MARS

A Foundational Map Auto-Regressor

Qi Zhang*, Suvam Bag*, Rupanjali Kukal*, Mikael Figueroa,
Rishi Madhok, Nikolaos Karianakis, Fuxun Yu

Microsoft

ICLR
International Conference On
Learning Representations

Motivation: Challenges in Map Generation

Complex Multi-Stage Pipeline Design

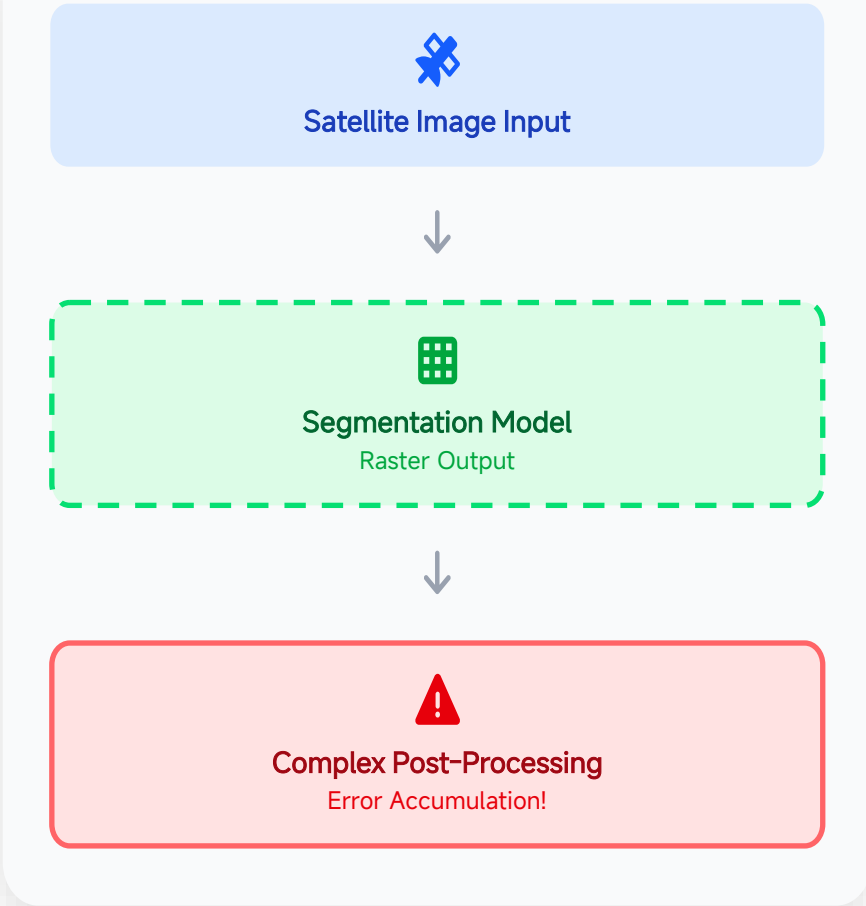
Current approaches use **segmentation + post-processing**, causing error accumulation at each stage

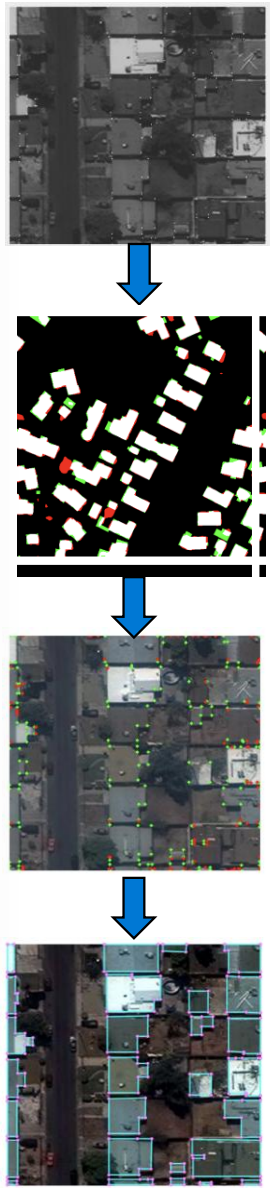
Single Map Feature Support

Existing methods only handle **roads OR buildings**, not both in a unified model

Heuristic Post-Processing Heavy

Post-processing logic differs substantially between feature classes with many hyperparameters





Motivation: Challenges in Map Generation

Our Insight

Treat vectorized map primitives as a **formal language** and use auto-regressive sequence-to-sequence learning

Sequence-to-Sequence Map Learning

1. **Map-to-Sequence Conversion:** Convert graph-structured map data into sequential tokens
2. **Auto-Regressive Learning:** Train model to predict next token given previous tokens
3. **Direct Vector Output:** Generate vectorized map data without any post-processing

Key Advantage

End-to-end learning **without any post-processing** for both roads and buildings

Our Contributions

1

Unified End-to-End Architecture

First foundational map auto-regressive model (**MARS**) that generates both roads and buildings in a single model without any post-processing

2

Chat with MARS

Novel **interactive human-in-the-loop** capability enabling users to prompt MARS with starting points for missing objects, correct drifting predictions, and add overlooked elements

3

MAP-3M Dataset

Curated the **largest multi-class map dataset** to date with 3M images —10× larger than existing datasets, with both road and building annotations

Method: Map-to-Sequence Conversion

Road Network Serialization

Before: Flat Graph

After: Sequential Polylines

[R, x_1^1,y_1^1, x_2^1,y_2^1,, ..., R, x_1^2,y_1^2, ..., ..., R, x_1^M,y_1^M, ...]

