CARTOGENS: GEOGRAPHICALLY CONDITIONED VECTOR SCENE GENERATION FOR BUILDING GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Building generalization across map scales requires coordinated changes in building shape, count, and boundary structure, altering both geometry and topology of the vector scene. These changes cannot be decided building by building, yet existing learning methods process individual buildings, local groups, or rasters that require vectorization. We introduce CartoGenS, an end-to-end framework that generates coarser-scale vector building scenes without predefined source-to-target building correspondences. A shared encoder represents building boundaries, road context, and source-scene physical extent. A single set of learned queries predicts vertices for the entire target scene, and directed successor prediction links them into cycles that form exterior and interior rings, so the number and nesting of output polygons are predicted rather than inherited from the source buildings. Successor scores include a learned relative-geometry term. For scene-level training and evaluation, we construct paired vector building scenes with road context for a fixed 1:10,000-to-1:25,000 transition. CartoGenS achieves the highest overall polygon IoU, panoptic quality, and boundary IoU among the evaluated methods, exceeding the strongest baseline by 9.48 percentage points in boundary IoU. Full factorial ablations highlight the importance of geographic conditioning, and relative-geometry scoring further improves overall performance.

1 INTRODUCTION

Maps at different scales represent geographic features at different levels of detail (Sester, 2005). Building generalization transforms detailed footprints into representations suited to smaller map scales through boundary simplification, aggregation of nearby buildings, and related operations (Feng et al., 2019). The generalized map should remain legible and preserve the spatial layout of building groups, so that it still conveys how buildings are organized when individual footprints can no longer be drawn in full detail (Takahashi et al., 2024).

Unlike curve simplification such as the Douglas–Peucker algorithm, which removes vertices from one polyline at a time (Douglas & Peucker, 1973), building generalization also changes how many polygons and holes a scene contains, and these changes cannot be decided one building at a time. Proximity and similarity criteria often admit several groupings, and choosing among them depends on both local shapes and the overall layout (Takahashi et al., 2024). Simplification, aggregation, and other cartographic operators must also be coordinated, which remains difficult to automate (Feng et al., 2019). Decisions about geometric detail and building combinations therefore depend on the features of each building and on the spatial relationships among buildings (Wang et al., 2025).

Rule-based methods implement cartographic requirements through predefined geometric thresholds and spatial constraints. For example, the *Aggregate Polygons* and *Simplify Building* tools in ArcGIS Pro[1], a widely used commercial geographic information system (GIS), use aggregation distances and simplification tolerances, respectively (Figure 1(a)). Raster models learn generalization from rasterized building scenes (Feng et al., 2019; Yu et al., 2026), but their outputs must be vectorized to obtain polygon geometry (Fan et al., 2026) (Figure 1(b)). Vector learning methods mainly address the

[1] https://pro.arcgis.com/

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.

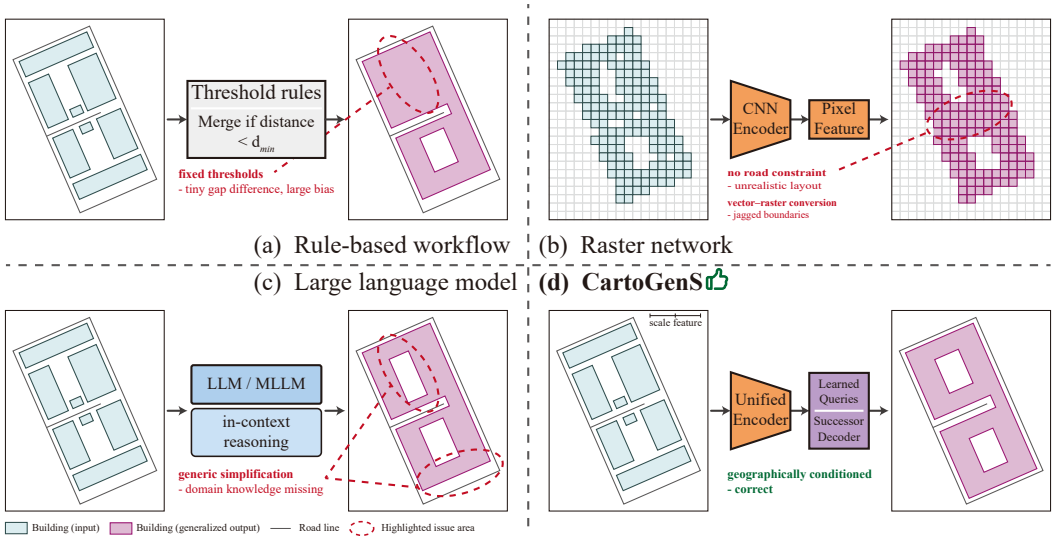


Figure 1: CARTOGENS generates vector building scenes at a coarser map scale from source vectors end to end. Panels (a)–(c) illustrate selected workflows and possible failure modes; (d) shows aggregation within groups while retaining road-separated gaps.

simplification of individual buildings (Zhou et al., 2023; Cui et al., 2025) or many-to-one aggregation within local building groups (Wang & Feng, 2024), and do not generalize complete scenes. General-purpose large language models such as GPT and Claude can generate vectors from prompts containing source data, cartographic instructions, and examples (Figure 1(c)), but their fine-grained geometric control remains limited without additional multiscale cartographic knowledge (Zhou et al., 2026). Using geographic context to coordinate changes in building shape with the spatial organization of the whole scene, directly in vector form, therefore remains an open modeling problem.

We introduce the Cartographic Scene Generator (CARTOGENS), an end-to-end framework for *geographically conditioned vector scene generation* (Figure 1(d)). Encoding buildings and roads together gives the model spatial context beyond individual footprints, and source-scene scale restores the physical extent removed by coordinate normalization. From this representation, the model predicts target vertices and the connections between them, adding a learned relative-geometry term to the connection scores. No vertex is assigned to a particular building in advance: the predicted connections form a permutation whose cycles are the exterior and interior rings, so the output topology is predicted rather than inherited from individual source buildings or local groups. Trained on complete paired scenes, the model can merge nearby buildings while keeping courtyards and the gaps between separate buildings.

Our contributions are:

- **A new scene-level vector generation task.** We formulate building generalization as a scene-level vector generation task without predefined source-to-target building correspondences. Conditioned on road context and source-scene scale, the task covers geometric and topological changes in building shape, count, and boundary structure across a paired map-scale transition, so that outputs reflect the spatial organization of the whole scene.
- **A new geographically conditioned generative model.** We propose CARTOGENS, which encodes building boundaries, road context, and source-scene scale in a shared representation. A single set of learned queries predicts vertices for the entire target scene, and the cycles of directed successor relations between these vertices define the output boundaries, so geometry and topology are predicted jointly. Road context and physical extent condition the input representation, while relative-geometry scoring guides how output boundaries are formed.
- **A new benchmark for vector scene generalization.** We construct paired vector scenes with road context, covering single and multiple target buildings with and without holes. A common evaluation protocol measures geometric agreement and target structure across methods,

and full factorial ablations assess the three model components. CARTOGENS achieves the highest overall polygon IoU, panoptic quality (PQ), and boundary IoU among the evaluated methods, and the ablations highlight the importance of geographic conditioning.

2 RELATED WORK

Building generalization and scene-level learning. Rule-based and optimization methods generalize buildings through geometric constraints and grouping objectives (Sester, 2005; Takahashi et al., 2024). Learning-based approaches estimate these transformations from paired maps. Raster models learn scene transformations (Feng et al., 2019; Yu et al., 2026), and diffusion models condition generation on the target map scale before vectorizing their outputs (Fan et al., 2026). Direct vector methods learn to edit or generate individual contours (Zhou et al., 2023; Cui et al., 2025; Chen et al., 2026b) or to aggregate local building groups (Wang & Feng, 2024). We instead learn from complete vector scenes, so that building geometry, count, and spatial organization change together.

Geographic context and spatial relation modeling. Geographic knowledge guides aggregation through constraints imposed by obstacles (Zhang et al., 2023) and through reasoning about proximity and spatial structure (Su et al., 2025). Raster models capture geometric structure with multiscale features and distance maps (Zhou et al., 2024), and explainable-AI analysis relates their predictions to building boundaries and to the spaces between neighboring buildings (Fu et al., 2024). Vector models encode contour neighborhoods for shape analysis (Wang et al., 2026) and relationships between buildings for grouping (Yang et al., 2025). CARTOGENS conditions scene encoding on road context and source-scene scale, and adds relative-geometry scoring to boundary prediction.

Structured vector generation. Polygonal representations describe vertices and their connections along boundaries. Set prediction provides a framework for unordered outputs (Carion et al., 2020) and has been extended to image-to-graph generation (Shit et al., 2022) and to floorplan reconstruction, where RoomFormer assigns separate query groups to individual polygons (Yue et al., 2023). PolyWorld and Re:PolyWorld form polygon boundaries by predicting a permutation over vertices (Zorzi et al., 2022; Zorzi & Fraundorfer, 2023). Sequence models offer an alternative: Pix2Poly combines vertex tokens with learned connectivity (Adimoolam et al., 2025), and MARS generates map vectors autoregressively (Zhang et al., 2026). These methods recover polygons from raster observations without polygon topology. We instead apply the vertex-and-successor representation to cartographic abstraction, where the input is already a vector scene whose polygon count and nesting change at a coarser scale; all target rings are therefore drawn from one shared vertex pool, and predicted successor cycles determine the output topology.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND OVERVIEW

Given a source building scene $\mathcal{B}_s$ and its geographic conditions $\mathcal{C}_s$, we learn the scene-to-scene mapping

$$F_\theta : (\mathcal{B}_s, \mathcal{C}_s) \longmapsto \widehat{\mathcal{B}}_t, \qquad \mathcal{C}_s = (\mathcal{R}_s, s), \tag{1}$$

where $\mathcal{R}_s$ is the contextual road geometry and s describes the physical extent of the source scene. Scene bounds are given during both training and inference, and source and target cover the same area under the fixed scale transition of the paired dataset. Complete paired scenes supervise training without predefined source-to-target building correspondences.

Two properties of the task shape the design (Figure 2). First, how a building should be generalized depends on its surroundings, so a shared geographic encoder represents all source buildings together with roads and source-scene scale. Second, the number and grouping of target buildings are not known in advance, so the decoder predicts a shared vertex pool whose successor assignment forms the target buildings and their holes. Coordinate and successor losses train both parts end to end.

