MSc thesis in Geomatics

Identifying Sidewalks from Crowdsourced SVI

Neelabh Singh

2026



MSc thesis in Geomatics

Identifying Sidewalks from Crowdsourced SVI

Neelabh Singh

June 2026

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of Master of Science in Geomatics

Neelabh Singh: *Identifying Sidewalks from Crowdsourced SVI* (2026)
This work is licensed under a Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The work in this thesis was carried out in the:



3D geoinformation group
Delft University of Technology

Supervisors:	Dr. Hugo Ledoux
	Lukas Beuster
Co-reader:	Dr. Ken Arroyo Ohori

Abstract

Sidewalks are fundamental urban infrastructure, yet large-scale, openly available geospatial sidewalk datasets remain scarce worldwide, hindering pedestrian routing, accessibility analysis, and urban planning applications. This thesis investigates the feasibility of transforming crowdsourced street view imagery from Mapillary into structured georeferenced sidewalk data using only openly licensed data and pre-trained models. Two complementary automated pipelines are developed and evaluated across three study areas in Amsterdam (Netherlands) and Boston (United States).

The first pipeline, Sidewalk Inventory Mapping, constructs a binary presence inventory (yes/no per roadside) by aggregating pre-computed semantic segmentation outputs and SfM-corrected camera metadata from the Mapillary API onto the OpenStreetMap road network, without any local model inference. The second pipeline, Sidewalk Geometry Reconstruction, downloads imagery and processes it locally using a vision foundation model (DINOv3) for semantic segmentation and monocular metric depth estimation (Depth Anything V3) to reconstruct sidewalk polygons and centrelines.

Results demonstrate that the inventory pipeline achieves algorithmic precision of 97.2% to 98.6% and algorithmic recall up to 86.5%, providing highly reliable sidewalk detection where imagery exists. The geometry reconstruction pipeline successfully produces sidewalk polygons but with limited spatial accuracy (IoU 0.100–0.329, width MAE 1.18–1.40 m), resulting from cascading errors in segmentation, depth estimation, and GPS positioning. Both pipelines are fundamentally bounded by Mapillary's spatial coverage, with system recall dropping substantially in areas lacking imagery. The inventory approach emerges as the practical, scalable solution for city-wide deployment, while the geometry reconstruction demonstrates technical feasibility but requires further refinement for production use. All outputs are compatible with open data standards and can support OpenStreetMap enrichment workflows.

Acknowledgments

This thesis marks the completion of a reformative chapter in my life, a milestone that would have been entirely impossible to reach without the guidance and support of the remarkable mentors and loved ones around me.

I am immensely grateful to my supervisor, Hugo Ledoux, for constantly pushing the boundaries of my academic thinking and providing me with the much-needed guidance and motivation during moments of doubt. Equal credit goes to my co-supervisor, Lukas Beuster, for creating such a welcoming environment to discover and learn, and for always offering his insights and time whenever it was needed most.

On a personal note, this academic endeavor would not have been possible without the support of my family, friends, and colleagues. Thank you for making things feel a little lighter while this intensive process transpired. The most heartfelt appreciation belongs to my amazing mother, whose unwavering faith in my capabilities and resilient spirit have always served as my ultimate foundation.

Contents

Abstract v

Acknowledgments vi

Acronyms xv

1 Introduction 1
- 1.1 Background and Motivation . . . 1
- 1.2 Research Objectives and Scope . . . 2
  - 1.2.1 Objective . . . 3
  - 1.2.2 Scope of research . . . 4
- 1.3 Results Overview . . . 5
- 1.4 Thesis Outline . . . 5

2 Theoretical Background 7
- 2.1 Sidewalk Definition . . . 7
- 2.2 OSM Sidewalks . . . 7
  - 2.2.1 OSM enrichment procedure and results . . . 8
- 2.3 Street View Imagery (SVI) Description . . . 9
  - 2.3.1 The Shift to crowdsourced SVI data: Mapillary API . . . 10
- 2.4 Deep Learning Models for Semantic Segmentation and Metric Depth Estimation 12
  - 2.4.1 Semantic Segmentation . . . 13
  - 2.4.2 Metric Depth Estimation . . . 15
- 2.5 Depth Estimation Strategy for Sidewalk Reconstruction . . . 18
  - 2.5.1 Multi-View Reconstruction and Its Limitations for Sidewalks . . . 18

3 Related work 22
- 3.1 Category 1: Traditional Semantic Segmentation from SVI . . . 22
- 3.2 Category 2: Aerial Imagery and Network Generation . . . 23
- 3.3 Category 3: Geometric Reconstruction and Metric Mapping from SVI . . . 24
- 3.4 Category 4: Participatory Approach and Smartphone-Based Mapping . . . 24
- 3.5 Synthesis: Positioning the Research . . . 25

4 Methodology **26**
4.1 Sidewalk Inventory Mapping . . . 26
4.1.1 Overview . . . 26
4.1.2 Data Sources . . . 27
4.1.3 Point-Level Detection . . . 30
4.1.4 Road-Level Aggregation . . . 33
4.2 Sidewalk Geometry Reconstruction . . . 36
4.2.1 Data Acquisition . . . 37
4.2.2 Perspective Reprojection . . . 39
4.2.3 Semantic Segmentation . . . 40
4.2.4 Depth Estimation . . . 41
4.2.5 Geometric Projection and Polygon Assembly . . . 42

5 Results and Accuracy **48**
5.1 Study Areas . . . 48
5.1.1 Image Coverage . . . 48
5.1.2 Ground Truth Datasets . . . 50
5.2 Segmentation Results . . . 52
5.2.1 Mapillary API and Local DINOv3 masks . . . 52
5.2.2 Segmentation Accuracy . . . 55
5.3 Pipeline 1 . . . 55
5.3.1 Sidewalk Inventory Mapping Results . . . 55
5.3.2 Accuracy Assessment . . . 57
5.4 Pipeline 2 . . . 62
5.4.1 Sidewalk Geometry Reconstruction Results . . . 62
5.4.2 Accuracy Assessment . . . 63
5.5 Computational Cost . . . 67
5.5.1 Pipeline 1: Sidewalk Inventory Mapping . . . 67
5.5.2 Pipeline 2: Sidewalk Geometry Reconstruction . . . 67
5.5.3 Synthesis . . . 68

6 Conclusions **69**
6.1 Conclusion . . . 69
6.2 Discussion . . . 71
6.3 Future Work . . . 73

Appendix : Declaration of AI/LLM usage **79**

Reproducibility self-assessment **80**

Current State of OpenStreetMap Sidewalk Coverage **82**

List of Figures

1.1 Illustration of the current gap in automated sidewalk mapping from crowdsourced SVI. **Left:** Diverse Mapillary images (vehicle top, pedestrian bottom) that are openly available but not automatically usable as geospatial data. **Right:** Sidewalk vectors (solid blue) that would enable routing and mapping if extracted automatically. Image credits: © Mapillary contributors (Creative Commons Attribution-ShareAlike (CC BY-SA) 4.0, accessed 2025-01-06); basemap © OpenStreetMap contributors Open Database License (ODbL). 3

2.1 Two mapping examples from the OSM wiki (`area:highway=footway`): (Left) A sidewalk mapped as a simple closed polygon outlined in red. (Right) A sidewalk mapped as a multipolygon featuring an inner ring outlined in red. Source: (OpenStreetMap Wiki, n.d.e). 8
2.2 Examples of street-view imagery from Mapillary: (a) panoramic image captured from a specialized vehicle and (b) perspective image captured by a pedestrian (Mapillary, 2025a). 9
2.3 Mapillary API sidewalk semantic segmentation 12
2.4 Examples from the Mapillary Vistas v2 dataset highlighting sidewalks (class 24, pink), clearly distinguished from roads and other surfaces. These annotations support semantic segmentation model training for automated sidewalk extraction. 13
2.5 The inference workflow for sidewalk presence detection. Raw street-level imagery is processed by the DINOv3 ViT-7B/16 foundation backbone to extract dense spatial features. These representations are subsequently decoded by the Mask2Former head to generate the final sidewalk segmentation mask. 14
2.6 Pose-conditioned DA3 inference pipeline. A sliding window of four consecutive street-level images and their known camera parameters (world-to-camera extrinsics from SfM; intrinsics from the crop geometry) are passed to the DA3-Giant model. Camera condition tokens inject geometric context into all attention layers. Cross-view self-attention exploits spatial parallax between views to produce geometrically consistent metric depth maps. The output is four metric depth maps. 16

2.7 3D points reconstructed from RAFT-based dense optical flow correspondences, generating a large number (20000) of matches across consecutive images (100 images); the resulting feature points on sidewalk surfaces are noisy and geometrically inconsistent with substantial vertical and lateral dispersion. 19
2.8 Sidewalk polygons generated by the geometric flat-ground projection method overlaid on a satellite basemap. 19
2.9 Examples from two consecutive image sequences showing the scarcity of feature correspondences on sidewalk surfaces. 21

3.1 Highlighted sidewalk occlusion by tree canopy and building shadows in Publieke Dienstverlening Op de Kaart (PDOK) orthophoto imagery (Utrecht, Netherlands). The figure is included for visualization purposes only. 23

4.1 Overview of the proposed four-phase methodology for automated sidewalk inventory mapping using Mapillary imagery and OpenStreetMap road centerlines. 28
4.2 Examples of four sample perspective camera images and their binary sidewalk masks provided by the Mapillary API 29
4.3 (a) Raw perspective image from Mapillary. (b) Binary sidewalk mask from segmentation (white = sidewalk). (c) The bottom 50% is bisected at W/2; white pixel counts in the left c_{left} and c_{right} halves determine sidewalk presence on each side of the road. 31
4.4 Panoramic 8-sector spatial analysis. (a) Segmented equirectangular image with sidewalk overlay (green). (b) Binary mask of the bottom 50% divided into 8 equal sectors corresponding to cardinal compass bearings. Pixel counts per sector determine sidewalk presence in each direction. Note: This segmented panoramic image is created using DinoV3, as Mapillary doesn't provide panoramic image segmentation. 32
4.5 Overview of the proposed five-phase methodology for automated sidewalk geometry reconstruction using segmentation and depth estimation models. . . 38
4.6 Perspective reprojection of an equirectangular panorama. (a) The source 360° panorama with colored bands indicating the four cardinal crop regions. (b.1-b.4) The four extracted perspective crops at yaw angles 0, 90°, 180°, and 270°, each with a downward pitch of $-20°$, 1024×1024 pixels, and a field of view of 90°. The downward tilt visibly increases the proportion of ground-level surfaces (sidewalks, roads) relative to a horizontal extraction. 40
4.7 Sidewalk semantic segmentation masks overlaid on panoramic imagery, the final output after inference, with all individual processed panoramic crops recombined. 41
4.8 Example outputs from the metric depth estimation phase. (a) A perspective crop from a 360° panorama in the De Pijp study area (pitch $= -20$). (b) The corresponding metric depth map produced by DA3, where colour encodes absolute distance in metres from the camera (clipped at 30 m). (c) The per-pixel confidence map; high-confidence regions (yellow/orange) correspond to nearby surfaces suitable for 3D reconstruction, while low-confidence regions (dark purple) are discarded during projection. 42

4.9 The back-projection pipeline from image space to metric 3D space. (a) Raw input image. (b) Binary sidewalk mask identifying pixels of interest. (c) Metric depth map providing the depth scaling factor d for each pixel. (d) The resulting depth-scaled 3D points plotted in a camera-relative plan view ($\mathbf{X}_c$), demonstrating how the 2D mask and depth perfectly reconstruct the lateral spatial layout of the sidewalk before applying the global ENU transformation. 44
4.10 Examples of the raw projected sidewalk point cloud. The overlapping coverage from multiple viewpoints provides redundant observations that are aggregated during cross-sectional analysis. 45
4.11 Cross-section geometry and width estimation for sidewalk polygon extraction. Road centerlines are sampled at 5 m intervals, and projected 3D points are classified by their lateral offset: the inner edge (P_{15}) and outer edge (P_{90}) define the sidewalk boundaries. Post-processing removes small, thin, or overlapping polygons. 46

5.1 Overview of the three study areas showing image coverage point distributions: (a) Boston South End, USA; (b) De Pijp, Amsterdam; and (c) Hobbemakade, Amsterdam. 49
5.2 Ground truth sidewalk polygons for the three study areas. (a) Boston South End, (b) De Pijp, (c) Hobbemakade. 51
5.3 Comparison of sidewalk detection results obtained from the Mapillary API and the locally deployed DINOv3 model across the study areas (continued on the next page). 53
5.3 Comparison of sidewalk detection results obtained from the Mapillary API and the locally deployed DINOv3 model across the three study areas. Green points indicate images in which a sidewalk was detected, while red points indicate images without sidewalk detection. 54
5.4 Examples of Mapillary imagery from the Boston South End study area. Dense on-street parking completely obscures the adjacent sidewalks from road-centre cameras, leading to missed detections. 56
5.5 Sidewalk inventory maps for Boston South End and De Pijp showing detected sidewalk presence on the left and right sides of the road. Green: sidewalk detected; red: no sidewalk detected; grey: insufficient data (`no_data`). (continued on the next page). 58
5.5 Sidewalk inventory maps for Hobbemakade showing detected sidewalk presence on the left and right sides of the road. Green: sidewalk detected; red: no sidewalk detected; grey: insufficient data (`no_data`). 59
5.6 3×2 expanded confusion matrices for each study area. Each cell counts individual roadsides (one side of a 5-metre road segment) with the percentage of the total shown in parentheses. Green cells indicate correct classifications, red cells indicate errors, and grey cells indicate roadsides with no imagery coverage. 61
5.7 Reconstructed sidewalk polygons (red) for Boston South End and De Pijp. For each study area, two maps are presented: the isolated reconstructed polygon (red) and the reconstructed polygons overlaid onto the authoritative ground truth (blue) for spatial comparison. (continued on the next page). 64
5.7 Reconstructed sidewalk polygons (red) for Hobbemakade. For each study area, two maps are presented: the isolated reconstructed polygon (red) and the reconstructed polygons overlaid onto the authoritative ground truth (blue) for spatial comparison. 65

1 Existing OpenStreetMap (OSM) sidewalk data for Utrecht. The map illustrates the sparsity of explicitly mapped sidewalk centerlines (`footway=sidewalk`) and confirms the complete absence of 2D polygon representations (area:highway =footway). 82
2 Existing OpenStreetMap (OSM) pedestrian network data for Amsterdam. Similar to Utrecht, sidewalk polygon representations are entirely missing. The current dataset is only a centerline and represents only a fraction of the city's actual pedestrian infrastructure. 83

List of Tables

2.1 ADE20K semantic segmentation performance (frozen backbone). Higher mIoU is better. Results from Siméoni et al. (2025), Tables 3 and 11. 15
2.2 3D reconstruction accuracy (F1↑, %) on standard geometry benchmarks. "w/o poses" = model estimates poses internally; "w/ poses" = ground-truth camera poses provided (matching this pipeline's usage). Results from Lin et al. (2025). 17

4.1 Bearing ranges for the eight panoramic compass sectors. 31
4.2 Data quality classification criteria based on image count. 36

5.1 Spatial dimensions of the three study areas. 49
5.2 Image acquisition statistics per study area. To process panoramic imagery using the local perspective model, each panorama is reprojected into four individual perspective crops. 50
5.3 Comparison of Mapillary API and local DINOv3 segmentation results. "Images Processed" reflects the number of original images submitted to each segmentation source, with panoramic crops aggregated back to their parent image. An image is considered to detect a sidewalk if at least one side exceeds the minimum coverage threshold ($\tau = 0.05\%$). 52
5.4 Sidewalk inventory mapping results per study area. Each 5-meter road segment is evaluated independently on its left and right sides, yielding two roadside assessments per segment. Each roadside receives one of three classifications: `true` (sidewalk detected), `false` (no sidewalk detected despite available imagery), or `no_data` (insufficient imagery coverage to make any determination). 56
5.5 Accuracy metrics derived from the expanded confusion matrix. Algorithmic metrics evaluate only road-sides where the model made a decision; System Recall includes coverage gaps. 60
5.6 Pipeline 2 geometry reconstruction output per study area. 62
5.7 Area-based accuracy metrics for Pipeline 2 geometry reconstruction. Ground truth is clipped to the imagery-covered area to exclude regions where no reconstruction was possible. 64

5.8 Dimensional and positional accuracy of reconstructed sidewalk polygons. Width errors reflect the difference between predicted and ground-truth width estimations. Centerline offset measures the lateral displacement between matched polygon centroids. 66

Acronyms

API Application Programming Interface . . . 2
BGT Basisregistratie Grootschalige Topografie . . . 50
CC BY-SA Creative Commons Attribution-ShareAlike . . . ix
CSVI Crowdsourced Street View Imagery . . . 3
ENU East-North-Up . . . 17
EXIF Exchangeable Image File Format . . . 11
GeoJSON Geographic JavaScript Object Notation . . . 9
GNSS Global Navigation Satellite System . . . 11
GSV Google Street View . . . 2
GSVD Google Street View Depth . . . 10
JOSM Java OpenStreetMap Editor . . . 9
LiDAR Light Detection and Ranging . . . 36
mIoU mean Intersection over Union . . . 14
MVS Multi-View Stereo . . . 18
MVT Mapbox Vector Tile . . . 12
ODbL Open Database License . . . ix
OSM OpenStreetMap . . . 7
PDOK Publieke Dienstverlening Op de Kaart . . . x
SfM Structure from Motion . . . 3
SLI Street-Level Imagery . . . 9
SVI Street View Imagery . . . vii
ViT Vision Transformer . . . 14
VRAM Video Random Access Memory . . . 67

1 Introduction

1.1 Background and Motivation

Sidewalks are a fundamental component of the urban infrastructure, providing essential protected space for pedestrian movement and access to urban facilities (Hamim et al., 2023; Morra et al., 2024). Beyond serving as transit corridors, they are vital for public health, social inclusion, and urban vitality (Hosseini et al., 2023). For individuals with disabilities, the elderly, and those with young children, the availability and quality of sidewalks directly impact independence, safety, and overall quality of life (Weld et al., 2019; Xia et al., 2023). Research reveals that walkability acts as a critical driver for sustainable urban development, positively influencing environmental resilience, pedestrian safety, and public health (Zuniga-Teran et al., 2017).

Despite their significance, pedestrian access has declined in many cities over the past century due to vehicle-centered urban planning, leaving many pedestrian networks fragmented or entirely absent (Hosseini et al., 2023). This lack of safe, dedicated space restricts the mobility of vulnerable populations and often forces them into shared roadways. This exposure significantly increases the risk of injury; roads without sidewalks experience pedestrian crashes at a rate 1.67 times higher than those with them (Abou-Senna et al., 2022; Aceves-González et al., 2020). For decades, urban planners have advocated for healthier urban designs centered on walkable communities to counter these risks and deficiencies (Forsyth and Southworth, 2008).

The primary obstacle to improving this infrastructure is a persistent lack of authoritative, standardized, and openly available sidewalk datasets (Zhou et al., 2025; Kang et al., 2021). While vehicular road networks have been digitally mapped with high precision, mapping

for pedestrian infrastructure remains significantly behind. This "data gap" hinders the development of advanced downstream applications, such as shade-based routing algorithms or accessibility assessments (Beuster et al., 2025; Hosseini et al., 2023). Consequently, vulnerable users, such as wheelchair travelers and the elderly, lack the accurate routing data required to navigate safely (Ning et al., 2022). Furthermore, the emergence of autonomous delivery robots creates a new demand for up-to-date and high-definition digital maps to operate effectively in complex urban environments (Figliozzi and Jennings, 2020).

Traditional methods for assessing sidewalk accessibility are increasingly as unsustainable for city-wide scale. Manual on-site inspections are labour-intensive, expensive, and prone to human error, often requiring years to complete for a single major city (Smith et al., 2013; Morra et al., 2024). While high-resolution satellite and aerial imagery offer scalability, they face significant occlusion limitations from building shadows and tree canopies, which often hide the narrow concrete surfaces of sidewalks (Hosseini et al., 2023; Beuster et al., 2025; Senlet and Elgammal, 2012). More accurate technologies such as mobile LiDAR are also available, but remain costly and technically complex for widespread use (Brunken and Gühmann, 2020; Ai and Tsai, 2016).

In recent years, Street View Imagery (SVI) has emerged as a promising alternative for mapping (Zhou et al., 2025). However, the dominant platform, Google Street View (GSV), operates under restrictive licensing that prohibits the extraction of features for open databases. Consequently, researchers are increasingly turning to crowdsourced platforms such as Mapillary. These open platforms provide high-resolution, ground-level perspectives capable of capturing fine-grained features such as pavement conditions and precise widths, often invisible to aerial sensors, and offer licenses that permit integration into OpenStreetMap (Ning et al., 2022; Lieu and Guhathakurta, 2025).

Although crowdsourced platforms host millions of street-level images, this raw visual data lacks the explicit geometric and semantic structure required for spatial analysis. To become actionable for applications such as pedestrian network routing, width estimation, and the creation of comprehensive infrastructure inventories, these images must be translated into structured vector geometries (e.g., linestrings or polygons). At present, a significant technological gap remains, as there is no fully automated, scalable method capable of performing this extraction with the geometric accuracy necessary to build and maintain reliable municipal sidewalk inventories.

1.2 Research Objectives and Scope

The goal of this MSc thesis is to bridge the disconnect between raw crowdsourced street-level imagery and actionable geospatial sidewalk data by developing and evaluating two complementary automated workflows that operate exclusively on openly licensed data. This research capitalizes on the substantial improvements in vision foundation models for semantic segmentation, such as DINOv3 (Siméoni et al., 2025), alongside the enhanced capabilities of monocular depth estimation architectures like Depth Anything V3 (Lin et al., 2025), which enable the transformation of 2D pixel-level masks and 3D depth data into georeferenced vector representations. The first workflow uses metadata and pre-computed outputs already available through the Mapillary Application Programming Interface (API) to produce a scalable sidewalk inventory (`yes/no` labels on road segments) without any local model inference. The second workflow downloads imagery and processes it locally using

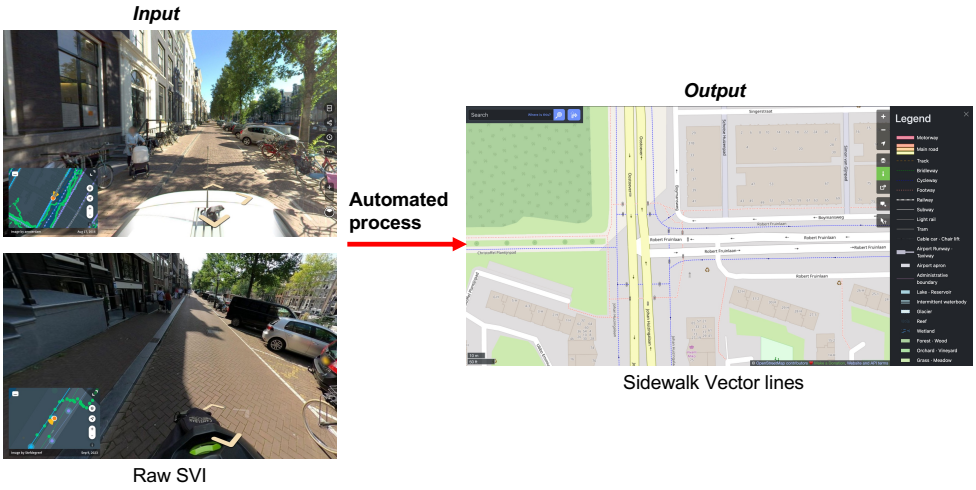


Figure 1.1: Illustration of the current gap in automated sidewalk mapping from crowdsourced SVI. **Left:** Diverse Mapillary images (vehicle top, pedestrian bottom) that are openly available but not automatically usable as geospatial data. **Right:** Sidewalk vectors (solid blue) that would enable routing and mapping if extracted automatically. Image credits: © Mapillary contributors (CC BY-SA 4.0, accessed 2025-01-06); basemap © OpenStreetMap contributors ODbL.

state-of-the-art segmentation and metric depth estimation models to reconstruct the physical geometry of sidewalk surfaces. This study aims to address a key global data gap and help in contributing to open data initiatives like OpenStreetMap (OpenStreetMap contributors, 2024), see Figure 1.1.

