-LinkNet (Zhou et al., 2018a), DeepRoadMapper (Mattyus et al., 2017), LDPoly (Jiao et al., 2025a), RoadCorrector (Li et al., 2024), farmland boundary vectorization methods include CT-HiffNet (Wu et al., 2025a), BSNet (Shunying et al., 2023), and CPVF (Liu et al., 2025b), as well as indoor floorplan reconstruction methods that estimate room masks, corner heatmaps, or junction maps prior to vectorization (Chen et al., 2019; Stekovic et al., 2021; Chen et al., 2022). Although these methods benefit from mature dense prediction frameworks, they rely on intermediate raster representations and heuristic vectorization, resulting in discretization errors and suboptimal geometric and topological consistency.

Sequential Vectorization. To explicitly model geometric structures, sequential vectorization methods progressively construct vector representations by iteratively predicting vertices, edges, or graph connections. Representative methods include Zhao et al. (2021); Ma et al. (2025) for buildings, as well as RoadTracer (Bastani et al., 2018), VecRoad (Tan et al., 2020), NETracer (Liu et al., 2025a), GLD-Road (Deng et al., 2025), and Curvi-Tracker (Heng et al., 2026) for road network reconstruction. Compared with segmentation-based pipelines, these methods directly optimize vector structures and better preserve topology. However, their autoregressive inference suffers from error accumulation, exposure bias, and limited inference efficiency, particularly for large-scale scenes.

Direct vector reconstruction. Recent advances leverage Transformer-based set prediction to directly generate structured vector primitives without relying on intermediate raster representations or iterative tracing. Representative approaches include PolyWorld (Zorzi et al., 2022), Re:PolyWorld (Zorzi & Fraundorfer, 2023), PolyR-CNN (Jiao et al., 2024), Line2Poly (Wei et al., 2024), RoIPoly (Jiao et al., 2025b) and Wei et al. (2023) for building reconstruction, Sat2Graph (He et al., 2020), PolyRoad (Hu et al., 2023b), AerialHDMapper (Xie et al., 2025) and DOGE (Sun et al., 2025a) for road extraction, and PLR-Net (Li et al., 2025) together with E2EVAP (Pan et al., 2023) for farmland boundary reconstruction, and RoomFormer (Yue et al., 2023), FRI-Net (Xu et al., 2024a),SLIBO-Net (Su et al., 2023) and CAGE (Liu et al., 2025c) for indoor floorplan reconstruction. These methods achieve more consistent end-to-end optimization and improved geometric representation. Nevertheless, they are typically designed for specific domains with task-dependent representations and objectives, limiting their generalization across heterogeneous spatial reconstruction tasks.

2.2 Unified Spatial Vector Reconstruction

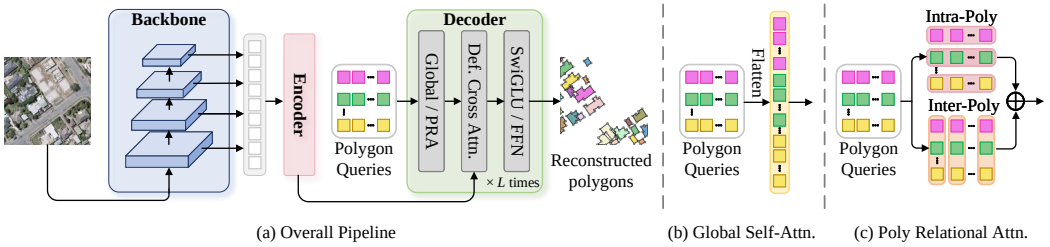
Recent efforts have attempted to unify different aspects of spatial vector reconstruction. Wang et al. (2023) reconstructs primitive graphs to jointly model building footprints and road networks. Yan et al. (2026b) unifies polygons, polylines, and line segments through instance–geometry interaction, while ACPV-Net (Jiao et al., 2026) extends polygon vectorization from single-class to all-class reconstruction with shared-topology constraints. SkySense V2 (Zhang et al., 2025) adopts a unified backbone for multimodal remote sensing understanding, and Yan et al. (2026a) reformulates heterogeneous vector mapping as structured language generation. These studies primarily unify the representation, category, or modeling paradigm of vector mapping, whereas we investigate a shared reconstruction framework across heterogeneous spatial applications.

3 METHOD

This section presents UniPoly, a unified framework for spatial vector reconstruction across diverse applications. UniPoly establishes a shared polygon-based reconstruction pipeline and provides configurable components to accommodate the characteristics of different reconstruction tasks.

3.1 SHARED POLYGON RECONSTRUCTION ARCHITECTURE

As illustrated in Fig.2, given an input image (e.g. satellite image, density map), the shared reconstruction pipeline first extracts hierarchical visual features using a Swin Transformer (Liu et al., 2022). Features from its four stages are denoted as $\{F_1, F_2, F_3, F_4\}$, corresponding to resolutions of 1/4, 1/8, 1/16, and 1/32 of the input size, respectively. These multi-scale features are projected into a common embedding space and processed by a Transformer encoder to obtain contextual representations. A Transformer decoder subsequently predicts a fixed set of polygons. Each polygon is represented by a group of edge queries, yielding the structured query representation $Q \in \mathbb{R}^{N_p \times N_e \times C}$, where N_p denotes the number of polygon hypotheses, N_e denotes the number of edge queries per polygon, and C is the embedding dimension. These grouped queries are jointly decoded into polygon edges to form the final vector representation.



(a) Overall Pipeline (b) Global Self-Attn. (c) Poly Relational Attn.

Figure 2: Overview of UniPoly. (a) The unified end-to-end R2V reconstruction pipeline, consisting of a multi-scale backbone, an encoder, and a query-based decoder for polygon reconstruction. (b) Global Self-Attention, which jointly models relations among all polygon queries. (c) PRA, which decomposes query interactions into intra-polygon and inter-polygon relations for more structured polygon modeling.

3.2 CONFIGURABLE DECODER DESIGN

Although the polygon representation and reconstruction pipeline can be shared, different tasks exhibit distinct geometric structures. Edges within a polygon are inherently coupled, whereas interactions across polygons provide contextual information but may also introduce irrelevant interference. A unified framework should therefore allow query interactions to adapt to the structural characteristics of different reconstruction tasks. We further allow the feature transformation module to be configured.

Global Self-Attention. In the global configuration, all edge queries are flattened into a single sequence and are jointly processed by self-attention. Thus, each query can directly interact with all other queries regardless of their polygon membership. This configuration provides unrestricted information exchange among the decoder queries.

PRA. While global self-attention enables unrestricted interactions among all queries, it can also introduce interference from visually similar but spatially unrelated regions in complex outdoor scenes. To preserve useful contextual interactions while suppressing such interference, we introduce PRA, which organizes query interactions according to polygon structure and models them along two complementary dimensions: intra- and inter-polygon.

Intra-polygon Attention performs self-attention along the edge dimension within each polygon group. Queries belonging to the same polygon exchange information to model internal geometric dependencies and maintain polygon-level coherence.

Inter-polygon Attention performs self-attention along the polygon dimension, enabling corresponding edge queries from different polygon hypotheses to exchange contextual information. It captures relationships among different polygon instances.

The two relational components are independently configurable and can be applied either individually or jointly. When both components are enabled, their outputs are aggregated to obtain the updated query representation.

Feature Transformation. The decoder FFN can also be configured for different reconstruction tasks. We use the standard FFN for indoor, building, and farmland reconstruction, while adopting a SwiGLU FFN for road reconstruction (Shazeer, 2020). The gated formulation provides an additional nonlinear feature transformation while preserving the same decoder interface and polygon-query representation.

3.3 Geometry-aware Multi-level Optimization

Following Liu et al. (2025c), UniPoly adopts a geometry-aware optimization strategy that jointly considers edge geometry, edge validity, and region consistency. The overall objective is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{edge} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{region}, \tag{1}$$

where $\mathcal{L}_{edge}$ supervises the coordinates of predicted polygon edges, $\mathcal{L}_{cls}$ determines the validity of edges, and $\mathcal{L}_{region}$ enforces region-level consistency through raster supervision.

While the edge-level objectives provide direct geometric constraints on polygon primitives, they do not fully capture region-level consistency. To improve region supervision, UniPoly incorporates differentiable rasterization (Lazarow et al., 2022). We further observe that the effectiveness of raster supervision is highly influenced by rasterization granularity. Low-resolution rasterization introduces quantization errors and weakens boundary gradients, particularly for thin structures and complex geometries. Therefore, UniPoly employs fine-grained raster supervision with higher rasterization resolution, providing more accurate region-level optimization signals.

To accommodate tasks with different optimization characteristics, UniPoly employs Adaptive Loss Balancing (ALB) to learn the relative contributions of different supervision terms during training. In our road reconstruction setting, ALB dynamically adjusts the weights of these supervision terms according to their training dynamics. The motivation, detailed formulation and implementation are provided in Appendix A.2.

4 Unified Evaluation Protocol

Spatial vector reconstruction aims not only to recover object regions but also to preserve accurate geometric structures. As reconstruction methods have moved from raster representations toward direct vector prediction, evaluation has moved beyond conventional pixel-level metrics to capture geometric accuracy and structural consistency. However, vector-specific metrics are often designed independently for individual applications, leading to fragmented evaluation protocols and making comparisons across different reconstruction tasks difficult.