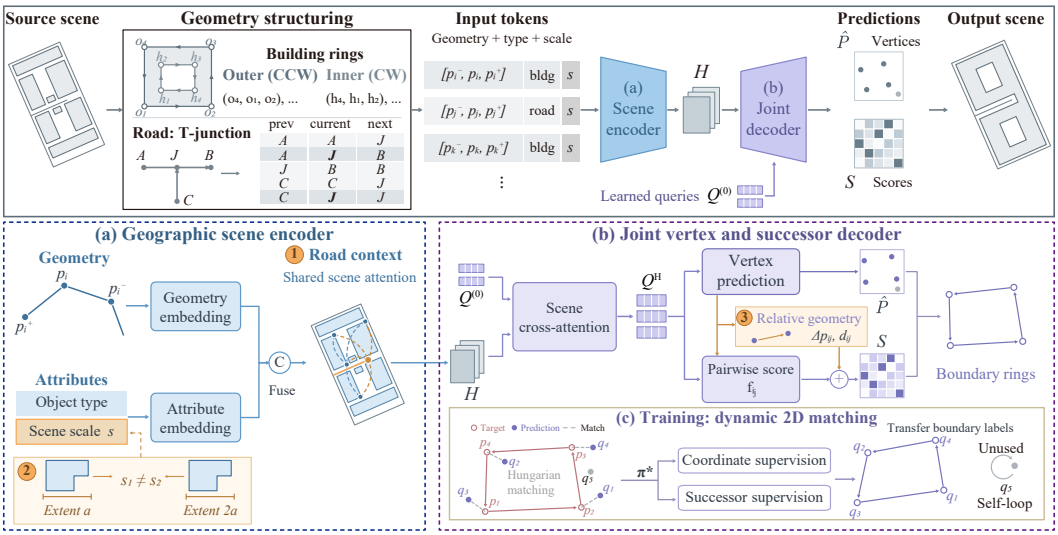


Figure 2: Overview of CARTOGENS. The geographic scene encoder (a) combines local geometry, feature type, and source-scene scale. The joint decoder (b) predicts vertices and successor scores. During training, dynamic matching (c) assigns target vertices to queries, and the same assignment supervises vertex coordinates and transfers target connections as successor labels, with unmatched queries labeled as self-loops; inference assigns successors and assembles boundary rings.

3.2 CARTOGENS

Source representation. We model each building boundary vertex as a token with ordered coordinates $p_i^-, p_i, p_i^+ \in \mathbb{R}^2$ for its predecessor, itself, and successor around a closed ring. This triple encodes local geometry and adjacency (Zhou et al., 2023; Wang et al., 2026). Vertices along directed road paths use the same representation: endpoints repeat their own coordinate for the missing neighbor, and junctions can have multiple tokens with different predecessors and successors.

Geographic scene encoding. The complete source scene determines the center μ and extent a that normalize source and target coordinates to $\bar{p} = 0.9(p - \mu)/a$, with a random rotation added during training (Appendix B.1). Normalization removes absolute size, whereas cartographic thresholds such as aggregation distances are defined in metres; scenes of different physical size can thus map to the same geometry (Figure 2(a)), so we retain the extent through the source-scene scale

$$s = \tanh\left(-\frac{\log_{10}(\max(a, 10^{-12}))}{6}\right). \tag{2}$$

Each token combines its normalized coordinate triple with a type indicator $b_i \in \{0, 1\}$, which distinguishes buildings from roads, and with the scale s shared by the scene:

$$x_i = [\bar{p}_i^-, \bar{p}_i, \bar{p}_i^+, b_i, s] \in \mathbb{R}^8. \tag{3}$$

Separate MLPs embed the six coordinate and two attribute channels, and their concatenation is projected to width d. Shared self-attention maps the source tokens X to the decoder memory $H = E_\theta(X)$, with padding masked. Because buildings and roads share the geometric embedding and differ only in the type indicator, attention can relate buildings to one another and to roads, and road tokens inform building features even though roads are not generated (Appendix B).

Target representation. We represent all target rings by vertex coordinates $P \in \mathbb{R}^{M\times 2}$ with rows p_k in the same normalized frame, and by a directed successor matrix $A \in \{0, 1\}^{M\times M}$, where M is the target vertex count. Each vertex has one predecessor and one successor, so A is a permutation matrix whose cycles form counterclockwise exterior and clockwise interior rings; ring topology is thus encoded combinatorially, separately from P. The model predicts a fixed budget of Q candidate vertices, and a self-loop $i \to i$ marks candidate i as unused, so the active vertex count varies across scenes. For example, a scene with two buildings and one courtyard has two exterior cycles and one interior cycle from the same candidate set, and successor prediction assigns candidates to boundaries.

Vertex and successor prediction. We treat scene vertex generation as unordered set prediction (Carion et al., 2020) with Q shared learned queries $Q^{(0)}$. Decoder layers combine self-attention, cross-attention to H, and feed-forward blocks, and produce the decoded query states Q^H. The final state $z_i \in \mathbb{R}^d$ of query i determines its normalized position

$$\widehat{p}_i = 0.9 \tanh(g_\theta(z_i)), \tag{4}$$

and these positions form $\widehat{P} \in \mathbb{R}^{Q\times 2}$.

To connect candidate vertices into building boundaries, the model combines contextual descriptors with explicit relative geometry. We project z_i to a lower-dimensional descriptor, concatenate $\widehat{p}_i$ following Zorzi et al. (2022), and use an attention GNN to obtain h_i. ScoreNet maps each ordered descriptor pair to a directional score $f_{ij} = f_\theta(h_i, h_j)$. A learned relative-geometry term g_{ij} over the relative displacement and distance d_{ij} makes the spatial arrangement of each vertex pair explicit:

$$S_{ij} = f_{ij} + g_{ij}, \qquad g_{ij} = w^\top \phi(\widehat{p}_i, \widehat{p}_j) + b, \tag{5}$$

$$\phi(\widehat{p}_i, \widehat{p}_j) = [\Delta x_{ij}, \Delta y_{ij}, |\Delta x_{ij}|, |\Delta y_{ij}|, d_{ij}], \qquad \Delta p_{ij} = \widehat{p}_i - \widehat{p}_j, \quad d_{ij} = \|\Delta p_{ij}\|_2. \tag{6}$$

Both terms depend differentiably on the predicted coordinates, so successor supervision updates vertex positions as well as successor scores.

3.3 Training and inference

Dynamic matching. Training aligns unordered candidate vertices with the target scene. Hungarian matching assigns each valid target vertex p_k to a distinct query using coordinate L_1 distance (Carion et al., 2020):

$$\pi^* = \arg\min_{\pi \text{ injective}} \sum_{k=1}^{M} \|\widehat{p}_{\pi(k)} - p_k\|_1. \tag{7}$$

The matching is recomputed from the final coordinates at every training forward pass. The same assignment transfers target connections to queries: with $\Pi \in \{0,1\}^{Q\times M}$ the matrix of π^*, the successor labels are $Y = \Pi A \Pi^\top + \mathrm{diag}(\mathbf{1} - \Pi\mathbf{1})$, with self-loops for unmatched queries. Sharing one assignment keeps both targets consistent as correspondences change; since Y depends only on the cycles of A, supervision needs no ring order, starting vertex, or source-to-target building assignment.

Training objective. Matched coordinates are supervised with the Smooth-L_1 loss $\mathcal{L}_{\text{coord}}$. For successor prediction, log-domain Sinkhorn normalization with temperature τ converts S/τ into an approximately doubly stochastic matrix $\widetilde{A}$, so that queries compete for successor vertices (Cuturi, 2013; Zorzi et al., 2022). Let y_i be the successor label given by row i of Y and ω_i its row weight. The per-scene structural loss and total objective are

$$\mathcal{L}_{\text{succ}} = -\frac{\sum_{i=1}^{Q} \omega_i \log \widetilde{A}_{i,y_i}}{\sum_{i=1}^{Q} \omega_i}, \qquad \mathcal{L} = \lambda \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{succ}}, \tag{8}$$

where λ weights the coordinate loss. Matched rows have weight 1 and unmatched rows a smaller weight. The total loss trains the encoder and decoder end to end (Appendix B).

Inference and ring assembly. At inference, maximum-weight assignment on S connects the predicted vertices without target geometry. The assignment produces a permutation matrix $\widehat{A}$ that gives each candidate vertex one predecessor and one successor. Cycles of length at least three form candidate rings, while self-loops and two-cycles are discarded. Retained rings are therefore closed and vertex-disjoint, with no dangling or branching edges. Ring orientation separates exterior from interior rings, from which building polygons with holes are reconstructed (Appendix C).

4 Experiments

4.1 Experimental setup

Dataset. We construct paired vector scenes from OS Open Zoomstack[2], using blocks enclosed by roads as scene bounds. Each input contains the buildings of the 1:10,000 layer and the roads

[2] https://www.ordnancesurvey.co.uk/products/os-open-zoomstack

of the same product, and the generalization target contains the 1:25,000 buildings within the same geographic extent. England, Wales, and Scotland provide the training, validation, and test scenes, respectively (Table 1). Across all 35,811 pairs, 98.0% have fewer target buildings than source buildings. Target building count and the presence of interior rings define four categories, which test whether a model can reduce building count while keeping the separations and holes that each target requires. Training is subsampled to reduce category imbalance, while validation and test keep their original proportions. Appendix A details contents, splits, and statistics.

Table 1: Paired-scene splits by target structure. Single and multiple refer to the number of target buildings; training is subsampled by category.

Target structure	Train (England)	Validation (Wales)	Test (Scotland)
Single, no holes	8,000	3,635	3,007
Single, with holes	7,794	337	800
Multiple, no holes	8,000	386	975
Multiple, with holes	2,579	100	198
Total	26,373	4,458	4,980

Baselines. Raster baselines are MSEONet (Huang et al., 2025), RoofX-Net (Chen et al., 2026a), and SAM-BG (Yu et al., 2026), all taking building masks as input. For prompted vector generation, we use GPT-6 Astra[3], a recently released multimodal model with strong general capabilities in understanding and generating vector data, run with high reasoning effort and denoted GPT-6-Astra-High. Its input comprises source buildings and roads, a preview image, and reference example pairs from the training set; weights are not updated, and failed responses count as empty outputs. ArcGIS Pro represents rule-based generalization: its tools apply fixed distance and tolerance thresholds directly to vector data, with roads as barriers. The main comparison evaluates complete systems under their respective input configurations, and Section 4.3 isolates component contributions within CARTOGENS. Appendix B specifies the experimental settings.

Evaluation. All methods are scored against the vector reference geometries and on a shared 256×256 grid whose bounds are defined by the source scene. For vector-domain evaluation, outputs are converted to polygonal regions with the reconstruction in Appendix C, and raster outputs are first thresholded and vectorized. For raster-domain evaluation, vector outputs are rasterized on the shared grid. Polygon IoU measures building overlap, and smallest bounding rectangle (SBR) IoU measures extent and orientation; both count missing reference buildings as zero. Turning-function distance (TFD) measures exterior shape over positive-overlap matches and is reported with match coverage (Arkin et al., 1991). Panoptic quality (PQ) combines overlap with penalties for missed and extra buildings at IoU > 0.5 (Kirillov et al., 2019). Raster metrics are IoU, boundary IoU (BIoU) (Cheng et al., 2021), F1, and structural similarity (SSIM). Category scores and hole IoU assess the recovery of different target structures. Appendix D gives metric definitions and denominators.

