Identifying Sidewalks from Crowdsourced SVI

MSc thesis in Geomatics

Thesis by : Neelabh Singh
Supervisors : Hugo Ledoux and Lukas Beuster
Co-reader : Ken Arroyo Ohori

Sidewalk Definition:

A sidewalk is a pedestrian facility associated with a road, not a standalone path in the real world. (OSM)

Presentation Structure

01 Introduction and Related work

02 Methodology

03 Results

04 Conclusions

01. Introduction and Existing Methods

Need
Existing Methods
Objective
Research Questions

Why Sidewalk mapping?

Pedestrian networks are the "Forgotten Layer" globally?

Road networks are mapped with high precision,
Authoritative and **openly** available sidewalk datasets are rare limiting the research for :

- **Accessibility:** **Wheelchair users** need precise width data
- **Routing and Mobility:** **Automated delivery** robots for navigation
- **Walkability:** To check availability of sidewalks on a street
- **Urban Planning:** Connectivity analysis, safe walkable streets
- **Pedestrian Environment:** Heat and air pollution exposure

Amsterdam OSM sidewalk as lines

Utrecht OSM sidewalk as lines

OSM : sidewalk as lines are rare and as polygons are non existent.

Existing Methods : Sidewalk Mapping

1. Traditional Semantic Segmentation (SVI)

- **Approach:** Using CNNs to classify pixels in street view images.

- **Limitation:** It is Binary and **Sidewalk presence,** Lacks geometric, **Models need training** struggle to generalize in different cities (e.g., detecting red brick vs. grey concrete). On **proprietary data**

2. Aerial & Satellite Mapping

- **Approach:** Using orthophotos (Satellite/Aerial) to extract city-wide pedestrian networks (e.g., TILE2NET).

- **Limitation:** **Occlusion** from **tree canopies** and building shadows, Global high resolution **Imagery availability,** Imagery resolution of 10cm is too low to **detect accurate sidewalk boundary.**

Occluded sidewalks from trees in Utrecht

Existing Methods : Sidewalk Mapping

3. Geometric Reconstruction (LiDAR & GSV)

- **Approach:** Mobile LiDAR scanners or Google Street View Depth to get precise 3D point clouds.
- **Limitation:** **Expensive LiDAR** ($50k+ hardware); **Google Street View:** Proprietary closed license, vehicle-only viewpoints, can t**ake years to get updated**

4. Participatory & Manual Auditing

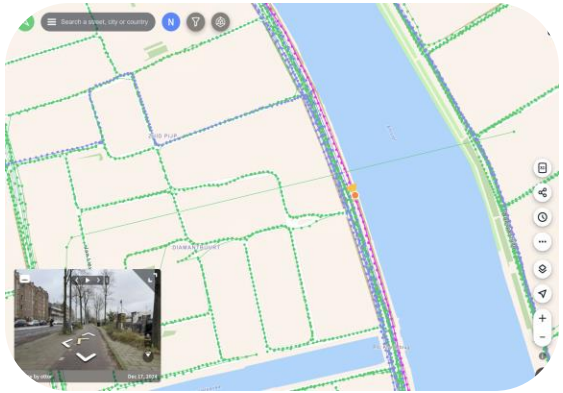
- **Approach:** Virtual Auditing by volunteering (e.g., MIT Project Sidewalk / smartphones on wheelchairs).
- **Limitation:** **Scalability & Reliability; Volunteer retention** is low.

Filling The Gap

Category 1 : A method that is **inventory** and **geometric,** No **training required** (better genrazibility foundation models)
Category 2 : Pedestrian-level imagery to overcome **aerial occlusion**
Category 3 : Uses Software-based depth (replaces LiDAR**), non proprietary data**, dynamic updates
Category 4 : Scalable and open data

Opportunity : Mapillary CSVI as the Solution

Mapillary has 3+ billion geotagged images
Global coverage
Open license



Images are stitched in a sequence using SfM-corrected camera poses
(More Precise than raw GPS cordinates)

Pedestrian and cyclist perspectives with better view of Sidewalks. Overcoming Tree occlusion.

Pre-computed semantic segmentation masks for sidewalks via API

The Disconnect:

Billions of images exist, but they lack the geometric/semantic structure needed for spatial analysis. How do we bridge this gap?

Objective : Pixels to Sidewalk

Aim : *Bridge the **disconnect between** raw SVI and geospatial **sidewalk data** by developing and evaluating two complementary automated workflows that operate exclusively on openly licensed data.*

Objective Develop Automated Mapping Pipeline/s:

Panoramic imagery (360 view)

Perspective Imagery

Input: Raw Crowdsourced Street View Images

Automated process

Output: Sidewalk presence (yes/no tags)
Sidewalk **Vector Data** (Lines or **Polygons**)

(Segment image pixels): Sidewalk class masks
(Position of pixels): Calculate the depth and reconstruct geometry.

Research questions

Main Research Question

How can crowdsourced street-level imagery be transformed into structured georeferenced sidewalk data using only open data and pre-trained models, and what are the accuracy and practical limitations?

RQ1 **Sidewalk Inventory Mapping ("What is there?")**
Can a city-wide sidewalk presence inventory by pre computed segmentation outputs and camera meta data already available through the Mapillary API, without performing any local model inference?

RQ2 **Sidewalk Geometry Reconstruction (The ”What does it look like”)**
Can an automated pipeline combining local semantic segmentation (DINOv3)and monocular metric depth estimation (Depth Anything V3) produce georeferenced sidewalk polygons and centerlines?

RQ3 **Accuracy, Coverage, and Trade-offs (The ”How well does it work”)**
How do the inventory-based and the geometry reconstruction approach compare in terms of detection accuracy, geometric precision, spatial coverage, and computational cost?