To enable a consistent and comprehensive comparison, we establish a unified evaluation protocol that retains raster-based metrics for compatibility with existing methods and complements them with vector-specific measures. The eight metrics are organized into four reconstruction perspectives: **Region Accuracy**, **Geometric Complexity**, **Boundary Fidelity**, and **Vector Fidelity**. Detailed metric formulations are provided in Appendix A.1, and the corresponding computation procedures are illustrated in Fig. 3.

Region Accuracy evaluates whether reconstructed polygons correctly recover object regions. We adopt **IoU** to measure the overlap between rasterized predictions and ground-truth regions, providing an overall assessment of region reconstruction accuracy.

Geometric Complexity evaluates whether reconstructed polygons maintain an appropriate level of geometric complexity while preserving reconstruction accuracy. We employ **C-IoU**, which jointly considers region overlap and vertex-count efficiency by penalizing

the relative deviation in the number of vertices, encouraging compact yet accurate polygon reconstruction (Zorzi et al., 2022). In addition, **N-ratio** directly compares the number of predicted vertices with the ground truth, reflecting whether reconstructed polygons are over-simplified or contain redundant vertices (Zorzi et al., 2022).

Boundary Fidelity evaluates the geometric fidelity of reconstructed contours. We employ **Boundary IoU(B-IoU)** to measure boundary overlap (Cheng et al., 2021), **PoLiS** to quantify boundary displacement (Avbelj et al., 2015).

Vector Fidelity evaluates the fidelity of the reconstructed vector representations. Specifically, **Polygon F1**, **Corner F1**, and **Angle F1** assess polygon detection, vertex localization, and local geometric consistency, respectively (Chen et al., 2022).

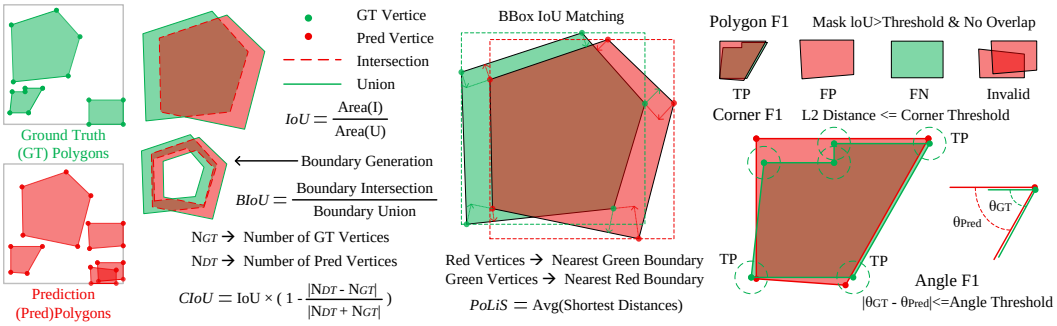


Figure 3: Illustration of the proposed unified evaluation metrics. The figure shows the computation of IoU, B-IoU, C-IoU, N-ratio, Polygon F1, Corner F1 and Angle F1 based on ground-truth and predicted polygons.

5 Experiments

5.1 Datasets and Implementation Details

Datasets. To evaluate the generality of UniPoly across diverse spatial vector reconstruction scenarios, we conduct experiments on four datasets covering indoor floorplan, building, farmland and road. These datasets exhibit diverse polygon structures and geometric characteristics, ranging from multiple adjacent polygons in indoor scenes to thin and elongated structures in road scenes, providing a comprehensive benchmark for evaluating unified vector reconstruction. The detailed statistics of these datasets are summarized in Table 1. We report the average number of polygons and vertices per scene to characterize the structural complexity of each dataset, together with polygon relations and geometric characteristics.

Table 1: Statistics of datasets with different geometric characteristics.

Dataset	Domain	Input	Resolution	Avg. #Polygons	Avg. #Vertices	Polygon Relation	Geometric Characteristics
Structured3D (Zheng et al., 2020)	Indoor	Density map	256×256	6.7	5.6	Strong	Multiple adjacent polygons
WHU-Building (Ji et al., 2019)	Building	Ortho photo	512×512	14.3	7.6	Weak	Irregular man-made boundaries
iFLYTEK (Zhao et al., 2022)	Farmland	Ortho photo	512×512	11.5	8.6	Moderate	Large irregular regions
Deventer (Jiao et al., 2025a)	Road	Ortho photo	256×256	2.6	24.8	Weak	Thin and elongated structures

Implementation Details. UniPoly is implemented in PyTorch and trained on a single NVIDIA RTX 5090 GPU with 32 GB of memory. We use the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 2×10^{-4}. Although the overall UniPoly framework remains unchanged, the configuration of its components can vary across datasets due to differences in geometric complexity and optimization behavior. Accordingly, Structured3D uses global attention, as its indoor floorplans exhibit relatively regular polygon layouts with well-defined spatial relationships, whereas the outdoor datasets use PRA to provide explicit polygon-level interactions for their more irregular polygon layouts. Standard FFN is used for most datasets, while SwiGLU is adopted for Deventer. Fixed loss weights are used in most cases, whereas adaptive loss weighting is adopted for Deventer to better balance

the reconstruction objectives. Label smoothing is applied to iFLYTEK. Detailed ablation studies of these configurable components are provided in Appendix A.3. Detailed training settings are summarized in Tab. 2.

Table 2: Training configurations used by UniPoly.

Dataset	Attn.	FFN	Epochs	Warmup (epochs)	LR Schedule	Final LR	Loss Weighting	Loss Weights			Queries	Other
								Loss_{cls}	Loss_{edge}	Loss_{region}	($N_{\text{poly}} \times N_{\text{edge}}$)	
Structured3D	Global	Standard	650	–	Step Decay	2×10^{-5}	Fixed	0.6	6.0	1.0	20×40	–
WHU-Building	PRA	Standard	350	10	Cosine Decay	1×10^{-6}	Fixed	0.6	6.0	1.0	40×30	–
iFLYTEK	PRA	Standard	320	10	Cosine Decay	1×10^{-6}	Fixed	2.0	5.0	1.0	40×30	Label Smoothing
Deventer	PRA	SwiGLU	650	10	Cosine Decay	2×10^{-6}	ALB	0.6	6.0	1.0	8×150	–

5.2 Comparison with State-of-the-art Methods

Table 3: Comparison with state-of-the-art methods under the unified evaluation protocol for spatial vector reconstruction. Best and second-best results are highlighted in **bold** and underlined, respectively.

Dataset	Methods	Venue	Params(M)	t(s)	Region Accuracy	Geometric Complexity		Boundary Fidelity		Vector Fidelity		
					IoU↑	C-IoU↑	$\lvert N\text{-ratio}-1\rvert$↓	B-IoU↑	PoLiS↓	Poly F1↑	Corner F1↑	Angle F1↑
Structured3D (Indoor floorplan)	HEAT	CVPR'22	48.9	0.11	85.8	79.7	**0.02**	39.4	1.7	96.0	44.9	42.9
	RoomFormer	CVPR'23	40.6	**0.01**	92.3	88.2	0.04	71.6	1.6	97.0	61.9	59.2
	FRI-Net	ECCV'24	58.6	0.09	92.3	86.8	0.09	68.7	1.7	98.8	56.7	56.4
	PolyRoom	ECCV'24	42.2	0.02	92.9	87.6	0.09	71.6	1.4	98.4	64.1	62.2
	CAGE	NeurIPS'25	211.3	**0.01**	93.0	89.0	0.06	72.1	1.3	98.7	70.0	68.2
	Ours	-	210.9	**0.01**	**95.7**	**91.8**	0.06	**85.1**	**0.8**	**98.9**	**85.4**	**83.6**
Whu-Building (Outdoor building)	DeepSnake	CVPR'20	41.8	**0.03**	71.4	3.4	67.80	64.1	2.4	42.2	2.9	0.5
	FFL	CVPR'21	76.6	0.04	80.9	46.1	2.68	74.6	2.3	64.4	45.9	12.0
	HiSup	ISPRS'23	74.3	**0.03**	**87.8**	71.7	1.10	**83.1**	**1.5**	90.3	65.5	44.7
	TopDiG	CVPR'23	41.0	0.08	83.8	62.3	0.98	77.7	2.2	84.2	51.4	20.9
	GCP	TGRS'25	50.6	0.05	83.3	62.3	0.96	77.7	1.9	40.6	32.3	23.4
	Ours	-	211.1	**0.03**	86.7	**79.9**	**0.05**	81.1	1.8	**90.5**	**77.8**	**46.8**
iFLYTEK (Outdoor farmland)	SEANet	ISPRS'23	28.8	**0.01**	68.1	46.0	1.84	46.2	11.4	39.6	16.7	3.1
	TFNet	IGARSS'25	1.6	0.02	60.5	36.0	3.22	39.5	11.2	34.2	12.0	2.0
	CLPsNet	TGRS'24	49.1	**0.01**	54.4	18.7	9.80	29.4	13.1	29.9	7.2	0.8
	E2EVAP	ISPRS'23	29.8	0.02	64.5	21.8	7.67	45.6	7.2	12.5	4.0	0.6
	B2PNet	ISPRS'26	57.9	**0.01**	65.0	52.5	0.45	**53.2**	**5.7**	55.6	27.6	2.5
	Ours	-	211.1	0.03	**72.2**	**60.0**	**0.02**	52.0	7.2	**58.0**	**32.5**	**11.5**
Deventer (Outdoor road)	D-LinkNet	CVPRW'18	31.1	0.07	71.7	54.5	0.68	63.0	3.3	65.4	35.2	9.6
	CoANet	TIP'21	59.1	0.02	71.9	51.4	1.08	61.6	3.9	61.5	34.7	8.5
	FFL	CVPR'21	76.6	0.04	57.1	28.4	4.78	47.1	5.5	20.0	15.7	2.9
	HiSup	ISPRS'23	74.3	0.03	71.5	56.3	0.24	63.1	3.0	64.6	37.5	12.1
	LDPoly	ISPRS'25	296.0	0.04	**75.0**	**62.0**	0.18	**67.4**	2.6	**81.2**	46.1	15.5
	Ours	-	210.9	**0.01**	74.5	59.3	**0.01**	64.5	**2.3**	75.9	**47.5**	**16.8**