1.2.1 Objective

The primary goal of this thesis is to evaluate the feasibility of converting Crowdsourced Street View Imagery (CSVI) into georeferenced sidewalk data, ranging from binary presence inventories to vector geometries, using only open data sources and pre-trained models. The research is guided by the following central question:

Main Research Question:

- How can crowdsourced street-level imagery be transformed into structured georeferenced sidewalk data, spanning from lightweight presence inventories to metric vector geometries by using only open data and pre-trained models, and what are the accuracy and practical limitations of each approach?

To address this goal, the following research questions have been established:

RQ1: Sidewalk Inventory Mapping (The "What is there")
Question: To what extent can a scalable, city-wide sidewalk presence inventory be constructed by aggregating pre-computed semantic segmentation outputs and Structure from Motion (SfM)-corrected camera metadata already available through the Mapillary API, without performing any local model inference?

RQ2: Sidewalk Geometry Reconstruction (The "What does it look like")
Question: How effectively 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")
Question: How do the inventory-based approach and the geometry reconstruction approach compare in terms of detection accuracy, geometric precision, spatial coverage, and computational cost?

1.2.2 Scope of research

The scope defines the technical and geographic boundaries of this study to ensure feasibility.

- **Data Sources:** The research is strictly limited to open-source, crowdsourced street view imagery CSVI from Mapillary. This excludes proprietary platforms like Google Street View to ensure the resulting geometries can be legally integrated into open-data initiatives.
- **Technical Framework:** The study uses specific segmentation and depth models, including self-supervised foundation models (e.g., DINOv3) alongside monocular depth estimation (Depth Anything). The project focuses on the integration of these existing state-of-the-art models rather than the development of new neural network architectures.
- **Geometric Output:** The workflow generates yes/no tags and georeferenced vector polygons specifically for sidewalks; it does not map secondary assets like street furniture.
- **Geographic Boundary:** The pipeline will be primarily tested in two cities, i.e, Amsterdam in the Netherlands and Boston in the USA. It is selected due to the researcher's familiarity with the study area and the availability of ground-truth datasets for validation to measure positional and metric accuracy.

Out of Scope

To ensure the project remains feasible within the thesis timeline, the following aspects are explicitly excluded from the scope of this research:

- **Global Applicability:** While foundation models are designed for broad robustness, this study does not evaluate performance across all global urban contexts (e.g., informal or unpaved pedestrian infrastructure in rural or non-Western regions) without region-specific adaptation or fine-tuning.
- **Severe and Persistent Occlusions:** While common partial occlusions are addressed, extreme cases in which sidewalks are fully obscured for entire street segments (e.g., continuous blockage by parked trucks or construction barriers) are treated as data limitations rather than algorithmic failures.

- **Model Development and Data Collection:** This research applies existing pre-trained models (segmentation and depth estimation) without developing new architectures, training procedures, or fine-tuning. The study relies exclusively on open crowdsourced street-view imagery (Mapillary) and does not involve primary data collection or dedicated capture campaigns.

1.3 Results Overview

The two pipelines were evaluated across three study areas: De Pijp and Hobbemakade in Amsterdam (Netherlands), and Boston South End (United States), using authoritative ground truth sidewalk datasets.

The Sidewalk Inventory Mapping pipeline achieves algorithmic precision exceeding 97% across all study areas, confirming that positive sidewalk detections are highly reliable. Algorithmic recall reaches up to 86.5% where sufficient imagery exists. A practical constraint is that the Mapillary API does not currently provide pre-computed segmentation for panoramic images in all regions, necessitating local model inference for semantic segmentation in such cases. As in Boston, relying solely on pre-computed API segmentation yields 98.6% precision without any local GPU computation. Also, the dominant limitation is not algorithmic but spatial: system recall is bounded by Mapillary's image coverage, with road segments lacking imagery receiving no assessment.

The Sidewalk Geometry Reconstruction pipeline successfully produces georeferenced polygons and centrelines, demonstrating that monocular metric depth estimation can serve as a software-based alternative to LiDAR. However, geometric accuracy remains constrained. Width estimation errors average approximately 1.2-1.4 m, lateral positional offsets range from 2.6-3.2 m, and area-based IoU values span from 0.100 to 0.329. These limitations are primarily driven by sparse spatial coverage and the compounding effects of segmentation imprecision, depth estimation uncertainty, and inter-sequence GPS drift.

Computationally, the inventory approach requires only minutes on a standard CPU, while the geometry reconstruction requires several hours of GPU computation per square kilometre. These results establish the inventory mapping approach as a practical, scalable tool for city-wide deployment, while geometry reconstruction represents a proof-of-concept requiring further refinement.

1.4 Thesis Outline

This thesis is composed of six chapters, including this introduction. The content is structured as follows:

Chapter 2 presents the theoretical background underpinning the research. It begins with a formal definition of sidewalks and their representation within OpenStreetMap, followed by a description of street-view imagery and the advantages of crowdsourced platforms such as Mapillary. The chapter then introduces the deep learning models used for semantic segmentation (DINOv3) and monocular metric depth estimation (Depth Anything V3), and concludes with a discussion of the depth estimation strategy, including the experimental

motivation for choosing monocular depth over multi-view reconstruction for sidewalk surfaces.

Chapter 3 reviews related work across four methodological categories: traditional semantic segmentation from SVI, aerial imagery and network generation, geometric reconstruction and metric mapping from SVI, and participatory smartphone-based approaches. It synthesises the identified limitations of each category and positions the current research within the existing landscape.

Chapter 4 details the methodology for both automated pipelines. The first, Sidewalk Inventory Mapping, describes the four-phase workflow for constructing binary sidewalk presence labels by aggregating Mapillary segmentation outputs onto the OpenStreetMap road network. The second, Sidewalk Geometry Reconstruction, presents the five-phase pipeline that combines semantic segmentation, metric depth estimation, and geometric back-projection to produce georeferenced sidewalk polygons, centerlines, and width estimates.

Chapter 5 presents the results and accuracy assessment. It introduces the three study areas and their ground truth datasets, reports the segmentation outputs from both the Mapillary API and the local DINOv3 model, evaluates the inventory mapping pipeline using a 3×2 expanded confusion matrix, assesses the geometry reconstruction pipeline through area-based overlap and dimensional accuracy metrics, and concludes with a comparison of computational cost and scalability between the two approaches.

Finally, Chapter 6 offers conclusions by directly answering the research questions, reflecting on the strengths and limitations of both pipelines in a discussion, and providing recommendations for future research directions.

2

Theoretical Background

2.1 Sidewalk Definition

A sidewalk is formally defined as *"a paved or improved path for pedestrians, usually at the side of a road"* (Merriam-Webster, 2024). In the context of geographic data and urban infrastructure mapping, the OpenStreetMap community defines a sidewalk as a pedestrian facility physically associated with a road segment rather than a standalone path, existing on one, both, or neither side of the carriageway (OpenStreetMap contributors, 2024). Structurally, these features may be separated from the road by curbs, verges, or pavement markings; capturing this specific spatial relationship is highly critical for accurate pedestrian routing applications.

2.2 OSM Sidewalks

In OpenStreetMap (OSM), three conceptual models are typically used by mappers, each with different trade-offs in terms of structural simplicity and geometric accuracy. At first, sidewalks can be recorded implicitly as `tag attributes` on the parent roadway itself (e.g., sidewalk=both, sidewalk:left=yes, sidewalk:right=yes), without any separate geometry. This approach is structurally simple but does not produce any independent spatial feature; the sidewalk's existence is recorded only as metadata on the road centerline and cannot be visualized or routed independently. In the first pipeline (inventory mapping), this tag-attribute model is used to represent sidewalk presence.

The second is as a `separate linear way` or sidewalk centerline, tagged with *highway=footway* and `footway=sidewalk`, which records the pedestrian path as an independent line geometry running parallel to the associated road. This line-based representation enables pedestrian routing to compute paths through a connected graph of edges and nodes. However, a simple line does not capture the variable width, exact curb geometry, or surface irregularities of the physical sidewalk surface. As shown in Figure 2.1, the third representation model is as an `area polygon`, tagged with *area:highway=footway* combined with *footway=sidewalk*, which records the full physical extent of the sidewalk surface as a closed way or multipolygon relation (OpenStreetMap Wiki, n.d.d).

Regarding the current state of sidewalk data in OSM for the study areas of this research: in Amsterdam, only a small fraction of sidewalks have been explicitly mapped as *footway=sidewalk* centerlines, and polygon-based sidewalk representations (*area:highway=footway*) are entirely absent. The situation in Utrecht is identical, relying exclusively on sparse 1D `LineStrings` with zero polygon geometries present. Figures 1 and 2 in Appendix 6.3 visually illustrate the sparsity of this existing data.



Figure 2.1: Two mapping examples from the OSM wiki (`area:highway=footway`): (Left) A sidewalk mapped as a simple closed polygon outlined in red. (Right) A sidewalk mapped as a multipolygon featuring an inner ring outlined in red. Source: (OpenStreetMap Wiki, n.d.e).

2.2.1 OSM enrichment procedure and results

Despite the potential value of enriching OSM with the sidewalk geometries produced by this research, direct upload to the OSM database is not pursued within this thesis for several procedural and practical reasons. The OSM community enforces strict import guidelines requiring community consultation through mailing lists and user groups, detailed documentation of data conversion and conflation procedures, and verification of license compatibility with the Open Database License (OpenStreetMap Wiki, n.d.a). The Automated Edits Code of Conduct further stipulates that imports of algorithmically derived data should only be conducted by contributors with experience in OSM community practices and only after local community approval (OpenStreetMap Wiki, n.d.b). Successful precedents illustrate the time investment involved: the Seattle sidewalk import was vetted line-by-line by the local community through phased mapathons rather than automated bulk upload (OpenStreetMap Wiki, n.d.c), and the San Jose import divided the city into approximately 640 census-block

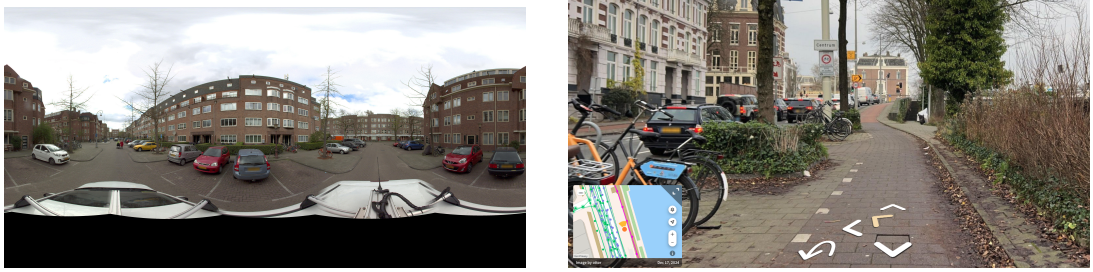
sections, each requiring individual reviewer validation. These procedural requirements extend well beyond what is achievable within the timeline of a master's thesis.

For these reasons, this research produces sidewalk data as `Geographic JavaScript Object Notation (GeoJSON) files`, and the output includes both polygon geometries representing the physical sidewalk footprint and skeletonised centerlines suitable for routing applications. GeoJSON is natively supported by Java OpenStreetMap Editor (JOSM), a widely used OSM editor, and can be loaded directly as a reference layer for manual mapping. By hosting these files on a publicly accessible repository with documentation, the data remains independently usable for urban planning and accessibility analysis.

2.3 SVI Description

Street view imagery SVI, also referred to as Street-Level Imagery (SLI), terms used interchangeably in the literature, consists of georeferenced, ground-level photographs captured along a trajectory such as a street or footpath. Unlike satellite or aerial imagery, which provides a top-down perspective, SVI captures the environment from a human vantage point, the same viewpoint experienced by a pedestrian, cyclist, or driver. This ground-level perspective makes SVI particularly well-suited for detecting urban features, such as pavement surfaces, kerb cuts, and street furniture, that are either invisible or ambiguous when viewed from above (Biljecki and Ito, 2021).

Each image in an SVI dataset is associated with a geographic position and a compass bearing, enabling spatial queries and directional analysis. Images are captured using a range of hardware, from professional 360° camera rigs mounted on survey vehicles to smartphones carried by everyday contributors. Panoramic cameras record a full equirectangular projection of the surrounding environment (360° × 180°), while perspective cameras record a single forward-facing view with a limited horizontal field of view. Both types are present in modern SVI datasets and require different analytical treatment when used for automated feature extraction, as shown in Figure 2.2.



(a) Panoramic street-view image.

(b) Perspective street-view image.

Figure 2.2: Examples of street-view imagery from Mapillary: (a) panoramic image captured from a specialized vehicle and (b) perspective image captured by a pedestrian (Mapillary, 2025a).

SVI is available through both commercial and crowdsourced platforms. Commercially operated services like Google Street View, Baidu Maps Total View, and Microsoft Bing Streetside

provide high-quality, globally consistent imagery captured by dedicated survey fleets but operate under restrictive licensing terms that prohibit the extraction of derived data for open mapping initiatives. Whereas platforms such as Mapillary and KartaView, by contrast, rely on crowdsourced imagery from smartphones and dashcams and distribute their content under open licenses (e.g., CC BY-SA 4.0) that explicitly permit the derivation of geospatial features for use in open databases such as OpenStreetMap. For this reason, this research relies exclusively on Mapillary as its SVI source.

2.3.1 The Shift to crowdsourced SVI data: Mapillary API

Mapillary is a crowdsourced street-level imagery platform owned by Meta Platforms, Inc., containing a global archive of more than three billion geotagged photographs contributed by volunteers, municipalities, and professional mapping organisations (Mapillary, 2025a). The following section outlines the main advantages of using Mapillary for this research.

Open Crowdsourcing vs. Proprietary Systems

To date, the majority of SVI-based studies have relied on Google Street View GSV; this preference stems from the refined nature of GSV data: images are captured using high-end, professionally calibrated sensors with consistent heights and orientations, and provided with proprietary depth maps Google Street View Depth (GSVD) that simplify 3D reconstruction (Ning et al., 2022). However, GSV operates as a "closed-loop" system; its Terms of Service prohibit feature extraction for open initiatives like OpenStreetMap OSM, creating a gap between research findings and map implementation (Mapillary, 2025a).

By contrast, crowd-source SVI platforms like Mapillary operate under CC BY-SA 4.0 licenses, granting explicit permission for users to derive metadata and geometry for OSM (Mapillary, 2025b; Biljecki and Ito, 2021). This ensures data license compatibility, guaranteeing that the resulting sidewalk datasets remain openly available for public planning.

Pedestrian-Centric Perspectives and Data Coverage

Unlike commercial fleets bound by rigid multi-year update cycles, CSVI represents a continuously growing, dynamic dataset. Contributors of CSVI can immediately map unrecorded areas or update imagery to reflect recent infrastructure changes and upload fresh imagery (Ito et al., 2025), effectively removing the dependency on a private corporation's collection. In contrast, proprietary platforms like Google Street View acknowledge their content can be "anywhere from a few months to a few years old," leading to significant latency between real-world changes and data availability (Google, 2023).

CSVI is frequently captured by everyday pedestrians and cyclists rather than vehicle-mounted cameras. This pedestrian-centric perspective offers significant technical advantages for sidewalk extraction: it is much less susceptible to occlusion from parked cars.

Geometric Integrity through SfM-Corrected Metadata

A core technical advantage of Mapillary is its use of Structure from Motion SfM via the OpenSfM library (Mapillary, 2023). Beyond sparse point clouds, this process generates refined camera pose metadata that provides a stable geometric foundation superior to raw Global Navigation Satellite System (GNSS) tags.

Correcting "Urban Canyon" Error: Standard GNSS sensors in consumer-grade devices, frequently used by the Mapillary community, are prone to multipath interference in dense urban environments, with positional errors often exceeding 10 meters (Merry and Bettinger, 2019). To mitigate this "urban canyon" effect, Mapillary applies Structure-from-Motion SfM reconstruction to image sequences. Starting from noisy GNSS-based camera priors, bundle adjustment jointly optimises camera poses and 3D scene structure by enforcing geometric consistency across overlapping images (Schönberger and Frahm, 2016). By minimising reprojection errors of shared visual features observed from multiple viewpoints, this process smooths out erratic GNSS jumps.

Crucially, this correction is not guaranteed for every image in the database. Successful SfM optimization requires sufficient visual overlap within an image sequence; isolated images or sequences with poor feature matching will fail this process. For images where reconstruction is successful, the refined camera locations are stored in the `computedLngLat` attribute. While Mapillary does not provide exact metric accuracy values for this attribute, these computed coordinates better reflect the true street trajectory and provide a substantially more reliable geospatial baseline than raw, uncorrected GNSS measurements (Mapillary, 2023).

Metadata Availability

Each image in the Mapillary archive is associated with a set of metadata fields that describe its spatial position, orientation, and capture context. Of particular importance for this research are the computed position and computed compass angle, which are derived from Mapillary's Structure-from-Motion SfM reconstruction pipeline rather than from the raw Exchangeable Image File Format (EXIF) metadata embedded in the image file. For the full information on available attributes, refer to Mapillary (2025c).

Images are captured using a wide variety of devices, ranging from smartphones mounted on bicycle handlebars to professional 360° panoramic camera systems installed on survey vehicles; see Figure 2.2. This diversity of capture hardware is a defining characteristic of the dataset and has direct methodological implications, and images from perspective cameras and panoramic cameras must be handled separately.

Mapillary API semantic segmentation: trained on Vistas Dataset

In addition to imagery and positional metadata, Mapillary provides pre-computed semantic segmentation results accessible through its Detections API, see Figure 2.3. This capability was first introduced in 2016, when Mapillary applied deep neural network-based pixel-level classification to over 20 million images on the platform, enabling automatic recognition of street-scene objects including roads, buildings, vegetation, and pedestrian surfaces (Mapillary, 2016). The segmentation results are served through the API endpoint

`/{image_id}/detections` as vector polygons encoded in the Mapbox Vector Tile Mapbox Vector Tile (MVT) format, where each detection polygon is associated with a semantic class from the Mapillary Vistas taxonomy. These MVT-encoded polygons can be decoded and rasterised into binary masks for any given semantic class. For this research, detections belonging to the `construction--flat--sidewalk` class are used to generate binary sidewalk masks.



Figure 2.3: Mapillary API sidewalk semantic segmentation

The semantic segmentation models are trained using the Mapillary Vistas Dataset (Neuhold et al., 2017), a widely used benchmark for street-level scene understanding.

- **Scale and Diversity:** The dataset contains approximately 25,000 high-resolution street-view images captured across six continents under diverse weather, seasonal, and illumination conditions. The imagery reflects the heterogeneity of crowdsourced data, including variations in camera devices, viewpoints, and image quality.
- **Annotations:** Images are annotated using dense, pixel-accurate polygon labels across 124 semantic object categories in the latest version, enabling fine-grained semantic segmentation.
- **Pedestrian Infrastructure:** Crucially for this study, the dataset explicitly labels pedestrian-related surfaces such as *sidewalk* and *curb*, clearly distinguishing them from visually similar classes (e.g., bike lanes, terrain, or road markings) (see Figure 2.4).
- **Usage:** The dataset is partitioned into training, validation, and testing splits to support supervised model end-to-end training.

2.4 Deep Learning Models for Semantic Segmentation and Metric Depth Estimation

This section provides the theoretical background for semantic segmentation and metric depth estimation models used in the pipeline. The segmentation model assigns a binary sidewalk label to every pixel; the depth model creates a depth map for each pixel in meters.

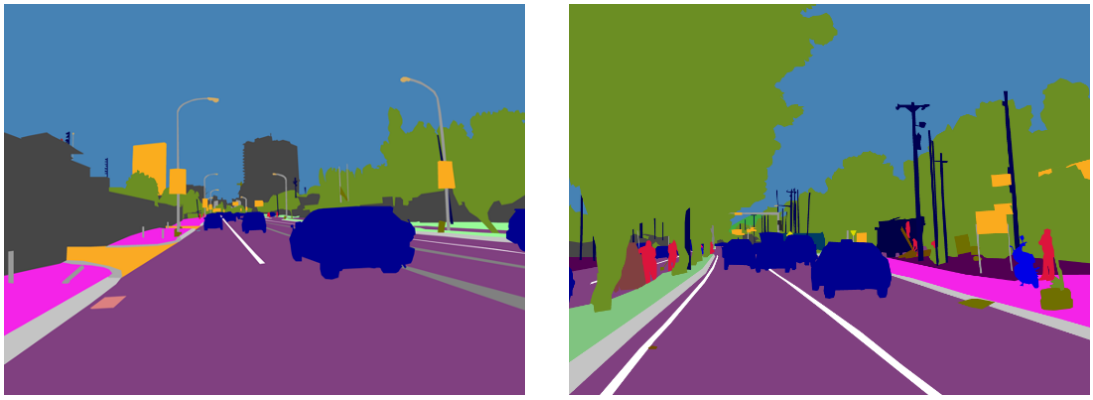


Figure 2.4: Examples from the Mapillary Vistas v2 dataset highlighting sidewalks (class 24, pink), clearly distinguished from roads and other surfaces. These annotations support semantic segmentation model training for automated sidewalk extraction.

2.4.1 Semantic Segmentation

Semantic segmentation assigns a class label to every pixel in an image. In this pipeline, the task is reduced to binary classification: each pixel is classified as a `sidewalk` or `background`. The input is a 1024 × 1024 perspective crop, and the output is a binary mask of identical dimensions.

Two segmentation approaches are used across the two pipelines. For the inventory mapping pipeline, pre-computed segmentation available through the Mapillary Detections API is used directly (Section 2.3.1), no local model inference is performed, and the segmentation model running server-side at Mapillary is treated as a black box trained on the Vistas dataset. For the geometry reconstruction pipeline, a local segmentation model is required to produce full-resolution masks from downloaded imagery. The following four state-of-the-art architectures were identified as candidates for this local inference task, representing the evolution from traditional supervised learning to foundation and open-vocabulary models. Each model embodies a distinct approach to learning and transferring semantic knowledge.

- **DeepLabV3+ (Supervised Baseline):** (Chen et al., 2018) A fully supervised convolutional segmentation model trained on the Mapillary Vistas dataset.
- **DINOv3 (Feature Foundation Model):** (Siméoni et al., 2025) A self-supervised vision transformer pretrained on large-scale unlabeled datasets.
- **SAM (Segment Anything Model – Promptable Foundation Model):** (Kirillov et al., 2023) A class-agnostic segmentation model that produces high-quality masks given spatial prompts.
- **ODISE (Open-Vocabulary Vision–Language Model):** (Xu et al., 2023) An open vocabulary segmentation model that leverages joint text–image embeddings to identify objects based on natural language descriptions.

For this research, we choose DINOv3, which also comes with a `Mask2Former` head (Cheng et al., 2021) segmentation head that has been trained on the large-scale labelled `ADE20K`

dataset (Zhou et al., 2019), including a sidewalk class, making it directly applicable to sidewalk extraction without any fine-tuning or additional training. The model weights are not publicly hosted, and access was obtained by submitting a request to Meta Platforms through their model release programme. At the time this research was conducted, DINOv3 achieved among one of the highest mean Intersection over Union (mIoU) scores on the ADE20K benchmark (see Table 2.1), positioning it as a strong candidate for dense segmentation tasks. Due to time constraints within the thesis timeline, a systematic comparison across all four candidate architectures was not feasible. DINOv3 was therefore selected as the primary segmentation model based on its benchmark performance and the availability of a ready-to-use sidewalk-class head.