03. Methodology

- ***Sidewalk Inventory Mapping (Pipeline 1)***
- ***Sidewalk Geometric Reconstruction (Pipeline 2)***

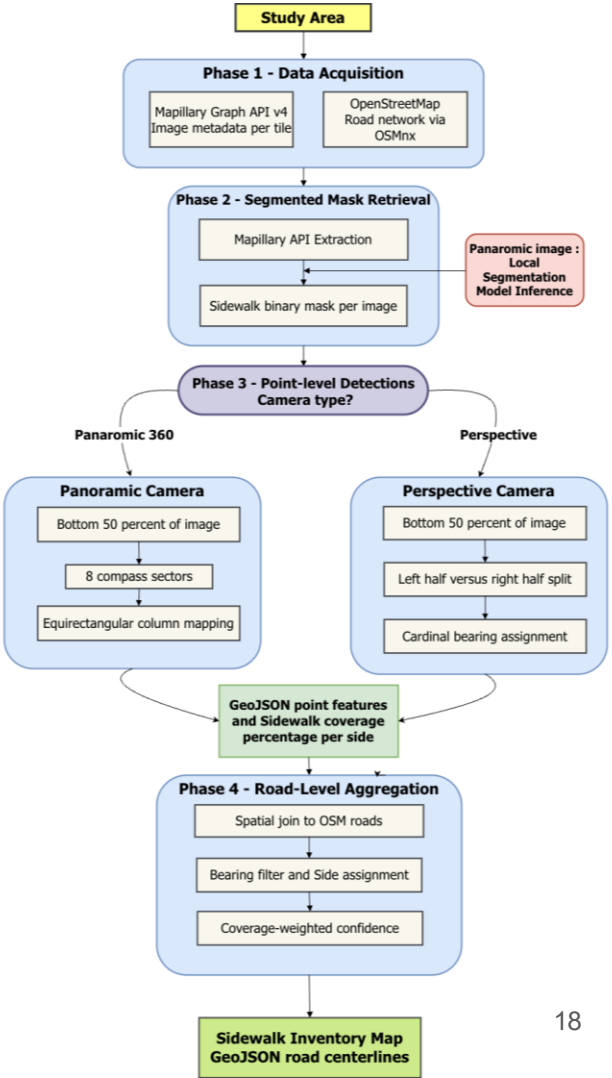
Pipeline 1 : Overview

Sidewalk Inventory Mapping : Binary yes/no per roadside

Four-Phase Overview

1. **Data acqusition :** Bounding box → Retrieve Mapillary image and metadata + OSM road network
2. **Segmentation Mask Retrieval :** Query Mapillary Detections API
3. **Point-Level Detection :** Analyze masks (left/right side) → Generate georeferenced point features with sidewalk presence information per side
4. **Road-Level Aggregation :** Spatial join of point feature information to OSM roads → Final inventory map

Key principle : Local model inference needed for segmentation (when API masks not available)



Pipeline 1 : Data Sources

Mapillary API

- Images
- Metadata (SfM corrected camera positions)
- Segmented Masks for Sidewalk (not for panoramic images in Amsterdam)

OSM road network

Mapillary

Pipeline 1 - Point-Level Detection

How We Determine Left/Right Sidewalk from a Single Image?

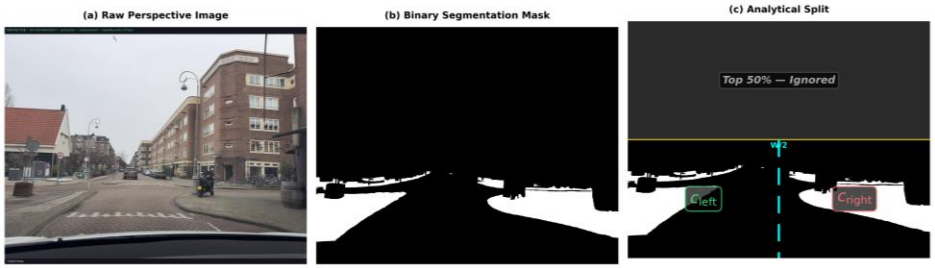
Create Point Features at camera position with sidewalk information for each image

Perspective images

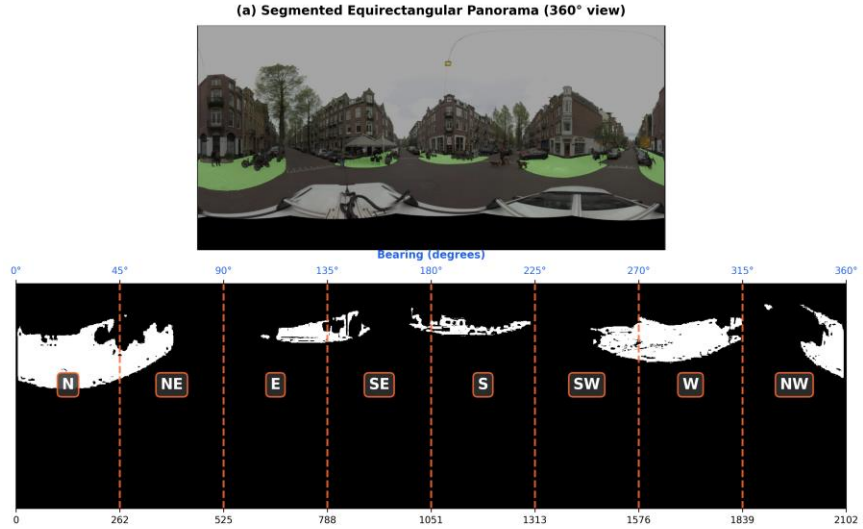
- Use bottom 50% of image (upper half = sky, facades)
- Divide into 2 parts
- Coverage: c_left = sidewalk_pixels_left / total pixels * 100

Panoramic images

- 360° field of view → divide bottom 50% into 8 compass sectors (45° each)
- Left Side are Sectors located counter-clockwise
- Right Side are Sectors located clockwise.