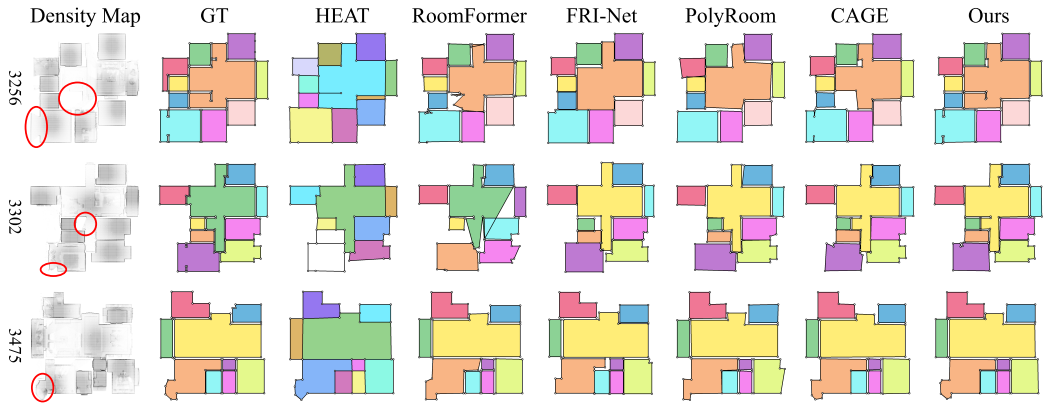


Figure 4: Qualitative comparison on Structured3D. The three rows illustrate challenging cases involving (a) floorplans with sparse point observations caused by occlusions, (b) diagonal room boundaries, and (c) regions with incomplete observations due to occlusions.

Indoor floorplan reconstruction. Indoor floorplans contain multiple adjacent polygons with strict topological relationships, requiring accurate region partitioning and boundary alignment. As shown in Table 3, UniPoly achieves the best overall performance on this task, improving IoU from 93.0% to 95.7% compared with CAGE (Liu et al., 2025c). Beyond region accuracy, UniPoly substantially improves B-IoU from 72.1% to 85.1% and reduces PoLiS

from 1.3 to 0.8. As shown in Figure 4, UniPoly better recovers complete room boundaries under sparse and incomplete observations, including occluded and diagonal structures. Note that, on this benchmark, we adopt a stricter corner-matching threshold for **Corner F1** than prior indoor floorplan works: a predicted corner is regarded as correctly localized only when it lies within 2 pixels of a GT corner, whereas earlier studies commonly allow a matching distance of 10 pixels (Yue et al., 2023; Chen et al., 2022; Xu et al., 2024a; Liu et al., 2025c). The angular-deviation threshold used for **Angle F1** is unchanged from those works. All methods in Table 3 are evaluated under this unified criterion on the indoor benchmark.

Building Contour Vectorization. Building contours contain diverse and irregular man-made boundaries. Compared with HiSup, UniPoly achieves comparable region accuracy while substantially improving polygon-level fidelity (Xu et al., 2023), increasing Corner F1 from 65.5% to 77.8% and Angle F1 from 44.7% to 46.8%. Figure 5 further shows that UniPoly produces more compact and accurately localized polygons under visually confusing, densely distributed, and partially occluded buildings.

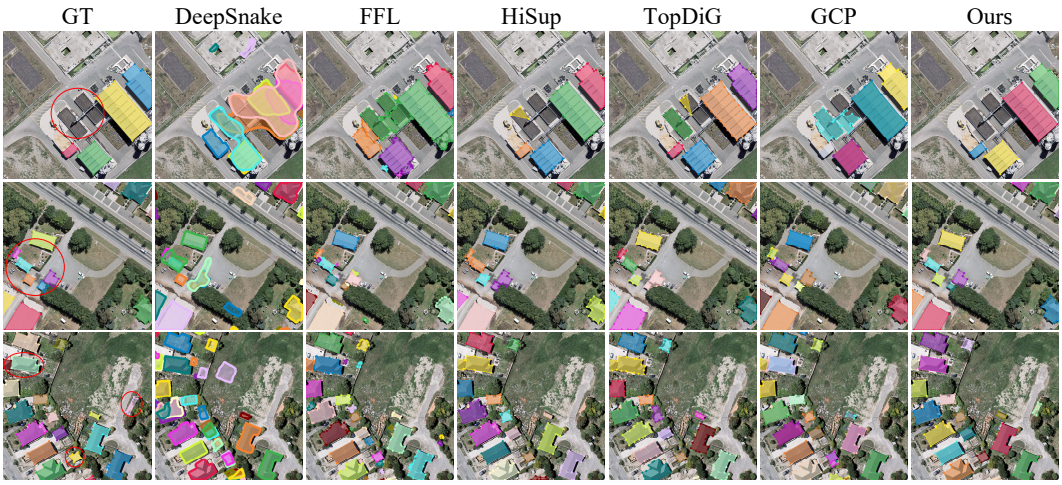


Figure 5: Qualitative comparison on WHU-Building. The three rows illustrate challenging cases involving (a) non-building objects with similar visual appearances to buildings, (b) buildings with fine-grained contour structures, and (c) densely distributed and partially occluded buildings.

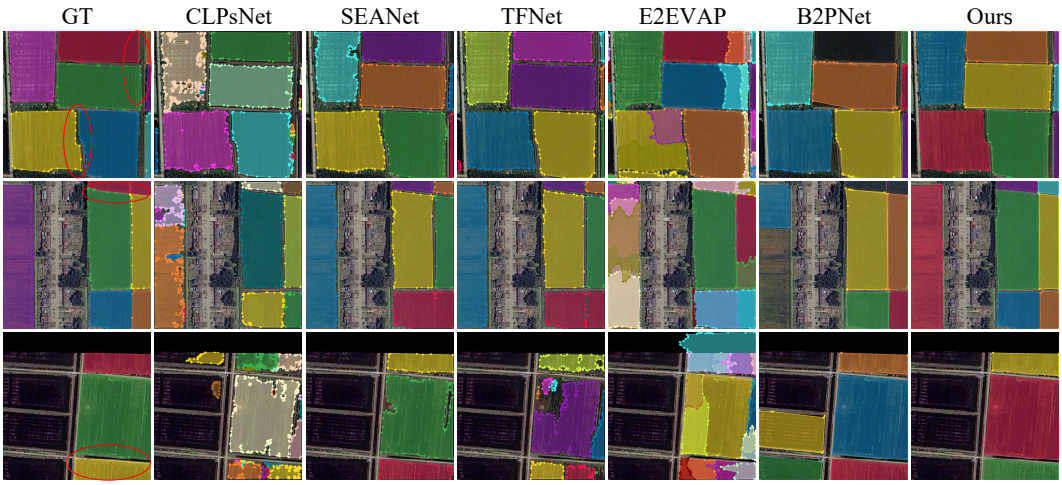


Figure 6: Qualitative comparison on iFLYTEK. The three rows illustrate cases involving (a) adjacent parcels with gaps and dissimilar boundaries, (b) closely distributed parcels, and (c) regularly distributed parcels with similar appearances.

Farmland Boundary Vectorization. Farmland contains large-scale and highly irregular polygon structures, making precise boundary reconstruction particularly challenging. Compared with B2PNet, UniPoly improves Corner F1 from 27.6% to 32.5% and Angle F1 from 2.5% to 11.5%. Figure 6 presents qualitative comparisons on iFLYTEK. UniPoly better preserves highly concave parcel boundaries and distinguishes closely distributed or visually similar parcels with compact polygon representations.

Road Boundary Vectorization. Roads are characterized by thin, elongated, and connected structures. Although LDPoly achieves slightly higher region accuracy, UniPoly obtains better vector-level reconstruction quality, improving Corner F1 and Angle F1 to 47.5% and 16.8%, respectively. Figure 7 shows that existing methods often produce redundant vertices and local boundary deviations around complex junctions, curves, and narrow road segments. In contrast, UniPoly generates more compact polygons with better-preserved connectivity and boundary geometry.