DINOv3 : Vision Transformer Backbone

The semantic segmentation model utilised is `DINOv3 ViT-7B/16` (Siméoni et al., 2025), a 7-billion parameter foundation model built upon the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020). Unlike traditional convolutional networks, the ViT architecture relies on Multi-Head Self-Attention mechanisms. This allows the model to capture long-range spatial dependencies across the full image in a single layer, a critical property for maintaining the continuity of elongated infrastructural features like sidewalks. Furthermore, DINOv3's self-supervised pre-training on LVD-1689M (1.69 billion curated web images) yields highly robust, dense feature representations that generalise across diverse visual domains (Siméoni et al., 2025), translating effectively to the unstructured, real-world street imagery utilised in this pipeline.

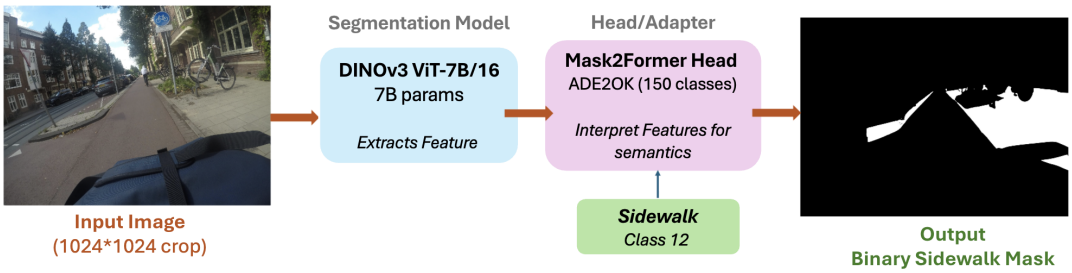


Figure 2.5: The inference workflow for sidewalk presence detection. Raw street-level imagery is processed by the DINOv3 ViT-7B/16 foundation backbone to extract dense spatial features. These representations are subsequently decoded by the Mask2Former head to generate the final sidewalk segmentation mask.

To isolate the sidewalk class from the dense feature maps of the DINOv3 backbone, the pipeline employs a Mask2Former head (`dinov3_vit7b16_ade20k_m2f_head`) provided by Meta. While the Vision Transformer (ViT) backbone extracts task-agnostic feature embeddings, the Mask2Former adapter is specifically tasked with interpreting these features to identify semantic classes, including the sidewalk category (class index 12 in the ADE20K dataset (Zhou et al., 2019)).

The complete inference workflow, from the raw input imagery through the DINOv3 backbone and Mask2Former head to the final semantic mask, is illustrated in Figure 2.5. For images exceeding the native training resolution, a sliding-window tiling strategy of 1024 is applied with a 128-pixel overlap, averaging predictions in the overlap regions to maintain boundary continuity.

Benchmark Performance and Inference Time

Table 2.1 reports DINOv3 performance on the ADE20K benchmark (Zhou et al., 2019). The metric is mean Intersection over Union (mIoU), measuring the ratio of overlap between the predicted mask and the ground truth on a scale of 0-100.

Table 2.1: ADE20K semantic segmentation performance (frozen backbone). Higher mIoU is better. Results from Siméoni et al. (2025), Tables 3 and 11.

Model	ViT	mIoU (linear)	mIoU (Mask2Former)	Cityscapes (linear)
DINOv2	g/14	49.5	—	—
DINOv3	7B/16	55.9	**63.0**	**81.1**

The Mask2Former head employed in this pipeline leverages the frozen DINOv3 backbone to achieve 63.0 mIoU on the 150-class ADE20K benchmark (Siméoni et al., 2025). This comprehensive multi-class capability means the sidewalk class benefits from joint training with all 150 categories, encouraging robust boundary discrimination against semantically similar surfaces (e.g., road, plaza, path). In the context of this pipeline, high boundary precision is the most critical property. Figure 2.5 visually confirms that the model produces the sharp, precise boundary delineations necessary for this task. Maintaining this accuracy is crucial to ensure that misclassified edge pixels do not propagate as systematic width bias during the geometric projection process of the sidewalk geometry reconstruction pipeline.

The 6.7-billion parameter ViT-7B/16 model processes a single 1024×1024 crop in approximately 0.3-0.5 seconds on an NVIDIA V100 GPU. Processing four crops per panorama sequentially yields a throughput of 5-8 panoramas per minute, allowing a study area comprising 1000 panoramas to be processed within 3-4 hours on a single node.

2.4.2 Metric Depth Estimation

Monocular metric depth estimation predicts the absolute distance in metres from the camera to the scene surface at every pixel. Metric depth, as opposed to scale-ambiguous relative depth, is strictly required in this pipeline to project sidewalk pixels into physically accurate world coordinates. The input is the same 1024×1024 perspective crops used by the segmentation model, and the output is a per-pixel depth map in metres. The theoretical motivation for choosing monocular depth over traditional multi-view reconstruction is detailed in Section 2.5.

While alternative metric depth estimation models, such as ML Depth Pro (Bochkovskii et al., 2025) and Metric Anything (Ma et al., 2026), were evaluated, Depth Anything 3 (Lin et al., 2025) was ultimately selected due to its native *pose-conditioned* inference mode, which accepts

multiple posed images in a single forward pass and enforces geometric consistency across views without any external post-processing.

Depth Anything 3

The model utilised is `DA3-Giant` (1.10 B parameters), the highest-capacity variant of the Depth Anything 3 (DA3) foundation model (Lin et al., 2025). DA3 is built on a single-plane vision transformer backbone (DINOv2 (Oquab et al., 2024)) without complex architectural specialisations. This design inherits the strong feature representations of large-scale self-supervised pre-training, which generalise effectively to the diverse urban scenes encountered in street-level imagery.

The key capability that distinguishes DA3 from single-image depth models is its *pose-conditioned* inference mode. When known camera parameters are provided, they are encoded into compact *camera condition tokens* via a lightweight MLP and concatenated with the patch tokens of each view. These tokens participate in all attention operations, including an *input-adaptive cross-view self-attention* in the final one-third of the transformer layers, where tokens from all input views are grouped together. This provides the network with explicit geometric context, the spatial arrangement of all input cameras, allowing it to reason about scene geometry jointly rather than estimate depth for each image in isolation.

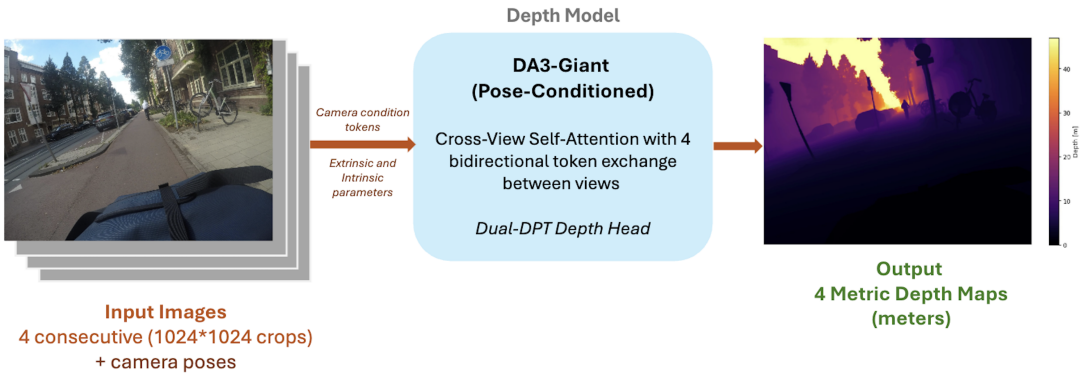


Figure 2.6: Pose-conditioned DA3 inference pipeline. A sliding window of four consecutive street-level images and their known camera parameters (world-to-camera extrinsics from SfM; intrinsics from the crop geometry) are passed to the DA3-Giant model. Camera condition tokens inject geometric context into all attention layers. Cross-view self-attention exploits spatial parallax between views to produce geometrically consistent metric depth maps. The output is four metric depth maps.

The complete pose-conditioned inference pipeline, from the input images and camera parameters through the DA3 backbone to the output metric depth maps, is illustrated in Figure 2.6.

Multi-View Parallax and Geometric Consistency

When the same scene is observed from multiple camera positions, objects at different distances shift by different amounts across views, a phenomenon known as spatial parallax.

This differential shift provides a direct geometric cue for absolute depth that a single image cannot access. In the sidewalk geometry reconstruction pipeline, a sequence window strategy is used where four consecutive images from the same Mapillary capture sequence (typically spaced 2-5 m apart) are passed to DA3 in each pose-conditioned inference call, providing substantial parallax at the 3-15 m distances relevant to sidewalk geometry (Figure 2.6).

For each image, the world-to-camera extrinsic matrix is constructed from the SfM rotation (`computed_rotation`) and the GPS-derived position in a local East-North-Up (ENU) coordinate system (in metres); the intrinsic matrix encodes the focal length and principal point of the perspective crop. By supplying these poses with the default *align_to_input_ext_scale* = *True*, DA3 anchors its depth predictions to the metric scale of the input extrinsics; the output is directly generated in metres with no post-hoc calibration. The cross-view attention exploits the parallax to enforce consistency, suppressing scale drift between overlapping views.

Benchmark Performance and Inference Time

Table 2.2 reports 3D reconstruction accuracy (F1 score) on standard geometry benchmarks. The F1 metric measures the proportion of reconstructed 3D points that lie within a distance threshold of the ground-truth surface (and vice versa); higher values indicate more geometrically accurate depth. Both pose-free inference (the model estimates its own camera poses) and pose-conditioned inference (ground-truth poses provided, as in this pipeline) are shown to demonstrate the benefit of supplying known camera parameters.

Table 2.2: 3D reconstruction accuracy (F1↑, %) on standard geometry benchmarks. "w/o poses" = model estimates poses internally; "w/ poses" = ground-truth camera poses provided (matching this pipeline's usage). Results from Lin et al. (2025).

Model	**HiRoom**	**ETH3D**	**7Scenes**	**ScanNet++**
Without poses (model estimates cameras internally)				
VGGT	56.7	57.2	47.9	66.4
Pi3	75.8	72.7	44.2	63.1
DA3-Giant	85.1	79.0	53.5	77.0
With known poses (pose-conditioned, as used in this pipeline)				
VGGT	70.2	66.7	51.4	70.7
Pi3	85.0	80.6	57.5	73.3
DA3-Giant	**95.6**	**87.1**	56.5	**79.3**

Two findings are relevant. First, DA3-Giant in pose-conditioned mode achieves the highest reconstruction accuracy on three of four benchmarks, including ETH3D, an outdoor dataset whose scenes are closest to the urban street setting of this pipeline. The F1 on ETH3D uses a distance threshold of 0.25 m: an F1 of 87.1 means that 87% of depth-projected 3D points land within 25 cm of the true surface. Second, supplying known poses substantially improves accuracy (e.g., DA3-Giant on ETH3D: 79.0 → 87.1). This confirms that the pose-conditioned approach adopted in this pipeline extracts meaningfully better depth than processing images independently.

Computationally, the DA3-Giant model processes a pose-conditioned batch of four 1024 × 1024 images in approximately 1-2 seconds on an NVIDIA V100 GPU. Processing a study

area of several hundred panoramas requires approximately two to four hours on a single GPU node.

2.5 Depth Estimation Strategy for Sidewalk Reconstruction

Recovering the physical geometry of sidewalk surfaces from street-level imagery requires per-pixel depth. Two paradigms exist: multi-view reconstruction, which triangulates 3D points from pixel correspondences across overlapping images, and monocular depth estimation, which predicts depth from a single image using learned geometric priors. This section motivates the choice of the monocular depth estimation approach discussed in the previous section for the geometry reconstruction pipeline.

2.5.1 Multi-View Reconstruction and Its Limitations for Sidewalks

Structure-from-Motion (SfM) and Multi-View Stereo (MVS) methods reconstruct 3D scene geometry by detecting corresponding features in overlapping images, matching these features across views, and triangulating their 3D positions (Schönberger and Frahm, 2016). For textured, geometrically complex surfaces, such as building façades, poles, and vegetation, these methods produce dense, accurate point clouds. The OpenSfM library (Mapillary, 2023) implements such a pipeline and was initially considered for this research as a means to reconstruct sidewalk surfaces from overlapping Mapillary sequences directly.

However, sidewalk surfaces present a particularly challenging case for feature-based reconstruction. Multi-view methods are known to produce sparse or unreliable reconstructions on textureless, planar surfaces (Aanæs et al., 2016). Paved sidewalks are characteristically:

- Flat and planar: minimal geometric variation provides few depth discontinuities for stereo matching to exploit.
- Low-texture: uniform concrete, asphalt, or tile surfaces lack the distinctive local patterns required for reliable feature detection and matching.
- Repetitive: where texture does exist (e.g., regular tile patterns), it creates ambiguous correspondences that lead to incorrect matches.
- Affected by illumination variation: shadows from buildings and trees create apparent texture that shifts between views captured at different times, producing false matches.

As a result, while the overall street scene typically reconstructs with good density, structures like buildings, poles, vehicles, and vegetation yield abundant feature points, the sidewalk region itself consistently produces sparse, unreliable reconstructions. To validate this limitation, preliminary experiments were conducted using the OpenSfM library on Mapillary sequences from the study area. These tests confirmed that the flat, untextured sidewalk surfaces yielded 3D points only in the single digits per sequence via triangulation; see Figure 2.9. This density is far too sparse to fit a reliable geometric plane, let alone extract meaningful width measurements or continuous polygon geometries.

As an alternative to sparse feature matching, dense optical flow methods were also evaluated. Specifically, RAFT (Recurrent All-Pairs Field Transforms for Optical Flow) (Teed and



Figure 2.7: 3D points reconstructed from RAFT-based dense optical flow correspondences, generating a large number (20000) of matches across consecutive images (100 images); the resulting feature points on sidewalk surfaces are noisy and geometrically inconsistent with substantial vertical and lateral dispersion.



Figure 2.8: Sidewalk polygons generated by the geometric flat-ground projection method overlaid on a satellite basemap.

Deng, 2020), a recurrent neural network designed to estimate dense per-pixel correspondences between consecutive images, was applied to overlapping image pairs to derive depth via triangulation. Although RAFT successfully produces accurate flow fields in highly textured regions, its resulting depth estimates on sidewalk surfaces proved noisy and spatially incoherent, often displacing sidewalk feature points into physically implausible vertical and

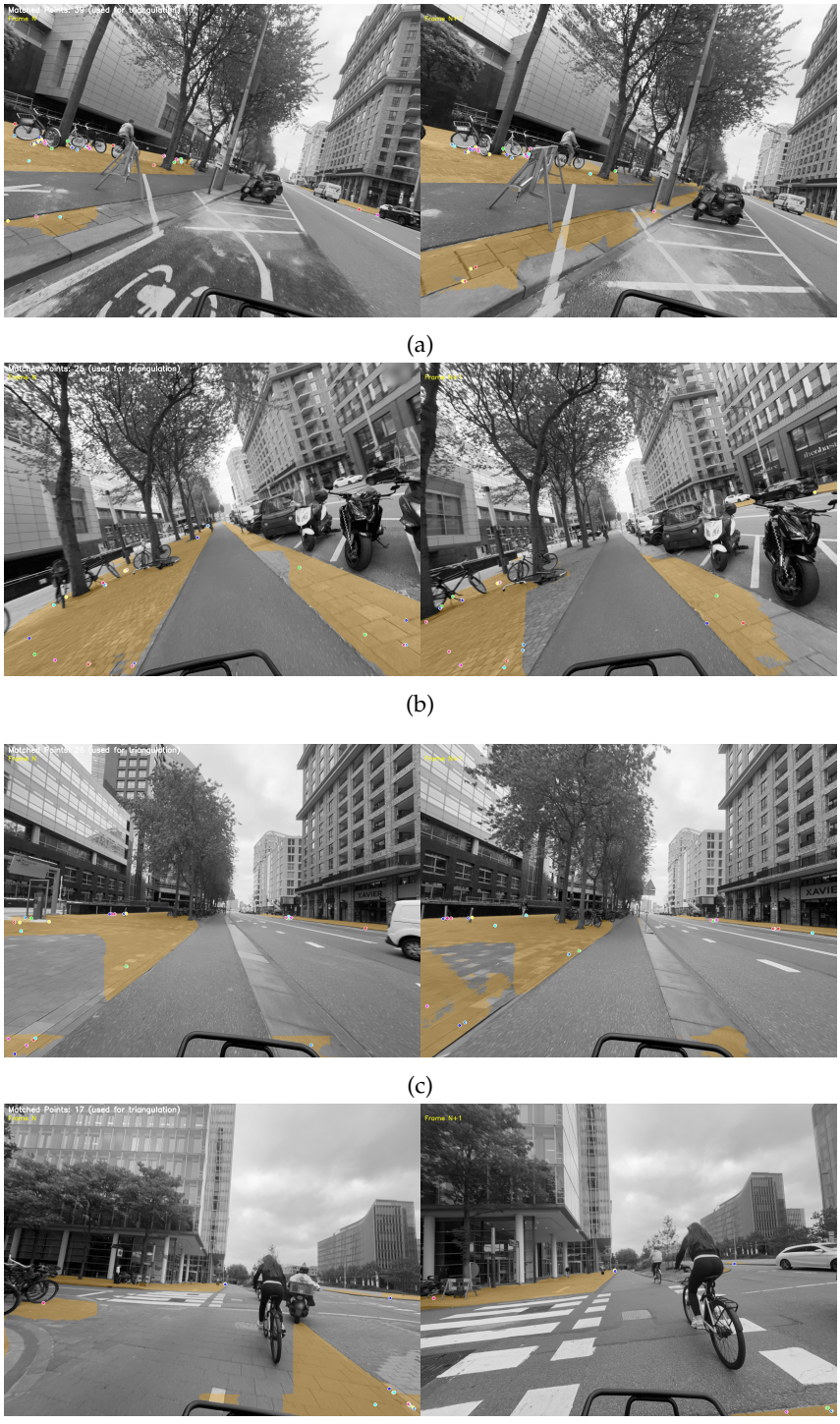
lateral positions (Figure 2.7). Ultimately, the fundamental limitation persists: as sparse matching, dense optical flow relies heavily on photometric consistency between views, a requirement that flat, untextured sidewalk surfaces fail to reliably satisfy.

Monocular depth estimation models overcome this limitation entirely. Rather than establishing pixel correspondences across views, these models predict depth from a single image by leveraging deep, learned priors about scene geometry, perspective cues, relative object sizes, occlusion relationships, and semantic context (Yang et al., 2024). Because the prediction is driven by this geometric intuition rather than local feature matching, monocular models estimate depth for flat, untextured surfaces with the same reliability as highly textured regions. This texture-independent property motivates the adoption of monocular metric depth estimation (Depth Anything 3) for the geometry pipeline (Lin et al., 2025). This ability to deduce 3D structure from learned geometric priors makes monocular metric depth estimation (Depth Anything 3) the ideal architectural choice.

Geometric Flat-Ground Projection

A conceptually simpler alternative was also investigated: a geometric flat-ground projection method that bypasses both feature matching and learned depth estimation entirely. Using the camera's GPS coordinates and the capture height h_{cam} reported in the Mapillary image metadata, rays are cast from the camera origin through each sidewalk-masked pixel and analytically intersected with the ground plane $Z = -h_{\text{cam}}$. The resulting 3D intersection points are then transformed to geographic coordinates using the camera's position and orientation and snapped longitudinally to the nearest OpenStreetMap road segment to regularise the output. While this approach is computationally inexpensive and does not require a learned depth model, its accuracy is fundamentally limited by the quality of the input GPS coordinates. Consumer-grade GNSS receivers, such as those in smartphones and action cameras used for Mapillary capture, exhibit vertical position errors on the order of several metres (Merry and Bettinger, 2019). Since the flat-ground assumption encodes all depth information in the camera height h_{cam}, even small vertical GPS error propagates directly into the lateral displacement of the projected sidewalk points.

As shown in Figure 2.8, the resulting sidewalk polygons are spatially inconsistent: projected boundaries remain susceptible to GPS errors, particularly in the vertical component (camera height), even after SfM correction, frequently overlapping with the road surface or adjacent buildings. Furthermore, the flat-ground assumption inherently cannot accommodate elevated or sloped road geometry, such as bridges, overpasses, or cambered streets, where the true ground surface deviates substantially from a single horizontal plane. These limitations rendered the geometric projection approach insufficiently reliable for generating the accurate, continuous sidewalk polygons required by the inventory pipeline.



(a)

(b)

(c)

(d)

Figure 2.9: Examples from two consecutive image sequences showing the scarcity of feature correspondences on sidewalk surfaces.

3

Related work

The integration of artificial intelligence (AI) into pedestrian infrastructure mapping has evolved along four primary methodological paths. These approaches range from fundamental image classification to sophisticated three-dimensional (3D) metric reconstruction, each possessing distinct data requirements and technical constraints.

3.1 Category 1: Traditional Semantic Segmentation from SVI

This methodological category traces the evolution of sidewalk detection from early feature-based machine learning to modern deep neural networks. Pioneering work by Smith et al. (2013) demonstrated one of the earliest automated approaches, utilising a Random Forest classifier to identify sidewalk segments based on hand-crafted local and global features, including geometric context, lane detection, and colour histograms. Similarly, Kang et al. (2021) applied a Random Forest classifier but enhanced the workflow by using graph-based segmentation to aggregate image-level results into street-level inventories.

These feature-engineering approaches have largely been superseded by end-to-end deep learning models, which achieve superior results by automatically learning complex feature representations, such as surface textures, material boundaries, and geometric edges—directly from raw pixel inputs of SVI rather than relying on manual feature definition. For instance, Weld et al. (2019) modified a ResNet-18 neural network and Hamim et al. (2023) implemented the HRNet+OCR semantic segmentation model, demonstrating that deep architectures can robustly identify sidewalk presence even in complex urban environments where traditional methods often fail.

Despite these improvements in detection accuracy, this category faces three critical limitations. First, it primarily generates descriptive tags rather than quantitative metric data; studies typically identify *whether* a sidewalk is present or absent, but do not reconstruct its measurable dimensions such as width or slope Hamim et al. (2023). Second, environmental factors pose a major challenge to segmentation consistency; it was Kang et al. (2021) found that large shadows often cause segmentations to merge multiple distinct objects into one. Finally, these models struggle with geographic generalisability. Hamim et al. (2023) emphasised that models trained on standard Western datasets perform poorly when applied to different geographic contexts (e.g., Bangladesh), requiring extensive retraining with local images to achieve acceptable performance.

3.2 Category 2: Aerial Imagery and Network Generation

These methodologies aim to map the broad connectivity and topological structure of city-wide pedestrian infrastructure using an overhead perspective (Hosseini et al., 2023; Ning et al., 2021). The process generally involves: (1) acquiring sub-metre resolution orthorectified aerial imagery, which is a data prerequisite for distinguishing narrow pedestrian pathways from surrounding surfaces; (2) applying semantic segmentation frameworks, such as TILE2NET, to extract georeferenced polygons for sidewalks and crosswalks; and (3) generating interconnected centrelines to produce topologically accurate pedestrian routing networks (Hosseini et al., 2023; Ning et al., 2021).



Figure 3.1: Highlighted sidewalk occlusion by tree canopy and building shadows in PDOK orthophoto imagery (Utrecht, Netherlands). The figure is included for visualization purposes only.

The main technical challenge is that sidewalks are often hidden from above by tree canopies, building shadows, or overhead structures (see Figure 3.1) (Hosseini et al., 2023). Earlier work by Senlet and Elgammal (2012) addressed this by filling in the missing parts of occluded sidewalks using surrounding road and building information. More recent hybrid models,

such as those by Ning et al. (2021), extend this approach by combining aerial and street-level imagery to recover sidewalk segments that remain invisible from above.