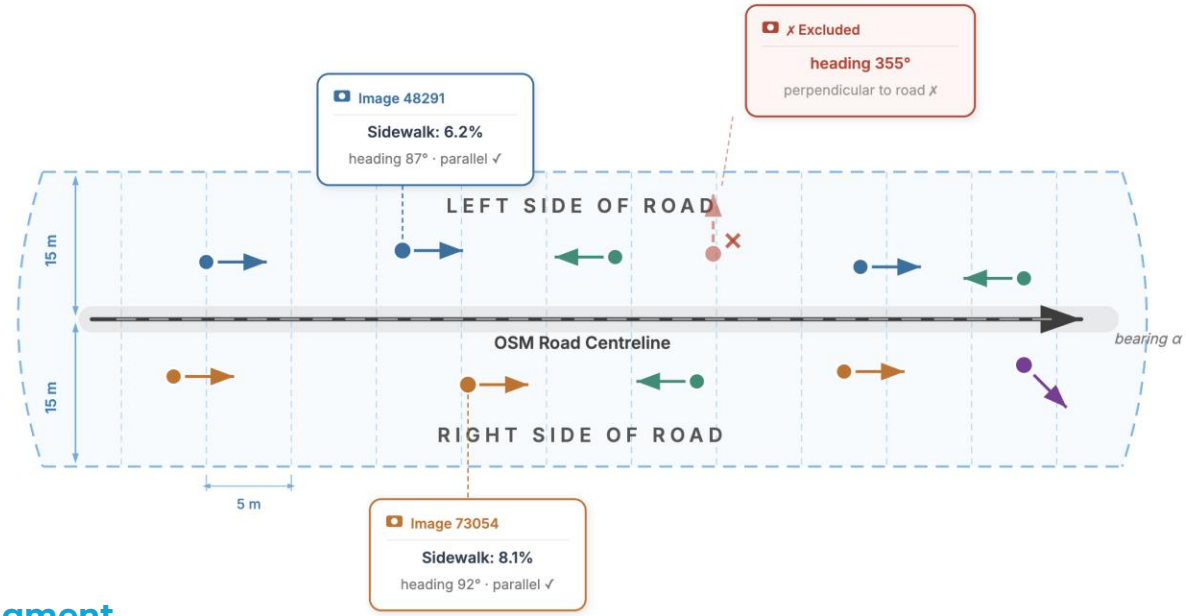
Perspective image



Panoramic image

Pipeline 1 – Road-Level Aggregation and Output

Road preparation	OSM roads split into **5m segments**; normalize bearing direction
Spatial association to OSM roads	**15m buffer** around centerline → identify candidate images point feature
Parallel heading filter	Only retain **images whose bearing aligns with the road direction.** (filters out perpendicular cameras which have ambiguous left/right sides)



Output per 5m (adjustable) road segment

- Sidewalk presence: **`true` / `false` / `no_data`** (for **left and right side**)
- Number of contributing images

→ GeoJSON LineString features with left/right sidewalk attributes, directly usable for OSM enrichment

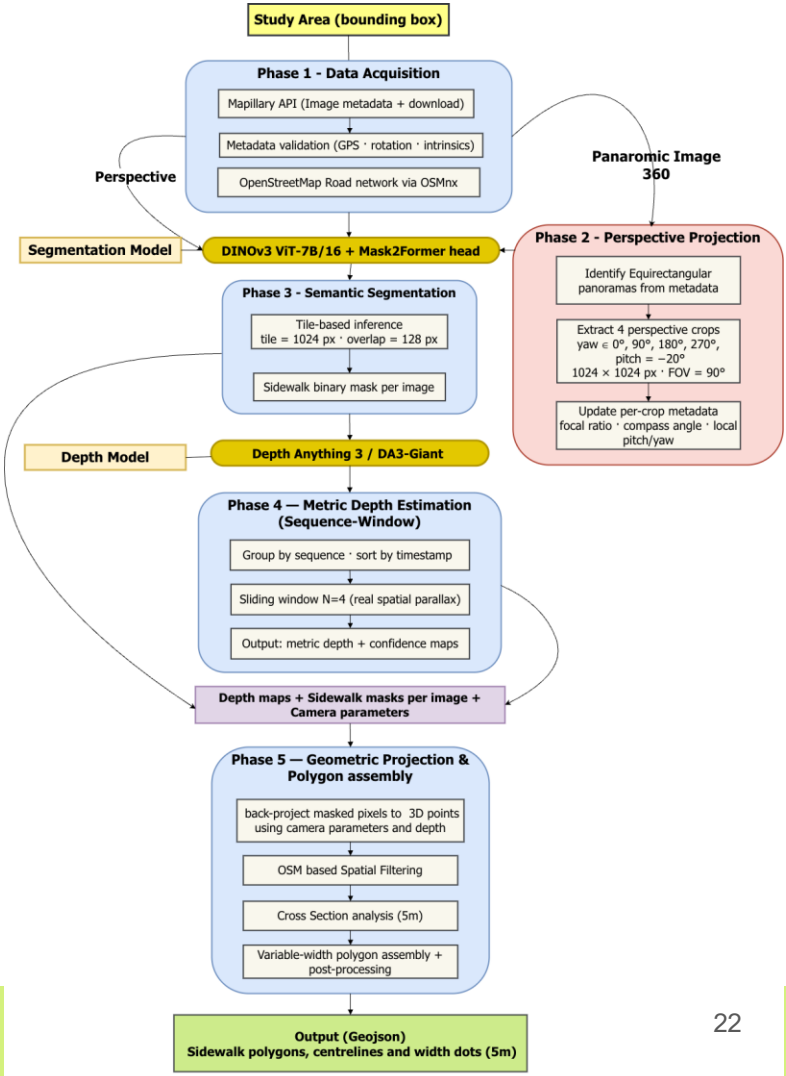
Pipeline 2 : Overview

Sidewalk Geometry Reconstruction : Sidewalk polygons, centerlines and width

Five-Phase Overview

1. **Data Acquisition :** Bounding box for imagery and metadata from Mapillary + OSM road network
2. **Perspective Reprojection :** Panoramas → perspective crops
3. **Semantic Segmentation :** DINOv3 → binary sidewalk masks
4. **Metric Depth Estimation :** DA3-Giant in pose-conditioned mode (4-image sliding window) → per-pixel depth in metres
5. **Geometric Projection & Polygon Assembly :** Back-projection of masked pixels→ spatial filtering → cross-section analysis → polygon construction

Only images with valid metadata are processed

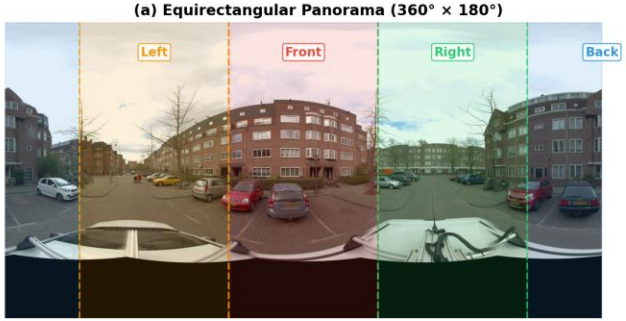


Pipeline 2 - Perspective Reprojection

Converting Panoramas to Perspective Crops for model inference

Panoramic images (equirectangular 360°) → cannot be processed by standard segmentation/depth models(assume pinhole camera)

Solution: Extract 4 perspective crops per panorama:

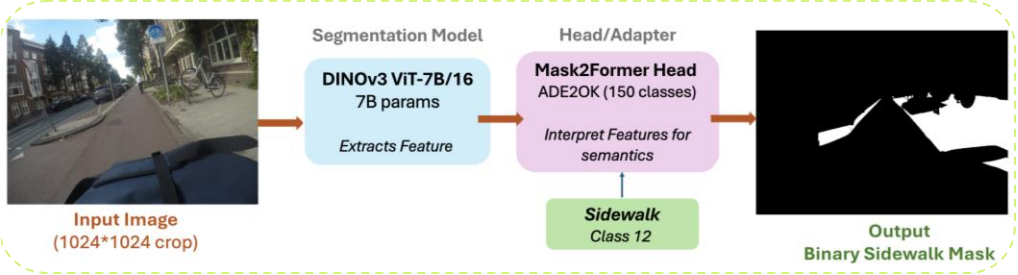


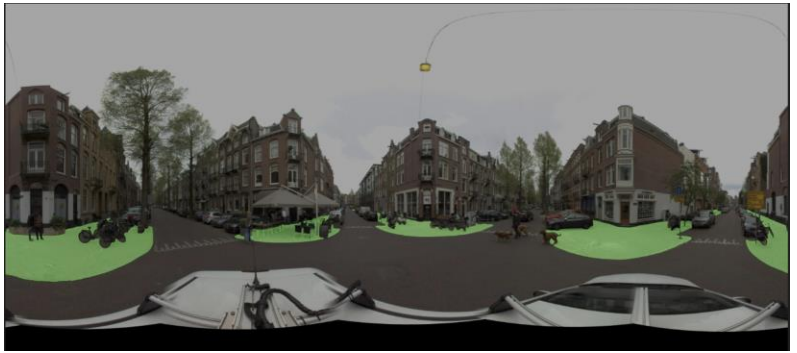
Metadata update for crops:
Each crop inherits parent panorama's GPS position but receives a rotated compass angle based on its relative yaw

Pipeline 2 - Segmentation & Depth

Phase 3: Semantic Segmentation
Identify Sidewalk pixels in an image

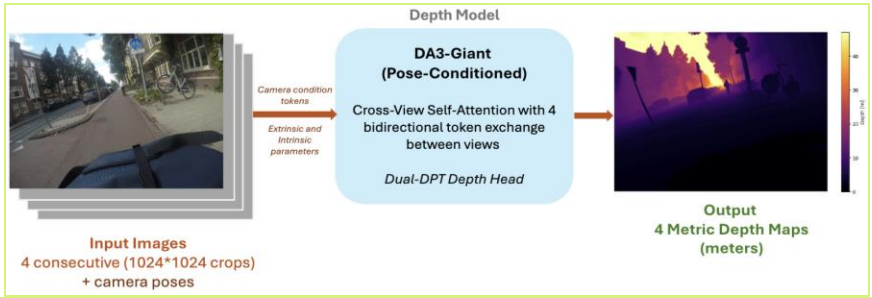
- DINOv3 + Mask2Former → binary sidewalk masks

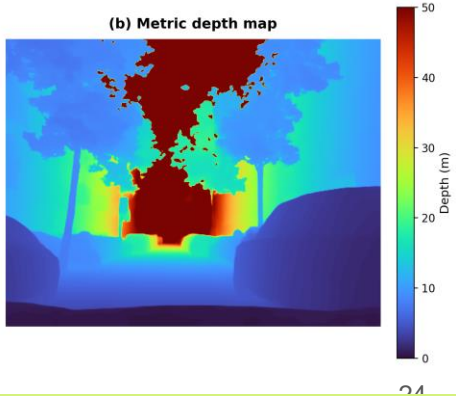




Phase 4: Metric Depth Estimation
Per-pixel depth in metres

- DA3-Giant in pose-conditioned mode
- Window of **4 consecutive images** per inference call





Why not MVS for depth? Finding feature correspondence on texture less surface is difficult and high computation cost

Pipeline 2 - Geometric Back-Projection

From Pixels to 3D World Coordinates

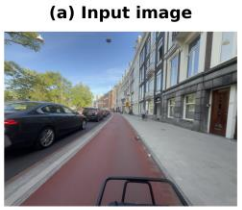
4-step Ray casting:

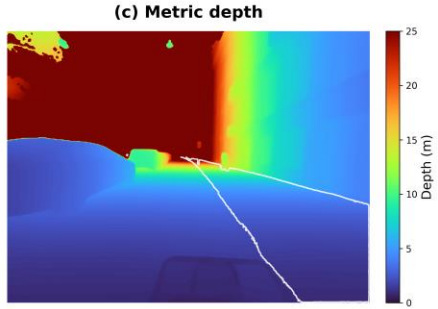
1. Extract the **2D image coordinates** of all sidewalk pixels:
2. Use Image intrinsic matrix to **find direction of ray** of sidewalk pixels
3. Multiply the ray by its estimated metric depth to get **3D coordinates relative to the camera**:
4. Use knows extrinsic matrix (camera poses) to **project the points**

→ Output: **3D point cloud with position**

Filtering of projected point:

- **Depth map range:** remove d > 20m (increasing estimation error)
- **Height filter:** remove points with Up > 1.5m (not ground-level)
- **Road exclusion zone:** buffer around road centerline





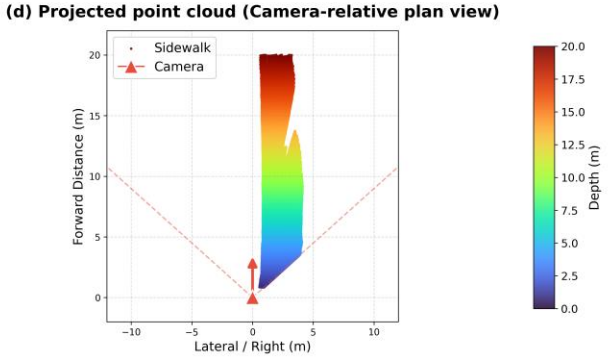
Back-Projection Formulation

Intrinsic Matrix ($\mathbf{K}$): focal length & principal pt

Camera-to-World ($\mathbf{T}_{c2w}$): Extrinsic matrix $\mathbf{E}^{-1}$

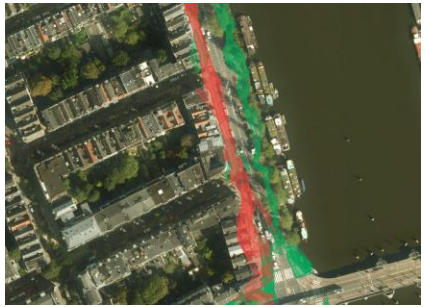
Ray Direction: $\mathbf{r} = \mathbf{K}^{-1}\mathbf{p}$