→ Stroke-based decomposition enables sequential representation

Basic Types

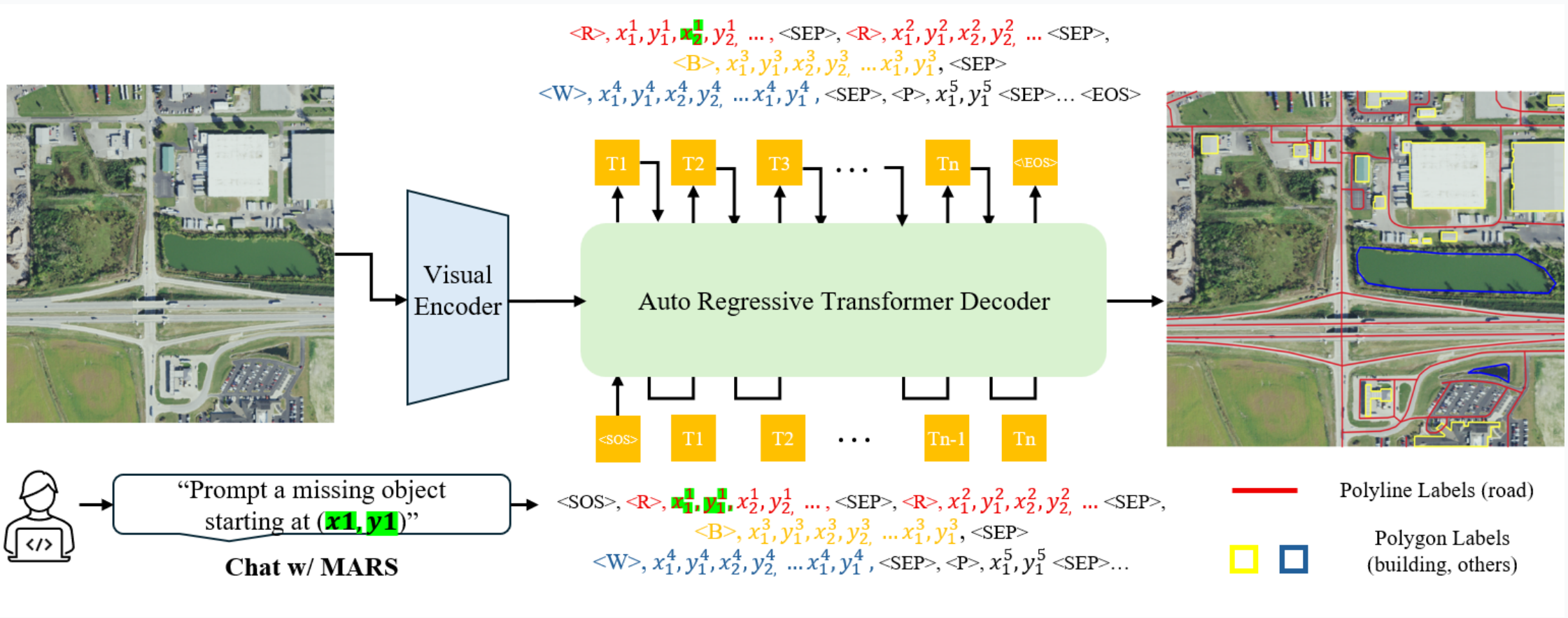
- **Point:** [x, y]
- **Polyline:** [x_1,y_1, x_2,y_2, ...]
- **Polygon:** [x_1,y_1, x_2,y_2, ..., x_1,y_1]

Unified Sequence

P, x_1^0,y_1^0, ..., B, x_1^1,y_1^1, x_2^1,y_2^1, ..., x_1^1,y_1^1, ..., R, x_1^i,y_1^i, x_2^i,y_2^i, ..., B, x_1^N,y_1^N, x_2^N,y_2^N, ..., x_1^N,y_1^N

Result: All map elements as sequences for **AR learning**

Method: MARS Architecture



Chat with MARS: Human-in-the-Loop

1 Start-of-Sequence

Provide starting point for better conditioning

[<SOS>, B, $\bar{x}_1^1$, $\bar{y}_1^1$, x_2^1, y_2^1, x_3^1, y_3^1, ..., x_1^1, y_1^1, ..., <EOS>]

B = User-provided category class token
$\bar{x}_1^1$, $\bar{y}_1^1$ = User-click coordinates
x_2^1, y_2^1, ... = Auto-regressively generated subsequent vertices