While effective for network connectivity, aerial mapping faces a fundamental resolution limit. Standard aerial imagery (typically 10–30 cm/pixel) lacks the fine-grained resolution required to detect qualitative accessibility barriers like small cracks (< 2 cm), surface irregularities, or thin utility poles, which are easily resolvable in street-level photography (Hosseini et al., 2023).

3.3 Category 3: Geometric Reconstruction and Metric Mapping from SVI

This category moves beyond simple detection to extracting measurable physical dimensions—such as width and cross-slope—required for regulatory compliance by relying on LiDAR-based depth maps (Ai and Tsai, 2016). While recent research utilises SVI as a scalable alternative to expensive mobile LiDAR, the methodology remains intrinsically tied to the proprietary Google ecosystem (Kang et al., 2021; Liu et al., 2025). Generating high-resolution Digital Terrain Models (DTMs) typically requires integrating RGB imagery with commercial Google Street View Depth (GSVD) maps, creating a significant dependency on paid data sources that are unavailable in open repositories like Mapillary (Ning et al., 2022). Consequently, researchers are restricted by the costs of data acquisition and the fixed temporal resolution of Google's collection schedules, which often lag behind actual infrastructural changes.

Finally, despite the precision of LiDAR-generated depth maps, the data acquisition angle presents a fundamental geometric limitation. Because data is captured from a vehicle moving along the roadway, the line of sight to the sidewalk is frequently interrupted by dynamic and static occlusions. Liu et al. (2025) and Smith et al. (2013) note that parked vehicles, moving traffic, and roadside barriers often block the camera's lateral view, resulting in significant data gaps where the sidewalk surface is completely invisible to the sensor, regardless of the depth map's resolution.

3.4 Category 4: Participatory Approach and Smartphone-Based Mapping

This category leverages citizen engagement to collect data using consumer-grade hardware. Approaches generally fall into two distinct workflows: "virtual auditing" which gamifies existing street view imagery, and "in-situ sensing" which captures fresh data from the field.

In the virtual approach, platforms like Project Sidewalk allow users to virtually navigate Google Street View (GSV) to label accessibility features; these labels are then aggregated using clustering algorithms to identify issues like missing curb ramps (Saha et al., 2019). Conversely, in-situ approaches utilise smartphone sensors attached to pedestrians or wheelchairs to capture real-time data. For instance, Morra et al. (2024) developed a Visual AI framework (Sidewalk AI Scanner) that employs a PSPUnet semantic segmentation model to detect sidewalk surfaces and embeds Augmented Reality (AR) markers into the video feed to estimate

width. Earlier work by Iwasawa et al. (2015) focused on non-visual sensing, using 3D accelerometer signals and Support Vector Machines (SVM) to classify surface roughness based on wheelchair vibration patterns.

While these participatory methods reduce the cost of professional surveys, they face unique reliability challenges. Virtual auditing suffers from data sparseness and volunteer retention; Saha et al. (2019) noted that volunteers drop off significantly faster than paid workers, and labelling precision for complex features (such as "missing curb ramps") can be as low as 20.5%. Furthermore, data quality is heavily constrained by technical limitations and user technique. The Visual AI approach in Morra et al. (2024) requires strict operational proximity, as measurement accuracy was found to drop significantly once the distance from the camera exceeds 5 metres. Similarly, sensor-based methods are sensitive to user movement patterns; Iwasawa et al. (2015) observed that skilled wheelchair users often perform maneuvers (e.g., "wheelies") to avoid impacts, which can inadvertently mask surface irregularities from smartphone accelerometers.

3.5 Synthesis: Positioning the Research

The literature review reveals a clear methodological gap. Category 1 (SVI Segmentation) excels at visual detection but is largely limited to binary classification, identifying the presence or absence of sidewalks without explicit spatial localisation. Critically, the studies in this category depend on proprietary imagery (predominantly Google Street View) and require researchers to train or fine-tune segmentation models locally, making them computationally expensive, difficult to reproduce, and legally incompatible with open mapping initiatives such as OpenStreetMap. Category 2 (Aerial Mapping) enables large-scale network connectivity but struggles with local resolution and persistent occlusions from tree canopies and building shadows. Category 3 (Geometric Reconstruction) provides high geometric precision yet remains constrained by proprietary data sources, vehicle-centric viewpoints, and reliance on costly LiDAR sensors or commercial depth maps (e.g., GSVD). Category 4 (Participatory Approaches) reduces survey costs but suffers from volunteer attrition, inconsistent data quality, and strict operational proximity requirements.

This thesis addresses these limitations through two complementary pipelines that operate exclusively on openly licensed crowdsourced data (Mapillary and OpenStreetMap). The first pipeline constructs a city-wide sidewalk presence inventory without any local model inference, leveraging pre-computed segmentation already available through the Mapillary Detections API and aggregating results onto the OSM road network. This replaces the local training requirements and proprietary data dependencies of Category 1 studies with a lightweight, reproducible, API-based workflow. The second pipeline extends beyond binary classification toward metric geometric reconstruction, a capability previously confined to Category 3's LiDAR-dependent methods, by combining a self-supervised vision transformer (DINOv3) for segmentation with monocular metric depth estimation (Depth Anything V3) as a software-based alternative to hardware-dependent LiDAR. The use of ground-level imagery overcomes the occlusion limitations of aerial mapping (Category 2), while pre-trained models available through academic access programs eliminate the cost barriers of traditional laser scanning (Category 3).

Together, the two pipelines span the spectrum from scalable inventory mapping to detailed geometric reconstruction, both maintaining full compatibility with open data standards.

4

Methodology

This chapter presents two methodologies for automated sidewalk mapping from crowdsourced street-level imagery. The first, **Sidewalk Inventory Mapping**, constructs a city-scale presence inventory by aggregating pre-computed semantic segmentation outputs and SfM-corrected camera metadata available through the Mapillary API, combined with the OpenStreetMap road network as a spatial reference. Because all segmentation is performed server-side by Mapillary and no local model inference is required, this approach is computationally lightweight and scalable to large urban areas. The second, **Sidewalk Geometry Reconstruction**, processes Mapillary panoramas locally using a vision foundation model for semantic segmentation (DINOv3) and a monocular metric depth estimation model (Depth Anything V3), projecting the resulting 2D sidewalk masks into georeferenced 3D geometries. This approach recovers the physical dimensions and spatial footprint of sidewalk surfaces, but is computationally intensive and requires dense panoramic image coverage. Each methodology is described in detail in the sections that follow.

4.1 Sidewalk Inventory Mapping

4.1.1 Overview

The methodology attempts to develop an automated pipeline for detecting and mapping sidewalk presence from crowdsourced street-level imagery by assigning binary (`yes/no`) labels for each road segment, indicating whether a sidewalk exists on the left side, right side, both sides, or neither side of the road. Figure 4.1 shows the overview of the four-phase methodology for automated sidewalk inventory mapping.

1. **Phase 1 - Setup Study Area:** The spatial extent of the study area is defined as a bounding box in geographic coordinates, and two datasets are retrieved: image metadata from the Mapillary Graph API and the road network from OpenStreetMap.
2. **Phase 2 - Downloading Data:** The Mapillary Detections API is queried for each image to retrieve the pre-computed semantic segmentation results. Detections belonging to sidewalk-related classes are decoded from their compressed vector format and rasterised into a binary mask in which white pixels indicate the presence of a sidewalk surface and black pixels indicate its absence.
3. **Phase 3 - Sidewalk Detection:** The binary masks are analysed to generate georeferenced point features indicating sidewalk presence. Depending on the camera type (perspective or panoramic), the region of interest within the mask is processed to determine sidewalk coverage. Using the camera's compass bearing, these pixel-level detections are translated into absolute geographic directions, recording whether a sidewalk exists on the left, right, or both sides of the camera's trajectory.
4. **Phase 4 - Aggregation:** The georeferenced point features from Phase 3 are spatially joined to the OSM road network. Individual image observations are geometrically filtered based on camera orientation and assigned to the corresponding left or right roadside. Finally, all valid detections along a road segment are aggregated into a single confidence score, producing the final continuous sidewalk inventory map.

4.1.2 Data Sources

The methodology relies on two complementary open data sources. First, the Mapillary platform provides SfM-corrected, geotagged street-level imagery contributed by volunteers and professional mapping organizations, alongside pre-computed semantic segmentations based on the Mapillary Vistas dataset (Neuhold et al., 2017). Second, the OpenStreetMap (OSM) road network serves as the geometric and topological reference onto which the visual detections are aggregated. By combining these sources within a four-phase computational pipeline, sidewalk presence is inferred for both sides of every road segment in the study area, yielding a confidence score derived from the volume and quality of the contributing images.

Mapillary

As discussed in Section 2.3.1, each image in the Mapillary archive is associated with metadata fields describing its spatial position, orientation, and capture context. Rather than relying on the raw EXIF data embedded in the image files, this research utilizes the computed position and computed compass angle. Furthermore, the dataset contains a mix of standard perspective photos and 360-degree panoramic images. Due to their fundamentally different optical geometries, these camera types are identified via their metadata and processed through distinct, parallel branches of the pipeline.

In addition to the visual imagery, Mapillary's pre-computed semantic segmentations (annotated with 124 classes, as detailed in Section 2.3.1) are accessed via the Mapillary Detections API. As shown in Figure 4.2, these detections are delivered as vector polygons encoded in the Mapbox Vector Tile (MVT) format. For the specific purpose of this research, the pipeline

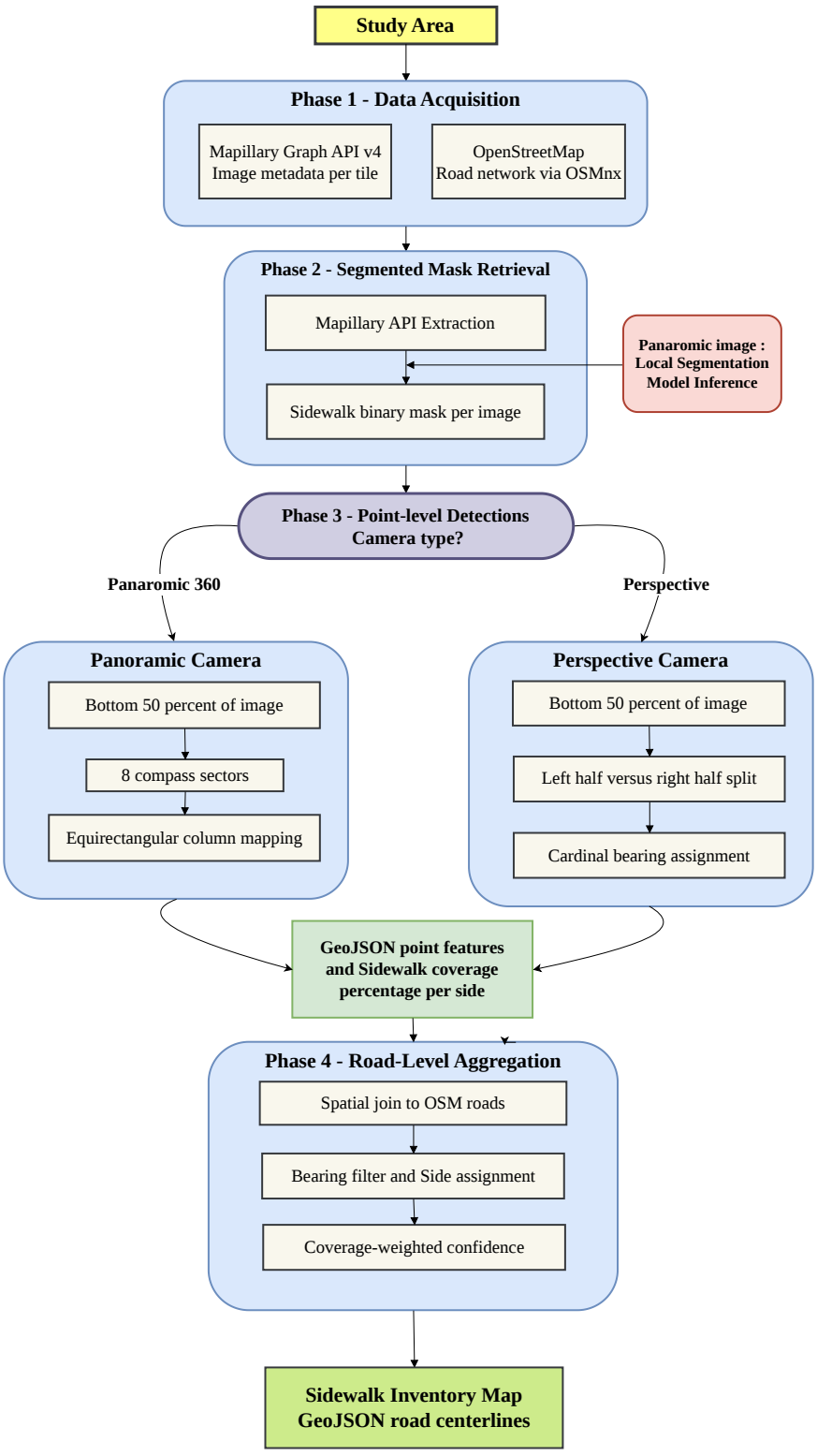


Figure 4.1: Overview of the proposed four-phase methodology for automated sidewalk inventory mapping using Mapillary imagery and OpenStreetMap road centerlines.



(a) (b)

(c) (d)

Figure 4.2: Examples of four sample perspective camera images and their binary sidewalk masks provided by the Mapillary API

filters and extracts only the `construction--flat--sidewalk` class, which isolates standard paved footway surfaces.

A key limitation encountered during data acquisition is that the Mapillary API only provides precomputed segmentation masks for perspective images in the Amsterdam region (Table 5.3). To resolve this bottleneck, this thesis implements a local segmentation model to generate the necessary masks for panoramic images. The methodology presented in this chapter outlines the complete, functional pipeline developed to process both image types.

Access to Mapillary data is provided through the Mapillary Graph API, which supports spatial queries via bounding boxes and per-image retrieval of metadata and detection results.

OpenStreetMap

For this research, OSM serves as the spatial reference framework for the sidewalk inventory: road centerline geometries from OSM define the linear features to which point-level detections are aggregated, and road attribute data, including street names and highway classification, are carried through to the final output.

The road network was downloaded using OSMnx, a Python library that provides a high-level interface to the OSM Overpass API and automates the construction of a topologically consistent directed graph from the retrieved data (Boeing, 2017). For the aggregation phase, non-vehicular path types, including dedicated footways, cycleways, pedestrian areas, and steps, were excluded from the road network before processing. Following the OSM spatial definition outlined in Section 2.1, non-vehicular paths were excluded to ensure the analysis focuses strictly on pedestrian facilities associated with the vehicular road network.

4.1.3 Point-Level Detection

The purpose of this step is to convert pixel-level sidewalk detections from each binary mask into georeferenced point features containing sidewalk observations. While the mask identifies the image regions containing sidewalks, determining whether these correspond to the left or right side of the road requires the camera's compass bearing at the moment of capture, i.e., the direction in which the camera was pointing and the associated image geometry. Using this information, Phase 2 generates a GeoJSON point feature for each image that records whether a sidewalk was detected on the left side, the right side, or both sides of the camera's direction of travel, together with a quantitative measure of the sidewalk coverage in each direction.

The inputs in this step are the binary masks produced by Mapillary and the per-image metadata, specifically the SfM-refined compass bearing (θ_{best}), the camera type flag ('is_pano'), and the SfM,-refined geographic position. The analysis procedure is based on the camera type; perspective images and panoramic images are processed by separate methods, as described in the following sections respectively.

Spatial Region of Interest: Based on various observations of the data used, the upper half of a street-level photograph typically contains sky, building facades, tree canopies, and other elevated features that are not relevant to ground-level sidewalk detection. Sidewalk surfaces, being ground-level features, are predominantly visible in the lower portion of the image. The region of interest is therefore defined as the set of pixels with row index $r \geq \lfloor H/2 \rfloor$, where H is the image height in pixels. This also reduces the influence of incorrectly segmented sidewalk masks in the upper part of image.

Perspective Camera Analysis

Left/Right Split: Within the lower half of the image, the left and right sides of the camera's field of view are separated by a vertical bisector at column $\lfloor W/2 \rfloor$, where W is the image width. Pixels to the left of this bisector correspond to the left side of the camera's field of view, and pixels to the right correspond to the right side. The sidewalk coverage for each side is computed as the proportion of white pixels in that half relative to the total number of pixels in the full image:

$$c_{\text{left}} = \frac{n_{\text{left}}}{W \times H} \times 100 \tag{4.1}$$

$$c_{\text{right}} = \frac{n_{\text{right}}}{W \times H} \times 100 \tag{4.2}$$

where n_{left} and n_{right} are the counts of white pixels in the left and right halves of the lower region respectively, and c_{left}, c_{right} are coverage percentages; see Figure 4.3. Within the processing pipeline, sidewalk detection is governed by a minimum coverage constraint, τ. A sidewalk is considered detected on a given side if its coverage percentage exceeds this parameter. Based on empirical observation, this threshold is currently set to $\tau = 0.05\%$, to effectively filter out noise and minor segmentation artifacts. At a spatial resolution of 1024×1024, this constraint ensures that any isolated sidewalk predictions containing fewer than approximately 524 pixels are strictly disregarded.

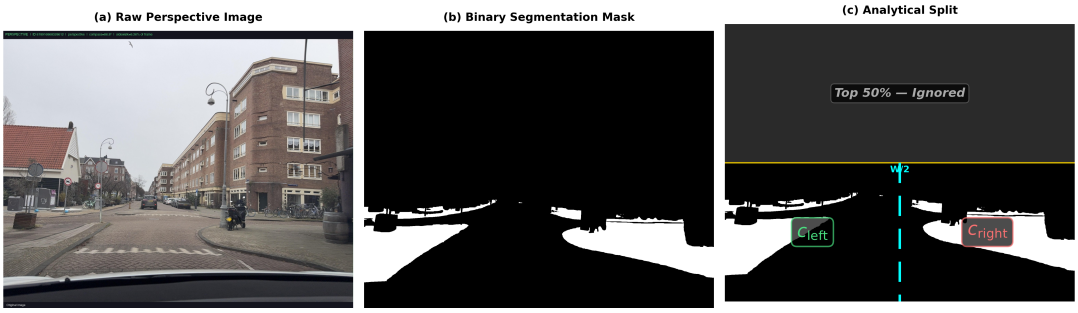


Figure 4.3: (a) Raw perspective image from Mapillary. (b) Binary sidewalk mask from segmentation (white = sidewalk). (c) The bottom 50% is bisected at W/2; white pixel counts in the left c_{left} and c_{right} halves determine sidewalk presence on each side of the road.

Cardinal Bearing Assignment : The geographic direction corresponding to the left and right sides of a perspective image is determined by the camera's compass bearing. If the camera is pointing in direction θ_{cam} (measured clockwise from north, in degrees), then the left side of the image corresponds to the direction perpendicular to the left of the camera's heading, and the right side corresponds to the direction perpendicular to the right:

$$\theta_{\text{left}} = (\theta_{\text{cam}} - 90^\circ) \bmod 360^\circ \tag{4.3}$$

$$\theta_{\text{right}} = (\theta_{\text{cam}} + 90^\circ) \bmod 360^\circ \tag{4.4}$$

These cardinal bearings are stored in the output feature and used in next Phase to assign each detection to the correct side of the road.

Panoramic Camera Analysis

Eight-Sector Decomposition : Because a panoramic camera captures all compass directions simultaneously, the left/right split used for perspective cameras is not applicable. Instead, the lower region of the panoramic mask is divided into eight compass sectors of 45° each, covering the full 360° field of view (see Figure 4.4):

Table 4.1: Bearing ranges for the eight panoramic compass sectors.

Sector	Bearing Range
N	337.5° – 22.5°
NE	22.5° – 67.5°
E	67.5° – 112.5°
SE	112.5° – 157.5°
S	157.5° – 202.5°
SW	202.5° – 247.5°
W	247.5° – 292.5°
NW	292.5° – 337.5°

Each sector corresponds to a range of pixel columns in the equirectangular image. Because compass bearing maps linearly to horizontal pixel position in this projection, the pixel column x for an absolute bearing β, given a camera heading θ_{cam}, is:

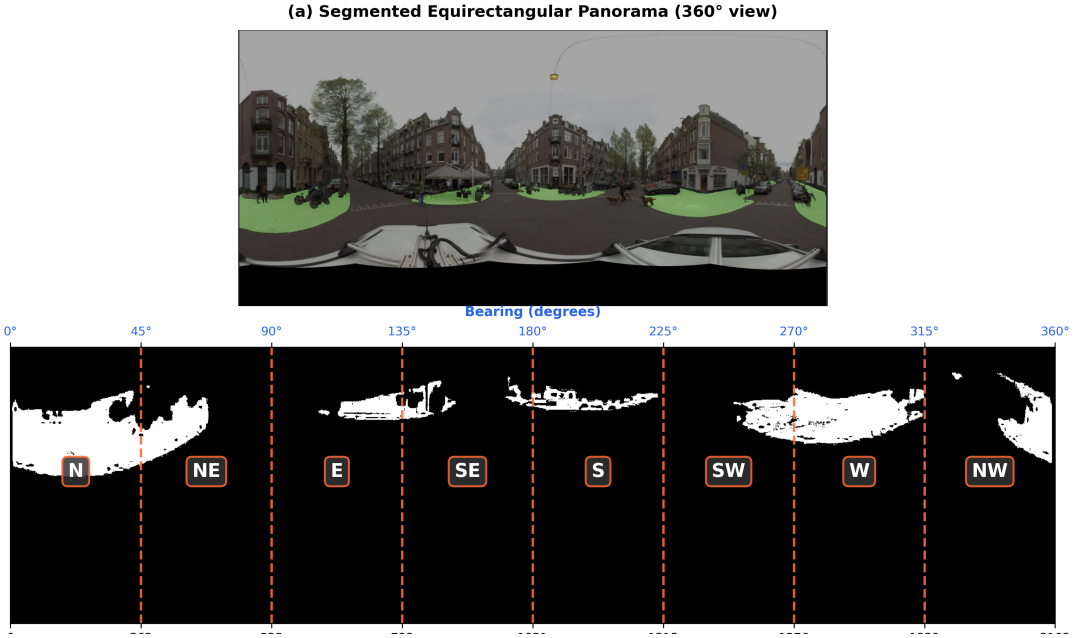


Figure 4.4: Panoramic 8-sector spatial analysis. (a) Segmented equirectangular image with sidewalk overlay (green). (b) Binary mask of the bottom 50% divided into 8 equal sectors corresponding to cardinal compass bearings. Pixel counts per sector determine sidewalk presence in each direction. Note: This segmented panoramic image is created using DinoV3, as Mapillary doesn't provide panoramic image segmentation.

$$x(\beta) = \left[\frac{(\beta - \theta_{\text{cam}} + 180^\circ) \bmod 360^\circ}{360^\circ} \right] \times W \tag{4.5}$$

This formula computes the relative bearing, centers the camera's forward direction in the image (column $W/2$), and scales the result to pixel coordinates. The column range $[x(\beta_{\text{start}}), x(\beta_{\text{end}})]$ for a given sector is found by applying Equation 4.5 to its start and end bearings. If a sector crosses the $0^\circ/360^\circ$ boundary (e.g., the North sector), its pixel columns are split into two non-contiguous ranges at the left and right image edges, which are concatenated before counting.

The sidewalk coverage for each sector s is computed as similar to formula 4.1 :

$$c_s = \frac{n_s}{W \times H} \times 100 \tag{4.6}$$

where n_s is the count of white pixels within the sector's column range in the lower region, and $W \times H$ is the total number of pixels in the full image.