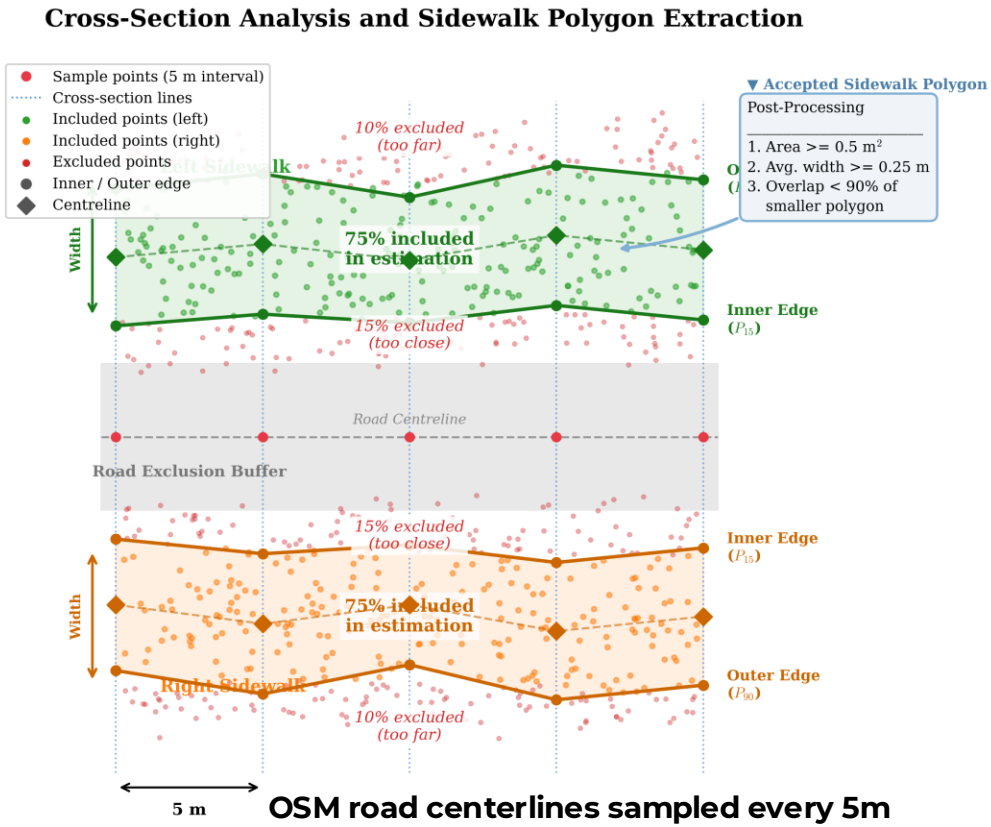
Depth Scaling: $\mathbf{X}_c = d \cdot \mathbf{r}$



Pipeline 2 - Polygon Assembly

From Point Cloud to Sidewalk Polygons





OSM road centerlines sampled every 5m

Output : Polygons + Centerlines + Width dots (all GeoJSON, WGS84)

04. Results and Accuracy

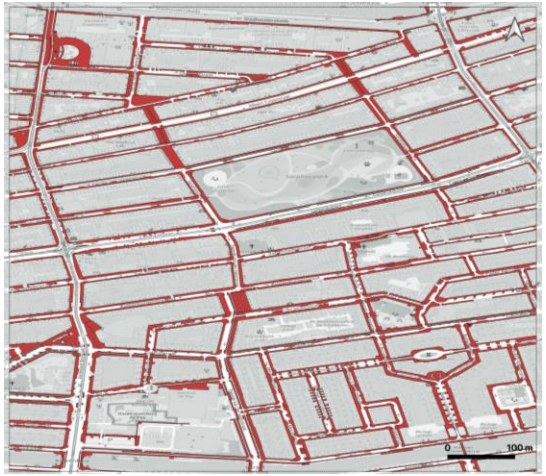
- *Study Area*
- *Sidewalk Inventory Mapping (Pipeline 1)*
- *Sidewalk Geometric Reconstruction (Pipeline 2)*
- *Computation Costs*

Study Area & Ground Truth

Ground Truth Dataset:

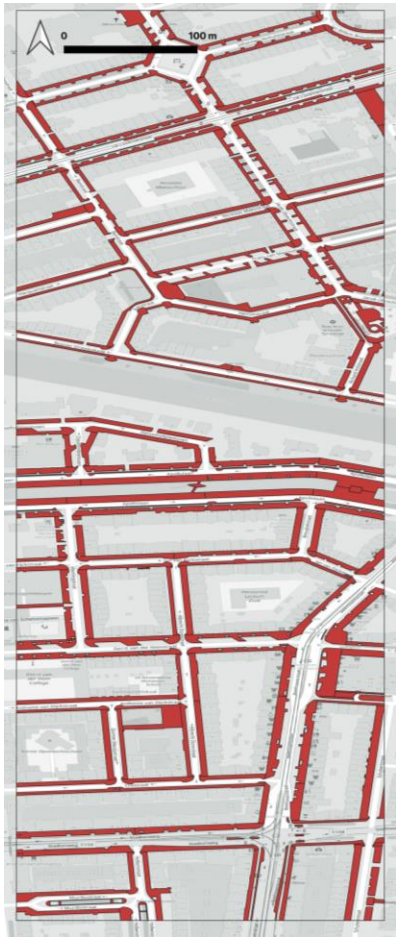
- Amsterdam: BGT (Basisregistratie Grootschalige Topografie) — official Dutch national registry
- Boston: City of Boston public sidewalk polygon dataset

Study Area	Dimensions (W × H)	Approx. Area	City
De Pijp	780 m × 1,120 m	0.87 km^2	Amsterdam, NL
Hobbemakade	275 m × 1,117 m	0.31 km^2	Amsterdam, NL
Boston South End	676 m × 714 m	0.48 km^2	Boston, USA