Prevents error accumulation

2 Mid-of-Sequence

Correct drifting predictions

[<SOS>, R, x_1^1, y_1^1, x_2^1, y_2^1, ~~$\bar{x}\_old$, $\bar{y}\_old$~~, $\bar{x}_3^1$, $\bar{y}_3^1$, x_4^1, y_4^1, ..., <EOS>]

x_1^1, y_1^1, x_2^1, y_2^1 = Previously generated (correct) tokens
$\bar{x}_3^1$, $\bar{y}_3^1$ = User-click to replace drifting vertex
x_4^1, y_4^1, ... = Subsequent tokens generated from new trajectory

Corrects trajectory drift

3 End-of-Sequence

Add missing objects

[<SOS>, R, x_1^1, y_1^1, x_2^1, y_2^1, ..., ~~<EOS>,~~ B, $\bar{x}_1^1$, $\bar{y}_1^1$, x_2^1, y_2^1, ..., <EOS>]

B, $\bar{x}_1^1$, $\bar{y}_1^1$ = New category object prompted
x_2^1, y_2^1, ... = Auto-completed building vertices
Final <EOS> = True end of extended sequence

Improves recall

MAP-3M: Large-Scale Dataset

Dataset Comparison

Table 1: MAP-3M provides 10× more images, and 100× more spatial coverages than literature datasets. GSD: meter/pixel. ∗: Cityscale contains 2K chips, we tile to 224x224 for comparison.

Dataset	# Images	Image Size	Coverage Area	GSD	Building Cls	Road Cls
Cityscale (He et al., 2020a)	49220	224x224*	2470 km^2	1.0	×	✓
SpaceNet3 (SpaceNet, 2018)	2541	400x400	407 km^2	1.0	×	✓
AICrowd (AICrowd, 2020)	258044	300x300	2090 km^2	0.3	✓	×
MAP-3M (Ours)	**∼3M**	512x512	**294069** km^2	0.6	✓	✓

Dataset Details

Source
NAIP aerial imagery via Microsoft Planetary Computer

Coverage
5,000 cities across 50 US states (2020–2024)

Resolution
512×512 chips at 0.6m/pixel GSD

Results: Road Network and Building Extraction

TOPO Metrics Comparison

Table 3: TOPO-based road performance comparison on Cityscale and SpaceNet.

Model	CITYSCALE			SPACENET		
	P	R	F1	P	R	F1
Seg-UNet (Ronneberger et al., 2015)	75.34	65.99	70.36	68.96	66.32	67.61
Seg-DRM (Máttyus et al., 2017)	76.54	71.25	73.80	82.79	72.56	77.34
Seg-Improved (Batra et al., 2019)	75.83	68.90	72.20	81.56	71.38	76.13
Seg-DLA (Yu et al., 2018)	75.59	72.26	73.89	78.99	69.80	74.11
RoadTracer (Bastani et al., 2018)	78.00	57.44	66.16	78.61	62.45	69.90
Sat2Graph (He et al., 2020b)	80.70	72.28	76.26	85.93	76.55	80.97
TD-Road (He et al., 2022)	81.94	71.63	76.43	84.81	77.80	81.15
RNGDet (Xu et al., 2022)	85.97	69.78	76.87	90.91	73.25	81.13
RNGDet++ (Xu et al., 2023)	85.65	72.58	78.44	91.34	75.24	**82.51**
SamRoad (Hetang et al., 2024)	**90.47**	67.69	77.23	**93.03**	70.97	80.52
MARS	84.28	**81.53**	**82.88**	79.68	**84.56**	82.05

AICrowd V1 Performance

Table 4: Building performance comparison on Aicrowd V1.

Model	AICROWD-V1										
	AP	AP50	AP75	AR	AR50	AR75	bAP	IoU	C-IoU	PoLiS	N-ratio
PolyMapper	55.7	86.0	65.1	62.1	88.6	71.4	22.6	77.6	65.3	2.215	1.29
FFL*	67.0	92.1	75.6	73.2	93.5	81.1	34.4	84.3	73.8	1.945	1.13
PolyWorld	63.3	88.6	70.5	75.4	93.5	83.1	50.0	91.2	88.2	0.962	0.93
PolyBuilding	78.7	96.3	89.2	84.2	97.3	92.9	-	94.0	88.6	-	0.99
HiSup	79.4	92.7	85.3	81.5	93.1	86.7	66.5	94.3	89.6	0.726	-
Pix2Poly	79.6	91.6	85.2	87.7	-	-	-	95.03	89.85	**0.479**	1.111
GeoFormer	**91.5**	**96.6**	**93.1**	97.8	98.8	98.1	**97.1**	**98.1**	**97.4**	0.913	1.01
MARS	87.30	95.20	90.46	**97.94**	**99.28**	**98.62**	92.44	97.32	96.31	0.997	**0.4542**

- Outperforms/Matches **specialized** models
- **Unified model** for both roads & buildings

Simplicity

No hyperparameter tuning (NMS IoU, confidence thresholds) required—pure end-to-end learning

Scalability

Demonstrates potential to scale from 2-class to multi-class with finer-grained classification (highway, pedestrian way, etc.)

Resource

Paper

Dataset

Demo