Output Format

The output of Phase 2 is a GeoJSON `FeatureCollection` in which each feature is a `Point` geometry located at the SfM-refined position of the camera. Each feature carries the following properties:

- `image_id`: the Mapillary image identifier
- `best_compass_angle`: the SfM-refined compass bearing used for analysis
- `is_pano`: a boolean flag indicating camera type
- `cardinal_left_bearing` and `cardinal_right_bearing`: the geographic bearings of the left and right sides (perspective cameras only)
- `sidewalk_left` and `sidewalk_right`: "Yes" or "No" indicating whether the coverage threshold was exceeded on each side (perspective cameras), or "unknown" if no compass bearing was available
- `left_coverage_pct` and `right_coverage_pct`: sidewalk coverage percentages for each side
- `pano_sector_coverage`: a dictionary of coverage percentages keyed by sector name (panoramic cameras only; null for perspective cameras)

4.1.4 Road-Level Aggregation

The last step is to aggregate the point-level sidewalk detections onto the OpenStreetMap road network, creating a confidence-scored assessment of sidewalk presence for each side of each road segment. A single road segment in a dense urban area may be observed from dozens of images captured at various times, from multiple angles, and with diverse camera types. In this step, these image observations are turned into a coherent summary, including the orientation of the camera relative to the road, as well as the quantity and quality of the available detections. The inputs are the GeoJSON point features from Phase 3 and the OSM road network from Phase 1.

Road Network Preparation

As discussed in Section 4.1.2, the road network is filtered to retain only vehicular road types, as the objective is to characterize sidewalk presence relative to the vehicular road network. Road segments with a bearing between 180° and 360° are reversed so that all segments point in the northward or eastward half of the compass (bearing 0° to 179.9°). This normalisation step ensures that the concepts of "left" and "right" are defined consistently for all segments, regardless of the arbitrary direction in which the road was digitised in OpenStreetMap.

The pipeline supports splitting each road segment into smaller sections before aggregation `--chunk-size-m`, which allows the output to capture spatial variation in sidewalk presence within multiple smaller sections along a road, rather than producing a single binary result for the entire segment or street.

Spatial Indexing and Point-to-Road Association

For each road segment, the images that could plausibly have observed that segment are identified using a spatial buffer. Currently, a buffer of radius $d_{\text{buf}} = 15$ metres is constructed around each road segment's centreline geometry, and all image points falling within this buffer are considered candidate observations for that segment. The 15-metre buffer from the road centerline is adjustable and chosen to accommodate the positional uncertainty of Mapillary's SfM-refined coordinates as well as the images from the pedestrian cameras. A spatial index is used to perform these buffer queries efficiently across the full road network.

Local Road Bearing

For each candidate image within a road segment's buffer, the local bearing of the road at the point nearest to the image is computed. Using the overall bearing of the road segment would be incorrect for curved roads, where the tangent direction varies continuously along the segment. Instead, the road geometry is sampled over a 2-metre window centred on the projection of the image point onto the road centreline.

Let it d be the distance along the road centreline to the point nearest to the image. Two points are interpolated on the centreline at distances $d - 1$ m and $d + 1$ m, respectively, clamped to the endpoints of the segment. The local bearing α is then the following:

$$\alpha = \left(\text{atan2}(x_2 - x_1, y_2 - y_1) \times \frac{180^\circ}{\pi} + 360^\circ \right) \bmod 360^\circ \tag{4.7}$$

where (x_1, y_1) and (x_2, y_2) are the projected coordinates of the two interpolated points. By passing the coordinate differences as $(\Delta x, \Delta y)$ rather than the standard mathematical $(\Delta y, \Delta x)$, the atan2 function correctly computes the angle clockwise from the positive y-axis. This ensures the result is a true geographic compass bearing, assuming a projected coordinate system where North corresponds to positive y and East to positive x. This local bearing is used for all subsequent filtering and side-assignment operations for that image-road pair.

Parallel Heading Filter

Images whose camera bearing is approximately perpendicular to the road provide ambiguous information about which side of the road a detected sidewalk is on: a camera pointing across the road may see sidewalks on both sides simultaneously or may see neither. Only images whose camera bearing is approximately parallel to the road are used for aggregation since images pointing across the road provide ambiguous left/right information. An image is retained if its bearing is within 40° of the road direction or its reverse; images near road endpoints (within 5 metres of an intersection) are always included regardless of bearing to preserve coverage at junctions.

Physical Side Determination

For each image that passes the parallel filter, the physical side of the road on which the camera is located is determined using the signed cross product of the road direction vector and the vector from the road to the image point. Let $\mathbf{p}_1 = (x_1, y_1)$ and $\mathbf{p}_2 = (x_2, y_2)$ be the two interpolated points defining the local road direction (as in Section 4.1.4), and let $\mathbf{q} = (x_q, y_q)$, be the projected position of the image. The signed cross product is calculated as:

$$D = (x_q - x_1)(y_2 - y_1) - (y_q - y_1)(x_2 - x_1) \tag{4.8}$$

If $D > 0$, the image is located to the right of the road (relative to the direction of travel); if $D < 0$, it is located to the left.

For each image passing the parallel filter, the side of the road on which the detected sidewalk lies is determined by combining two signals. The physical side, whether the camera is geometrically to the left or right of the road centerline, is computed using the signed cross product of the road direction vector and the vector from the road to the image point. For images captured from a vehicle on the road (within 3 metres of the centreline), the physical position is ambiguous, and the camera bearing relative to the road direction is used instead. This hybrid rule accounts for the two dominant image sources in the dataset: dashcam sequences that drive along the road centreline, and pedestrian sequences captured from the footway.

Coverage-Weighted Confidence

After all images have been assigned to a roadside, the individual sidewalk detections are aggregated into a single confidence score. Rather than using a simple binary proportion, which treats all images equally regardless of detection size, a coverage-weighted confidence measure is employed. This is to ensure that a strong detection from a clear, unobstructed view carries more evidential weight than a marginal detection.

Let $\mathcal{A}$ be the set of all images assigned to a given roadside, and let $\mathcal{S} \subseteq \mathcal{A}$ be the subset in which a sidewalk was detected (i.e., coverage $c_i > \tau$). The confidence score is computed as:

$$\text{conf} = \frac{\sum_{i \in \mathcal{S}} c_i}{\sum_{i \in \mathcal{A}} (c_i + \varepsilon)} \tag{4.9}$$

Where c_i is the sidewalk coverage percentage for image i, minimum coverage constraint τ and $\varepsilon = 0.1$ is a small regularisation constant. This constant prevents division by zero when all images exhibit zero coverage and prevents images with negligible coverage from dominating the denominator.

Table 4.2: Data quality classification criteria based on image count.

Quality Class	Criterion	Interpretation
`no_data`	0 images	No Mapillary coverage for this road side
`low_sample`	1–2 images	Insufficient evidence for reliable assessment
`medium`	3–9 images	Moderate evidence; results should be interpreted with caution
`high`	$\geq$ 10 images	Sufficient evidence for confident assessment

Data Quality Classification

In addition to the confidence score, each roadside is assigned a data quality classification based on the number of images that contributed to the assessment:

A sidewalk is classified as present on a given roadside if the confidence score meets or exceeds a threshold of conf ≥ 0.5 and the data quality is at least `low_sample`. Roadsides with `no_data` quality are assigned a null presence value, indicating that no assessment could be made.

Output Format

The output file constitutes the final sidewalk inventory map. The output of Phase 3 is a GeoJSON `FeatureCollection` in which each feature is a `LineString` geometry (divided into 5m chunks) representing a road segment. Each feature carries the following properties:

- **Road attributes** (from OSM): street name, highway classification, normalised road bearing, and segment length in metres.
- **Image statistics**: the total number of images within the 15-metre buffer (*n_images_in_buffer*) and the number that passed the parallel filter (*n_images_parallel*).
- **Per-side assessment** (repeated for left and right): sidewalk presence (boolean or null), confidence score (0-1), number of contributing images, number of images in which a sidewalk was detected, mean coverage percentage, and data quality classification.

4.2 Sidewalk Geometry Reconstruction

The second methodology reconstructs the sidewalk's physical geometry, its spatial extent, width, and position relative to the road from crowdsourced street-level imagery. Unlike the inventory mapping pipeline in Section 4.1, which determines only the binary presence of sidewalks, this pipeline recovers vector geometries: variable-width polygons representing the sidewalk footprint, centrelines suitable for pedestrian routing, and per-location width estimates. The approach achieves this without Light Detection and Ranging (LiDAR) proprietary depth data by combining a vision foundation model for semantic segmentation (DINOv3 ViT-L/14; see Section 2.4.1) with a pose-conditioned monocular depth estimation

model (`DA3-Giant`; see Section 2.4.2) to project pixel-level sidewalk detections into a georeferenced three-dimensional coordinate frame. Figure 4.5 shows the overview of the five-phase methodology for the automated sidewalk geometry reconstruction pipeline.

- **Phase 1 - Data Acquisition:** This step retrieves the two spatial datasets on which the pipeline depends: geotagged street-level imagery with associated SfM metadata from Mapillary, and the vehicular road network from OpenStreetMap.
- **Phase 2 - Perspective Reprojection:** This process converts equirectangular panoramic images into perspective crops (image dimensions (1024×1024) and assigns required metadata values, suitable for the pinhole-camera-model assumptions of both the segmentation model (Phase 2) and the depth estimation model (Phase 3).
- **Phase 3 - Semantic Segmentation:** This produces a binary sidewalk mask for each image, both native perspective images and the reprojected panoramic crops generated in Phase 2, by running inference with the DINOv3 ViT-L/14 segmentation model and Mask2Former head with sidewalk class understanding.
- **Phase 4 - Metric Depth Estimation:** This produces a per-pixel metric depth map for every image that has a segmentation mask, using the `DA3-Giant` model in pose-conditioned mode, by establishing a sequence-window approach for consecutive images in a sequence.
- **Phase 5 - Geometric Projection and Polygon Assembly:** This is the last step and core of the reconstruction methodology. It transforms the 2D sidewalk masks and metric depth maps into georeferenced 3D point clouds. Sidewalk geometry and width are estimated using the lateral distribution of sidewalk points relative to each road segment, which are then assembled into smoothed, variable-width polygons and centerlines.

The following subsections describe each phase of the pipeline and its role in the overall workflow. Phase 1 shares substantial methodological overlap with Section 4.1; these shared elements are noted concisely and cross-referenced rather than repeated. Phases 3 and 4 invoke the deep learning models whose theoretical background is presented in Chapter 2 (Sections 2.4.1 and 2.4.2, respectively). Phase 5 contains the geometric reconstruction logic and is described in full detail.

4.2.1 Data Acquisition

Image retrieval from Mapillary follows a procedure similar to the inventory mapping pipeline detailed in Section 4.1. For each image within the study area, the following metadata fields are requested: geographic position (both raw and SfM-refined), compass bearing (both raw and computed), camera type, panoramic flag, image dimensions, camera intrinsic parameters, sequence identifier, and capture timestamp.

Metadata validation is done before downloading image data; each image record is validated against the minimum metadata requirements for successful 3D projection in Phase 4:

- A computed geographic position (GPS coordinates from SfM reconstruction) must be present.
- A computed rotation vector must be available.
- Camera intrinsic parameters and the focal length ratio must be present.

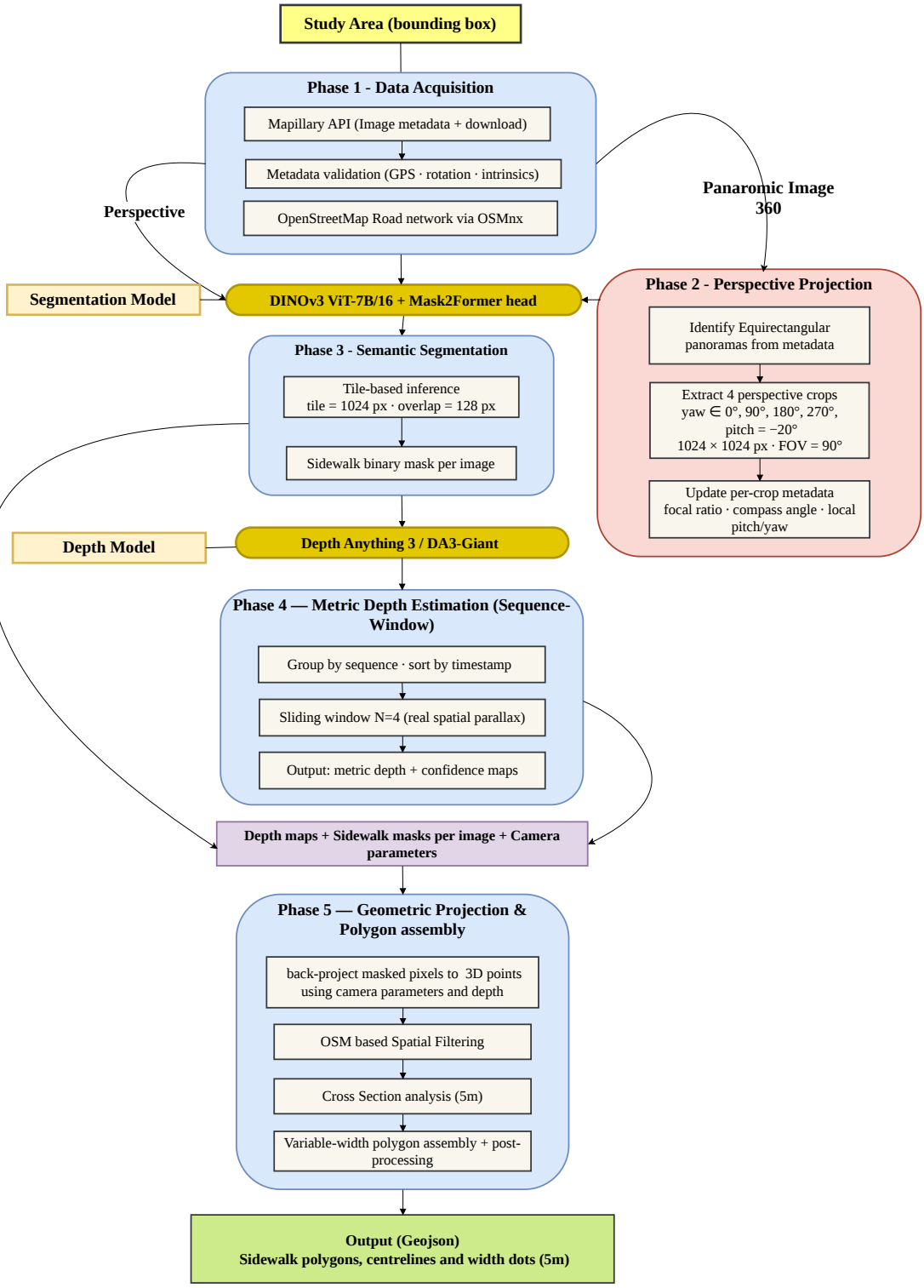


Figure 4.5: Overview of the proposed five-phase methodology for automated sidewalk geometry reconstruction using segmentation and depth estimation models.

Images failing any of these strict metadata checks are excluded prior to download, as they would inevitably cause errors during the subsequent depth estimation or 3D projection phases. Valid images are then downloaded at a configurable maximum width (e.g., 1024 pixels, 2048 pixels, or original resolution) via the Mapillary API. During the subsequent inference phase, these base images are divided into 1024x1024 pixel tiles to accommodate the computational and memory constraints of the segmentation and depth estimation models.

Simultaneously, the OpenStreetMap road network is downloaded using OSMnx (Boeing, 2017) with the `network_type=drive` parameter, similar to the inventory mapping pipeline; see Section 4.1.

4.2.2 Perspective Reprojection

In this phase, panoramic images are cropped into perspective views that meet the pinhole-camera-model assumptions of both the segmentation and depth estimation models. Each image's metadata is inspected for the `camera_type` field, and images classified as `is_pano` are identified as panoramas requiring reprojection.

For each panorama, four perspective crops are extracted at cardinal yaw angles ($0°, 90°, 180°, 270°$) with a constant downward pitch of $-20°$. Each crop is 1024×1024 pixels with a field of view (FOV) of $90°$. The equirectangular-to-perspective remapping is achieved by projecting 3D rays from the crop's local frame, rotating them by the target yaw and pitch, and mapping them to spherical coordinates on the source panorama. This is implemented via OpenCV's `remap` function, utilizing bilinear interpolation and border wrapping to seamlessly handle the continuous $360°$ boundary.

For each generated perspective crop, a new metadata record is created containing the following attributes:

- The focal length ratio, derived from the 1024×1024 pixel image dimensions and a $90°$ FOV. Given the image width $W = 1024$, the focal length f is calculated as:

$$f_{\text{ratio}} = \frac{f}{\max(W, H)} = \frac{\frac{W}{2\tan(45°)}}{\max(W, H)} = \frac{512}{1024} = 0.5 \tag{4.10}$$

- The compass angle, rotated by the crop's specific yaw offset:

$$\theta_{\text{crop}} = (\theta_{\text{base}} + \text{yaw}) \pmod{360°} \tag{4.11}$$

- The local yaw and pitch values, which are strictly required in subsequent phases, to accurately reconstruct the full camera-to-world rotation.

At typical Mapillary camera mounting heights of 1.5-2.5 m, a horizontal view (pitch $= 0$) causes the sidewalk surface to occupy only the lower quarter of the resulting perspective crop. $-20°$ pitch tilts the virtual camera downward, increasing the proportion of the ground plane visible in each crop; see Figure 4.6. This pitch value is stored in the per-crop metadata and is incorporated into the extrinsic rotation matrices during depth estimation and 3D projection. This angle was determined empirically: larger tilts begin to exclude the building façades and scene context needed by the depth model, while smaller tilts leave the sidewalk at the frame's extreme lower edge, where both segmentation boundary precision and depth accuracy are lowest. Native perspective images (i.e., non-panoramic images with



Figure 4.6: Perspective reprojection of an equirectangular panorama. (a) The source 360° panorama with colored bands indicating the four cardinal crop regions. (b.1-b.4) The four extracted perspective crops at yaw angles 0, 90°, 180°, and 270°, each with a downward pitch of $-20°$, 1024×1024 pixels, and a field of view of 90°. The downward tilt visibly increases the proportion of ground-level surfaces (sidewalks, roads) relative to a horizontal extraction.

valid camera parameters) bypass this phase entirely and proceed directly to Phase 3 without modification.

4.2.3 Semantic Segmentation

During this phase, semantic segmentation is performed to generate binary sidewalk masks for all perspective images and panoramic crops. This is achieved by running inference through the DINOv3 ViT-L/14 model, where the dedicated Mask2Former head interprets the visual features to identify and isolate the "sidewalk" semantic class. A comprehensive theoretical background of this model, including its benchmark performance, is presented in Section 2.4.1.

Inference Strategy: Images exceeding 2048 pixels in their largest dimension are downsampled to fit within a 2048-pixel bounding box to prevent GPU memory exhaustion. For images larger than the model's native input resolution of 1024×1024 pixels, a sliding-window tiling strategy is applied: the image is covered by overlapping 1024×1024 tiles with 128 pixels of overlap between adjacent tiles.

Mask Output: The model produces a per-pixel probability map for the sidewalk class. Pixels exceeding a probability threshold of 0.5 are assigned to the sidewalk class; all others are assigned to the background. The final binary mask is stored as an 8-bit grayscale PNG file

(where white represents the sidewalk and black represents the background). The resulting segmentation masks, overlaid on the original panoramic images with all individual crops seamlessly recombined, are illustrated in Figure 4.7.

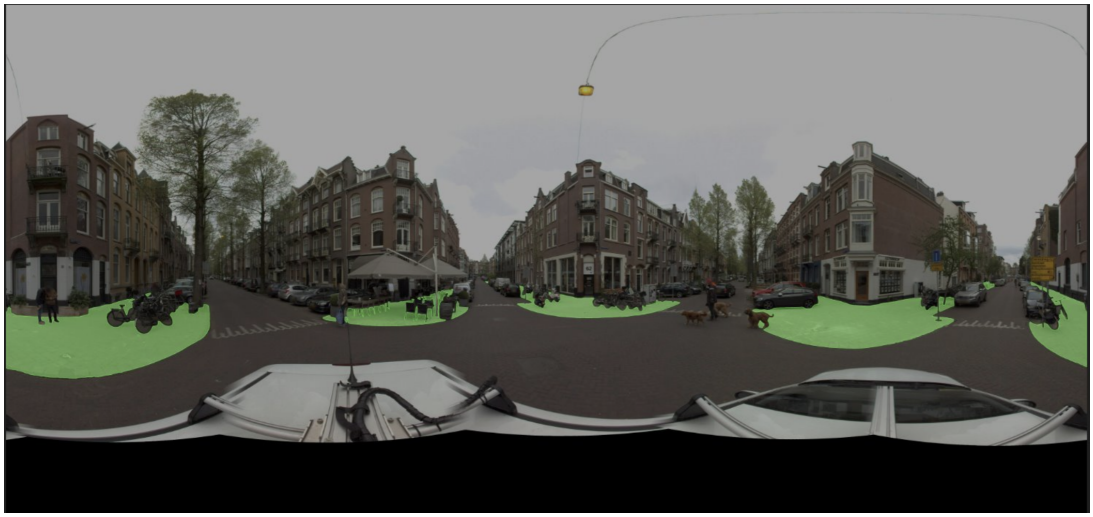


Figure 4.7: Sidewalk semantic segmentation masks overlaid on panoramic imagery, the final output after inference, with all individual processed panoramic crops recombined.

4.2.4 Depth Estimation

Phase 3 generates metric depth maps for images containing segmentation masks using `DA3-Giant` (theoretical background in Section 2.4.2). The DA3 pose-refined constraint requires parallax (objects shifting between views taken from different positions) to enforce cross-view depth consistency and recover metric scale.

The approach groups images by their Mapillary `sequence` field (a set of images captured consecutively by a moving camera) and sorts them chronologically by their `captured_at` timestamp. Consecutive images within a sequence are typically 2-5 m apart, providing essential spatial parallax. To leverage spatial parallax, inference relies on a sequence-window strategy where images are grouped chronologically by Mapillary sequence.

Window Construction: A sliding window of $N = 4$ images is utilized. For perspective images, the window consists of N consecutive frames centred on the target. However, because panoramic crops from the same origin lack parallax, their window is constructed using the target crop alongside front-facing crops from adjacent panoramas:

$$[\text{front_crop}_{k-1},\ \text{target_crop}_{k},\ \text{front_crop}_{k+1},\ \text{front_crop}_{k+2}]$$

Coordinate System and Extrinsics: Camera positions are converted from geographic coordinates (WGS 84) into a local East-North-Up (ENU) metric frame centred on the dataset's

first panorama. For each image k, a 4×4 world-to-camera matrix is constructed as:

$$\mathbf{E}_k = \begin{bmatrix} \mathbf{R} & -\mathbf{R}\,\mathbf{t}_k \\ \mathbf{0}^T & 1 \end{bmatrix}, \qquad \mathbf{R} = \mathbf{R}_{\text{crop}}\,\mathbf{R}_{\text{heading}}$$

where $\mathbf{R}_{\text{heading}}$ encodes the camera's compass heading and $\mathbf{R}_{\text{crop}}$ applies the yaw and pitch of the perspective crop (identity for non-panoramic images). $\mathbf{t}_k$ is the camera position in ENU metres. A pinhole intrinsic matrix $\mathbf{K}$ is constructed with focal length $f = f_{\text{ratio}} \cdot \max(W, H)$ and principal point at the image centre.

The ENU reference origin (latitude, longitude) is stored as `enu_reference.json`. Phase 5 must load this identical origin when back-projecting depth to 3D coordinates; using a different origin would systematically offset all projected points.