De Pijp

Boston South End



Hobbemakade

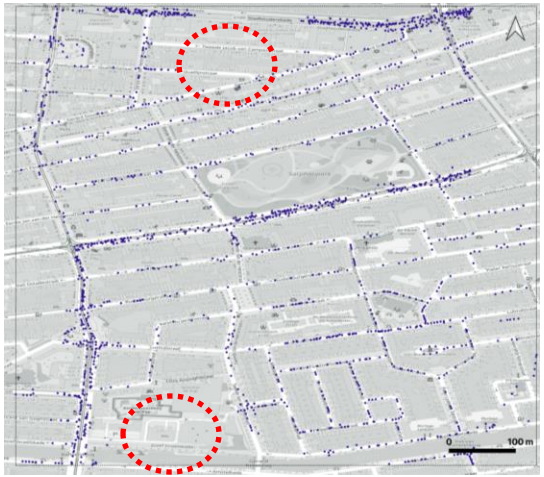
Note : *Due to job runtime limits on the Delftblue HPC cluster, testing the pipeline to a larger study area was not possible.*

Image Acquisition

Mapillary Bounding box Image acquisition :

- **De Pijp and Hobbemakade panaromic imagery dominated** & 30% perspective images which are pedestrian views
- Boston: 5,396 images (mostly perspective) from car dash boards rather than by pedestrian.
- Spatial coverage : Multiple road no imagery in Boston

Metric	De Pijp	Hobbemakade	Boston South End
Perspective images	569	471	4,856
Panoramic images	988	1,020	540
Total Raw Images	**1,557**	**1,491**	**5,396**
Perspective crops (4× Panoramas)	3,952	4,080	0
Total Processed Images	**4,521**	**4,551**	**5,396**



De Pijp

Boston South End

Hobbemakade

Note : *Each panoramic image is divided in 4 crops for model inference*

No imagery

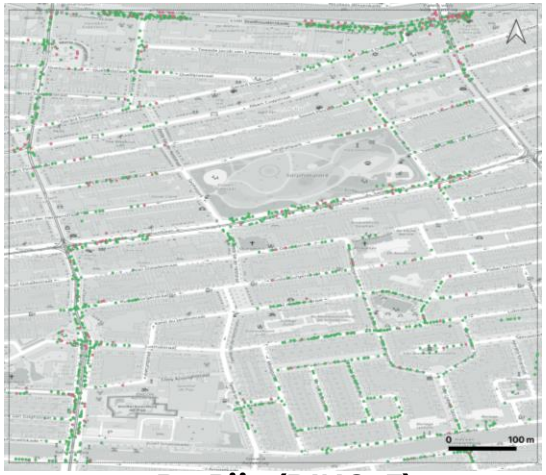
Segmentation Results

Key findings:

- **Amsterdam API limitation:** as panoramic images (70%) not segmented → only 20-24% detection
- Local DINOv3 processes all imagery → 88-92% detection rate,
- Highest detection in Hobbemakade

- Boston: both image type masks available; only 2% difference in detection

Study Area	Source	Images Processed	Sidewalk Detected	% Images with Sidewalk
De Pijp	Mapillary API	1,557	374	24.0%
	Local DINOv3	1,557	1,376	88.3%
Hobbemakade	Mapillary API	1,491	296	19.9%
	Local DINOv3	1,491	1,372	92.0%
Boston South End	Mapillary API	5,396	3,723	69.9%
	Local DINOv3	5,396	3,861	71.6%



De Pijp (DINOv3)

Boston South End (Mapillary API)

Hobbemakade (DINOv3)

Note: *The low Mapillary API detection rate in Amsterdam reflects an API data-availability gap for panoramic imagery segmentation, not a difference in model quality.*

Segmentation Accuracy

Mapillary v\s DINOv3 local Segmentation :

- Local DINOv3 segmentation masks are much sharper and has less artefacts compared to Mapillary.
- Mapillary doesn't provide any accuracy assessment: based on old architecture models. Still good for inventory mapping

Mapillary masks : small artefacts

DINOv3 masks sharp edges

Less Sidewalk Detections in Boston than Amsterdam:

- The Amsterdam 90% sidewalk detection whereas Boston dataset 70%
- **20 % difference in Boston** primarily driven by severe visual **occlusion by on street parking**

Occlusion by parked vehicles in Boston South End

Note: *As roadside occlusions artificially* ***inflate False Negative (FN) rates****, this pipeline is reliable only for confirming sidewalk presence,* ***a non-detection outcome should not be interpreted as definitive proof of sidewalk absence****.*

Pipeline 1 : Results

Inventory Maps

Metric	De Pijp	Hobbemakade	Boston(Mappilary API)
Road segments (5 m)	5,565	2,776	19,260
Road-sides (left + right)	11,130	5,552	38,520
Sidewalk detected (`true`)	4,827 (43.4%)	2,714 (48.9%)	2,759 (7.2%)
Not detected (`false`)	747 (6.7%)	1,046 (18.8%)	4,260 (11.1%)
No data (`no_data`)	5,556 (49.9%)	1,792 (32.3%)	31,501 (81.8%)
Data quality of decided sides (`true` + `false`):			
High (≥10 image crops)	3,983 (71.5%)	3,647 (97.0%)	3,506 (50.0%)
Medium (3–9 image crops)	1,496 (26.8%)	111 (3.0%)	1,940 (27.6%)
Low sample (1–2 image crops)	95 (1.7%)	2 (0.1%)	1,573 (22.4%)

5m road segments : 15m road buffer

Key insight:

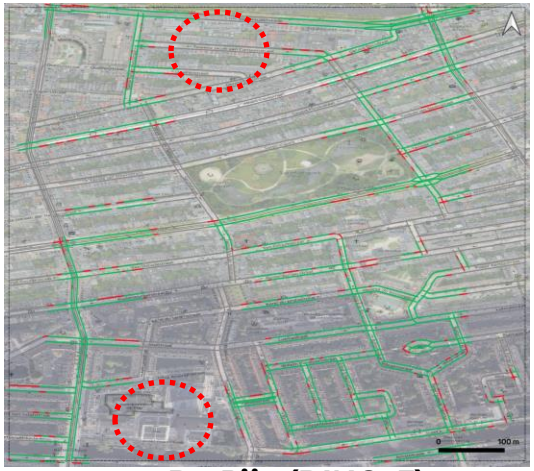
At Cross sections : Dense imagery leads to better chance of detections

Dominant No data segments: limitation is spatial coverage, not algorithmic failure.