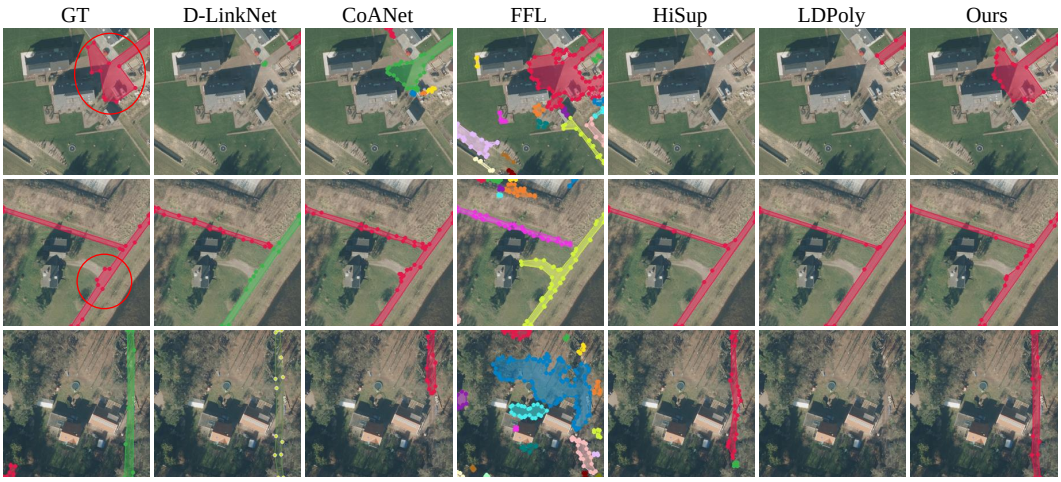


Figure 7: Qualitative comparison on Deventer. The three rows illustrate challenging cases involving (a) complex road junctions with multiple directional changes, (b) branching road structures, and (c) narrow roads with strong background interference and partial occlusion.

Across the four benchmarks, UniPoly consistently balances reconstruction accuracy and computational efficiency, while existing methods excel only under specific spatial structures or at the expense of efficiency. HiSup performs strongly on building reconstruction but is less effective for roads. Moreover, unlike segmentation-based methods such as LDPoly and HiSup that rely on parameter-sensitive post-processing, UniPoly provides an end-to-end reconstruction pipeline, leading to more robust polygon reconstruction across diverse spatial scenarios.

6 Conclusion

We presented UniPoly for unified R2V reconstruction across indoor and outdoor spatial scenes. UniPoly adopts a shared polygon-based reconstruction paradigm with configurable components, together with a unified evaluation protocol for assessing vector reconstruction quality. Experiments on four benchmarks covering floorplan, building, farmland, and road reconstruction show that UniPoly achieves the best or second-best performance on 31 of 32 evaluated metrics while maintaining competitive inference efficiency.

Despite these results, several directions remain to be explored. Future work will investigate joint training across multiple datasets to better exploit shared geometric knowledge, and extend the polygon representation to support holes and more complex topological structures. These extensions could further broaden the applicability of unified R2V reconstruction to more diverse real-world spatial scenes.

AI USE STATEMENT

In this work, we used generative AI tools to aid writing and polish the manuscript, including improving grammar, language clarity, readability, and presentation. It is important to note that the generative AI tools was not involved in the ideation, research methodology or experimental design. The authors independently developed, implemented, and validated the final methodology and experiments, and all experimental results were produced and verified by the authors. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of the paper, including all scientific claims, methodology, experimental results, analysis, and conclusions.

REFERENCES

Janja Avbelj, Rupert Müller, and Richard Bamler. A Metric for Polygon Comparison and Building Extraction Evaluation. *IEEE Geoscience and Remote Sensing Letters*, 12(1): 170–174, 2015. ISSN 1558-0571. doi: 10.1109/LGRS.2014.2330695.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4720–4728, Salt Lake City, UT, 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00496.

Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2661–2670, 2019. doi: 10.1109/ICCV.2019.00275.

Jiacheng Chen, Yiming Qian, and Yasutaka Furukawa. Heat: Holistic edge attention transformer for structured reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16448–16457, 2022. doi: 10.1109/CVPR52688.2022.01601.

Jiacheng Chen, Ruizhi Deng, and Yasutaka Furukawa. Polydiffuse: Polygonal shape reconstruction via guided set diffusion models. *Advances in Neural Information Processing Systems*, 36:1863–1888, 2023.

Qi Chen, Lei Wang, Steven L. Waslander, and Xiuguo Liu. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170:114–126, 2020. ISSN 09242716. doi: 10.1016/j.isprsjprs.2020.10.008.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation, 2021.

Tao Cheng, Xusheng Ji, Gaoxiang Yang, Hengbiao Zheng, Jifeng Ma, Xia Yao, Yan Zhu, and Weixing Cao. DESTIN: A new method for delineating the boundaries of crop fields by fusing spatial and temporal information from WorldView and Planet satellite imagery. *Computers and Electronics in Agriculture*, 178:105787, 2020. ISSN 01681699. doi: 10.1016/j.compag.2020.105787.

Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, Salt Lake City, UT, USA, 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00781.

Ligao Deng, Yupeng Deng, Yu Meng, Jingbo Chen, Zhihao Xi, Diyou Liu, and Qifeng Chu. GLD-Road:A global-local decoding road network extraction model for remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:741–755, 2025. ISSN 09242716. doi: 10.1016/j.isprsjprs.2025.07.026.

David H. Douglas and Thomas K. Peucker. *Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature*, chapter 2, pp. 15–28. John Wiley & Sons, Ltd, 2011. ISBN 9780470669488. doi: https://doi.org/10.1002/9780470669488.ch2.

Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5887–5896, 2021. doi: 10.1109/CVPR46437.2021.00583.

Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M. Elshrif, Samuel Madden, and Amin Sadeghi. Sat2Graph: Road Graph Extraction through Graph-Tensor Encoding, 2020.

Wei He, Jiepan Li, Weinan Cao, Liangpei Zhang, and Hongyan Zhang. Building Extraction from Remote Sensing Images via an Uncertainty-Aware Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2024.3361211.

Zhan Heng, Maurice Pagnucco, Erik Meijering, and Yang Song. Curvi-Tracker: Curvilinear structure segmentation refinement by iterative tracking. *Pattern Recognition*, 173:112797, 2026. ISSN 00313203. doi: 10.1016/j.patcog.2025.112797.

Anna Hu, Liang Wu, Yongyang Xu, and Zhong Xie. SANET: A Shape-Aware Building Footprints Extraction Method in Remote Sensing Images by Integrating Fourier Shape Descriptors. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3425708.

Jingtao Hu, Junyu Gao, Yuan Yuan, Jocelyn Chanussot, and Qi Wang. LGNet: Location-Guided Network for Road Extraction From Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023a. ISSN 1558-0644. doi: 10.1109/TGRS.2023.3305031.

Yuan Hu, Zhibin Wang, Zhou Huang, and Yu Liu. Polyroad: polyline transformer for topological road-boundary detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2023b.

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. ISSN 1558-0644. doi: 10.1109/TGRS.2018.2858817.

Weiqin Jiao, Claudio Persello, and George Vosselman. Polyr-cnn: R-cnn for end-to-end polygonal building outline extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:33–43, 2024. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2024.10.006.

Weiqin Jiao, Hao Cheng, George Vosselman, and Claudio Persello. LDPoly: Latent diffusion for polygonal road outline extraction in large-scale topographic mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:820–842, 2025a. ISSN 09242716. doi: 10.1016/j.isprsjprs.2025.10.005.

Weiqin Jiao, Hao Cheng, George Vosselman, and Claudio Persello. RoIPoly: Vectorized building outline extraction using vertex and logit embeddings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:317–328, 2025b. ISSN 09242716. doi: 10.1016/j.isprsjprs.2025.03.030.

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)*, pp. 13244–13253, June 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Justin Lazarow, Weijian Xu, and Zhuowen Tu. Instance segmentation with mask-supervised polygonal boundary transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4372–4381, New Orleans, LA, USA, 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00434.

Jinpeng Li, Jun He, Weijia Li, Jiabin Chen, and Jinhua Yu. RoadCorrector: A Structure-Aware Road Extraction Method for Road Connectivity and Topology Correction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3380914.

Mengmeng Li, Jiang Long, Alfred Stein, and Xiaoqin Wang. Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 200:24–40, 2023. ISSN 09242716. doi: 10.1016/j.isprsjprs.2023.04.019.

Mengmeng Li, Chengwen Lu, Mengjing Lin, Xiaolong Xiu, Jiang Long, and Xiaoqin Wang. Extracting vectorized agricultural parcels from high-resolution satellite images using a point-line-region interactive multitask model. *Computers and Electronics in Agriculture*, 231:109953, 2025. ISSN 01681699. doi: 10.1016/j.compag.2025.109953.

Chao Liu, Yangbo Jiang, and Nenggan Zheng. NETracer: A Topology-Aware Iterative Tracing Approach for Tubular Structure Extraction. *In ICCV*, 2025a.

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, 2025b. ISSN 09242716. doi: 10.1016/j.isprsjprs.2025.05.019.

Yiyi Liu, Chunyang Liu, Bohan Wang, Weiqin Jiao, Bojian Wu, Lubin Fan, Yuwei Chen, Fashuai Li, and Biao Xiong. Cage: Continuity-aware edge network unlocks robust floorplan reconstruction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132460–132485, 2025c.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

Jiang Long, Hang Zhao, Mengmeng Li, Xiaoqin Wang, and Chengwen Lu. Integrating segment anything model derived boundary prior and high-level semantics for cropland extraction from high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. ISSN 1558-0571. doi: 10.1109/LGRS.2024.3454263.

Xiaoyan Lu, Yanfei Zhong, Zhuo Zheng, and Junjue Wang. Cross-domain road detection based on global-local adversarial learning framework from very high resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180:296–312, 2021. ISSN 09242716. doi: 10.1016/j.isprsjprs.2021.08.018.

Xiaodong Ma, Lingjie Zhu, Yuzhou Liu, Zexiao Xie, Xiang Gao, and Shuhan Shen. BPN: Building Pointer Network for Satellite Imagery Building Contour Extraction. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025. ISSN 1558-0571. doi: 10.1109/LGRS.2024.3514109.

Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. DeepRoadMapper: Extracting Road Topology from Aerial Images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3458–3466, Venice, 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.372.

Jie Mei, Rou-Jing Li, Wang Gao, and Ming-Ming Cheng. Coanet: Connectivity attention network for road extraction from satellite imagery. *IEEE Transactions on Image Processing*, 30:8540–8552, 2021. doi: 10.1109/TIP.2021.3117076.

Yu Meng, Ligao Deng, Zhihao Xi, Jiansheng Chen, Jingbo Chen, Anzhi Yue, Diyou Liu, Kai Li, Chenhao Wang, Kaiyu Li, Yupeng Deng, and Xian Sun. IRSAMap: Toward Large-Scale, High-Resolution Land Cover Map Vectorization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. ISSN 1558-0644. doi: 10.1109/TGRS.2025.3600249.

Yang Pan, Xinyu Wang, Liangpei Zhang, and Yanfei Zhong. E2EVAP: End-to-end vectorization of smallholder agricultural parcel boundaries from high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203:246–264, 2023. ISSN 09242716. doi: 10.1016/j.isprsjprs.2023.08.001.

Noam Shazeer. Glu variants improve transformer, 2020. URL `https://arxiv.org/abs/2002.05202`.

Wang Shunying, Zhou Ya'nan, Yang Xianzeng, Feng Li, Wu Tianjun, and Luo Jiancheng. BSNet: Boundary-semantic-fusion network for farmland parcel mapping in high-resolution satellite images. *Computers and Electronics in Agriculture*, 206:107683, 2023. ISSN 01681699. doi: 10.1016/j.compag.2023.107683.

Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. Montefloor: Extending mcts for reconstructing accurate large-scale floor plans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16014–16023, 2021. doi: 10.1109/ICCV48922.2021.01574.

Jheng-Wei Su, Kuei-Yu Tung, Chi-Han Peng, Peter Wonka, and Hung-Kuo James Chu. Slibo-net: Floorplan reconstruction via slicing box representation with local geometry regularization. *Advances in Neural Information Processing Systems*, 36:48781–48792, 2023.

Jiahui Sun, Junran Lu, Jinhui Yin, Yishuo Xu, Yuanqi Li, and Yanwen Guo. DOGE: Differentiable Bezier Graph Optimization for Road Network Extraction, 2025a.

Qian Sun, Chenrong Fang, Shuang Liu, Yidan Sun, Yu Shang, and Ying He. Polygraph: a graph-based method for floorplan reconstruction from 3d scans. *IEEE Transactions on Visualization and Computer Graphics*, 2025b.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8907–8915, Seattle, WA, USA, 2020. IEEE. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00893.

Chenhao Wang, Jingbo Chen, Yu Meng, Yupeng Deng, Kai Li, and Yunlong Kong. SAM-PolyBuild: Adapting the Segment Anything Model for polygonal building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:707–720, 2024. ISSN 09242716. doi: 10.1016/j.isprsjprs.2024.09.018.

Chenhao Wang, Zhihao Xi, Diyou Liu, Yuman Feng, Yupeng Deng, Kai Li, Jingbo Chen, Jiansheng Chen, and Yu Meng. PCP: A Prompt-Based Cartographic-Level Polygonal Vector Extraction Framework for Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. ISSN 1558-0644. doi: 10.1109/TGRS.2025.3614872.

Lei Wang, Min Dai, Jianan He, and Jingwei Huang. Regularized primitive graph learning for unified vector mapping. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16771–16780, Paris, France, 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01542.

Shiqing Wei, Tao Zhang, Shunping Ji, Muying Luo, and Jianya Gong. BuildMapper: A fully learnable framework for vectorized building contour extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:87–104, 2023. ISSN 09242716. doi: 10.1016/j.isprsjprs.2023.01.015.

Shiqing Wei, Tao Zhang, Dawen Yu, Shunping Ji, Yongjun Zhang, and Jianya Gong. From lines to Polygons: Polygonal building contour extraction from High-Resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209:213–232, 2024. ISSN 09242716. doi: 10.1016/j.isprsjprs.2024.02.001.

Hao Wu, Junyang Xie, Weihao Deng, Anqi Lin, Abdul Rashid Mohamed Shariff, Shamshodbek Akmalov, Wenbin Wu, Zhaoliang Li, Qiangyi Yu, Qunming Wang, Jian Zhang, Xin Mei, and Qiong Hu. CT-HiffNet: A contour-texture hierarchical feature fusion network for cropland field parcel extraction from high-resolution remote sensing images. *Computers and Electronics in Agriculture*, 239:111010, 2025a. ISSN 01681699. doi: 10.1016/j.compag.2025.111010.

Wei Wu, Shiyu Li, Haiping Yang, Yingpin Yang, Kun Li, Liao Yang, and Zuohui Chen. EEDNet: Edge and Edge Direction Network for simple and regular land parcel vectorization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229:223–238, 2025b. ISSN 09242716. doi: 10.1016/j.isprsjprs.2025.08.008.

Haofeng Xie, Xiangyun Hu, Huiwei Jiang, Hengming Dai, Pengwei Dong, and Pengfei Wang. AerialHDMapper: High-Definition Map Construction From Aerial Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–12, 2025. ISSN 1558-0644. doi: 10.1109/TGRS.2025.3540439.

Biao Xiong, Yusheng Jin, Fashuai Li, Yuwei Chen, Yiquan Zou, and Zhize Zhou. Knowledge-driven inference for automatic reconstruction of indoor detailed as-built bims from laser scanning data. *Automation in Construction*, 156:pp. 105097, 2023. doi: 10.1016/j.autcon.2023.105097.

Biao Xiong, LongJun Zhang, Ruiqi Huang, Junwei Zhou, Syed Riaz un Nabi Jafri, Bojian Wu, and Fashuai Li. Vf-plan: Bridging the art gallery problem and static lidar scanning with visibility field optimization. In *Pacific Graphics Conference Papers, Posters, and Demos*. The Eurographics Association, 2025. ISBN 978-3-03868-295-0. doi: 10.2312/pg.20251289.

Biao Xiong, Bohan Wang, Yiyi Liu, Liangliang Wang, Yanchao Yang, Liang Zhou, and Qiegen Liu. Roof-aware indoor bim reconstruction from lidar via graph-attention for residential buildings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:408–420, 2026. doi: 10.1016/j.isprsjprs.2025.12.024.

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, 2023. ISSN 09242716. doi: 10.1016/j.isprsjprs.2023.03.006.

Honghao Xu, Juzhan Xu, Zeyu Huang, Pengfei Xu, Hui Huang, and Ruizhen Hu. Fri-net: Floorplan reconstruction via room-wise implicit representation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024a.

Yongyang Xu, Zhihao Zhu, Mingqiang Guo, and Ying Huang. Multiscale edge-guided network for accurate cultivated land parcel boundary extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024b. doi: 10.1109/TGRS.2023.3344122.

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, 2026a.

Yinglong Yan, Jun Yue, Shaobo Xia, Hanmeng Sun, Tianxu Ying, Chengcheng Wu, Sifan Lan, Min He, Pedram Ghamisi, and Leyuan Fang. UniVector: Unified vector extraction via instance-geometry interaction. *Fundamental Research*, pp. S2667325826002505, 2026b. ISSN 26673258. doi: 10.1016/j.fmre.2026.03.025.

Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. Connecting the dots: Floorplan reconstruction using two-level queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 845–854, 2023.

Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. Skysense v2: A unified foundation model for multi-modal remote sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9136–9146, October 2025.

Wufan Zhao, Claudio Persello, and Alfred Stein. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:119–131, 2021. ISSN 09242716. doi: 10.1016/j.isprsjprs.2021.02.014.

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 *2022 14th International Conference on Advanced Computational Intelligence (ICACI)*, pp. 376–380, 2022. doi: 10.1109/ICACI55529.2022.9837765.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535, 2020. ISBN 3-030-58544-1.

Lichen Zhou, Chuang Zhang, and Ming Wu. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 192–1924, 2018a. doi: 10.1109/CVPRW.2018.00034.

Lichen Zhou, Chuang Zhang, and Ming Wu. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 192–1924, 2018b. doi: 10.1109/CVPRW.2018.00034.

Yu Zhu, Yaozhong Pan, Dujuan Zhang, Hanyi Wu, and Chuanwu Zhao. A deep learning method for cultivated land parcels' (CLPs) delineation from high-resolution remote sensing images with high-generalization capability. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–25, 2024a. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3425673.

Yunhui Zhu, Buliao Huang, Yizhan Fan, Muhammad Usman, and Huanhuan Chen. Iterative Polygon Deformation for Building Extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024b. ISSN 1558-0644. doi: 10.1109/TGRS.2024.3396813.

Stefano Zorzi and Friedrich Fraundorfer. Re:PolyWorld - A Graph Neural Network for Polygonal Scene Parsing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16716–16725, Paris, France, 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01537.

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1938–1947, New Orleans, LA, USA, 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00189.

A Appendix

A.1 Unified Evaluation Protocol

This section provides detailed definitions of the metrics used in our unified evaluation protocol. Given a predicted polygon set and a ground-truth polygon set, all metrics are calculated between matched polygon pairs and averaged over the test set. The evaluation protocol consists of four aspects: Region Accuracy, Geometric Complexity, Boundary Fidelity, and Vector Fidelity.