Inference and Output: The image window, alongside $\mathbf{E}$ and $\mathbf{K}$, is passed to the depth model. The parameter `align_to_input_ext_scale=True` instructs the model to rescale its internal predictions to the metric scale of the ENU extrinsics, converting relative monocular depth into absolute meter-scale depth without an external calibration target. The resulting metric depth map and per-pixel confidence map are saved as float32 NumPy arrays.

Figure 4.8 shows an example depth map and its corresponding confidence map for a perspective crop from the De Pijp study area. The depth map encodes absolute distance in metres from the camera, while the confidence map indicates the model's certainty at each pixel. The sidewalk surface in the foreground receives close-range depth values (2-8 m) and high confidence scores, making it suitable for accurate 3D reconstruction. Distant objects and sky regions receive low confidence and are discarded during projection.

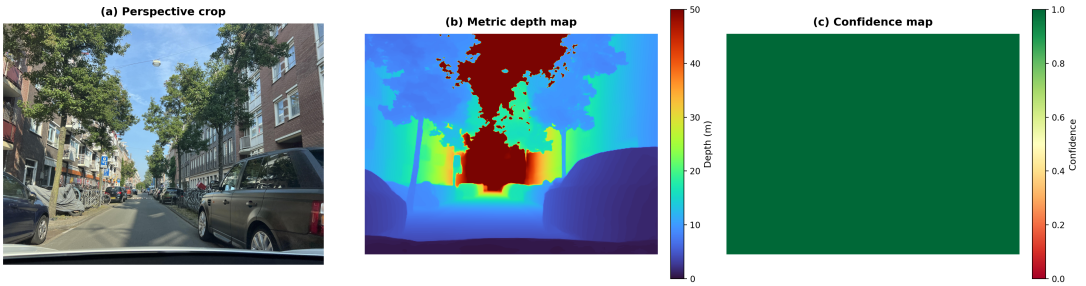


Figure 4.8: Example outputs from the metric depth estimation phase. (a) A perspective crop from a 360° panorama in the De Pijp study area (pitch $= -20$). (b) The corresponding metric depth map produced by DA3, where colour encodes absolute distance in metres from the camera (clipped at 30 m). (c) The per-pixel confidence map; high-confidence regions (yellow/orange) correspond to nearby surfaces suitable for 3D reconstruction, while low-confidence regions (dark purple) are discarded during projection.

4.2.5 Geometric Projection and Polygon Assembly

The final phase transforms the per-image outputs of Phases 3 and 4—binary sidewalk masks and metric depth maps—into georeferenced sidewalk geometry. The process consists of four stages: (1) back-projection of masked sidewalk pixels into a 3D point cloud expressed

in East-North-Up (ENU) world coordinates; (2) spatial filtering to remove unreliable or geometrically impossible points; (3) road-aligned cross-section analysis to measure sidewalk width and lateral extent; and (4) polygon construction and post-processing to produce the final GeoJSON output.

Back-Projection to 3D World Coordinates

For each image possessing both a segmentation mask and a depth map, the sidewalk pixels are projected into world coordinates using the standard pinhole camera back-projection. The intrinsic matrix $\mathbf{K}$ and the world-to-camera extrinsic matrix $\mathbf{E}$ are reconstructed from the same metadata used in Phase 4, loading the identical ENU reference origin from `enu_.reference.json` to ensure coordinate consistency.

The projection proceeds in four steps. First, the image coordinates (u, v) of the masked sidewalk pixels are collected as homogeneous vectors $\mathbf{p} = [u,\ v,\ 1]^\top$. Second, camera-space ray directions are computed:

$$\mathbf{r} = \mathbf{K}^{-1}\,\mathbf{p} \tag{4.12}$$

Third, each ray is scaled by its corresponding metric depth value d to produce 3D coordinates in the camera frame:

$$\mathbf{X}_c = d \cdot \mathbf{r} \tag{4.13}$$

Fourth, homogeneous camera-space coordinates $[\mathbf{X}_c;\ 1]$ are transformed to world coordinates via the camera-to-world matrix $\mathbf{T}_{c2w} = \mathbf{E}^{-1}$:

$$\mathbf{X}_w = \mathbf{T}_{c2w} \begin{bmatrix} \mathbf{X}_c \\ 1 \end{bmatrix} \tag{4.14}$$

yielding ENU coordinates (east, north, up) in metres relative to the shared reference origin. Figure 4.9 illustrates this transformation process, taking the raw imagery, mask, and depth map to produce the camera-relative spatial layout.

Spatial Filtering

Not all projected points represent valid sidewalk geometry; see Figure 4.10. Multiple filtering criteria are applied sequentially, and one can iterate over different configurations of spatial filtering and postprocessing according to the study area:

- **Confidence threshold:** If a confidence map was produced by DA3, points with confidence below 0.4 are discarded. This removes depth estimates on sky, reflective surfaces, and far-field objects where the model is uncertain.
- **Depth range:** Points with metric depth $d > 20$ m are removed as the distant points suffer from increasing depth estimation error.
- **Height filter:** Points whose ENU up-coordinate exceeds 1.5 m are removed. Sidewalks are ground-level features; elevated points indicate projections onto building façades or vegetation.

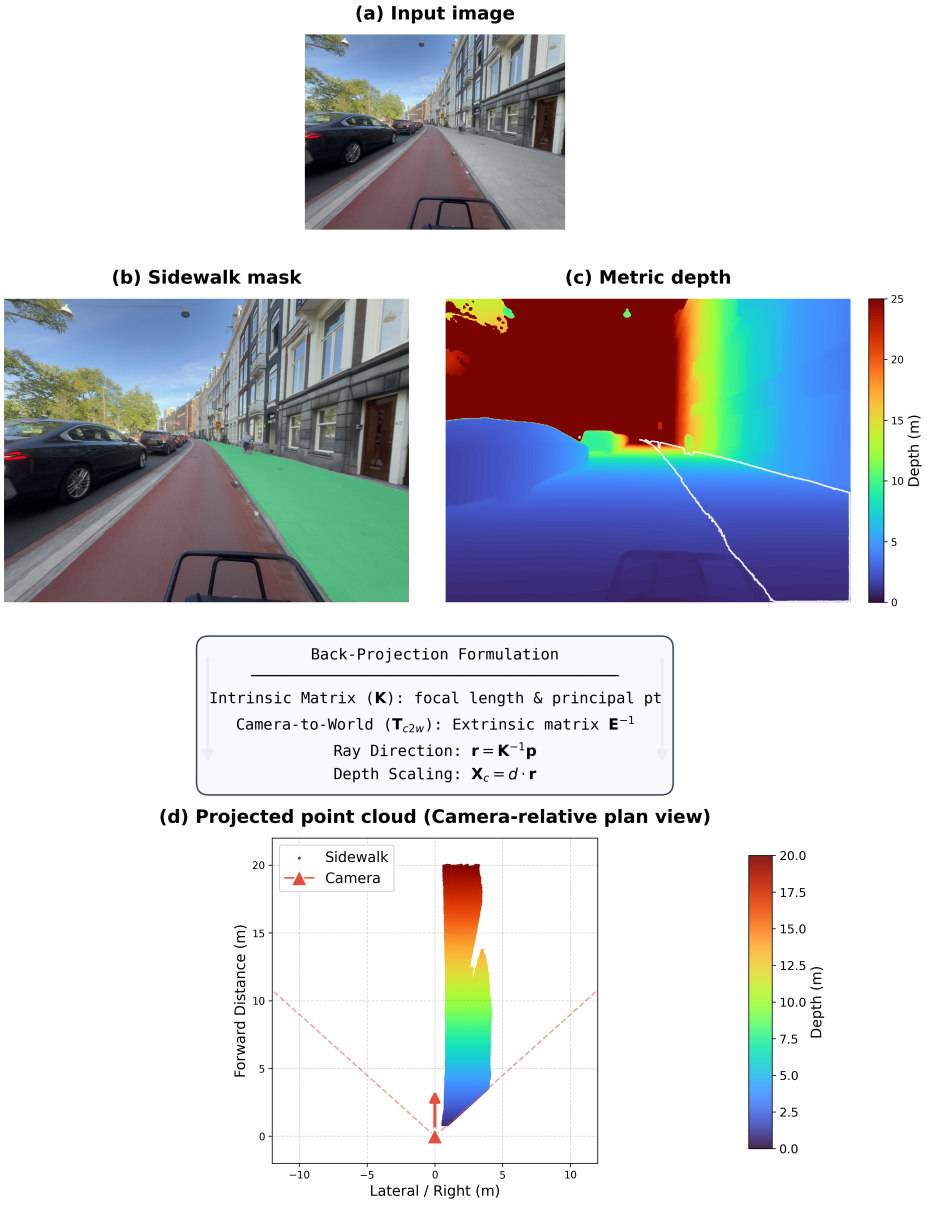


Figure 4.9: The back-projection pipeline from image space to metric 3D space. (a) Raw input image. (b) Binary sidewalk mask identifying pixels of interest. (c) Metric depth map providing the depth scaling factor d for each pixel. (d) The resulting depth-scaled 3D points plotted in a camera-relative plan view ($\mathbf{X}_c$), demonstrating how the 2D mask and depth perfectly reconstruct the lateral spatial layout of the sidewalk before applying the global ENU transformation.

- **Road exclusion zone:** Using the OSM road network, a buffer is constructed around each road centreline with a half-width proportional to the highway classification (e.g., 1.0 m for primary roads, 0.8 m for residential streets). Points falling within this buffer

are discarded to prevent sidewalk points from being erroneously placed on the road surface.

Figure 4.10 shows the raw projected point cloud, demonstrating how multiple viewpoints contribute to overlapping coverage of the same sidewalk area.



Figure 4.10: Examples of the raw projected sidewalk point cloud. The overlapping coverage from multiple viewpoints provides redundant observations that are aggregated during cross-sectional analysis.

Cross-Sectional Geometry and Width Estimation

To delineate the physical geometry and perform sidewalk width estimation, the filtered point cloud is analysed relative to the OpenStreetMap (OSM) road network. As shown in Figure 4.11, the road centerlines are uniformly sampled at regular 5-meter intervals to establish perpendicular cross-sections. At each sample point, localized 3D sidewalk points within a defined spatial boundary are orthogonally projected onto the cross-section to calculate their lateral offset from the road centerline.

The spatial distribution of these offsets is evaluated independently for both the left and right sides of the road to derive four structural parameters:

- The **inner edge** of the sidewalk: the 15th percentile of the lateral distance distribution.
- The **outer edge**: the 90th percentile of the lateral distance distribution, constrained to a maximum distance of 25 meters.
- The **centre**: the geometric midpoint between the established inner and outer edges.
- The **width**: the absolute difference between the outer and inner edges.

Post-Processing

Three post-processing steps remove geometric artifacts from the assembled polygons:

1. **Minimum area filter:** Polygons smaller than 0.5 m^2 are removed. These typically represent isolated noise clusters rather than continuous sidewalk segments.

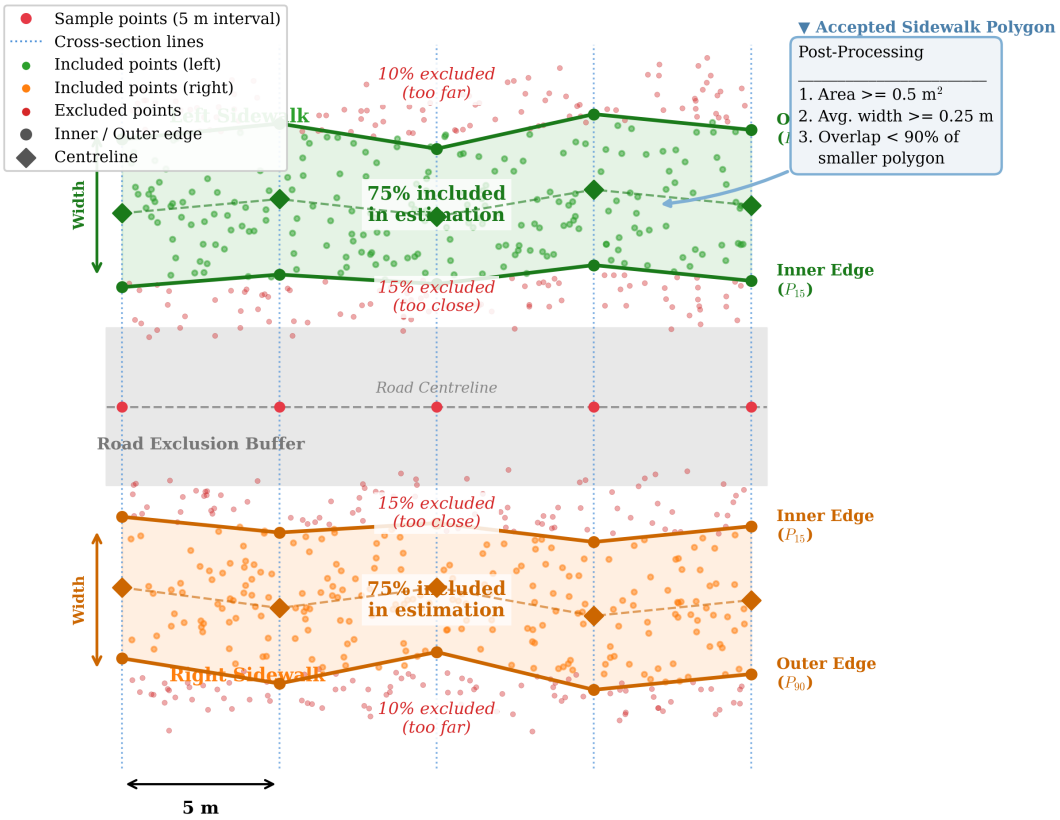


Figure 4.11: Cross-section geometry and width estimation for sidewalk polygon extraction. Road centerlines are sampled at 5 m intervals, and projected 3D points are classified by their lateral offset: the inner edge (P_{15}) and outer edge (P_{90}) define the sidewalk boundaries. Post-processing removes small, thin, or overlapping polygons.

2. **Thin polygon removal:** Polygons whose average width falls below 0.25 m are removed, preventing the generation of impractically narrow slivers that do not correspond to real sidewalk surfaces.
3. **Overlap deduplication:** Pairwise intersections between all polygons are computed. Where the intersection area exceeds 90% of the smaller polygon's area, the smaller polygon is discarded as a duplicate arising from overlapping image coverage of the same physical sidewalk.

Output

The sequence of valid cross-section measurements along each road edge is synthesized into a continuous polygonal footprint. The resulting inner and outer boundary point sequences are smoothed using a Savitzky-Golay filter to suppress high-frequency spatial noise while preserving the overall structural shape of the sidewalk. These smoothed boundaries are then concatenated to form a closed vector polygon. Additionally, for each valid polygon,

a smoothed centreline is generated by connecting the midpoints of all contributing cross-sections. The final output of the pipeline consists of three GeoJSON files:

- **Polygons** ({section}_polygons.geojson): Variable-width closed polygons representing the spatial footprint of each detected sidewalk segment
- **Centrelines** ({section}_centrelines.geojson): Smoothed LineString geometries running along the centre of each polygon, intended for future use in pedestrian routing applications.
- **Width dots** ({section}_dots.geojson): Point features placed at each 5-metre cross-section location, carrying a width_m attribute recording the estimated local sidewalk width. These provide a human-readable summary of width variation along each road segment.

All coordinates are stored in WGS84 (EPSG:4326) by converting the ENU metre offsets back to geographic coordinates using the inverse local tangent-plane approximation referenced to the shared ENU origin.

5

Results and Accuracy

This chapter presents and evaluates the outputs of the two automated sidewalk mapping pipelines developed in this thesis. Section 5.1 introduces the three study areas and their associated ground truth datasets. Section 5.2 reports the results of the Sidewalk Inventory Mapping pipeline (Pipeline 1), including segmentation outputs, the generated inventory maps, and their accuracy assessment against authoritative ground truth. Section 5.4 presents the results of the Sidewalk Geometry Reconstruction pipeline (Pipeline 2).

5.1 Study Areas

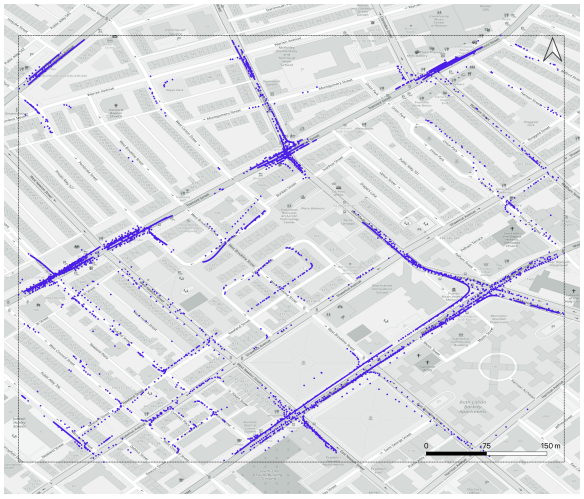
To evaluate the pipelines, three geographically bounded study areas were defined across two cities. These areas do not cover entire neighborhoods, but rather represent specific, localized subsections within two neighborhoods in Amsterdam (the Netherlands) and one in Boston (United States). The primary selection criteria for these targeted extents included the public availability of authoritative ground-truth sidewalk datasets and the opportunity to test the models across diverse international urban contexts. The spatial extents of each study area are summarized in Table 5.1.

5.1.1 Image Coverage

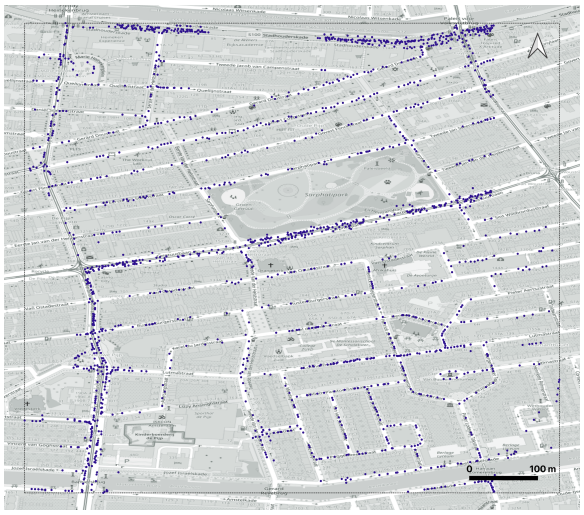
The spatial distribution of Mapillary image locations across all three study areas is visualized in Figure 5.1. Notably, the Hobbemakade region exhibits near-complete street coverage, whereas the De Pijp neighborhood lacks imagery for a few minor street segments. More

Table 5.1: Spatial dimensions of the three study areas.

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



(a) Boston South End



(b) De Pijp



(c) Hobbemakade

Figure 5.1: Overview of the three study areas showing image coverage point distributions: (a) Boston South End, USA; (b) De Pijp, Amsterdam; and (c) Hobbemakade, Amsterdam.

significantly, the Boston South End study area contains several streets entirely devoid of current image coverage. These unmapped segments will inevitably introduce localized data gaps in the subsequent spatial analysis. Table 5.2 summarises the image acquisition statistics for each study area.

Table 5.2: Image acquisition statistics per study area. To process panoramic imagery using the local perspective model, each panorama is reprojected into four individual perspective crops.

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**

As Table 5.2 shows, the Amsterdam study areas are heavily reliant on vehicle-captured panoramic imagery. Because the Mapillary API only provides pre-computed semantic segmentation masks for standard perspective imagery in the Amsterdam region, these panoramic images must be reprojected into four separate perspective crops to be processed by a local segmentation model. Mapillary's data pipeline generates and updates pre-computed features incrementally; while the Boston South End area has already received official semantic segmentation coverage for its panoramic images, the Amsterdam region has not yet been processed by Mapillary for this imagery type. Therefore, Boston's 540 panoramas were processed directly without reprojection using native Mapillary masks, whereas the Amsterdam regions required local crop extraction.

For the Amsterdam study areas (De Pijp and Hobbemakade), the perspective imagery consisting of 569 and 471 images, respectively, was confirmed to be captured via mobile phones or handheld devices, providing clear, unobstructed pedestrian-level views of the sidewalks. In contrast, while the Boston study area contains a large volume of perspective imagery, close manual inspection revealed that these sequences were almost entirely captured from vehicle dashboard cameras rather than by pedestrians.

5.1.2 Ground Truth Datasets

Accuracy assessment requires authoritative sidewalk datasets against which the pipeline outputs can be compared. Two independent ground truth sources were used, each corresponding to a different geographic context.

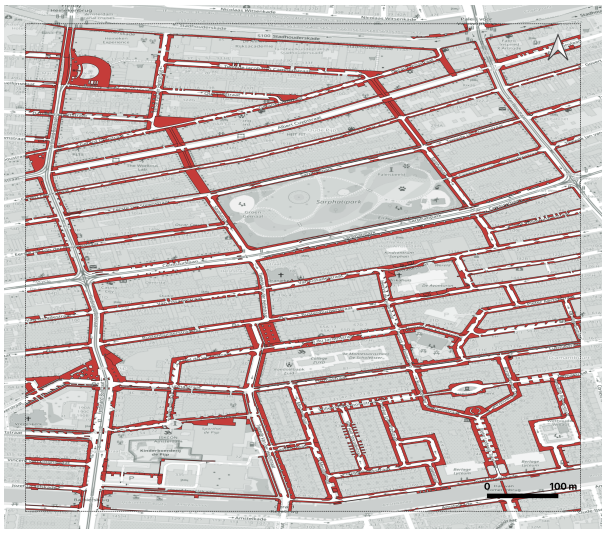
Amsterdam Basisregistratie Grootschalige Topografie (BGT) Dataset: The Dutch national large-scale topographic registration maintained by municipalities, providing polygon geometries for all paved surfaces. Sidewalk polygons were extracted by filtering on `Gebruiks functie = Voetpad` (footpath), yielding 651 polygons for De Pijp (total area: 133,009 m^2) and 228 polygons for Hobbemakade (52,623 m^2). These polygons represent the physical footprint of all registered sidewalk surfaces within each study area extent.

Boston Sidewalk Dataset: The city of Boston maintains publicly available sidewalk polygon geometries. Where 310 surveyed sidewalk polygons within the South End study area (total area: 53,333 m^2) were extracted.

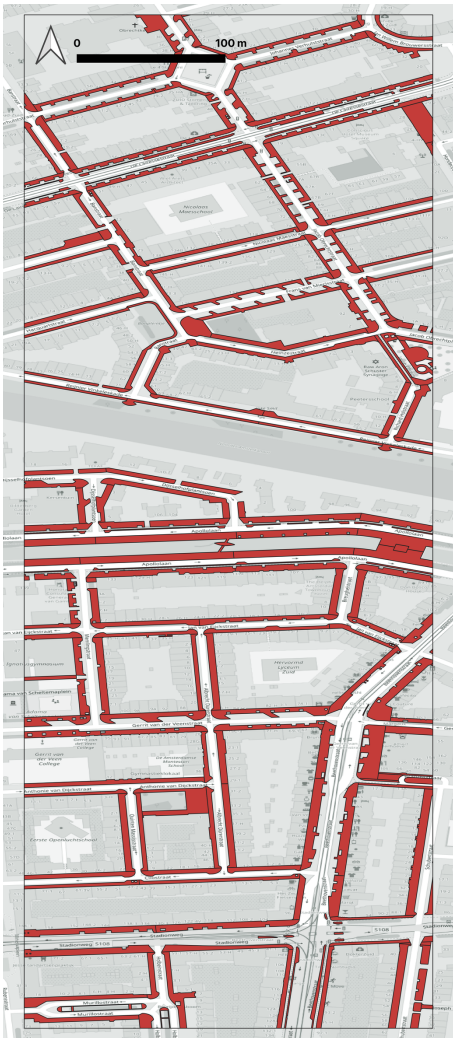
Both ground truth datasets confirm that the study areas have near-complete sidewalk coverage on both sides of virtually all roads, consistent with the dense pedestrian infrastructure characteristic of urban neighborhoods. Figure 5.2 visualizes the spatial extent of the ground truth polygons for all three areas.