4.2 COMPARISON WITH GENERALIZATION METHODS

CARTOGENS achieves the highest polygon IoU, PQ, and BIoU (Table 2), exceeding GPT-6-Astra-High by 0.0289 in PQ and 0.0948 in BIoU. Gains in PQ and BIoU indicate closer agreement with individual target buildings and their outlines.

The category results show where these gains arise (Table 3). The largest margin appears for single buildings with holes, where several source buildings must merge into one outline: CARTOGENS exceeds GPT-6-Astra-High by 0.0803 in polygon IoU and 0.0880 in PQ. Its hole IoU in this category, 0.3494, remains below the 0.3614–0.3925 of the raster methods (Table 8), so the gain lies mainly in the merged exterior outline. For multiple buildings without holes, CARTOGENS also has the highest polygon IoU and PQ, 0.4531 and 0.4238. Both cases require an output structure that differs from the input, one outline merged from several source buildings or separate buildings kept apart, and successor assignment over the shared vertex pool forms either without editing individual source buildings.

[3] https://openai.com/index/gpt-6-astra/

Table 2: Comparison on all 4,980 test scenes.

	Vector-domain metrics				Raster-domain metrics			
Method	Poly. IoU ↑	SBR IoU ↑	TFD ↓	PQ ↑	IoU ↑	BIoU ↑	F1 ↑	SSIM ↑
Raster-domain methods								
MSEONet	0.5440	0.5915	0.6620	0.4924	0.7170	0.3734	0.8352	0.8460
RoofX-Net	0.5702	0.6116	0.6314	0.5272	0.7324	0.3955	0.8455	0.8550
SAM-BG	0.5680	0.6113	0.6843	0.3802	0.7222	0.3821	0.8387	0.8497
Vector-domain methods								
ArcGIS Pro	0.5021	0.5615	0.6663	0.4982	0.7022	0.3529	0.8250	0.8435
GPT-6-Astra-High	0.6039	0.6541	**0.5004**	0.5866	0.7935	0.5658	0.8849	0.9032
CARTOGENS (ours)	**0.6319**	**0.6666**	0.5087	**0.6155**	**0.8220**	**0.6606**	**0.9023**	**0.9125**

Bold: best; underline: second best in each metric column. TFD is in radians; positive-overlap match coverage: MSEONet 83.56%, RoofX-Net 84.45%, SAM-BG 85.92%, ArcGIS Pro 76.72%, GPT-6-Astra-High 81.38%, and CARTOGENS 82.21%.

Table 3: Cross-method polygon IoU / PQ by target structure. Columns contain 3,007, 800, 975, and 198 test scenes. Each metric retains its within-category reference-building or pooled-count denominator.

Method	Single, no holes	Single, with holes	Multiple, no holes	Multiple, with holes
MSEONet	0.6916 / 0.5853	0.7271 / 0.6770	0.4008 / 0.3652	0.3655 / 0.3331
RoofX-Net	0.7117 / 0.6199	0.7321 / 0.6804	0.4357 / 0.4102	0.4016 / **0.3794**
SAM-BG	0.6983 / 0.3884	0.7253 / 0.5477	0.4407 / 0.3317	**0.4194** / 0.3265
ArcGIS Pro	0.6980 / 0.6445	0.6769 / 0.6524	0.3184 / 0.3099	0.3151 / 0.3270
GPT-6-Astra-High	**0.8358** / 0.7720	0.7439 / 0.7018	0.4024 / 0.3928	0.3838 / 0.3504
CARTOGENS (ours)	0.8253 / **0.8109**	**0.8242 / 0.7898**	**0.4531 / 0.4238**	0.4114 / 0.3671

Bold: best; underline: second best for each metric within each category.

The raster baselines match more reference buildings than CARTOGENS, yet have higher TFD and lower BIoU, indicating that their additional matches come with larger outline errors. GPT-6-Astra-High has slightly lower TFD than CARTOGENS but also lower coverage, 81.38% against 82.21%, so its TFD is computed over fewer matched buildings.

Among the baselines, GPT-6-Astra-High ranks second on every metric except TFD, but its lead over the raster methods comes mainly from single-building scenes, and for multiple buildings its PQ falls below that of RoofX-Net (Table 3). SAM-BG reaches polygon IoU close to that of RoofX-Net, yet its PQ is 0.1470 lower and its count MAE is the highest at 1.3289 (Table 8). ArcGIS Pro, which applies the same aggregation distance and simplification tolerance to every scene, recovers the exact building count most often but has the lowest polygon IoU and match coverage.

Figure 3 traces these errors to individual generalization decisions. Simplification must remove some holes and keep others: ArcGIS Pro keeps a hole that is absent from the target in Case 1, whereas RoofX-Net, SAM-BG, and GPT-6-Astra-High omit the reference courtyard in Case 2. In both cases, the raster outputs also have rounded and uneven outlines where the reference has straight edges and right angles, consistent with the higher TFD of the raster methods. Separation adds a further requirement. In Case 3, the raster methods merge the two reference buildings and GPT-6-Astra-High introduces a hole. CARTOGENS recovers the required outlines and separation more closely in these cases, and in Case 4 it produces two buildings with one hole.

4.3 ABLATION STUDIES

Table 4 evaluates all eight combinations of road context, source-scene scale, and relative-geometry scoring, labeled Road context, Scene scale, and Score bias. Base disables all three factors, and Full is CARTOGENS with all three enabled, reusing its main-comparison predictions. Adding each factor to Base and removing it from Full measure its contribution under shared training, initialization, and inference settings.

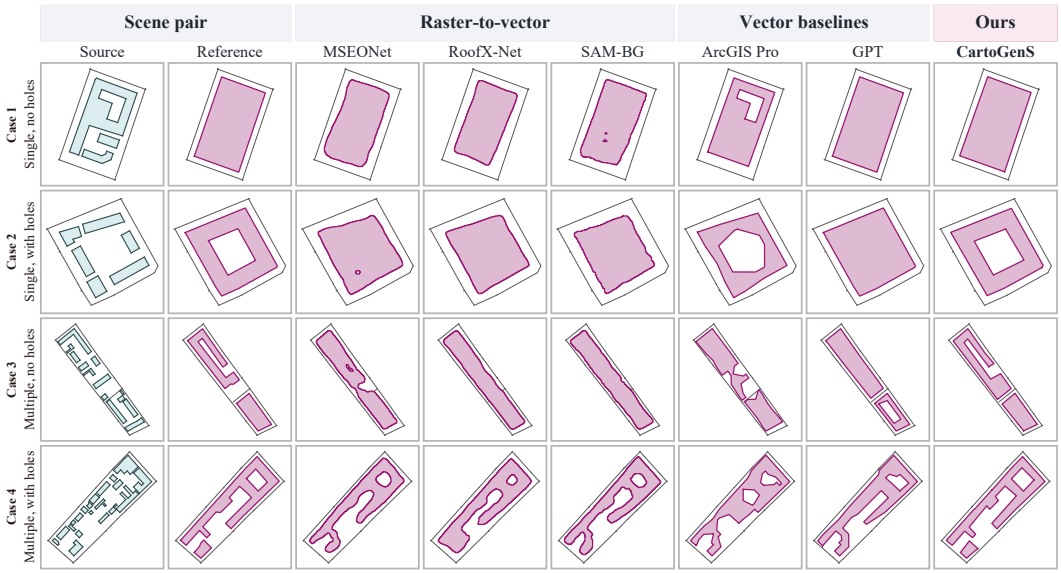


Figure 3: Qualitative comparison on selected test scenes for the four target structures: single and multiple buildings with and without holes. The results show differences in boundary simplification, building separation, and interior-ring recovery. Each row uses a common geographic extent across methods. Pale teal denotes source buildings, pale rose denotes reference and predicted buildings, and dark gray denotes roads.

All configurations retain local boundary features, attention, and descriptors containing predicted coordinates. Roads are masked after normalization over the full source scene, scale is disabled by zeroing its channel, and removing relative-geometry scoring drops g_{ij} from Equation 5. Figure 4(b) summarizes these contrasts.

Table 4: Full factorial ablation on 4,980 test scenes. Road context, Scene scale, and Score bias indicate enabled (1) or disabled (0) components.

Configuration	Road context	Scene scale	Score bias	Poly. IoU ↑	PQ ↑	BIoU ↑
Base	0	0	0	0.1156	0.0799	0.1092
Base + road	1	0	0	0.5002	0.5120	0.5218
Base + scale	0	1	0	0.1568	0.1295	0.1382
Base + score bias	0	0	1	0.1331	0.0938	0.1230
Full − road	0	1	1	0.1609	0.1346	0.1371
Full − scale	1	0	1	<u>0.5818</u>	<u>0.5738</u>	0.5822
Full − score bias	1	1	0	0.5621	0.5607	<u>0.6091</u>
CartoGenS (full)	1	1	1	**0.6319**	**0.6155**	**0.6606**

Bold: best; <u>underline</u>: second best in each metric column.

Road context. Roads give the largest gains in both the addition and removal comparisons (Table 4). Without roads, all four configurations leave over half of the scenes empty, even when scale and relative-geometry scoring are enabled, and adding roads to Base reduces empty outputs from 60.9% to 19.4% (Table 10). Empty outputs concentrate in scenes with few target vertices: Base leaves 72.7% of single-building scenes without holes empty but 11.6% of multiple-building scenes with holes, and the former also remains the most often empty for Base + road and Full (Table 11). Unlike raster methods, which label each pixel of an aligned input mask, CartoGenS generates every vertex and connection, so a scene yields no building when no closed ring forms. The road gains therefore combine more reliable ring formation with fewer omitted buildings and better separation across roads (Figure 6).

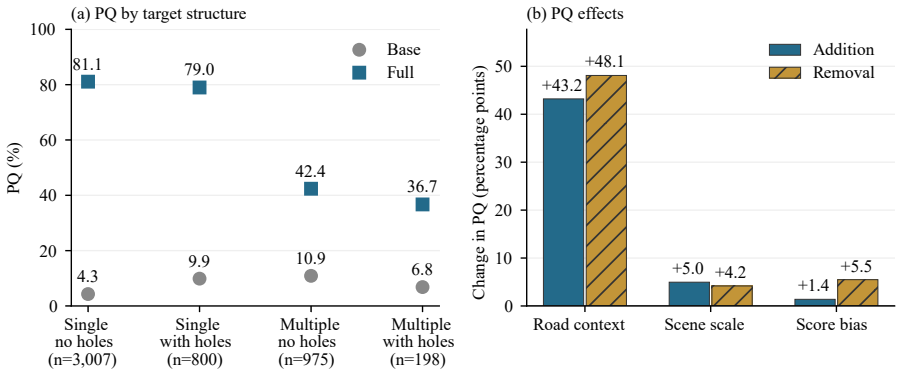


Figure 4: PQ by target structure and model component. (a) Base and Full in all four target categories, with scene counts. (b) Gains from adding each factor to Base or retaining it in Full; exact polygon IoU, PQ, and BIoU contrasts appear in Table 13.