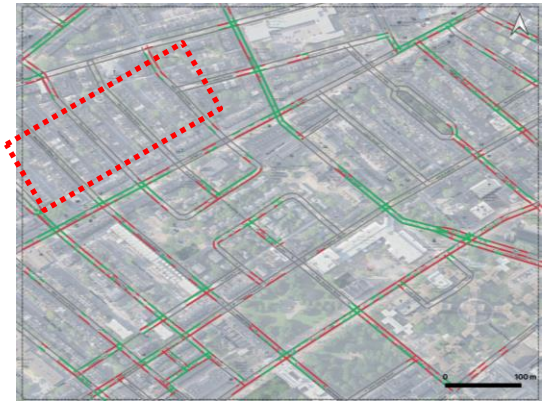
- Boston: 82% of roadsides has zero imagery
- Hobbemakade: best coverage → only 32% no-data

Sidewalk Yes

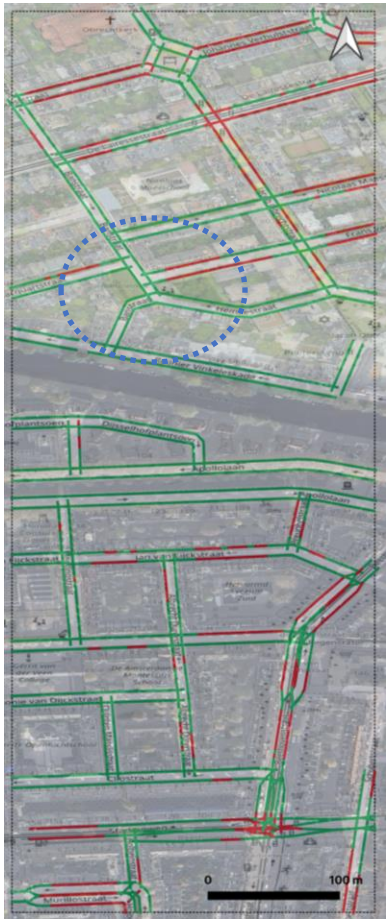
Sidewalk No



De Pijp (DINOv3)



Boston South End (Mapillary API)



Hobbemakade (DINOv3)

Note : *Hobbemakade shows we could detect sidewalks under tree*

No imagery

Pipeline 1 : Accuracy Assessment

Standard 2×2 matrices unfairly penalize the model for coverage gaps, 3×2 matrix separates System & Algorithmic accuracy

Pipeline 1: High Precision, Coverage-Bounded Recall

Algorithmic precision: 97.2–98.6% across all study areas

- Pipeline reports a sidewalk, it is correct >97% of the time
- Lack of FP in study area

Algorithmic recall gap (Boston: 40.8% vs Amsterdam: 86.5%) (includes FN)

- Dense on-street parking in Boston obscures sidewalks from road-centre cameras
- Amsterdam more pedestrian/cyclist imagery → less occlusion

System recall limitation (18%) (includes No Data 80% in Boston)

- Bounded by Mapillary coverage - Not a model failure

5m road segments : 15m road buffer

Study Area	TP	FP	FN	TN	ND^+	ND^-
De Pijp	4,702	125	736	11	3,256	2,300
Hobbemakade	2,637	77	1,030	16	183	1,609
Boston South End	2,719	40	3,953	307	12,040	19,461

Study Area	Algo. Prec.	Algo. Rec.	Sys. Rec.	Coverage	N_{total}
De Pijp	0.974	0.865	0.541	50.1%	11,130
Hobbemakade	0.972	0.719	0.685	67.7%	5,552
Boston South End	0.986	0.408	0.145	18.2%	38,520

Algorithmic failures by False Positives (FP) - No sidewalk according to Ground truth, but the algorithm says yes :
*Correct false positives are highly unlikely as s**tudy area features nearly 100% actual sidewalk coverage.***

Pipeline 2 : Results

Ground Truth
Reconstructed

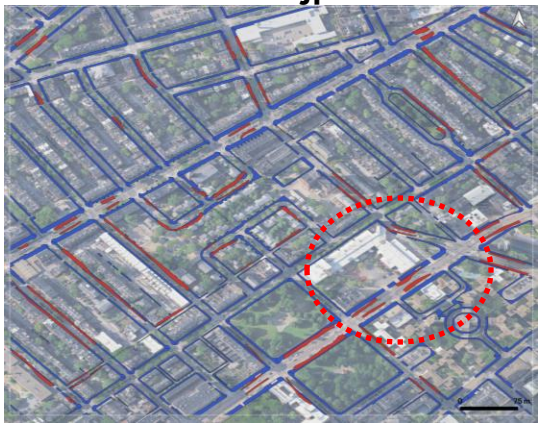
Sidewalk Geometry Reconstruction

Metric	De Pijp	Hobbemakade	Boston
Reconstructed polygons	179	116	67
Total reconstructed area	50,891 m^2	31,424 m^2	11,253 m^2
Ground truth sidewalk area	133,009 m^2	52,623 m^2	53,333 m^2
Ratio (predicted / GT)	0.38	0.60	0.21

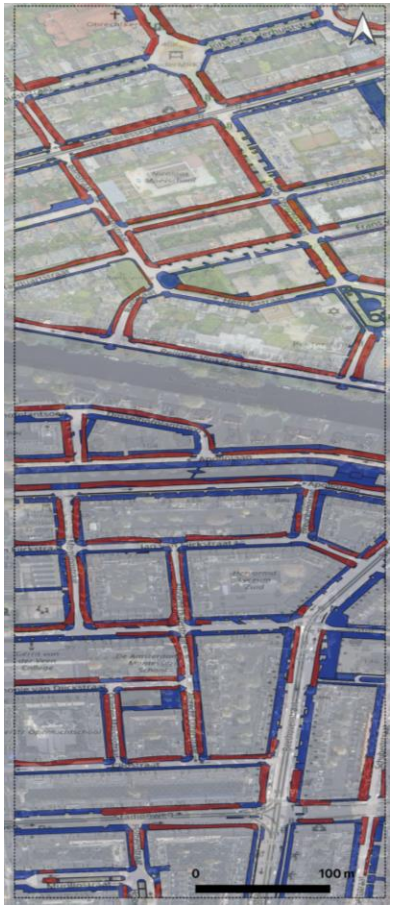
Observations :

- **Hobbemakade achieves highest completeness** (60%) — correlates with best image coverage
- Boston lowest (21%) — same coverage limitations
- Polygon shapes show GPS drift effects: irregular/curved boundaries rather than sharp edges
- Polygon displacement onto road/buildings