(a) Boston South End



(b) De Pijp



(c) Hobbemakade

Figure 5.2: Ground truth sidewalk polygons for the three study areas. (a) Boston South End, (b) De Pijp, (c) Hobbemakade.

5.2 Segmentation Results

5.2.1 Mapillary API and Local DINOv3 masks

The sidewalk inventory pipeline relies on binary sidewalk masks obtained through semantic segmentation. Two segmentation sources were evaluated: (1) pre-computed masks from the Mapillary Detections API and (2) locally generated masks using the DINOv3 segmentation model. Table 5.3 presents the detection statistics for both sources across all three study areas, and Figure 5.3 shows the spatial distribution of local segmentation results and their comparison for both segmentation approaches.

Table 5.3: Comparison of Mapillary API and local DINOv3 segmentation results. "Images Processed" reflects the number of original images submitted to each segmentation source, with panoramic crops aggregated back to their parent image. An image is considered to detect a sidewalk if at least one side exceeds the minimum coverage threshold ($\tau = 0.05\%$).

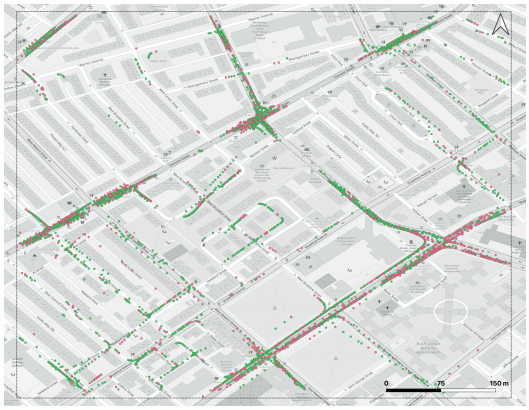
Study Area	Source	Images Processed	Sidewalk Detected	% Images with Sidewalk
De Pijp	Mapillary API	1,557	374	24.0%
	Local DINOv3	1,557	1,376	88.3%
Hobbemakade	Mapillary API	1,491	296	19.9%
	Local DINOv3	1,491	1,372	92.0%
Boston South End	Mapillary API	5,396	3,723	69.9%
	Local DINOv3	5,396	3,861	71.6%

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

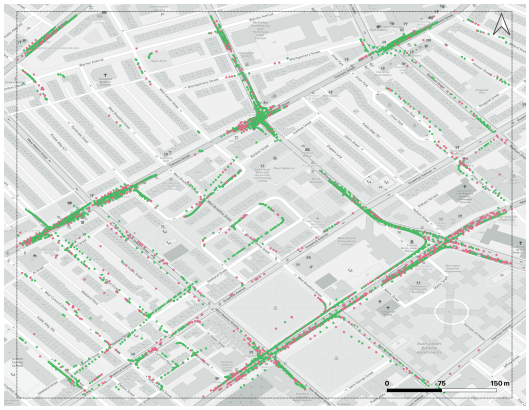
As detailed in Table 5.2, the Amsterdam study areas contain a significant volume of panoramic imagery, with 988 panoramas in De Pijp and 1,020 in Hobbemakade. In fact, panoramic formats constitute approximately 60% of the images across both Amsterdam datasets. Table 5.3 reveals that the Mapillary API returned pre-computed sidewalk segmentation masks for only 24% of the De Pijp images and 20% of the Hobbemakade images. This low retrieval rate stems from a current API limitation for the Amsterdam region, where Mapillary only returns segmentation outputs for perspective images and entirely excludes panoramas. Conversely, in the Boston South End study area, the API successfully provides complete segmentation for both panoramic and perspective imagery. When comparing the local DINOv3 segmentation against these complete Mapillary API results for Boston, there is only a 2% difference in overall detection rates, though this quantitative similarity does not account for the qualitative precision of the boundaries.

The local segmentation pipeline processes all available imagery, including reprojected panoramic crops, substantially expanding the number of usable observations. As Table 5.3 shows, while both sources evaluated the same 1,557 original images in De Pijp, DINOv3 successfully identified sidewalks in 1,376 images (3.7× more than the API), achieving an 88.3% detection rate. Similarly, in Hobbemakade, DINOv3 detected sidewalks in 1,372 images (4.6× more than the API), yielding 92% of images with sidewalks detected.

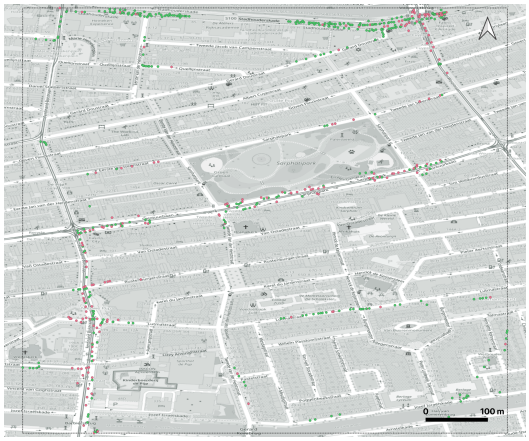
For Boston, where both sources process the exact same 5,396 images, the local DINOv3



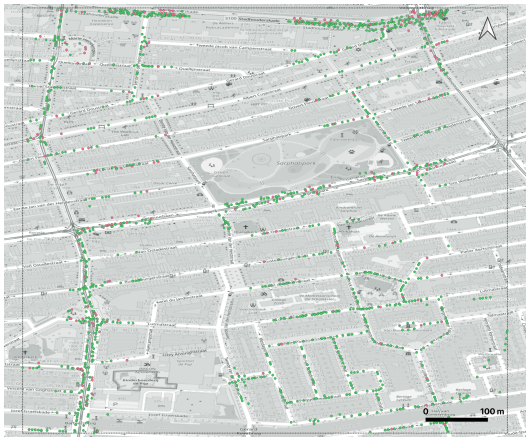
(a) Boston South End – Mapillary API



(b) Boston South End – DINOv3



(c) De Pijp – Mapillary API



(d) De Pijp – DINOv3

Figure 5.3: Comparison of sidewalk detection results obtained from the Mapillary API and the locally deployed DINOv3 model across the study areas (continued on the next page).



(e) Hobbemakade – Mapillary API

(f) Hobbemakade—DINOv3

Figure 5.3: Comparison of sidewalk detection results obtained from the Mapillary API and the locally deployed DINOv3 model across the three study areas. Green points indicate images in which a sidewalk was detected, while red points indicate images without sidewalk detection.

model achieves a detection rate 1.7 percentage points higher than the Mapillary API. This indicates that DINOv3's segmentation model produces more sensitive sidewalk detections.

The comparison between the two segmentation sources (Table 5.3) highlights two key findings. First, a fundamental limitation of the API-only approach is exposed: in certain regions (such as Amsterdam), panoramic images, which often provide the densest and most systematic street coverage in European cities, are simply not segmented by the Mapillary API.

Second, even when both sources process the same imagery (as in Boston), local DINOv3 segmentation achieves a higher detection rate (71.6% vs 69.9%), demonstrating superior sensitivity in identifying sidewalk surfaces. However, because Boston's perspective imagery already possesses dense pre-computed Mapillary API segmentation data, the downstream inventory mapping for Boston leveraged the API segmentation to eliminate the massive computational cost of local processing. For the Amsterdam study areas, which rely heavily on panoramic imagery unsupported by the API, local DINOv3 segmentation was exclusively used as the preferred approach.

5.2.2 Segmentation Accuracy

Mapillary does not officially report the quantitative accuracy of its API segmentation masks. Visual inspection of the masks overlaid on the original imagery (Figure 2.3) reveals imperfect boundary delineation, characterized by overshooting and missing patches. These masks remain suitable for the inventory mapping pipeline, which solely requires a binary 'yes/no' classification of sidewalk presence rather than exact geometric boundaries. This API-driven approach will also become viable for the Amsterdam study areas in the future once Mapillary introduces panoramic segmentation support. The geometry reconstruction pipeline, however, requires highly precise boundary localization, making these pre-computed masks inadequate for that specific task.

The quantitative accuracy of the local DINOv3 segmentation model is detailed in Section 2.4.1. From Table 5.3, a notable discrepancy exists in the overall detection rates across the study areas. The Amsterdam datasets exhibit approximately 90% sidewalk detection, while the Boston dataset achieves only around 70% across both segmentation sources.

A qualitative investigation reveals that this gap is primarily driven by severe visual occlusion. Several streets in the Boston South End feature dense on-street parking along both curbs. Imagery captured from vehicle-mounted cameras travelling down the centre of the road cannot physically observe the sidewalks hidden behind these parked cars. The Boston dataset also lacks the supplementary pedestrian- and cyclist-captured imagery that helps mitigate this issue in Amsterdam. Sidewalks physically present behind these continuous rows of vehicles will remain entirely undetected by the pipeline, as illustrated in Figure 5.4.

5.3 Pipeline 1

5.3.1 Sidewalk Inventory Mapping Results

The final sidewalk inventory was generated using local DINOv3 segmentation for the two Amsterdam areas (De Pijp and Hobbemakade), and Mapillary API segmentation for Boston



Figure 5.4: Examples of Mapillary imagery from the Boston South End study area. Dense on-street parking completely obscures the adjacent sidewalks from road-centre cameras, leading to missed detections.

South End. Road centrelines from OpenStreetMap were divided into 5-metre segments, and per-side and binary 'yes/no' classification of sidewalk presence was determined through the methodology described in Section 4.1. Table 5.4 summarizes the output.

Table 5.4: Sidewalk inventory mapping results per study area. Each 5-meter road segment is evaluated independently on its left and right sides, yielding two roadside assessments per segment. Each roadside receives one of three classifications: `true` (sidewalk detected), `false` (no sidewalk detected despite available imagery), or `no_data` (insufficient imagery coverage to make any determination).

Metric	**De Pijp**	**Hobbemakade**	**Boston (Mapillary 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%)

As Table 5.4 shows, the dominant output category in both De Pijp and Boston is `no_data`, meaning no Mapillary imagery was available to assess those roadsides; this is a coverage limitation rather than a limitation of the applied method. Hobbemakade performs best in this regard, with only 32.3% of road-sides lacking coverage, whereas Boston suffers from severe coverage limitations: only 18.2% of road-sides receive any assessment, with the vast majority (81.8%) lacking imagery entirely.

Where imagery is available, the data quality also differs substantially. Hobbemakade's systematic captures result in 97% of its assessed sides having 10 or more contributing image crops, providing robust multi-observation evidence. Boston presents the opposite pattern: even where imagery exists, over a fifth of assessed sides rely on just 1-2 image crops, making confident classification more difficult.

The resulting inventory maps are visualised in Figure 5.5, where each roadside is colour-coded by its detection outcome. A comparison between Figure 5.5 and Figure 5.3 reveals a clear spatial correlation: regions with a higher volume of segmented images exhibit correspondingly higher coverage in the final sidewalk inventory. Specifically, localized areas with a dense cluster of positive segmentations have a significantly higher probability of yielding a confirmed sidewalk detection. This relationship is quantitatively supported by the data. As shown in Table 5.3, Hobbemakade demonstrates the highest image-level segmentation rate (92%), which directly translates to the highest overall sidewalk detection rate of 48.9% (represented by the `Sidewalk detected (true)` attribute) across its road network (see Table 5.4).

Figure 5.5 highlights a distinct advantage of utilizing Street-Level Imagery (SVI) in vegetated urban environments. In the northern section of Hobbemakade, specifically along Heinzstraat, a dense tree canopy heavily obscures the sidewalks from standard aerial and satellite perspectives. However, by using the street-level perspective from SVI, the pipeline successfully detects these sidewalk surfaces beneath the trees.

5.3.2 Accuracy Assessment

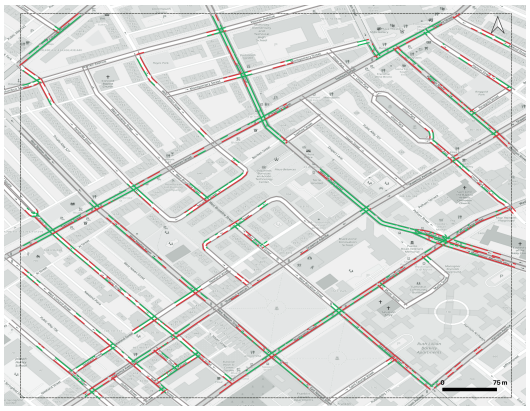
Methodology

The inventory results in Figure 5.5 are evaluated against the authoritative ground truth presented in Figure 5.2 using a spatial matching approach. This ground truth dataset covers sidewalks alongside nearly all the roads within the current study areas. For each roadside (left or right of a 5-meter segment), a one-sided buffer of 15 m is constructed perpendicular to the centerline. The ground truth label is then considered positive if any authoritative sidewalk polygon intersects this buffer zone.

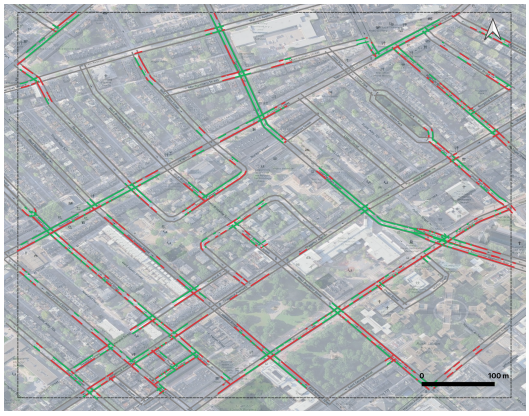
The 3×2 Expanded Confusion Matrix: A standard 2×2 confusion matrix treats all false negatives identically, unfairly penalising the deep learning model for Mapillary's incomplete spatial coverage. In this system, a road-side classified as `no_data` received zero imagery, where no model inference was ever performed, and thus no algorithmic error occurred. Conflating these coverage gaps with genuine model failures would inflate the false negative count and depress recall.

To address this, a 3×2 expanded confusion matrix is created that reports three prediction states *(Detected, Not Detected, No Data)* against the two ground truth states *(Sidewalk Exists, No Sidewalk)*. Each cell counts the number of individual roadsides (i.e., one side of a 5-meter road segment). When a prediction of `true` aligns with ground truth sidewalk presence, this counts as a True Positive (TP); when a prediction of `false` aligns with ground truth absence, it is a True Negative (TN). False Positives (FP) occur where the model detects a sidewalk but the ground truth disagrees, while False Negatives (FN) occur where imagery was available and the model predicted `false` despite a sidewalk existing.

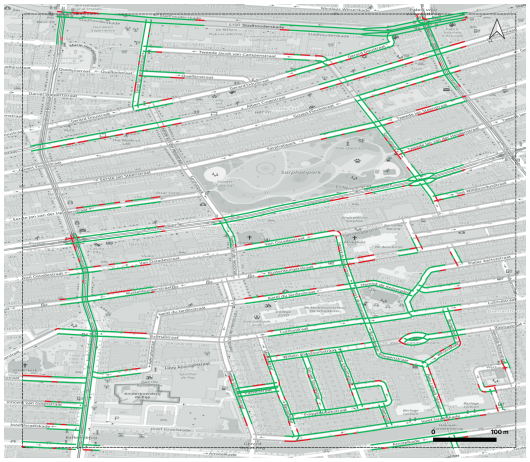
The remaining category accounts for regions lacking imagery, where ND^+ captures instances where a sidewalk exists but the lack of Mapillary imagery prevented model execution (representing a coverage gap rather than a model failure), whereas ND^- denotes instances where no imagery exists and no sidewalk is present, making it structurally irrelevant to model error. This explicit separation enables the following metrics:



(a) Boston South End inventory map



(b) Boston South End inventory map with satellite

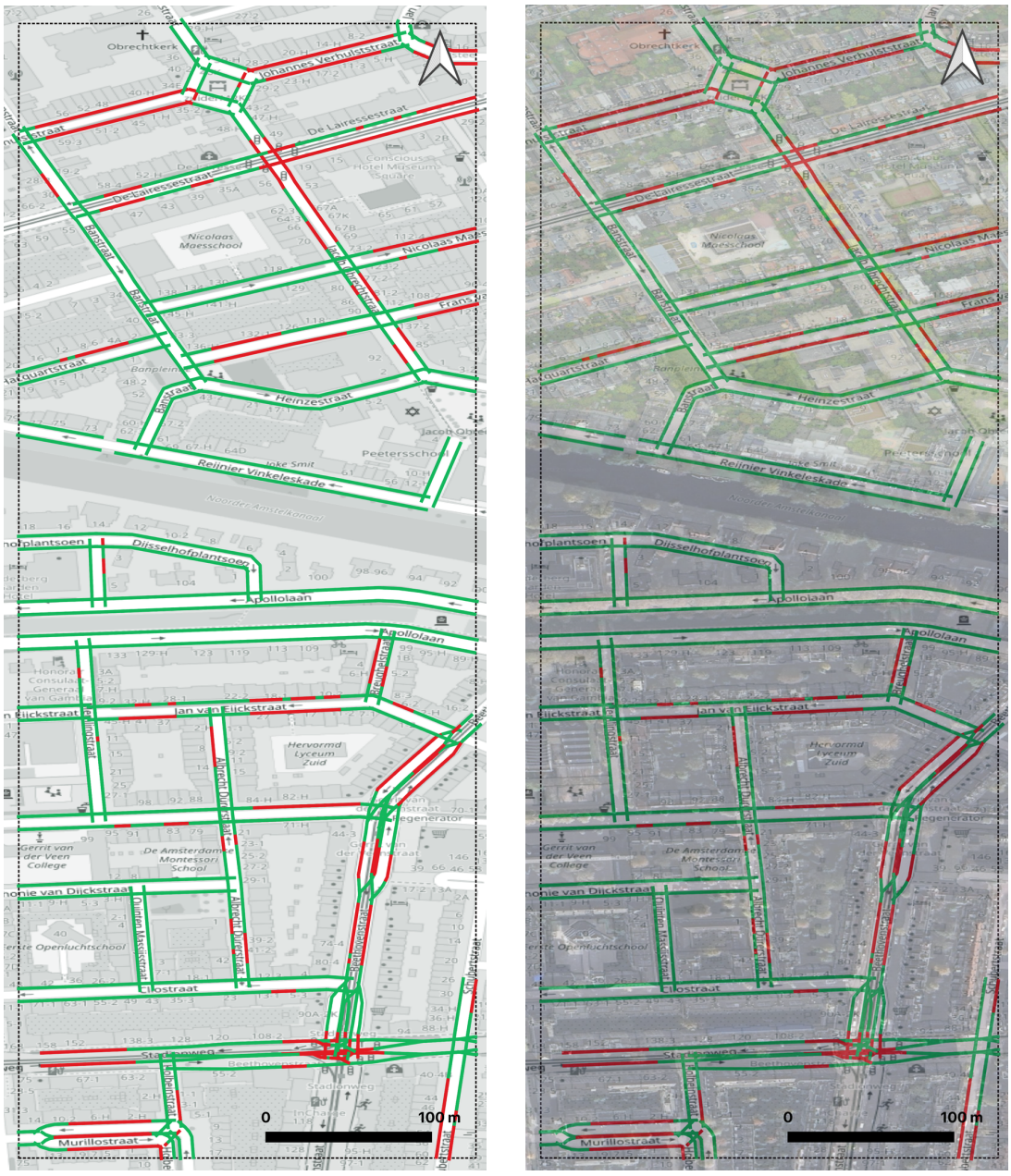


(c) De Pijp inventory map



(d) De Pijp inventory map with satellite

Figure 5.5: Sidewalk inventory maps for Boston South End and De Pijp showing detected sidewalk presence on the left and right sides of the road. Green: sidewalk detected; red: no sidewalk detected; grey: insufficient data (`no_data`). (continued on the next page).



(e) Hobbemakade inventory map

(f) Hobbemakade inventory map with satellite

Figure 5.5: Sidewalk inventory maps for Hobbemakade showing detected sidewalk presence on the left and right sides of the road. Green: sidewalk detected; red: no sidewalk detected; grey: insufficient data (`no_data`).

- **Algorithmic Recall** = $\frac{\text{TP}}{\text{TP} + \text{FN}}$: This evaluates the segmentation model on only those roadsides where it had imagery and made a decision. This isolates model performance from data availability.
- **System Recall** = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{ND}^+}$: this evaluates the *end-to-end system* including coverage limitations. It answers: "of all sidewalks that exist, how many did the complete pipeline detect?"
- **Algorithmic Precision** = $\frac{\text{TP}}{\text{TP} + \text{FP}}$: reliability of positive detections.
- **Algorithmic F1** = harmonic mean of algorithmic precision and recall.
- **System Coverage** = $\frac{\text{decided sides}}{\text{total sides}}$: proportion of roadsides receiving any assessment.

Results

Figure 5.6 presents the 3×2 expanded confusion matrix for each study area, and Table 5.5 summarises the derived metrics.

Table 5.5: Accuracy metrics derived from the expanded confusion matrix. Algorithmic metrics evaluate only road-sides where the model made a decision; System Recall includes coverage gaps.

Study Area	Algo. Prec.	Algo. Rec.	Sys. Rec.	Coverage	$\mathbf{N_{total}}$
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

Interpretation

Separating Algorithmic Performance from Data Gaps The primary advantage of the expanded 3×2 matrix (see Figure 5.6) is its ability to conclusively separate two distinct failure modes. First, it isolates algorithmic failures where imagery is available but sidewalks remain undetected (FN), which is primarily driven by physical occlusion from parked vehicles. Second, it explicitly quantifies the system's unmapped sidewalk deficits (ND^+). By isolating these states, a critical insight is revealed: the gap between algorithmic recall and total system recall across the study areas is entirely attributable to these missing imagery coverage gaps (ND^+) where real sidewalks exist, whereas regions lacking both imagery and sidewalks (ND^-) represent empty spaces that do not affect model accuracy.

Model Prediction	Ground Truth: Sidewalk Exists	Ground Truth: No Sidewalk
Detected	TP: 4,702 (42.2%)	FP: 125 (1.1%)
Not Detected	FN: 736 (6.6%)	TN: 11 (0.1%)
No Data	ND^+: 3,256 (29.3%)	ND^-: 2,300 (20.7%)

(a) De Pijp (n = 11,130 roadsides)

Model Prediction	Ground Truth: Sidewalk Exists	Ground Truth: No Sidewalk
Detected	TP: 2,637 (47.5%)	FP: 77 (1.4%)
Not Detected	FN: 1,030 (18.6%)	TN: 16 (0.3%)
No Data	ND^+: 183 (3.3%)	ND^-: 1,609 (29.0%)

(b) Hobbemakade (n = 5,552 roadsides)

Model Prediction	Ground Truth: Sidewalk Exists	Ground Truth: No Sidewalk
Detected	TP: 2,719 (7.1%)	FP: 40 (0.1%)
Not Detected	FN: 3,953 (10.3%)	TN: 307 (0.8%)
No Data	ND^+: 12,040 (31.3%)	ND^-: 19,461 (50.5%)

(c) Boston South End (n = 38,520 roadsides)

Figure 5.6: 3×2 expanded confusion matrices for each study area. Each cell counts individual roadsides (one side of a 5-metre road segment) with the percentage of the total shown in parentheses. Green cells indicate correct classifications, red cells indicate errors, and grey cells indicate roadsides with no imagery coverage.