Source-scene scale. Removing scale from Full lowers overall PQ by 0.0417 and BIoU by 0.0784, and Figure 7 shows changes in building extent and concavities. PQ nevertheless improves slightly for multiple buildings without holes and more for those with holes (Table 12). Within the fixed transition, restoring metric information thus appears to help set the extent and detail of generalized shapes, whereas separating multiple buildings relies more on relationships between buildings.

Relative-geometry scoring. It raises PQ by 0.0549 when retained in Full but by only 0.0139 when added to Base (Figure 4(b)). Since every configuration already passes predicted coordinates to successor scoring, the gain concerns how explicitly spatial relationships enter connection decisions. Figure 8 shows fewer crossings and better building separation when the term is present.

4.4 Scene structure and geometric quality

In multiple-building scenes, the main difficulty is dividing occupied regions into the correct buildings. For these categories, raster IoU stays above 0.72 while PQ falls to 0.3671–0.4238 (Tables 7 and 3), since an incorrect merge can preserve most of the occupied area while losing the separation between buildings. CartoGenS has the highest PQ in three categories, but for multiple buildings with holes RoofX-Net has higher PQ and SAM-BG higher polygon IoU (Table 3). This category requires both instance separation and holes, and its scenes also have more vertices and larger source extents (Appendix A.3), so the gap cannot be attributed to holes alone.

In scene topology, CartoGenS has the lowest count MAE and spurious-hole rate, yet its hole IoU is below that of the raster methods (Table 8) and falls from 0.3494 for single buildings to 0.2051 for multiple buildings (Table 7). Recovering holes together with the buildings that enclose them, while keeping the gaps between separate buildings, therefore remains the main open problem.

5 Conclusion

CartoGenS treats building generalization as scene-level vector generation. Learned queries predict vertices for the entire target scene, and successor relations among these vertices form exterior and interior rings. Because no vertex is assigned to a building in advance, building count and boundary structure are predicted jointly with building shape. On a paired benchmark for a fixed map-scale transition, CartoGenS achieves the highest overall polygon IoU, PQ, and boundary IoU among the evaluated methods, with the largest margin where several source buildings merge into one outline, and the ablations confirm the importance of geographic conditioning. Predicting topology as successor cycles may also suit other vector generation tasks in which polygon count and nesting change between input and output. The evaluation covers one map product and one transition, ring formation does not enforce geometric validity, and separating multiple buildings while recovering their holes remains difficult. Future work will condition generation on the target scale across several transitions, build validity constraints into ring formation, and improve this joint recovery.

AI USE STATEMENT

We used ChatGPT to assist with polishing the manuscript text. The authors are responsible for the methods, analyses, citations, and claims reported in the paper, including verification of material produced with AI assistance.

ETHICS STATEMENT

Automated map generalization can reduce map-maintenance effort, but its errors can alter the spatial relationships a map conveys, for example by joining buildings across a road or filling a courtyard. The evaluation therefore reports instance, boundary, and hole measures alongside area agreement (Appendix D), and its conclusions apply to the studied map product and scale transition (Appendix G). The work involves no human participants or personal-attribute inference. Source data are distributed under the Open Government Licence, which permits the redistribution and adaptation performed here with the attribution given in Appendix A.

REPRODUCIBILITY STATEMENT

Anonymized materials are available at Figshare: https://figshare.com/s/204eb0979eff75432372; the page provides access instructions. Sections 3 and 4 describe the method and evaluation design. Appendix A documents data provenance, splits, and preprocessing; Appendix B lists model, training, baseline, hardware, and software configurations; Appendix C specifies successor assignment and geometry reconstruction; and Appendix D gives the metric formulas. Appendices E and F provide complete component results and qualitative analyses.

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 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Esther M. Arkin, L. Paul Chew, Daniel P. Huttenlocher, Klara Kedem, and Joseph S. B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):209–216, 1991. doi: 10.1109/34.75509.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

Keyu Chen, Tianlei Wang, Zhiyou Yang, Fan Li, Ma Luo, Ruoning Zhang, Hong Qu, and Wenyu Chen. RoofX-Net: A tailored approach to accurate multi-type rooftop segmentation in remote sensing images using edge and scale awareness. *ISPRS Journal of Photogrammetry and Remote Sensing*, 234:46–59, 2026a. doi: 10.1016/j.isprsjprs.2026.02.015.

Yue Chen, Haizhong Qian, Xiao Wang, and Limin Xie. DiffSimplify: A building continuous simplification model based on conditional diffusion models. *International Journal of Digital Earth*, 19(1):2660426, 2026b. doi: 10.1080/17538947.2026.2660426.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15334–15342, 2021.

Longfei Cui, Junkui Xu, Lin Jiang, and Haizhong Qian. End-to-end vector simplification for building contours via a sequence generation model. *ISPRS International Journal of Geo-Information*, 14 (3):124, 2025. doi: 10.3390/ijgi14030124.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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. *Cartographica*, 10(2):112–122, 1973. doi: 10.3138/FM57-6770-U75U-7727.

Keying Fan, Ning Qian, Zhiyong Zhou, Monika Sester, and Yu Feng. A scale-adaptive framework for building generalization via instruction-guided diffusion and geometry-aware vectorization. *Journal of Geovisualization and Spatial Analysis*, 10:44, 2026. doi: 10.1007/s41651-026-00282-3.

Yu Feng, Frank Thiemann, and Monika Sester. Learning cartographic building generalization with deep convolutional neural networks. *ISPRS International Journal of Geo-Information*, 8(6):258, 2019. doi: 10.3390/ijgi8060258.

Cheng Fu, Zhiyong Zhou, Yanan Xin, and Robert Weibel. Reasoning cartographic knowledge in deep learning-based map generalization with explainable AI. *International Journal of Geographical Information Science*, 38(10):2061–2082, 2024. doi: 10.1080/13658816.2024.2369535.

Wubiao Huang, Fei Deng, Haibing Liu, Mingtao Ding, and Qi Yao. Multiscale semantic segmentation of remote sensing images based on edge optimization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. doi: 10.1109/TGRS.2025.3553524.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Monika Sester. Optimization approaches for generalization and data abstraction. *International Journal of Geographical Information Science*, 19(8-9):871–897, 2005. doi: 10.1080/13658810500161179.

Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes C. Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, and Bjoern Menze. Relationformer: A unified framework for image-to-graph generation. In *European Conference on Computer Vision (ECCV)*, 2022.

Youneng Su, Qing Xu, Xinming Zhu, Fubing Zhang, Xin Chen, and Yi Liu. Automatic building aggregation supported by knowledge graph. *Transactions in GIS*, 29(2):e70007, 2025. doi: 10.1111/tgis.70007.

Shigeo Takahashi, Ryo Kokubun, Satoshi Nishimura, Kazuo Misue, and Masatoshi Arikawa. Interactive optimization for cartographic aggregation of building features. *Computer Graphics Forum*, 43 (3):e15090, 2024. doi: 10.1111/cgf.15090.

Haitao Wang, Yongyang Xu, Anna Hu, Xuejing Xie, Siqiong Chen, and Zhong Xie. Building pattern recognition by using an edge-attention multi-head graph convolutional network. *International Journal of Geographical Information Science*, 39(4):732–757, 2025.

Haitao Wang, Yongyang Xu, Hao Li, Liang Wu, and Zhong Xie. Spatial context-aware weakly-supervised transfer learning: A framework for measuring the similarity of building shapes. *Transactions in GIS*, 30(4):e70314, 2026. doi: 10.1111/tgis.70314.

Yanning Wang and Yu Feng. Simplification and aggregation of vector building footprints using deep learning models. *Abstracts of the International Cartographic Association*, 7:184, 2024. doi: 10.5194/ica-abs-7-184-2024. Conference abstract.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xincheng Yang, Xukang Xie, and Dingming Liu. LA-GATs: A multi-feature constrained and spatially adaptive graph attention network for building clustering. *ISPRS International Journal of Geo-Information*, 14(11):415, 2025. doi: 10.3390/ijgi14110415.

Junbo Yu, Guanglei Pan, Yu Feng, Yi Xiao, Aji Gao, Li Liu, Huafei Yu, Tinghua Ai, and Dong Ren. A visual transformer and multi-task learning framework for building generalisation from raster maps. *International Journal of Digital Earth*, 19(2):2724180, 2026. doi: 10.1080/17538947.2026.2724180.

Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. Connecting the dots: Floorplan reconstruction using two-level queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Fubing Zhang, Qun Sun, Jingzhen Ma, Zheng Lyu, and Zhekun Huang. A polygonal buildings aggregation method considering obstacle elements and visual clarity. *Geocarto International*, 38 (1):2266672, 2023. doi: 10.1080/10106049.2023.2266672.

Qi Zhang, Suvam Bag, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Fuxun Yu. MARS – a foundational map auto-regressor. In *International Conference on Learning Representations*, 2026.

Zhiyong Zhou, Cheng Fu, and Robert Weibel. Move and remove: Multi-task learning for building simplification in vector maps with a graph convolutional neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:205–218, 2023. doi: 10.1016/j.isprsjprs.2023.06.004.

Zhiyong Zhou, Cheng Fu, and Robert Weibel. SpaGAN: A spatially-aware generative adversarial network for building generalization in image maps. *International Journal of Applied Earth Observation and Geoinformation*, 135:104236, 2024. doi: 10.1016/j.jag.2024.104236.

Zhiyong Zhou, Cheng Fu, Yu Feng, Guillaume Touya, Monika Sester, and Robert Weibel. GeoAI for map generalization in multi-scale cartography: Foundations, a research agenda, and interdisciplinary perspectives. *International Journal of Geographical Information Science*, 40(10): 3383–3411, 2026. doi: 10.1080/13658816.2026.2626935.

Stefano Zorzi and Friedrich Fraundorfer. Re:PolyWorld – a graph neural network for polygonal scene parsing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16762–16771, 2023. 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

APPENDIX

Overview. Appendix A describes dataset content, splits, and structural variation; Appendices B and C specify the model, training protocol, and geometry reconstruction; Appendix D defines the metrics and their aggregation; Appendices E and F extend the coverage, structure, and component analyses; Appendix G summarizes limitations.

A DATASET DESCRIPTION

A.1 SOURCE DATA AND SCENE CONTENT

The dataset pairs source buildings from the 1:10,000 (local) layer of OS Open Zoomstack with published generalized buildings from its 1:25,000 (district) layer over the same scene area. Roads from the same product provide source context. Coordinates use EPSG:27700 (British National Grid, metres). The layers are distributed under the Open Government Licence.[4]

Each scene covers a block enclosed by roads and contains at least two source buildings, with source-building coverage of at least 25%. Its target is the complete generalized building scene within the same extent, with one or multiple buildings and possible interior rings. Scenes are filtered to at most 768 source tokens and 256 target vertices. Boundary preprocessing uses tolerances of 0.3 m for straight segments and 0.5 m for circular rings, retaining other curves.