A.1.1 Region Accuracy

IoU (Intersection over Union) measures the region-level agreement between predicted and ground-truth polygons. Given the predicted region M_P and ground-truth region M_G, IoU is defined as:

$$IoU(M_P, M_G) = \frac{|M_P \cap M_G|}{|M_P \cup M_G|}, \tag{2}$$

where $|\cdot|$ denotes the area of a region. Higher IoU indicates better region reconstruction accuracy.

A.1.2 Geometric Complexity

Geometric Complexity evaluates whether reconstructed polygons preserve an appropriate number of vertices while maintaining reconstruction accuracy.

N-ratio directly measures the relative vertex number between predicted polygon P and ground-truth polygon G (Zorzi et al., 2022):

$$N\text{-ratio}(P, G) = \frac{N_P}{N_G}, \tag{3}$$

where N_P and N_G denote the numbers of vertices in predicted polygon P and ground-truth polygon G, respectively. A value closer to 1 indicates more consistent geometric complexity. Values larger than 1 indicate redundant vertices, while values smaller than 1 indicate insufficient geometric representation. For the final results, we report $|N\text{-ratio} - 1|$ to quantify the deviation from the ground-truth vertex count, where a smaller value indicates lower deviation in geometric complexity and a closer match to the true number of vertices.

C-IoU jointly evaluates region accuracy and vertex efficiency (Zorzi et al., 2022). The relative deviation of vertex numbers is first calculated as:

$$RD(P, G) = \frac{|N_P - N_G|}{N_P + N_G}. \tag{4}$$

Then, C-IoU is defined as:

$$C\text{-}IoU(P, G) = IoU(P, G)(1 - RD(P, G)). \tag{5}$$

Compared with IoU, C-IoU penalizes polygons with excessive or insufficient vertices and favors compact vector representations.

A.1.3 Boundary Fidelity

Boundary Fidelity evaluates the alignment between predicted and ground-truth polygon boundaries.

Boundary IoU (B-IoU) measures the overlap between the boundary regions of the predicted and ground-truth polygons (Cheng et al., 2021). Given the ground-truth polygon G and predicted polygon P, we define their boundary regions G_d and P_d as the sets of mask pixels within a distance of d pixels from their respective contours. B-IoU is then computed as:

$$B\text{-}IoU(G, P) = \frac{|G_d \cap P_d|}{|G_d \cup P_d|}. \tag{6}$$

Higher B-IoU indicates better alignment between the predicted and ground-truth boundaries.

PoLiS measures the symmetric distance between two polygon boundaries (Avbelj et al., 2015). Given matched polygons A and B with vertex sets $\{a_i\}_{i=1}^{N_A}$ and $\{b_j\}_{j=1}^{N_B}$, PoLiS is computed as:

$$PoLiS(A,B) = \frac{1}{2}\left(\frac{1}{N_A}\sum_{i=1}^{N_A} dist(a_i, \partial B) + \frac{1}{N_B}\sum_{j=1}^{N_B} dist(b_j, \partial A)\right), \tag{7}$$

where ∂A and ∂B denote polygon boundaries, and $dist(\cdot, \partial B)$ represents the minimum Euclidean distance from a vertex to the opposite polygon boundary. Lower PoLiS indicates smaller boundary displacement and better geometric alignment. A predicted polygon is considered matched to a ground-truth polygon only when their IoU exceeds 0.5, this is also used as the matching criterion for Poly F1.

A.1.4 Vector Fidelity

Vector Fidelity evaluates whether reconstructed polygons preserve structural properties in vector space, including polygon completeness, vertex accuracy, and geometric consistency.

Poly F1 evaluates polygon-level reconstruction accuracy. A predicted polygon is considered matched with a ground-truth polygon when their IoU exceeds a predefined threshold. Based on matched polygons, precision and recall are calculated as:

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

and the Poly F1 score is:

$$Polygon\ F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{10}$$

Corner F1 evaluates vertex localization accuracy. Predicted corners are matched with ground-truth corners according to a distance threshold, and the F1 score is computed based on the number of correctly matched, missing, and redundant corners. Following the vertex-matching protocol of previous indoor floorplan methods (Yue et al., 2023; Chen et al., 2022; Xu et al., 2024a; Liu et al., 2025c), a predicted corner is considered correctly localized when its Euclidean distance to a ground-truth corner is within a threshold. For outdoor scenes, we use a threshold of 10 pixels, while a threshold of 2 pixels is used for the indoor Structured3D (Zheng et al., 2020) dataset. However, whereas those works commonly adopt a matching threshold of 10 pixels, we use a stricter threshold of 2 pixels when evaluating Corner F1 on the Structured3D indoor benchmark to more rigorously assess vertex localization. All compared methods on the indoor benchmark are evaluated under this unified protocol, so the reported indoor Corner F1 values are obtained under a stricter standard and should not be directly compared with those previously reported in the literature.

Angle F1 measures local geometric consistency by evaluating edge orientation agreement. For a matched edge pair, the angular deviation is defined as:

$$\Delta\theta = |\theta_p - \theta_g|, \tag{11}$$

where θ_p and θ_g denote the orientations of predicted and ground-truth edges, respectively. Edges with angular deviations smaller than a predefined threshold are regarded as correctly reconstructed, and the corresponding F1 score is reported. We use a threshold of 5 degree for all benchmarks.

A.2 Additional Implementation Details

All models are trained using the dataset-specific configurations reported in Tab. 2. For datasets using cosine decay, a 10-epoch warm-up is applied before the learning rate decay. For iFLYTEK, label smoothing (Szegedy et al., 2016) is applied to improve training stability. For Deventer, we further employ ALB strategy to automatically adjust the relative contributions of different loss terms during training.

ALB. For Deventer reconstruction, we adopt ALB to alleviate the imbalance among different supervision terms. Fixed weighting requires manually determining the relative contributions of the classification, edge and region losses, which may cause certain objectives to dominate the optimization or become under-emphasized. Inspired by uncertainty-based multi-task learning (Cipolla et al., 2018), we introduce learnable parameters to estimate the contribution of different objectives. Given multiple loss terms $\{\mathcal{L}_i\}_{i=1}^{N}$, the overall objective is formulated as:

$$\mathcal{L} = \sum_{i=1}^{N} \left(w_i \mathcal{L}_i + \frac{1}{2} \log \sigma_i^2 \right) + \mathcal{L}_{reg}, \tag{12}$$

where σ_i denotes a learnable uncertainty parameter associated with the i-th loss term and the corresponding weight is defined as:

$$w_i = \exp(-\log \sigma_i^2). \tag{13}$$

During training, the uncertainty parameters are optimized jointly with the network parameters, allowing the model to automatically adjust the relative importance of different supervision signals. To prevent excessive variation of the learned weights, we further introduce a regularization term:

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{i} \text{softplus}(\log \sigma_i^2), \tag{14}$$

which constrains the adaptive weights and improves optimization stability. This adaptive strategy is only applied to the Deventer dataset, where different supervision terms exhibit more significant convergence discrepancies due to the characteristics of road structures.

Vectorization of Raster-based Baselines. For raster-output baselines without a dedicated vectorization procedure, we first extract polygon boundaries from the predicted raster masks and then simplify the resulting contours using the Douglas-Peucker (DP) algorithm (Douglas & Peucker, 2011) with a tolerance of 1 pixel, following the convention adopted in previous works such as HiSup (Xu et al., 2023) and LDPoly (Jiao et al., 2025a). This procedure is used for D-LinkNet (Zhou et al., 2018b) and CoANet (Mei et al., 2021). For methods with dedicated polygonization procedures, we follow the vectorization strategies described in their original papers. Specifically, SEANet (Li et al., 2023), TFNet (Long et al., 2024) and CLPSNet (Zhu et al., 2024a) generate parcel polygons using their respective polygonization procedures. We then further simplify the resulting polygons using the DP algorithm with a tolerance of 1 pixel to reduce the excessive number of vertices. FFL (Girard et al., 2021) uses its Active Skeleton Model (ASM), while HiSup (Xu et al., 2023) uses the Mask-and-Vertices Attraction (MaV-Attr) method, and LDPoly (Jiao et al., 2025a) uses its proposed polygonization method.The resulting polygons are used to compute C-IoU, $|N\text{-ratio} - 1|$, Poly F1, Corner F1, and Angle F1.

A.3 Ablations

We conduct extensive experiments to analyze the contribution of each design component in UniPoly.

A.3.1 Effect of PRA

The proposed PRA introduces explicit structural priors into Transformer decoding by reorganizing edge queries according to their corresponding polygons, while global self-attention treats all edge queries as independent tokens. We first investigate whether global self-attention can be directly applied to polygon reconstruction in outdoor scenes. As shown in

Fig. 8, the decoder with global self-attention suffers from unstable optimization and fails to achieve meaningful convergence during training. We conjecture that the complex information in ourdoor scenes introduces substantial interference during polygon formation, making it difficult for edge queries to identify and aggregate the geometric information belonging to the same polygon. By reorganizing queries according to polygon structures, the proposed PRA contrains the interactions among edge queries within their corresponding polygon structures, helping the decoder better localize individual polygons and extract edge information to form coherent polygon boundaries, and consequently leads to more stable optimization.