De Pijp



Boston South End



Hobbemakade

*Note: To ensure maximum reliability, all sidewalk polygons were constructed **using local segmentation** and depth models rather than the native Mapillary segmentation masks.*

Pipeline 2 : Accuracy Assessment

Dimensional Accuracy:

- IoU (Boundary matching) (8 to 30%)
- Area Precision : **35.3% to 62.3% predicted shapes are actually sidewalks**
- Area Recall : **12% to 41% of actual sidewalk are reconstructed**

Positional Accuracy:

- **Typical sidewalk width:** 2–3m → Width MAE of 1.2–1.4m is significant
- **Centerline offset:** 2.6–3.2m → polygonsfrequently overlap with road or buildings

Area based accuracy

Study Area	IoU	Area Precision	Area Recall
De Pijp	0.257	0.515	0.340
Hobbemakade	0.329	0.623	0.411
Boston South End	0.100	0.353	0.122

Width and Centerline accuracy

Study Area	Width MAE (m)	Width RMSE (m)	Mean Centerline Offset (m)
De Pijp	1.18	1.64	3.00
Hobbemakade	1.40	1.72	2.64
Boston South End	1.32	1.67	3.15

Note : *By clipping the ground truth dataset to the imagery-covered area, these IoU values isolate geometric reconstruction quality independently of regional data gaps*

Pipeline 2 : Error Analysis

Why Geometry Accuracy is Limited — Cascading Error Sources

1. Segmentation boundary imprecision: (DINOv3 mIoU = 63.0)
- **Pixels misclassified → during projection**

2. Depth estimation degradation : (DA3 F1 = 87.1%)
- **~13% of points placed >25cm from true surface**

3. GPS positioning / Inter-sequence drift :
- **Different devices, different times → different absolute positions in different sequence**
- **Fused polygons from multiple sequences = irregular boundaries**

4. Occlusion
- **Parked vehicles block lateral views**
- **Width estimation unreliable with partial visibility**

5. Spatial coverage : Hobbemakade with 30% road segments lack coverage whereas Boston 80%.

01. Conclusions

- *Answer RQ's*
- *Future Work*
- *Hybrid Approach*

Conclusions : Answering RQ's

RQ3 : *How do the inventory-based approach and the geometry reconstruction approach compare in terms of detection accuracy, geometric precision, spatial coverage, and computational cost?* ***(How well does it work)***

Criterion	RQ1 - Pipeline 1 (Inventory)	RQ2 - Pipeline 2 (Geometry)
Detection accuracy	**High precision 97–99%** where imagery exists.	**Low area precision** 38–65% of reconstructed area is outside actual boundaries.
Geometric output / Precision	Binary yes/no only (per roadside). **No geometric** shapes produced.	**Geometry : Polygons + centerlines.** Width MAE of 1 to 1.5 m; centerline positional offsets of 2.5–3. m, limited by. **Error sources : Data** - GPS drift, occlusion and coverage. **Models :** Segmentation (mIoU 63), depth (F1 87.1),
Coverage tolerance / Spatial Coverage	**Partial occlusion OK** (small visible region enough to detect). Bounded by Mapillary image coverage No-data 81.8% in Boston.	**Requires clear, full views of the sidewalk surface for width estimation.**
Computational cost	Minutes (**2–3 minutes for 1,500 images on standard CPU**). Zero GPU requirement.	**7–9 hours per km² (~1,500 images)** on NVIDIA A100 GPU. Segmentation alone consumes ~4.5 hours per 1,000 panoramas.
Deployment readiness	**Production-ready. Highly efficient for city-scale deployment.**	**Needs data filtration and error handling**
Best use case	City-wide inventory at scale.	Focused streets with pedestrian views and complete coverage.

Discussion

Strengths:
- **Open data (CC BY-SA)** for both inventory AND geometry
- **Zero-shot inference:** no training or fine-tuning needed → fully reproducible
- **Pedestrian-captured imagery** → detects sidewalks **under tree canopies** (invisible from aerial)
- **Complementary to aerial approaches** (TILE2NET etc.) which suffers from Dense tree occlusion

Key Limitations:
- Platform dependency: System recall bounded by Mapillary's community **coverage**
- Presence detection only: Cannot distinguish "no sidewalk" from "sidewalk occluded"
- GPS drift : Polygon accuracy limited by inter-sequence positional inconsistency
- Urban context : Low FP rate partially due to near-100% sidewalk coverage in study areas (needs suburban testing)

This thesis reveals a fundamental trade-off in crowdsourced sidewalk mapping:
Pipeline 1 delivers reliable but limited binary data
Pipeline 2 offers richer metric geometry at the cost of reduced accuracy.

Future Work

1. Evolving Platform : Once Mapillary extends panoramic segmentation globally → Pipeline 1 API-based globally no local computation

2. For Pipeline 2 Image source filtering : **Retain only pedestrian/cyclist imagery** → less occlusion but lower coverage; high urban canyon effect and smartphones have less GPS precision which has to be corrected, **trade-off to evaluate**

3. As Mapillary doesn't provide SfM corrected GPS accuracy, one can evaluate and try to **minimize GPS drift** within sequences for accurate polygon locations

4. Model improvement : DINOv2 (49.5 mIoU) → DINOv3 (55.9 mIoU) in one generation; **future models will directly improve Pipeline 2** without architectural changes it,

5. Broader geographic testing : Suburban areas with sparse sidewalks → **properly evaluate false-positive** rate and test segmentation model generalization

This thesis reveals a fundamental trade-off in crowdsourced sidewalk mapping:
Pipeline 1 delivers reliable but limited binary data
Pipeline 2 offers richer metric geometry at the cost of reduced accuracy.

Hybrid Approach

Condition	Aerial (TILE2NET)	(SVI)
Open streets, no canopy	Excellent	Good
Dense tree canopy	Occluded	Ground-level view
Building shadows	Ambiguous	Unaffected
Dense on-street parking	Visible from above	Blocked laterally
Suburban/rural coverage	Complete satellite	Sparse Mapillary

Hybrid Approach by combining Aerial (TILE2NET) and SVI pipeline:

- Use **aerial segmentation** for broad urban coverage (Good in occlusion from street parking)
- Use **SVI-based mapping** for targeted infill in vegetated/occluded zones

- Combined approach outperforms either in isolation.

Thank you! for listening