A.2 DATASET SPLITS AND CATEGORY COMPOSITION

The dataset contains 35,811 scene pairs: 26,373 training pairs from England, 4,458 validation pairs from Wales, and 4,980 test pairs from Scotland (Table 1; Figure 5(a)).

The four structural categories combine target building count (one or multiple) with the presence of interior rings. Table 1 gives category counts, and Figure 5(c) shows their proportions in each split.

The test references contain 154,426 vertices, 7,761 exterior rings, and 1,155 interior rings.

A.3 SCENE STRUCTURE AND PAIRED EXAMPLES

Figure 5(b) illustrates the changes in building count and boundary geometry that scene-level generalization must capture. Each category is represented by one test pair, showing aggregation, retained separation, or holes at a common input–target extent.

Across the three splits, source scenes contain 512,132 exterior rings and targets contain 67,367 exterior rings and 16,488 interior rings. In 35,080 of 35,811 pairs (98.0%), the target has fewer buildings; 711 pairs have equal counts and 20 have more target buildings (Figure 5(d)). Source graphs contain 18–764 tokens and targets contain 3–256 vertices (Figure 5(e)), spanning different input and output complexities within the fixed map-scale transition.

Table 5: Structural statistics over the 35,811 scene pairs, grouped by target category. Entries are scene medians except the explicitly marked mean hole count. Source vertices count building boundaries only; road tokens count directed traversal states and may repeat a geographic junction. Extent is the unaugmented source normalization factor a in Equation 9, in metres. Category counts by split are given in Table 1.

Target structure	Source polygons	Target polygons	Target holes (mean)	Source vertices	Target vertices	Road tokens	Extent (m)
Single, no holes	5	1	0.00	30	9	44	75.6
Single, with holes	16	1	1.38	78	25	49	112.2
Multiple, no holes	11	2	0.00	68	44	73	138.1
Multiple, with holes	26	3	1.45	147	93	99	177.5

[4]Contains OS data © Crown copyright and database right.

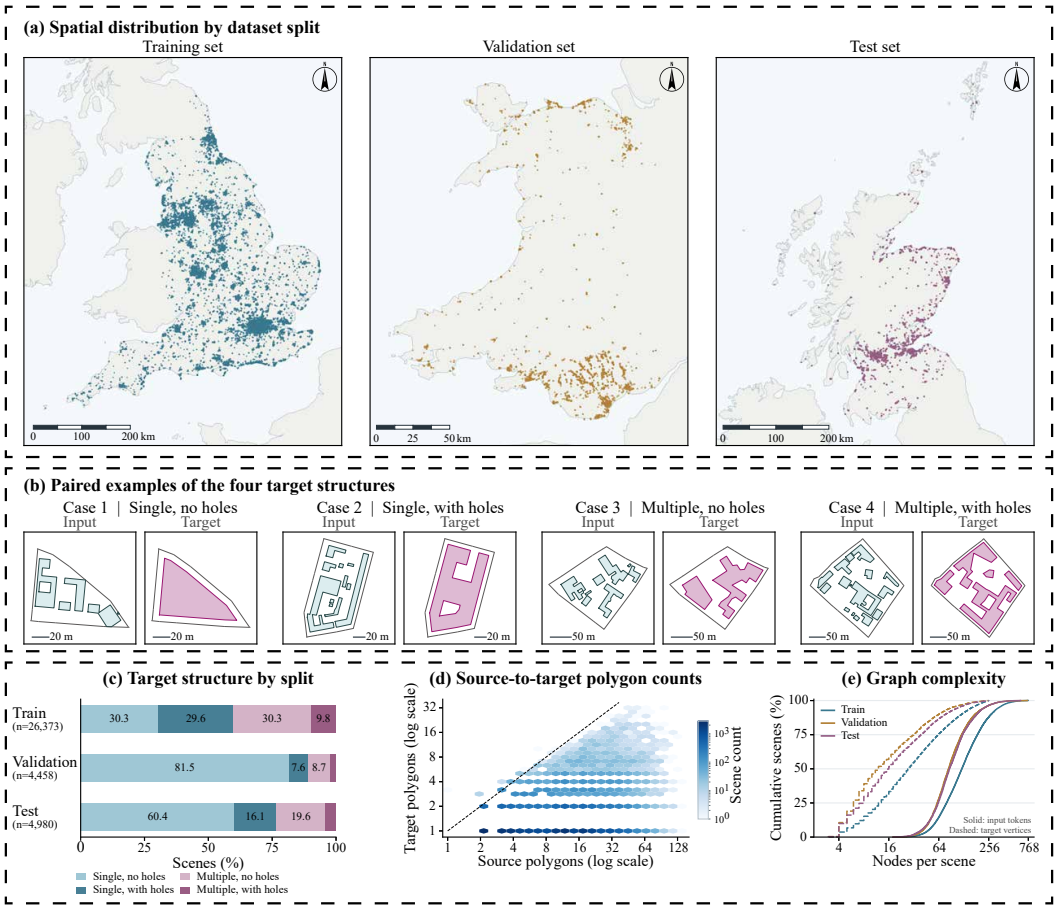


Figure 5: Paired vector scene dataset. (a) Geographic distribution of the three splits. (b) Four target structures illustrated by test pairs: pale teal denotes source buildings, pale rose target buildings, and dark gray roads. Input and target share the same extent in each case. (c) Category proportions in the training, validation, and test splits. (d) Source and target building counts over all 35,811 scene pairs; the dashed line denotes equal counts. (e) Cumulative distributions of input tokens (solid) and target vertices (dashed), by split.

Table 5 summarizes geometry and context within each category, pooling the three splits with equal weight per scene. Multiple-building scenes with holes have medians of 26 source buildings, 93 target vertices, and 99 road tokens, compared with 5, 9, and 44 for single-building scenes without holes. Median source extent also rises from 75.6 m to 177.5 m. These differences place the category-level results in Section 4.4 in context: building count, boundary complexity, and physical extent vary together.

B Model and Experimental Configuration

B.1 Source Tokens and Normalization

The eight-channel source representation is defined in Equation 3. Building tokens follow directed boundaries; road tokens follow directed traversal states. In the T-junction example in Figure 2, junction J occurs in both (A, J, B) and (C, J, J); repeated endpoint coordinates complete the triples. Normalization uses the complete source token set, including repeated junction coordinates.

For current source coordinates p_i, let $\mu = N^{-1} \sum_i p_i$ and let R be a sampled planar rotation during training, or the identity at evaluation. Define

$$a_0 = \max_{i,d} |[R(p_i - \mu)]_d|, \qquad a = \begin{cases} a_0, & a_0 > 0, \\ 1, & a_0 = 0, \end{cases} \qquad \bar{p} = \frac{0.9}{a} R(p - \mu). \tag{9}$$

All source coordinate triples and target vertices use this transformation. The feature s is computed from a using Equation 2, so extent is the normalization factor of the augmented source view. Target geometry never determines the center, extent, or scale feature. Removal of contextual tokens in an ablation occurs after normalization, preserving the input frame across configurations.

Target graphs store coordinates, ring classes, and directed successors. Exterior and interior classes are supervision metadata; there is no separate ring-class prediction head. Reordering polygons, changing ring starting vertices, or jointly permuting the coordinate and adjacency arrays leaves the represented target unchanged. Orientation reversal changes the exterior/interior interpretation under the adopted convention.

B.2 Architecture and optimization

Table 6 lists the CartoGenS configuration. Coordinate and attribute MLPs each map their inputs to 256 channels, and their concatenation is projected back to 256. Encoder and decoder feed-forward blocks have hidden width 512. The coordinate head is a three-layer MLP with GELU activations and a bounded $\tanh$ output.

The pairwise-score block f_{ij} in Figure 2(b) comprises the successor descriptor network and ScoreNet. Query states are projected to 192 channels, concatenated with the two predicted coordinate channels, and processed by six attention-based message-passing layers. ScoreNet evaluates ordered descriptor pairs with pointwise convolutional widths 256, 128, 64, and 1. The relative-geometry term g_{ij} linearly projects the five relative-geometry features in Equation 6 and is added to the final successor scores.

Each CartoGenS or ablation configuration trains on one NVIDIA GeForce RTX 5090. The inference environment uses Python 3.12.3, PyTorch 2.11.0 with CUDA 12.8, and NumPy 2.4.4.

Table 6: Model and training configuration of CartoGenS.

Component	Configuration
Input / encoder width d / attention heads	8 / 256 / 8
Encoder / query decoder / successor GNN layers	4 / 6 / 6
Candidate queries Q / successor descriptor width	256 / 192
Trainable parameters	9,918,601
Encoder and successor GNN attention	Standard attention, no added geometric bias
Relative-geometry term g_{ij}	Linear projection of five relative-geometry features
Optimizer / learning rate / weight decay	AdamW / 2×10^{-4} / 10^{-4}
Training epochs / batch size / seed	800 / 32 / 0
Warmup / schedule / gradient norm limit	1,000 steps / cosine decay / 5
Augmentation	Continuous paired planar rotation
Coordinate loss weight λ / Smooth-L_1 β	200 / 0.01
Successor row weights	1 matched, 0.1 unmatched
Sinkhorn iterations / temperature τ	100 / 2
Inference batch size / ordering	4 / fixed scene order

B.3 Objective reductions

For a batch of B scenes, let M_b be the target vertex count and π_b^* the injective coordinate assignment from Equation 7. With scalar Smooth-L_1 penalty ρ_β, the coordinate loss is

$$\mathcal{L}_{\text{coord}} = \frac{\sum_{b=1}^{B} \sum_{k=1}^{M_b} \sum_{d=1}^{2} \rho_\beta(\widehat{p}_{b,\pi_b^*(k),d} - p_{b,k,d})}{2 \sum_{b=1}^{B} M_b}. \tag{10}$$

It averages matched coordinate components over the batch and is zero if no vertices are matched. Only final coordinates receive point supervision. The per-scene successor losses in Equation 8 are averaged equally across scenes, and exterior and interior rows use the same foreground weight.

Log-domain Sinkhorn alternates row and column normalization for 100 iterations, then ends with row normalization. Row sums are one and column sums approximately one. The soft permutation objective follows Zorzi et al. (2022); non-self-loop successors indicate foreground membership.

B.4 Inference and checkpoint settings

CartoGenS and all ablations train for 800 epochs and use the final checkpoint; raster methods use final exponential-moving-average weights. Checkpoint selection does not use validation scores. Component ablations share widths, loss coefficients, and initialization: disabled projections are removed after constructing the full network.

B.5 Raster baselines and conversion

The three raster baselines, including the adapted segmentation architectures MSEONet and RoofX-Net, learn the paired transformation from binary building masks on a 256×256 grid. Their common 300-epoch training schedule uses Adam with learning rate 4×10^{-4}, no weight decay, batch size 8, cosine decay to 10^{-8}, exponential moving average 0.9992, gradient norm clipping at 20, seed 0, BF16 training, and no early stopping. Continuous rotation and horizontal/vertical flips are shared across these baselines. This schedule replaces the training schedules of the original methods.