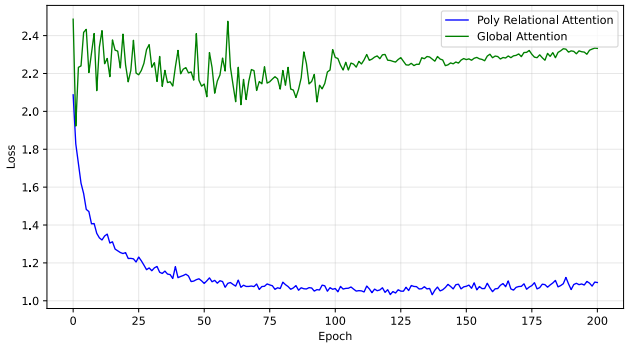


Figure 8: Comparison of training loss curves of global attention and PRA.

The proposed PRA consists of two complementary components. Intra-polygon attention captures geometric dependencies among edges within the same polygon, while inter-polygon attention models structural relationships across different polygon instances. We conduct an ablation study on the iFLYTEK benchmark to investigate the individual contribution of each component.

As shown in Table 4, intra-polygon attention provides a stronger contribution to polygon reconstruction than inter-polygon attention. Compared with inter-polygon attention alone, adding intra-polygon attention consistently improves the reconstruction performance, increasing IoU from 71.7% to 71.9%, C-IoU from 59.6% to 60.0%, and Poly F1 from 54.9% to 56.6%. Similar improvements are observed in boundary and vector fidelity, with B-IoU increasing from 50.7% to 51.3%, Poly F1 from 54.9% to 56.6%, Corner F1 from 29.6% to 31.2%, and Angle F1 from 10.0% to 10.8%. These results highlight the importance of explicitly modeling geometric dependencies among edges within the same polygon for accurate polygon reconstruction.

Furthermore, incorporating inter-polygon attention on top of intra-polygon attention yields additional performance gains. The combined intra+inter configuration achieves the best overall performance, improving IoU from 71.9% to 72.2%, B-IoU from 51.3% to 52.0%, and Poly F1 from 56.6% to 58.0%, while also achieving the best Corner F1 and Angle F1. Although C-IoU remains unchanged at 60.0%, the additional gains in boundary and vector fidelity suggest that inter-polygon interactions complement local polygon modeling by providing cross-instance structural information.

Table 4: Ablation study of intra- and inter-polygon attention components on iFLYTEK dataset.

Attention	Region Accuracy	Geometric Complexity		Boundary Fidelity		Vector Fidelity		
	IoU↑	C-IoU↑	$\lvert N\text{-ratio}-1\rvert\downarrow$	B-IoU↑	PoLiS↓	Poly F1↑	Corner F1↑	Angle F1↑
Inter	71.7	59.6	0.02	50.7	7.7	54.9	29.6	10.0
Intra	71.9	**60.0**	**0.01**	51.3	7.3	56.6	31.2	10.8
Intra+Inter	**72.2**	**60.0**	0.02	**52.0**	**7.2**	**58.0**	**32.4**	**11.5**

A.3.2 Effect of Raster Supervision Resolution

The differentiable rasterization module provides region-level supervision to complement sparse edge-level geometric constraints. Since raster supervision is generated through rasterization, its resolution directly affects the quality of optimization signals.

We investigate the influence of raster supervision resolution on the polygon reconstruction. As shown in Table 5, increasing the rasterization resolution generally improves reconstruction quality, especially for boundary-related metrics. This improvement is attributed to the reduced quantization error and more accurate boundary gradients provided by fine-grained raster supervision. However, excessively high rasterization resolution introduces additional computational overhead while providing diminishing improvements. Therefore, UniPoly adopts 256×256 for the ourdoor benchmarks and 128×128 for the indoor Structured3D benchmark to achieves a favorable trade-off between geometric accuracy and optimization efficiency.

Table 5: Ablation study on the rasterization resolution (X×Y) of L_{region} on the 20% WHU-Building dataset.

Resolution	Training time	Region Accuracy	Geometric Complexity		Boundary Fidelity		Vector Fidelity		
	(h)	IoU↑	C-IoU↑	$\lvert N\text{-ratio}-1\rvert$↓	B-IoU↑	PoLiS↓	Poly F1↑	Corner F1↑	Angle F1↑
64*64	53.7	81.3	62.7	0.29	73.4	2.6	86.5	66.1	23.1
128*128	55.7	82.1	63.6	0.27	74.5	2.3	86.5	66.8	25.4
256*256	57.8	82.8	64.4	0.26	**76.2**	**2.1**	86.9	68.5	29.8
512*512	63.6	**83.0**	**64.7**	**0.25**	**76.2**	2.2	**87.0**	**68.7**	**30.5**

A.3.3 Effect of SwiGLU FFN and ALB

We compare the standard FFN with SwiGLU FFN and fixed loss weights with ALB on the Deventer dataset, while keeping the remaining components unchanged. This ablation study examines the contribution of the two configurable componentsand examines the configuration adopted for road reconstruction. As show in the Tab. 6, replacing the standard

Table 6: Ablation study of SwiGLU FFN and ALB on Deventer dataset.

FFN	Loss Weighting	Region Accuracy	Geometric Complexity		Boundary Fidelity		Vector Fidelity		
		IoU↑	C-IoU↑	$\lvert N\text{-ratio}-1\rvert$↓	B-IoU↑	PoLiS↓	Poly F1↑	Corner F1↑	Angle F1↑
Standard	Fixed	67.8	44.9	0.15	56.0	4.0	55.7	30.4	9.8
SwiGLU	Fixed	71.2	55.0	0.06	60.6	4.0	64.7	39.8	14.1
SwiGLU	ALB	**74.5**	**59.3**	**0.01**	**64.5**	**2.3**	**75.9**	**47.5**	**16.8**

FFN with SwiGLU substantially improves reconstruction performance across all evaluated metrics. IoU increases from 67.8% to 71.2%, while C-IoU increases from 44.9% to 55.0%, $|N\text{-ratio}-1|$ decreases from 0.15 to 0.06, and B-IoU improves from 56.0% to 60.6%. The improvements in Poly F1, Corner F1, and Angle F1 further indicate more accurate polygon and vertex. These results suggest that the gated feature transformation of SwiGLU FFN is particularly beneficial for capturing geometric and boundary information in road reconstruction.

We further replace the fixed loss weights with ALB. ALB further improves the results over the SwiGLU + Fixed configuration: IoU increases from 71.2% to 74.5%, C-IoU from 55.0% to 59.3%, and B-IoU from 60.6% to 64.5%. Moreover, $|N\text{-ratio}-1|$ decreases to 0.01 and PoLiS to 2.3, indicating more accurate geometric complexity and boundary reconstruction. Vector fidelity also improves substantially, with Poly F1, Corner F1, and Angle F1 reaching 75.9%, 47.5%, and 16.8%, respectively. Overall, the results indicate that ALB complements the architectural improvement by better balancing the different reconstruction objectives.

A.3.4 Effect of Label Smoothing

Label smoothing is applied to the iFLYTEK dataset to improve training stability. We further investigate the influence of the label smoothing coefficient by evaluating three settings, i.e.,

0.0, 0.05, and 0.1, while keeping all other training configurations unchanged. As shown in Table 7, introducing label smoothing consistently improves the reconstruction performance. Compared with no label smoothing, both 0.05 and 0.1 achieve higher IoU and substantially better vector-level performance. The setting of 0.1 achieves the best overall performance, obtaining the highest C-IoU, B-IoU, and PoLiS, while maintaining competitive Poly F1, Corner F1, and Angle F1. Therefore, we adopt 0.1 as the default label smoothing coefficient for iFLYTEK.

Table 7: Ablation study of the label smoothing coefficient on iFLYTEK dataset.

Label Smoothing	Region Accuracy	Geometric Complexity		Boundary Fidelity		Vector Fidelity		
	IoU↑	C-IoU↑	$\lvert N\text{-ratio}-1\rvert\downarrow$	B-IoU↑	PoLiS↓	Poly F1↑	Corner F1↑	Angle F1↑
0.00	71.6	59.6	**0.02**	51.2	7.5	56.7	32.5	11.3
0.05	**72.2**	59.5	**0.02**	51.7	7.4	**58.0**	**32.8**	**11.8**
0.10	**72.2**	**60.0**	**0.02**	**52.0**	**7.2**	**58.0**	32.4	11.5

A.4 Limitations

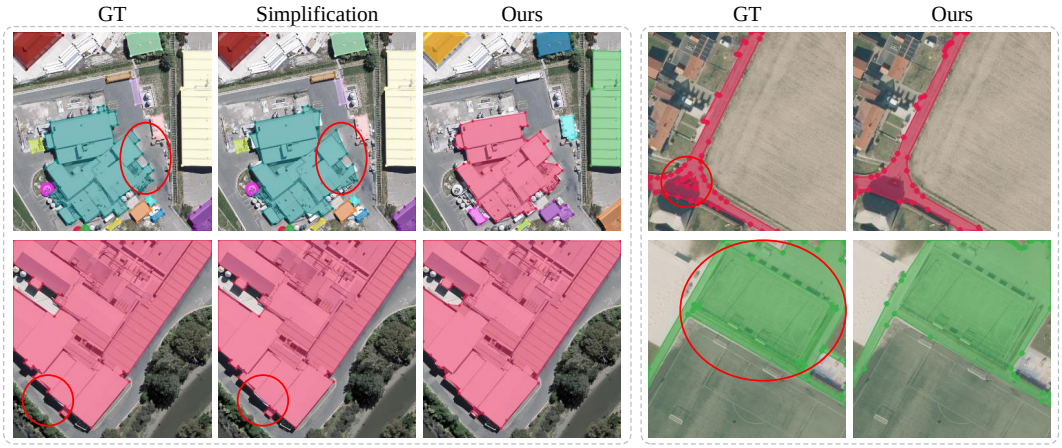


Figure 9: Illustration of the limitations of UniPoly: (a) limited capacity for scenes requiring more polygon and edge queries than predefined, and (b) limited support for instances composed of multiple disconnected polygons.