MSEONet[5] uses a ResNet101-V1c backbone, multiscale context aggregation, and boundary-point refinement. RoofX-Net[6] uses its MobileNetV2 configuration, ASPP, and the published multibranch edge objectives. Both start from random initialization. SAM-BG (Yu et al., 2026) uses the authors' adapted SAM ViT-B with pretrained weights, LoRA, boundary feature enhancement, and a multitask decoder.

The raster grid is a square centered on the source-token bounding box. Its side length is 1.1 times the larger coordinate span, adding a 5% margin on each side; source and target share these bounds. Binary predictions use $\text{probability} > 0.5$ and are converted to boundary rings by four-connected polygonization. Saved bounds restore polygon coordinates for the common vector-domain evaluation (Appendix C). Vector predictions use the same grid, with a pixel marked as building when its center lies inside a predicted polygon.

B.6 Vector baselines

GPT-6-Astra-High. This baseline uses the gpt-6-astra model identifier with high reasoning effort, treating each scene independently without weight updates on the paired-scene dataset. Each request contains source buildings and road segments as JSON, a coordinate-to-metre conversion factor, an input preview, and reference example pairs from the training set, with previews. Exterior rings are counterclockwise and holes clockwise; coordinates are scaled to approximately $[-1000, 1000]$.

The shared prompt requests closed, non-self-intersecting building rings with holes strictly inside one exterior, excludes roads from the output, and imposes no fixed vertex budget. It prohibits tools, generic geometry algorithms, and reuse of other test predictions. Parsed polygon rings enter the common vector-domain evaluator. The 17 failed responses among 4,980 test scenes count as empty outputs.

ArcGIS Pro. The workflow runs per scene in ArcGIS Pro 3.6 with two tools from its Cartography toolbox. *Aggregate Polygons* combines polygons whose boundaries lie within the aggregation distance into new polygons and does not aggregate across barrier features. *Simplify Building* simplifies building boundaries by removing small details while maintaining their essential shape and size. In our configuration, *Aggregate Polygons* merges source buildings with an aggregation distance of 25 m, a minimum hole size of 25 m^2, no minimum area, non-orthogonal output, and the source roads as barrier features. *Simplify Building* then simplifies the aggregates with a 5 m tolerance, no minimum area, and conflict checking against the same barriers. Outputs are exported in the original coordinates for the common vector-domain evaluation.

[5] https://github.com/HuangWBill/MSEONet

[6] https://github.com/K-Y-Chen/RoofX-Net

C SUCCESSOR ASSIGNMENT AND GEOMETRY RECONSTRUCTION

At inference, the successor scores S in Figure 2(b) determine connections between predicted vertices by maximizing the total connection score:

$$\widehat{A} = \arg\max_{A \in \mathcal{P}_Q} \langle A, S \rangle, \qquad \mathcal{P}_Q = \{A \in \{0,1\}^{Q \times Q} : A\mathbf{1} = \mathbf{1}, A^\top \mathbf{1} = \mathbf{1}\}. \tag{11}$$

The assignment uses raw logits and requires no target geometry. Each query has one predecessor and one successor, so the permutation decomposes into disjoint cycles. Self-loops mark unused candidates, two-cycles are discarded, and longer cycles define candidate building boundaries.

C.1 RELATIVE-GEOMETRY SCORING

The implementation uses all five features in Equation 6. For any complete successor permutation σ, the displacement components Δx_{ij} and Δy_{ij} sum to zero and the intercept sums to a constant:

$$\sum_{i=1}^{Q} (\widehat{p}_i - \widehat{p}_{\sigma(i)}) = 0, \qquad \sum_{i=1}^{Q} b = Qb. \tag{12}$$

These terms therefore add the same value to every assignment score in Equation 11, so absolute coordinate differences and Euclidean distance determine the relative-geometry term's effect on hard connection selection. The features are translation invariant, and rotation augmentation exposes the model to different scene orientations. This equivalence applies to hard decoding. During training, all five features enter the finite-iteration Sinkhorn objective, and successor supervision backpropagates through the successor scores and Sinkhorn normalization to update vertex positions and descriptors jointly.

C.2 REGION RECONSTRUCTION

For a cycle $C = (v_1, \ldots, v_n)$, the signed area is

$$a(C) = \frac{1}{2} \sum_{i=1}^{n} (v_{i,x} v_{i+1,y} - v_{i+1,x} v_{i,y}), \qquad v_{n+1} = v_1. \tag{13}$$

Positive and negative signs identify exterior and interior candidates, respectively.

For vector-domain evaluation, boundary rings from every compared method and ablation configuration follow the same region-reconstruction procedure. Each ring defines a polygonal region $\Omega(C)$; zero-width buffering (`buffer(0)`) converts invalid rings into valid regions, and empty or zero-area results are omitted. Exterior regions are unioned and interior regions subtracted to obtain

$$\widehat{\Omega} = \left(\bigcup_{C \in \widehat{\mathcal{R}}_{\text{ext}}} \Omega(C) \right) \setminus \left(\bigcup_{C \in \widehat{\mathcal{R}}_{\text{int}}} \Omega(C) \right). \tag{14}$$

Its connected polygon components define the predicted buildings used in the metrics. Reference geometry retains original exterior identities and assigns each interior ring to its unique strictly containing exterior.

D EVALUATION DEFINITIONS

D.1 POLYGON CORRESPONDENCE AND SHAPE

For predicted and reference building polygons $\widehat{B}_i$ and B_j in scene s, define

$$J_{ij} = \frac{|\widehat{B}_i \cap B_j|}{|\widehat{B}_i \cup B_j|}, \qquad \mathcal{M}_s = \arg\max_{\text{one-to-one matchings}} \sum_{(i,j) \in \mathcal{M}_s} J_{ij}. \tag{15}$$

Let $\mathcal{M}_s^+ \subseteq \mathcal{M}_s$ contain the pairs with $J_{ij} > 0$, $\mathcal{M}_+ = \bigcup_s \mathcal{M}_s^+$, and $N_{\text{ref}} = 7{,}761$ be the number of reference buildings in the test split. With $R(\cdot)$ the minimum-area enclosing rectangle,

$$\text{IoU}_{\text{poly}} = \frac{1}{N_{\text{ref}}} \sum_s \sum_{(i,j)\in\mathcal{M}_s} J_{ij}, \qquad \text{Cov} = \frac{|\mathcal{M}_+|}{N_{\text{ref}}}, \tag{16}$$

$$\text{IoU}_{\text{SBR}} = \frac{1}{N_{\text{ref}}} \sum_s \sum_{(i,j)\in\mathcal{M}_s^+} \frac{|R(\widehat{B}_i) \cap R(B_j)|}{|R(\widehat{B}_i) \cup R(B_j)|}. \tag{17}$$

Missing reference buildings contribute zero to both IoU sums. SBR IoU describes extent and orientation with reduced sensitivity to boundary detail. Extra predicted buildings enter PQ as false positives.

TFD compares accumulated boundary directions over normalized arc length (Arkin et al., 1991). Both exteriors are oriented counterclockwise and resampled at $K = 100$ equally spaced positions. Let $\widehat{\theta}_t$ be the unwrapped cumulative direction of the t-th chord on the predicted exterior, and $\theta_t^{(k)}$ the corresponding direction on the reference exterior after a cyclic shift by k positions. Then

$$\text{TFD}(\widehat{B}_i, B_j) = \min_{0\le k<K} \Big(\frac{1}{K} \sum_{t=1}^{K} \big(\widehat{\theta}_t - \theta_t^{(k)} - \bar{\alpha}_k\big)^2 \Big)^{1/2}, \qquad \bar{\alpha}_k = \frac{1}{K} \sum_{t=1}^{K} \big(\widehat{\theta}_t - \theta_t^{(k)}\big). \tag{18}$$

The reported TFD averages over $\mathcal{M}_+$ in radians and measures exterior shape independently of absolute position and scale; holes are evaluated separately. Coverage and the empty-output rate describe which outputs contribute.

D.2 Building counts, delineation, and interior regions

For scene s, let $\widehat{n}_s$ and n_s count predicted building components after reconstruction and reference buildings, respectively. Over all $N = 4{,}980$ test scenes,

$$\text{MAE}_{\text{count}} = \frac{1}{N} \sum_s |\widehat{n}_s - n_s|, \qquad \text{Acc}_{\text{count}} = \frac{1}{N} \sum_s \mathbf{1}[\widehat{n}_s = n_s], \tag{19}$$

$$\text{Empty} = \frac{1}{N} \sum_s \mathbf{1}[\widehat{n}_s = 0]. \tag{20}$$

These measures assess scene cardinality independently of spatial overlap; PQ additionally evaluates instance agreement.

PQ uses an independent one-to-one matching that first maximizes the number of pairs with $J_{ij} > 0.5$ and then their total IoU (Kirillov et al., 2019). With matched pairs TP and unmatched predicted and reference polygons FP and FN,

$$\text{PQ} = \frac{\sum_{(i,j)\in\text{TP}} J_{ij}}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}. \tag{21}$$

Counts and overlap sums are aggregated over the split. PQ penalizes missed and extra buildings as well as overlap error; it evaluates target delineation, not source-building group membership.

Let U_P^s and U_T^s be the unions of predicted and reference buildings in scene s, and H_P^s and H_T^s the unions of their interior regions. With $\mathcal{S}_h$ the 998 test scenes with reference holes and $\mathcal{S}_0$ the other 3,982 scenes,

$$\text{IoU}_{\cup} = \frac{1}{N} \sum_s \frac{|U_P^s \cap U_T^s|}{|U_P^s \cup U_T^s|}, \qquad E_{\text{area}} = \frac{1}{N} \sum_s \frac{\big||U_P^s| - |U_T^s|\big|}{|U_T^s|}, \tag{22}$$

$$\text{IoU}_{\text{hole}} = \frac{1}{|\mathcal{S}_h|} \sum_{s\in\mathcal{S}_h} \frac{|H_P^s \cap H_T^s|}{|H_P^s \cup H_T^s|}, \qquad R_{\text{sp}} = \frac{1}{|\mathcal{S}_0|} \sum_{s\in\mathcal{S}_0} \mathbf{1}[H_P^s \neq \emptyset]. \tag{23}$$

An empty output gives union IoU zero and area error one, and a missed interior region gives hole IoU zero. Matched centroid displacement averages $\|c(\widehat{B}_i) - c(B_j)\|_2$ over $\mathcal{M}_+$ after restoring coordinates to metres.

D.3 Raster measures and aggregation

Raster measures compare binary prediction and target masks P_s and T_s on the shared grids. Pixel counts $\mathrm{TP}_{\mathrm{px}}$, $\mathrm{FP}_{\mathrm{px}}$, and $\mathrm{FN}_{\mathrm{px}}$ are summed over the split, and $\partial_w M = M \setminus \mathrm{erode}_w(M)$ is the inner boundary band obtained by w iterations of 3×3 erosion, treating the exterior of the canvas as background:

$$\mathrm{IoU}_{\mathrm{ras}} = \frac{\mathrm{TP}_{\mathrm{px}}}{\mathrm{TP}_{\mathrm{px}} + \mathrm{FP}_{\mathrm{px}} + \mathrm{FN}_{\mathrm{px}}}, \qquad \mathrm{F1} = \frac{2\,\mathrm{TP}_{\mathrm{px}}}{2\,\mathrm{TP}_{\mathrm{px}} + \mathrm{FP}_{\mathrm{px}} + \mathrm{FN}_{\mathrm{px}}}, \tag{24}$$

$$\mathrm{BIoU} = \frac{\sum_s |\partial_w P_s \cap \partial_w T_s|}{\sum_s |\partial_w P_s \cup \partial_w T_s|}, \qquad w = \max\left(1, \mathrm{round}(0.02\sqrt{H^2 + W^2})\right). \tag{25}$$

BIoU follows Cheng et al. (2021) with pooled band counts. SSIM averages over scenes using a Gaussian window with $\sigma = 1.5$, data range 1, and population covariance on the binary masks (Wang et al., 2004).

Aggregation follows the evaluation unit of each metric: scene means weight scenes equally, reference-normalized measures use the number of reference buildings, and pooled measures accumulate counts before division. Category scores apply the same definitions within each structural category. The category-balanced summary then averages the four category scores equally; it is reported separately from the overall test score.

E Supplementary quantitative results

The supplementary results report CartoGenS by target structure, compare methods on building counts and hole recovery, then examine coverage and component behavior. Metrics follow Appendix D; Full uses the same final-epoch predictions as the main comparison and ablation study.

E.1 Results of CartoGenS by target structure

Table 7 reports CartoGenS by target structure, adding raster IoU, BIoU, and hole IoU to the polygon IoU and PQ in Table 3. Hole IoU is 0.3494 for single buildings with holes and 0.2051 for multiple buildings with holes. Boundary agreement varies less across structures than instance recovery. For scenes without and with holes, BIoU moves from 0.7034 and 0.6686 for single buildings to 0.5788 and 0.6028 for multiple buildings, and raster IoU stays above 0.72, whereas PQ falls from 0.8109 and 0.7898 to 0.4238 and 0.3671. Multiple-building outputs thus retain much of the outline quality of single-building outputs, and their lower PQ arises mainly from dividing the occupied region into the correct buildings.

Table 7: CartoGenS results by target structure. Overall pools all 4,980 scenes using each metric's denominator (Appendix D).

Target structure	Scenes	Poly. IoU ↑	PQ ↑	Raster IoU ↑	BIoU ↑	Hole IoU ↑
Single, no holes	3,007	0.8253	0.8109	0.8535	0.7034	–
Single, with holes	800	0.8242	0.7898	0.8274	0.6686	0.3494
Multiple, no holes	975	0.4531	0.4238	0.7327	0.5788	–
Multiple, with holes	198	0.4114	0.3671	0.7252	0.6028	0.2051
Overall	4,980	0.6319	0.6155	0.8220	0.6606	0.3208

Hole IoU is computed over the 998 scenes with reference holes, and missed regions score zero; – marks categories without reference holes.

E.2 Building counts and holes across methods

CartoGenS has the lowest count MAE, 0.3534, and the lowest spurious-hole rate, 0.1057 (Table 8). The raster methods reach hole IoU of 0.3400–0.3611 but have spurious-hole rates of 0.2572–0.5829. Among vector methods, CartoGenS has the highest hole IoU, 0.3208, compared with 0.2946 for GPT-6-Astra-High and 0.2632 for ArcGIS Pro. For single buildings with holes, its hole IoU is 0.3494, below the three raster methods.

Table 8: Building-count and hole comparison. Count MAE and accuracy use all 4,980 scenes; hole IoU uses 998 scenes with reference holes, of which 800 are single-building scenes, and spurious-hole rate uses the other 3,982 scenes.

Method	Count MAE ↓	Count accuracy ↑	Hole IoU ↑	Single-building hole IoU ↑	Spurious holes ↓
MSEONet	0.4104	0.7267	**0.3611**	**0.3925**	0.3453
RoofX-Net	<u>0.3878</u>	0.7518	<u>0.3538</u>	<u>0.3711</u>	0.2572
SAM-BG	1.3289	0.5241	0.3400	0.3614	0.5829
ArcGIS Pro	0.3890	**0.7861**	0.2632	0.2768	<u>0.1570</u>
GPT-6-Astra-High	0.3924	<u>0.7711</u>	0.2946	0.3033	0.1730
CARTOGENS (ours)	**0.3534**	0.7594	0.3208	0.3494	**0.1057**

Bold: best; <u>underline</u>: second best in each metric column.

E.3 COVERAGE AND GEOMETRIC AGREEMENT

Full reaches a union IoU of 0.8073 with an absolute relative area error of 0.1705 (Table 9). Positive-overlap matches cover 82.21% of the 7,761 reference buildings with a mean centroid displacement of 8.60 m, and only 3.86% of scenes have empty output. On the shared raster grids, F1 is 0.9023 and SSIM is 0.9125.

Table 9: Additional quality measures for Full at its final training epoch.

Measure	Evaluation population	Value
Union IoU / area error	4,980 scenes	0.8073 / 0.1705
Match coverage	7,761 reference buildings	0.8221
Matched centroid displacement	Positive-overlap matches	8.6037 m
Hole IoU	998 reference-with-hole scenes	0.3208
Spurious-hole rate	3,982 reference-without-hole scenes	0.1057
Empty-output rate	4,980 scenes	0.0386
Raster F1 / SSIM	Shared binary test grids	0.9023 / 0.9125

E.4 COMPLETE COMPONENT RESULTS

All eight configurations follow the shared training and inference protocol in Appendix B. Tables 10–13 provide the diagnostics, category scores, and exact contrasts underlying the main-text analysis and Figure 4.

Relative to Base + road, adding scale and relative-geometry scoring raises match coverage from 0.6957 to 0.8221 and hole IoU from 0.1205 to 0.3208, and reduces the empty-output rate from 0.1940 to 0.0386 (Table 10). The two components thus help recover scene content as well as refine geometry.

Table 10: Additional ablation diagnostics. Coverage is positive-overlap polygon matching divided by 7,761 reference buildings; TFD (radians) is conditional on those matches. Hole IoU averages over 998 reference-with-hole scenes. Empty-output rate uses all 4,980 scenes.

Configuration	SBR ↑	TFD ↓	Coverage ↑	Hole IoU ↑	Empty ↓
Base	0.1309	0.8288	0.3576	0.0314	0.6092
Base + road	0.5409	0.5679	0.6957	0.1205	0.1940
Base + scale	0.1731	0.8027	0.3962	0.0423	0.5976
Base + score bias	0.1514	0.8222	0.3875	0.0305	0.5701
Full − road	0.1743	0.7858	0.3960	0.0524	0.6050
Full − scale	<u>0.6222</u>	0.5335	<u>0.7834</u>	0.2011	<u>0.1002</u>
Full − score bias	0.6001	<u>0.5316</u>	0.7524	<u>0.2297</u>	0.1114
CARTOGENS (full)	**0.6666**	**0.5087**	**0.8221**	**0.3208**	**0.0386**

Bold: best; <u>underline</u>: second best in each metric column.

Table 11: Empty-output rate by target structure. Columns contain 3007, 800, 975, and 198 test scenes.

Configuration	Single, no holes	Single, with holes	Multiple, no holes	Multiple, with holes
Base	0.7273	0.5975	0.3549	0.1162
Base + road	0.2820	0.0725	0.0615	0.0000
Base + scale	0.7423	0.5663	0.2851	0.0657
Base + score bias	0.6774	0.5813	0.3272	0.0909
Full − road	0.7566	0.5813	0.2656	0.0707
Full − scale	0.1520	0.0238	0.0236	0.0000
Full − score bias	0.1736	0.0088	0.0267	0.0000
CARTOGENS (full)	0.0635	0.0000	0.0010	0.0000

Full remains best among the eight configurations when the four categories receive equal weight (polygon IoU/PQ: 0.6285/0.5979), so its aggregate lead also holds when rare structures receive the same weight as common ones. Scale gives the main category-level exception: removing it raises polygon IoU and PQ for multiple buildings with holes from 0.4114 and 0.3671 to 0.4260 and 0.3911, and slightly raises PQ for multiple buildings without holes (Table 12). In eight of the nine comparisons in Table 13, all but PQ for scale, removing a factor from Full changes the score more than adding it to Base. Because more than half of the Base outputs are empty (Table 11), removal contrasts from Full give the more informative estimate of each contribution.

Table 12: Polygon IoU / PQ by target structure. Columns contain 3,007, 800, 975, and 198 test scenes, respectively. Scores retain their reference-building and pooled-count denominators within each category.

Configuration	Single, no holes	Single, with holes	Multiple, no holes	Multiple, with holes
Base	0.0597 / 0.0431	0.1297 / 0.0988	0.1530 / 0.1089	0.1662 / 0.0683
Base + road	0.6037 / 0.6670	0.7063 / 0.6933	0.3791 / 0.3564	0.3823 / 0.3255
Base + scale	0.0832 / 0.0911	0.1729 / 0.1522	0.2034 / 0.1569	0.2394 / 0.1138
Base + score bias	0.0835 / 0.0640	0.1450 / 0.1136	0.1658 / 0.1200	0.1808 / 0.0710
Full − road	0.0729 / 0.0743	0.1638 / 0.1494	0.2262 / 0.1834	0.2322 / 0.1047
Full − scale	0.7151 / 0.7242	0.7591 / 0.7209	0.4483 / **0.4246**	**0.4260 / 0.3911**
Full − score bias	0.7180 / 0.7513	0.7987 / 0.7654	0.3990 / 0.3690	0.3798 / 0.3224
CARTOGENS (full)	**0.8253 / 0.8109**	**0.8242 / 0.7898**	**0.4531** / 0.4238	0.4114 / 0.3671

Bold: best; underline: second best for each metric within each category.

Table 13: Factor effects in score units. Addition compares a single-factor model with Base; removal subtracts the corresponding disabled-factor model from Full. Positive differences favor including the factor.

Factor	Contrast	Δ Poly. IoU	Δ PQ	Δ BIoU
Road context	Addition	+0.3845	+0.4321	+0.4126
Road context	Removal	+0.4710	+0.4810	+0.5235
Scene scale	Addition	+0.0412	+0.0496	+0.0290
Scene scale	Removal	+0.0502	+0.0417	+0.0784
Score bias	Addition	+0.0174	+0.0139	+0.0138
Score bias	Removal	+0.0698	+0.0549	+0.0515

F QUALITATIVE ANALYSIS

Figures 6–8 illustrate component effects in selected test scenes. Each figure compares source, reference, Base, Base plus one component, Full minus that component, and Full. Paired configurations differ in one component. The displayed predictions retain their original boundaries, with a common geographic extent within each row. Full-test effects are reported in Tables 10–13.