Although UniPoly achieves strong performance across diverse spatial reconstruction tasks, several limitations remain. In the current benchmarks, the predefined numbers of polygon and edge queries are generally sufficient to represent most instances, and thus have a limited impact on reconstruction quality. However, for scenes with substantially more polygons or edges than the allocated query capacity, the representation may require instance simplification or a reduction in the number of vertices, potentially leading to degraded reconstruction fidelity(Fig. 9(a)). In addition, the current representation assumes that each instance corresponds to a single polygon and therefore cannot directly represent complex polygons with holes. For the road dataset, we consequently reconstruct only the outer boundary for such cases, which may leave some disconnected components unreconstructed(Fig. 9(b)).

A.5 Additional Qualitative Results

We provide additional qualitative comparisons in this section. These results complement the quantitative evaluation in the main paper and show the reconstruction quality under diverse geometric conditions, including complex boundaries, dense object distributions, and incomplete observations.

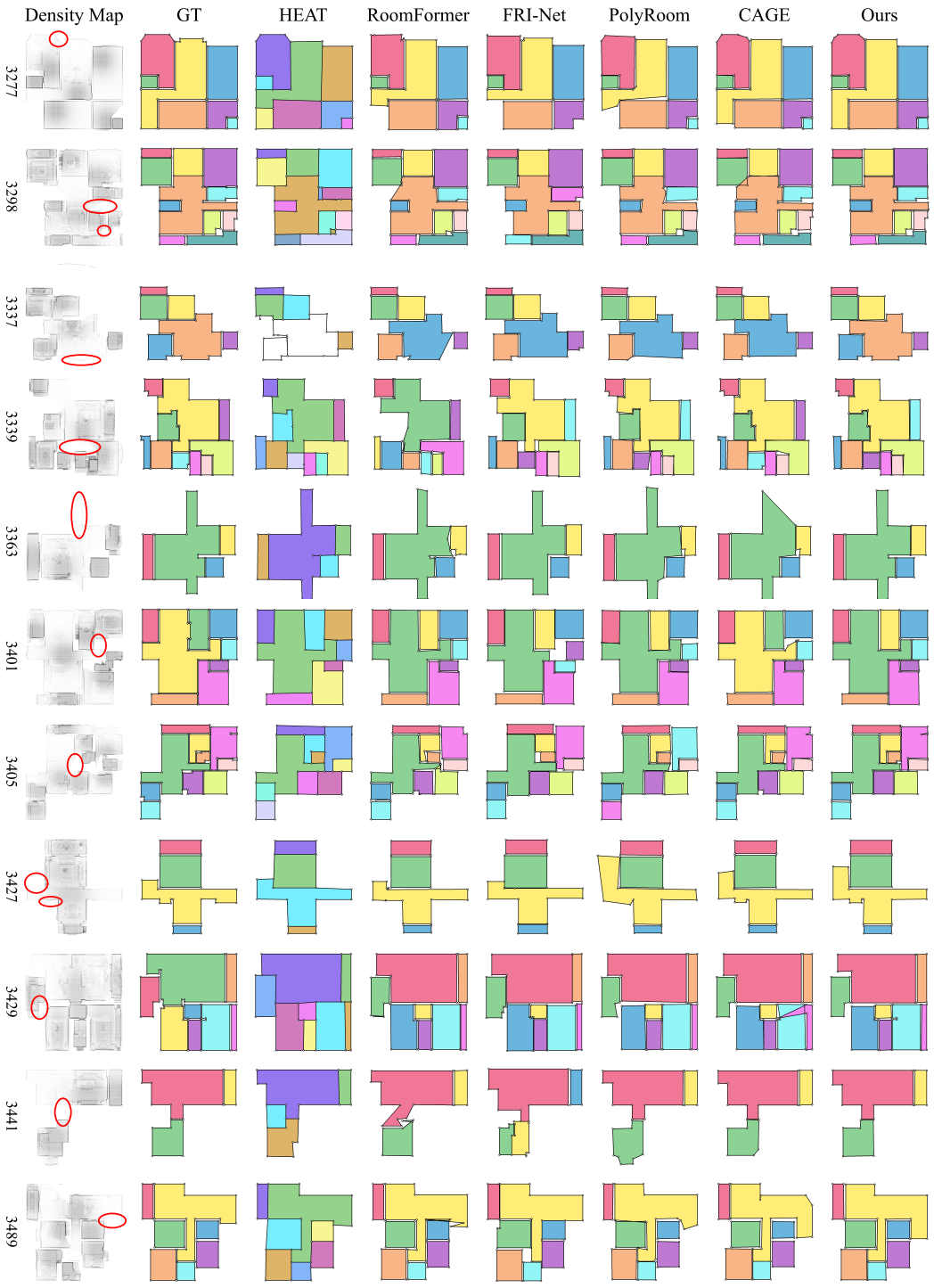


Figure 10: Additional qualitative results of UniPoly on the Structured3D benchmark for indoor floorplan reconstruction.



Figure 11: Additional qualitative results of UniPoly on the Whu-Building benchmark for building contour vectorization.

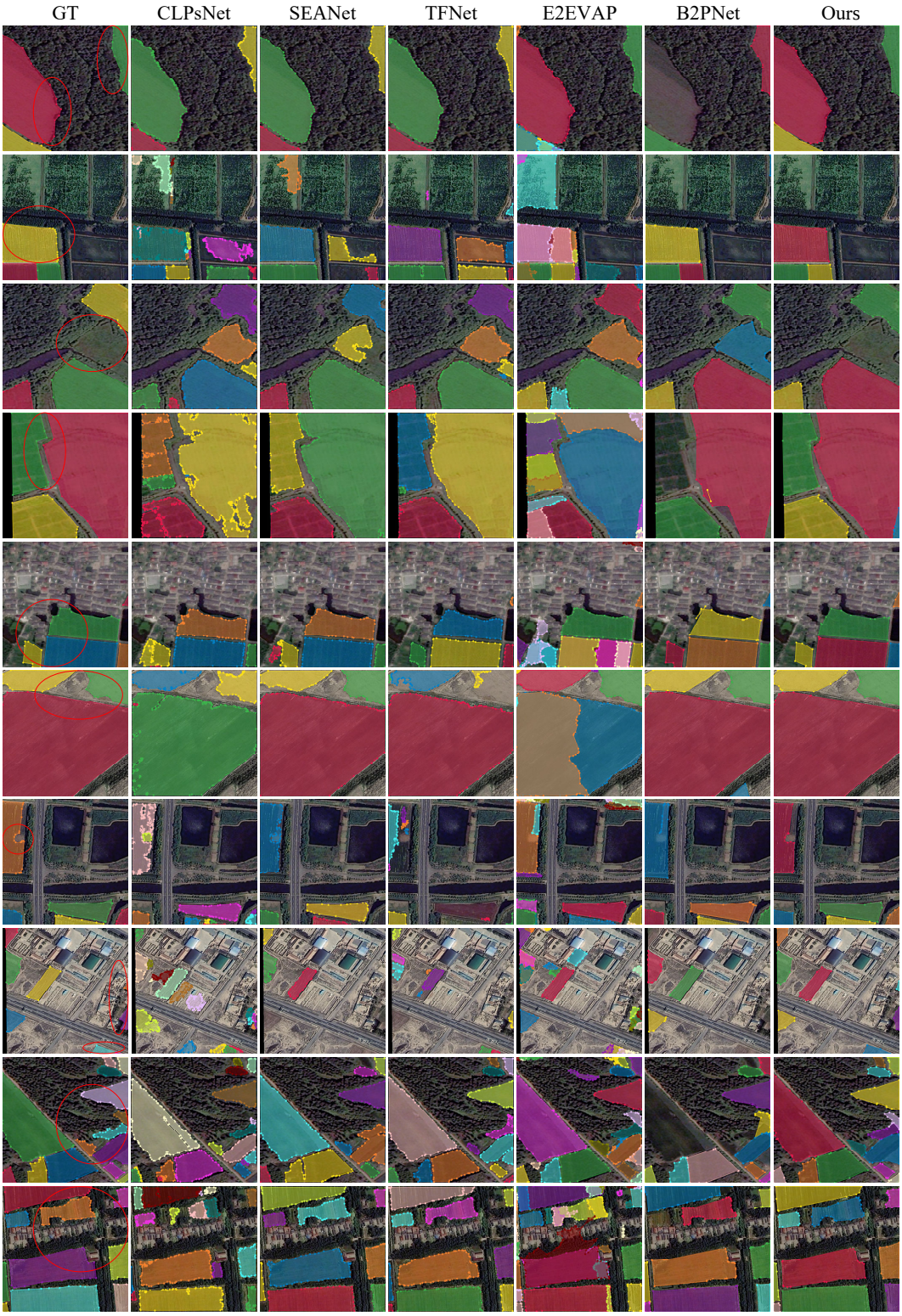


Figure 12: Additional qualitative results of UniPoly on the iFLYTEK benchmark for farmland boundary vectorization.



Figure 13: Additional qualitative results of UniPoly on the Deventer benchmark for road boundary vectorization.