F.1 ROAD CONTEXT

Omission and merging reveal two ways to lose scene organization (Figure 6). Omission removes a required building, while merging fills a gap that should separate buildings. Without road tokens, these errors occur in Cases 2–4; road-conditioned outputs recover two buildings, although Base + road still has shape and hole errors. The central road in Case 1 makes the separation error explicit: Full − road joins buildings across it, whereas Base + road and Full preserve the gap.

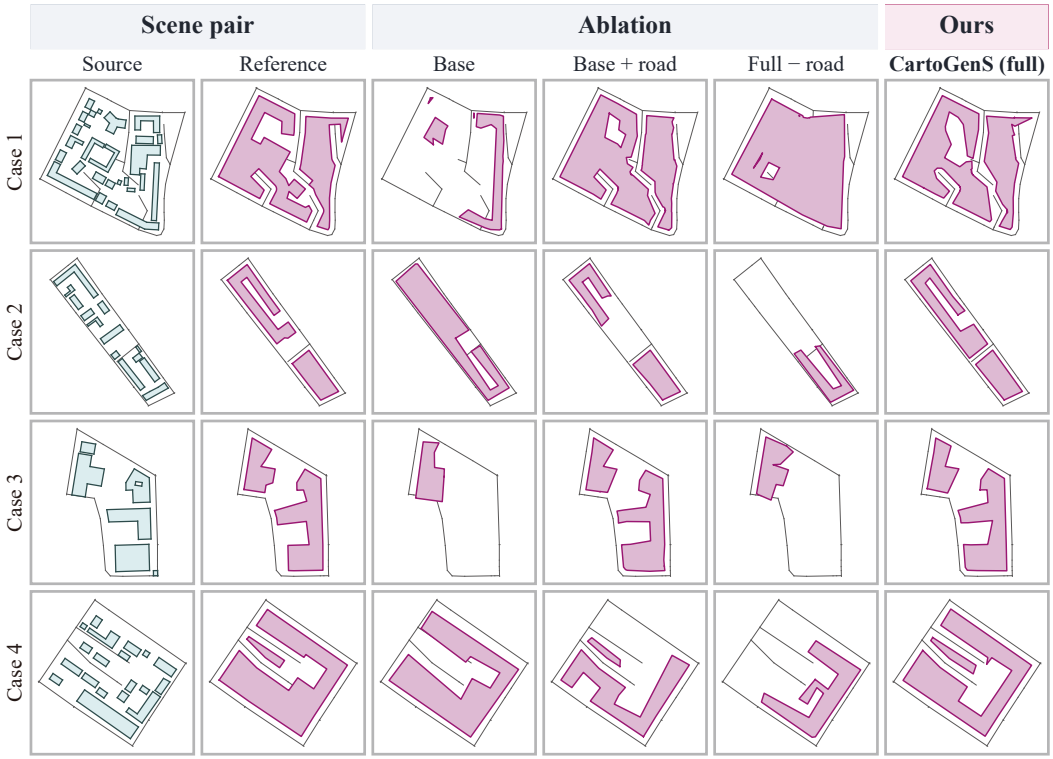


Figure 6: Road-context ablation. Adding roads to Base and retaining them in Full improve building separation and completeness in these four test scenes. Roads are drawn in every panel for spatial reference, including configurations that do not receive road inputs.

F.2 SOURCE-SCENE SCALE

Extent recovery and local detail respond differently to scale (Figure 7). Adding scale to Base recovers the elongated region in Case 1 and much of the missing extent in Cases 3–4; removing it from Full shrinks the right-hand building in Case 3 and deepens the indentation in Case 4. These changes suggest that physical extent informs how much of a shape is retained. Local concavities are recovered less consistently: Base + scale and Full − scale both fill the reference concavity in Case 2, while Full preserves it more closely.

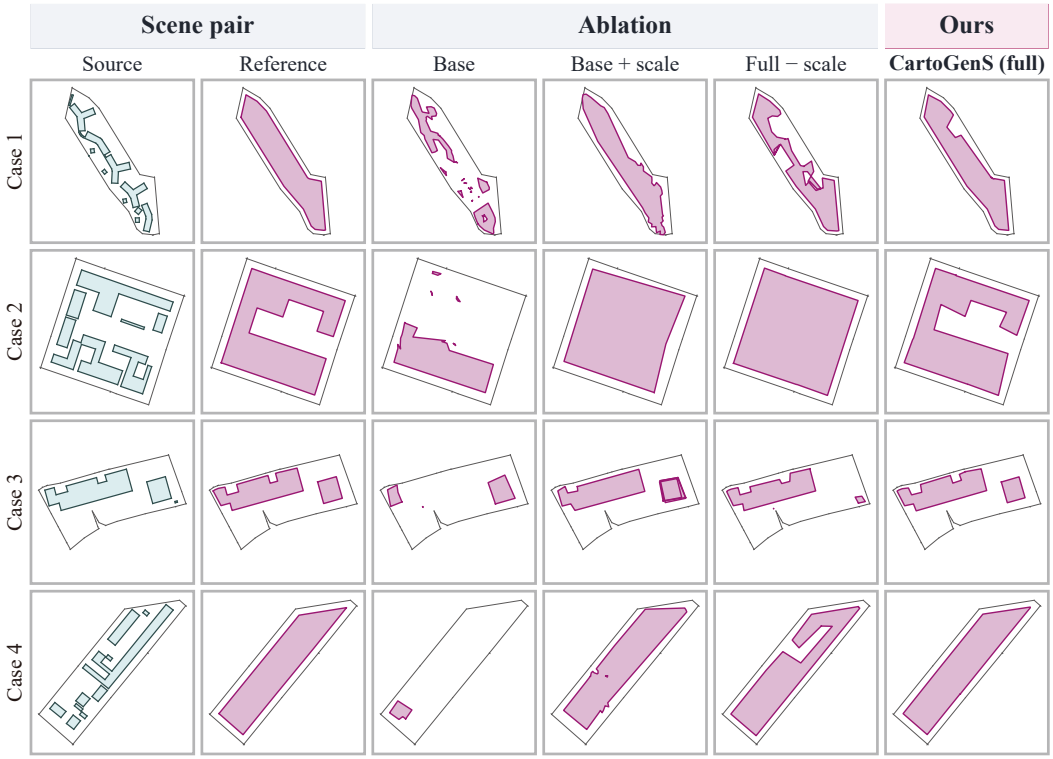


Figure 7: Source-scene scale ablation. The paired comparisons show changes in building coverage, concavities, and local extent. The scale feature describes the source scene; all configurations are evaluated on the same map-scale transition.

F.3 RELATIVE-GEOMETRY SCORING

Connection errors can change both local boundary shape and the number of separate buildings (Figure 8). Including the relative-geometry term avoids the crossings in Case 1; in Case 4, Full recovers four separate buildings, compared with three after removal. Yet a closed boundary can still encode the wrong structure. Base + score bias encloses an unwanted hole in Case 2, while Full keeps the reference concavity open. In Case 3, Full − score bias closes the concavity, and Full also fills much of it. These cases separate geometric compatibility from the decision about which spaces should remain open, consistent with the larger PQ gain of relative-geometry scoring within Full than within Base (+0.0549 against +0.0139; Table 13).

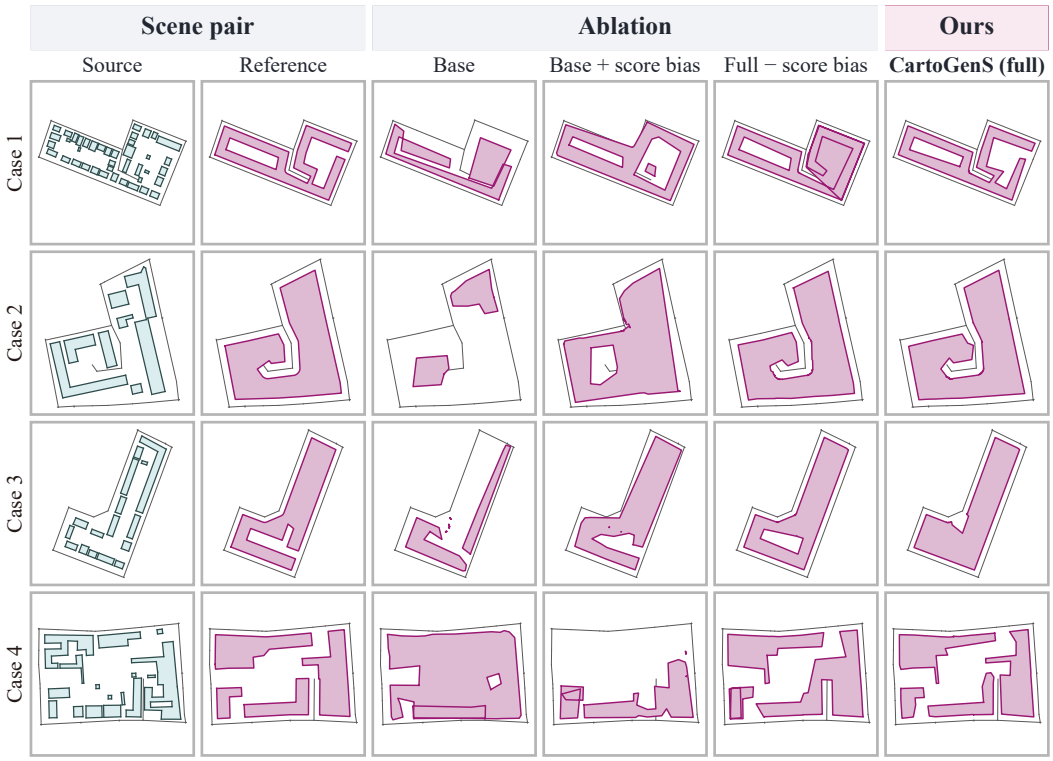


Figure 8: Relative-geometry scoring ablation. The comparisons expose crossing connections, unintended closures, and changes in building separation. Base + score bias and Full include the relative-geometry term; Base and Full − score bias omit it.

G SCOPE AND LIMITATIONS

The evaluation covers one paired map-scale transition and road-bounded scenes meeting the building-coverage and capacity criteria in Appendix A. Sparse scenes, larger outputs beyond the 256-query budget, and other target scales remain untested. England, Wales, and Scotland share a map product; transfer to different building patterns and cartographic conventions requires further evaluation. Ring formation does not enforce geometric validity; all methods are evaluated on the regions produced by the shared reconstruction in Appendix C.

Multiple-building scenes with holes have lower instance and hole recovery than single-building scenes with holes (Table 7). Improving the joint recovery of separate buildings and their holes remains a priority for complex scene generation.