The Impact of Coverage on Recall (Amsterdam vs. Boston) This distinction is clearly illustrated when comparing the study areas. In De Pijp, the model achieves 86.5% algorithmic recall, meaning when imagery exists, it detects sidewalks reliably. However, system recall drops to 54.1% because 3,256 roadsides with actual sidewalks received no imagery at all (ND^+). Conversely, in Hobbemakade, system recall (68.5%) remains much closer to algorithmic recall (71.9%) because its dense panoramic survey coverage resulted in only 183 positive roadsides falling into the no-data zone.

Boston provides the clearest illustration of the distinction between algorithmic and system performance. Relying on Mapillary API segmentation, the pipeline achieves an algorithmic precision of 0.986, the highest of all three areas. However, system recall is only 14.5% because 12,040 roadsides with confirmed ground-truth sidewalks received zero imagery coverage (ND^+). Of the 38,520 total roadsides, only 7,019 (18.2%) received any assessment.

Algorithmic Precision: High and Consistent As Table 5.5 shows, algorithmic precision is remarkably consistent across all three study areas (0.972-0.986), indicating that when the pipeline reports a sidewalk detection, it is correct over 97% of the time regardless of geographic context. This demonstrates the robustness of the confidence aggregation approach, which maintains near-perfect precision whether powered by local DINOv3 segmentation (in Amsterdam) or the Mapillary API (in Boston). The small number of remaining false positives (40-125 per area) is attributable to two causes: (1) road segments where the ground truth sidewalk polygon lies just beyond the 15 m matching buffer on wider arterial roads, and (2) segments at study area boundaries where the ground truth coverage is incomplete.

Low False Positives: Urban Area Ground Truth It is critical to contextualize these low false positive rates in our study area, as the authoritative ground truth for these dense urban environments features near-100% sidewalk coverage; there are very few true negative locations where a false positive could mathematically occur. To rigorously evaluate the baseline false positive rate and better understand potential semantic segmentation anomalies, future research must test this pipeline in non-urban or suburban contexts where pedestrian infrastructure is naturally sparse.

5.4 Pipeline 2

5.4.1 Sidewalk Geometry Reconstruction Results

This section represents the results of the sidewalk geometry reconstruction pipeline from Section 4.2, which creates sidewalk polygon geometries by projecting depth-estimated segmentation masks into geographic coordinates. Table 5.6 summarizes the reconstruction output across the three study areas.

Table 5.6: Pipeline 2 geometry reconstruction output per study area.

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

As detailed in Table 5.6, the predicted sidewalk area is substantially smaller than the ground truth across all study areas. This discrepancy is driven by the same data coverage limitations that constrain the inventory mapping pipeline; geometry can only be reconstructed where sidewalks are explicitly visible within the available imagery. Hobbemakade exhibits the highest reconstruction completeness, successfully capturing 60% of the ground truth area, whereas the Boston South End yields the lowest at just 21%. This also strongly correlates with the spatial coverage trends previously observed in the inventory mapping results (see Table 5.4).

Impact of GPS Accuracy on Polygon Shape A further source of geometric error visible in Figure 5.7 is the positional accuracy of the input imagery. Each reconstructed polygon is the union of individually back-projected masks from multiple image sequences, often captured by different instruments with varying GPS precision. Although Mapillary applies Structure-from-Motion (SfM) corrections to improve spatial consistency within individual sequences, these adjustments might not fully resolve absolute positional inaccuracies across different sequences. When masks from sequences with differing GPS drift are fused, the resulting polygon boundaries appear irregular and curved rather than the sharp, straight edges one would expect from actual sidewalk geometry. In extreme cases, entire polygon fragments are displaced onto the road surface or adjacent buildings, reflecting systematic GPS drift in a particular capture sequence that SfM correction was unable to fully resolve. This explains a portion of the false positive area: geometrically valid sidewalk detections that are spatially shifted from their true location.

The sidewalk inventory mapping pipeline is not significantly affected by this GPS drift error, but the sidewalk geometry reconstruction pipeline, which requires projecting and combining sidewalk masks from different devices, introduces substantial spatial inaccuracy.

5.4.2 Accuracy Assessment

The reconstructed polygons are evaluated against ground truth using area-based spatial overlap metrics. To avoid penalising the pipeline for coverage gaps (the same limitation identified in Pipeline 1), the ground truth is first clipped to the *imagery-covered area* defined as a 15 m buffer around all road segments that received at least one image assessment (i.e., segments where the Pipeline 1 prediction is `true` or `false`, not `no_data`). All geometries are then dissolved into single union geometries per study area before computing intersection and union areas.

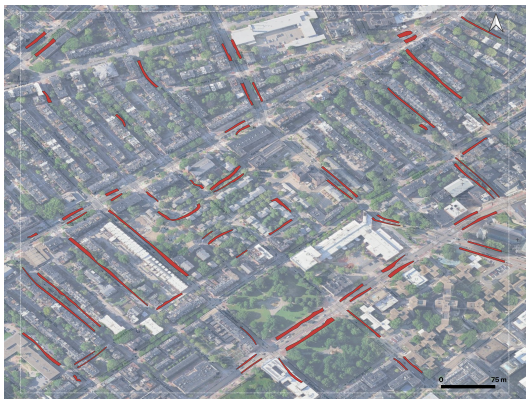
Three metrics are reported, where P is the predicted polygon area and G is the coverage-clipped ground truth area:

- **IoU** (Intersection over Union) = $\frac{|P \cap G|}{|P \cup G|}$: the standard spatial overlap measure.
- **Area Precision** = $\frac{|P \cap G|}{|P|}$: of the total area reconstructed by the pipeline, what proportion actually overlaps a ground truth sidewalk.
- **Area Recall** = $\frac{|P \cap G|}{|G|}$: of the total ground truth sidewalk area, what proportion is covered by the reconstruction.

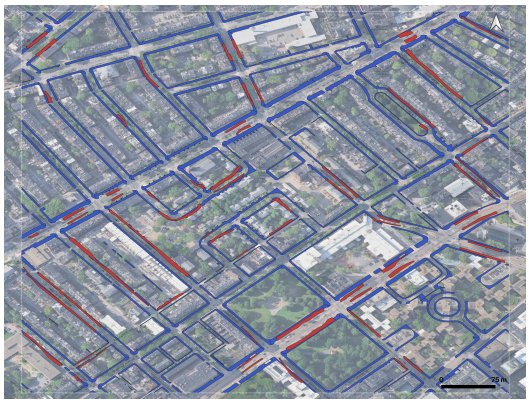
Table 5.7 presents the results.

Dimensional and Positional Accuracy

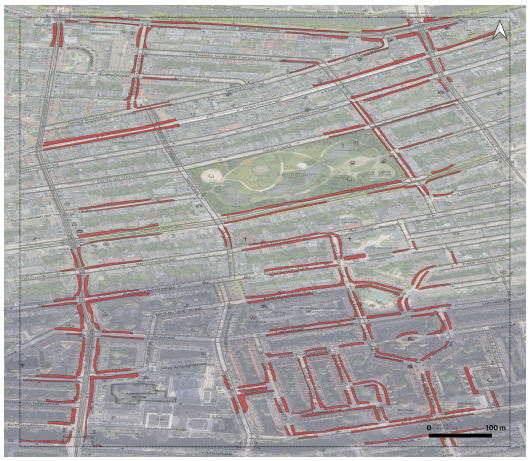
In addition to area-based metrics, the geometric quality of the reconstructed polygons was assessed by evaluating their width estimation and lateral positioning compared to the ground truth. To evaluate dimensional accuracy, the average width of each polygon was derived every 5m, providing a proxy for width. Positional drift was measured by calculating the



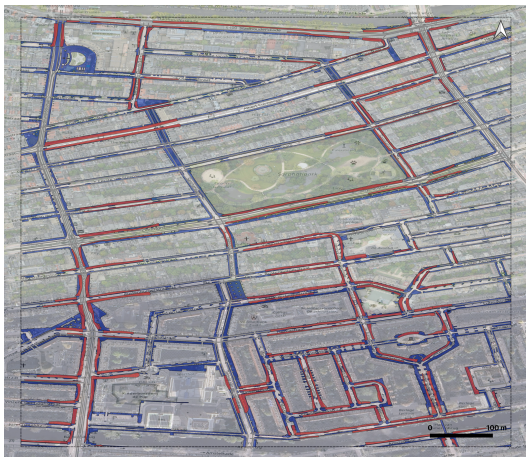
(a) Boston South End reconstructed polygons (red)



(b) Boston South End reconstructed polygons overlaid on authoritative ground truth (blue)



(c) De Pijp reconstructed polygons (red)

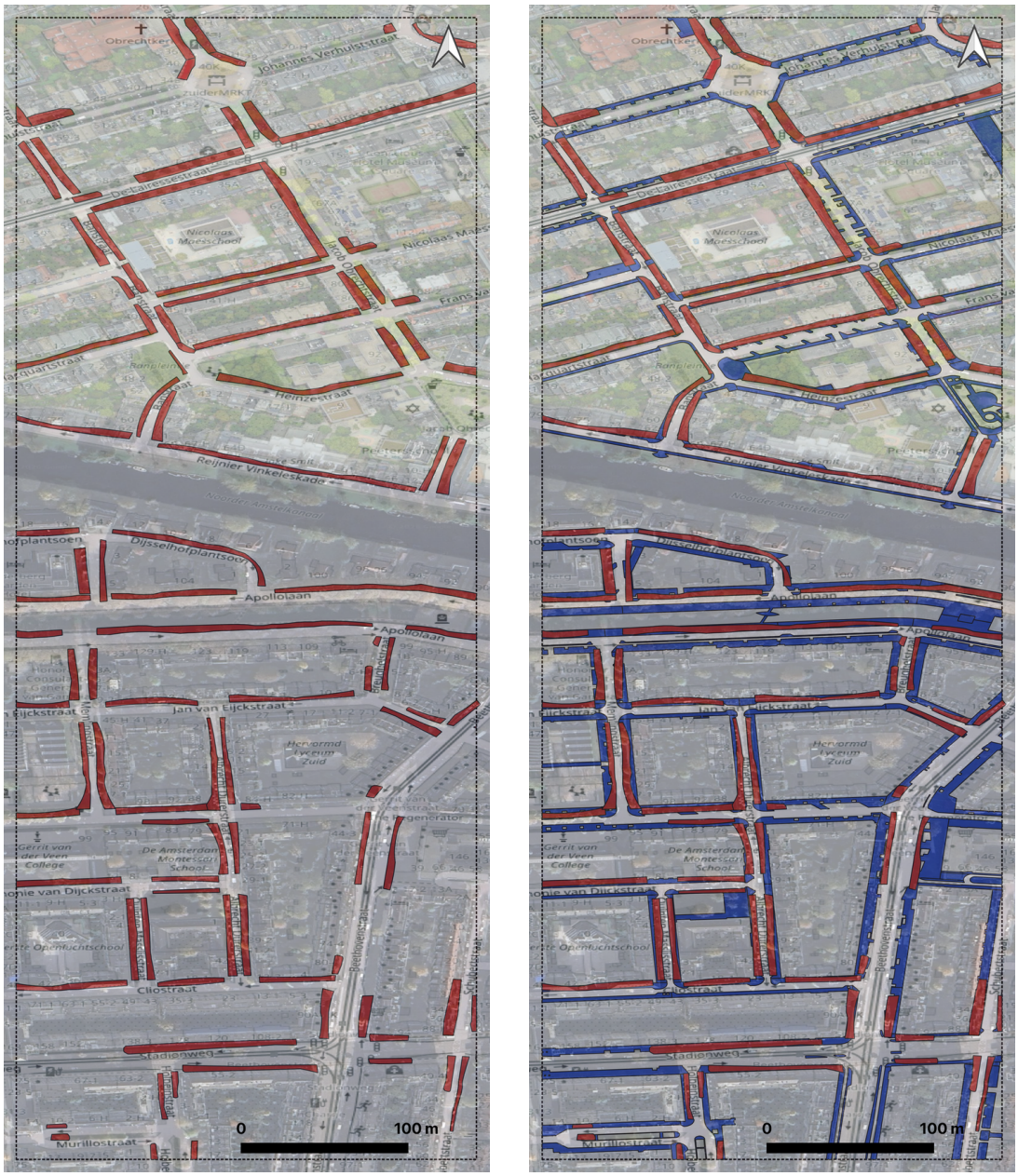


(d) De Pijp reconstructed polygons (red) overlaid on authoritative ground truth (blue)

Figure 5.7: Reconstructed sidewalk polygons (red) for Boston South End and De Pijp. For each study area, two maps are presented: the isolated reconstructed polygon (red) and the reconstructed polygons overlaid onto the authoritative ground truth (blue) for spatial comparison. (continued on the next page).

Table 5.7: Area-based accuracy metrics for Pipeline 2 geometry reconstruction. Ground truth is clipped to the imagery-covered area to exclude regions where no reconstruction was possible.

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



(e) Hobbemakade reconstructed polygons

(f) Hobbemakade reconstructed polygons (red) overlaid on authoritative ground truth (blue)

Figure 5.7: Reconstructed sidewalk polygons (red) for Hobbemakade. For each study area, two maps are presented: the isolated reconstructed polygon (red) and the reconstructed polygons overlaid onto the authoritative ground truth (blue) for spatial comparison.

centerline offset, defined as the Euclidean distance between the centroid of the predicted polygon and its best-matching ground truth counterpart. Table 5.8 summarizes these results.

Table 5.8: Dimensional and positional accuracy of reconstructed sidewalk polygons. Width errors reflect the difference between predicted and ground-truth width estimations. Centerline offset measures the lateral displacement between matched polygon centroids.

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

The width metrics reveal that the depth-projection approach introduces a substantial average width error (MAE) between 1.18 m and 1.40 m. Given that a typical urban sidewalk is only 1.5 to 3.0 metres wide, an error of this magnitude indicates that the pipeline's width estimations are not reliable for precise infrastructure planning. The RMSE values (1.64 m to 1.72 m) further highlight this variance, stemming from various error sources that cause reconstructed polygons to be drastically over- or under-sized.

The centerline offset error, ranging from 2.64 m in Hobbemakade to 3.15 m in Boston South End, quantifies significant lateral drift. These 2-3 metre shifts frequently cause reconstructed sidewalks to overlap with the road carriage or adjacent buildings.

Interpretation

As Table 5.7 shows, the spatial overlap remains low even when excluding no-data regions, with IoU values ranging from 0.100 to 0.329. The area precision (0.353 to 0.623) indicates that a substantial portion of the reconstructed polygon area falls outside actual sidewalk boundaries. This geometric inaccuracy is the compounding result of multiple error sources: segmentation model bleeding, depth estimation inaccuracies at varying distances, reprojection artefacts, inherent GPS drift, and varying device source quality.

Area recall (12.2–41.1%) represents the pipeline's ability to reconstruct sidewalk geometry where imagery is available. In Hobbemakade, where 67.7% (see Table 5.4) of the ground truth falls within the covered zone, the pipeline captures 41.1% of the accessible sidewalk area. This demonstrates that as the input density of the imagery improves, the area recall increases accordingly.

Ultimately, these results demonstrate that while Pipeline 2 can successfully identify and project the general spatial presence of sidewalks, the current methodology struggles to produce geometrically precise reconstructions. The combination of hardware-level GPS drift and software-level depth estimation limits the output's utility for strict dimensional analysis; currently generated polygons are better suited as indicative spatial markers rather than authoritative property boundaries.

5.5 Computational Cost

In addition to evaluating spatial accuracy, computational demands must be considered to assess viability for city-scale deployment. To this end, all experiments were benchmarked on the DelftBlue Supercomputer (Delft High Performance Computing Centre , DHPC).

5.5.1 Pipeline 1: Sidewalk Inventory Mapping

Conversely, Pipeline 1 offers a substantially more efficient approach:

- **Phase 1 (Image Retrieval):** Downloading of required data, dependent on internet speed.
- **Phase 2 (Segmentation):** The computational cost depends entirely on the data source. If pre-computed Mapillary API segmentation masks are available (e.g., Boston perspective imagery), local segmentation is skipped entirely, reducing the processing time for this phase to zero. If local segmentation is required (e.g., Amsterdam panoramas), it incurs the same 4.5-hour penalty per 1,000 panoramas as in Pipeline 2.
- **Phase 3 (Spatial Aggregation):** The final mapping phase is exceptionally fast. It takes only 2 to 3 minutes to project and aggregate the detections of 1,000 images onto the road network.

5.5.2 Pipeline 2: Sidewalk Geometry Reconstruction

Pipeline 2 is computationally intensive. Its execution times are primarily driven by the deep learning models:

- **Phase 1 (Retrieval):** Bound by internet speed. Panoramic images are split into four crops, multiplying downstream volume.
- **Phase 2 (Segmentation):** The pipeline's bottleneck. On an NVIDIA A100 GPU (80 GB VRAM), DINOv3 requires $\sim$4.1 s per crop. Processing 1,000 panoramas (4,000 crops) takes $\sim$4.5 hours. The model's heavy Video Random Access Memory (VRAM) footprint forces small batch sizes, restricting throughput.
- **Phase 3 (Depth Estimation):** `DA3-Giant` is more lightweight, taking $\sim$0.46 s per image ($\sim$30 mins per 1,000 panoramas). It only runs on frames where Phase 2 detected sidewalks, significantly boosting speed.
- **Phase 4 (3D Projection):** Using Python multiprocessing on a 32-core CPU, spatial projection takes $\sim$0.15 s per image, completing 1,000 panoramas in 5-10 minutes.

Overall, processing a small 1 km^2 area (e.g., De Pijp, 1,500 images) takes 7-9 hours on an A100 GPU, making city-scale deployment exceedingly difficult.

5.5.3 Synthesis

Given the high computational cost and constrained geometric accuracy of explicit 3D reconstruction (Pipeline 2), the inventory mapping approach (Pipeline 1) is a far more practical, scalable solution. It provides sidewalk assessment at a fraction of the computational time and cost. As demonstrated in Boston, relying on pre-computed API segmentation, where available, eliminates the local segmentation bottleneck while maintaining 98.6% algorithmic precision, cementing Pipeline 1 as the optimal strategy for massive urban networks.

Additionally, as discussed in Section 5.2.2, the input data has inherent coverage limitations and physical occlusions; both sidewalk inventory and sidewalk geometry reconstruction implementations must be classified strictly as sidewalk presence detection methods. A negative result only indicates that a sidewalk could not be visually confirmed; it cannot definitively verify absence, as the system is unable to distinguish between a genuinely missing sidewalk and one that is physically obscured.

6

Conclusions

6.1 Conclusion

The primary objective of this thesis was to empirically evaluate the feasibility of converting raw, Crowdsourced Street View Imagery (CSVI) into structured, georeferenced sidewalk data utilizing strictly openly licensed data and pre-trained models. By developing two complementary extraction pipelines-one for lightweight sidewalk presence (the **Sidewalk Inventory Mapping** pipeline 1) and another for georeferenced surface polygons (the **Sidewalk Geometry Reconstruction** pipeline 2) this research successfully bridges the technical gap between unstructured crowdsourced imagery and actionable urban infrastructure mapping. Crucially, this methodology functions strictly as an inference-only pipeline. By utilizing pre-trained foundation models, the approach completely eliminates the need for any form of model training or fine-tuning while simultaneously bypassing the traditional reliance on proprietary commercial datasets. The primary findings regarding the accuracy and practical limitations of these approaches are presented below by answering the three core research questions.

RQ1: Sidewalk Inventory Mapping (Pipeline 1): *To what extent can a scalable, city-wide sidewalk presence inventory be constructed by aggregating pre-computed semantic segmentation outputs and SfM-corrected camera metadata already available through the Mapillary API, without performing any local model inference?*

This research demonstrates that constructing a reliable, city-wide sidewalk presence inventory pipeline from Mapillary's pre-computed outputs is not only feasible but precise too.

Across two different types of urban areas, the pipeline achieves algorithmic precision between 0.972 and 0.986, meaning that when the pipeline reports a sidewalk as present, it is correct over 97% of the time. Algorithmic recall reaches 86.5% in De Pijp and 71.9% in Hobbemakade when local DINOv3 segmentation supplements the API for panoramic imagery.

The immediate scalability of this approach remains constrained by two primary limitations in the Mapillary data ecosystem. First, the Mapillary API currently lacks pre-computed segmentation for panoramic images in certain regions, including Amsterdam. This restricts the highly efficient, API-only workflow to areas dominated by perspective imagery. In contrast, regions like the Boston South End, where pre-computed masks are universally available across all image types, allowed for inventory generation with zero local computational overhead. Augmenting the pipeline with local DINOv3 inference successfully resolved this first deficit, elevating algorithmic recall to 86.5% in De Pijp and 71.9% in Hobbemakade. Second, while the algorithmic validity of the methodology is robust, total system effectiveness is fundamentally bounded by the Mapillary community's ability to provide comprehensive spatial coverage.

RQ2: Sidewalk Geometry Reconstruction (Pipeline 2): *How effectively can an automated pipeline combining local semantic segmentation (DINOv3) and monocular metric depth estimation (Depth Anything V3) produce georeferenced sidewalk polygons and centerlines?*

The research proves that a fully automated, software-driven geometric reconstruction pipeline is functionally possible, and the geometric reconstruction pipeline successfully produces georeferenced sidewalk polygons and centerlines, demonstrating that monocular metric depth estimation can serve as a software-based alternative to LiDAR for sidewalk surface reconstruction. By exploiting the implicit spatial and structural priors of the `DA3-Giant` foundation model, our approach overcomes a tested limitation of Multi-View Stereo (MVS), which consistently struggles to find feature correspondences on flat, textureless concrete surfaces.

However, while variable-width polygons and pedestrian routing centerlines can be mathematically projected, the geometric accuracy of these outputs is limited. The area-based spatial overlap (Intersection over Union) reflected these geometric misalignments, ranging from 0.100 to 0.329, and width estimation exhibited a mean absolute error of 1.18–1.40 m with centerline offsets of 2.64–3.15 m, anomalies that frequently resulted in reconstructed geometries projecting into adjacent vehicular roadways or building footprints.

These errors result from the compounding accumulation of multiple error sources: segmentation boundary imprecision, depth estimation degradation at distal ranges, uncorrected GPS multipath drift across different capture sequences, and occlusion by parked vehicles. While the pipeline proves that the concept is technically viable, the current accuracy is insufficient for precise infrastructure planning or direct integration into authoritative spatial databases.

RQ3: Comparison of Approaches: *How do the inventory-based approach and the geometry reconstruction approach compare in terms of detection accuracy, geometric precision, spatial coverage, and computational cost?*

- **Detection Accuracy:** Pipeline 1 achieves consistently high algorithmic precision (0.972 - 0.986) with moderate-to-high recall where imagery exists. Pipeline 2 achieves an area precision of 0.353 - 0.623, indicating that 38–65% of the reconstructed area falls outside the actual sidewalk boundaries due to various sources of error.
- **Geometric Precision:** Pipeline 1 produces no geometry, only binary yes/no labels per roadside. Pipeline 2 produces polygons with a width MAE of 1.18 – 1.40 m and centerline positional offsets of 2.64 – 3.15 m, reflecting the accumulated inaccuracies of the segmentation model (mIoU 63), the depth model (F1 of 87.1), and inter-sequence GPS drift.
- **Spatial Coverage:** Both pipelines are fundamentally bounded by Mapillary's image availability. Where imagery is absent, neither pipeline can produce output (49.9% no-data in De Pijp, 81.8% in Boston for Pipeline 1). Pipeline 1 is more tolerant of partial occlusion because even a small visible sidewalk region suffices for binary detection, whereas Pipeline 2 requires clear, unobstructed views of the full sidewalk surface for accurate width estimation.
- **Computational Cost:** Finally, the computational asymmetry between the two approaches dictates their practical utility. Pipeline 1 is orders of magnitude more efficient when Mapillary API segmentation is available (e.g., Boston); the aggregation phase completes in 2–3 minutes for 1,000 images on a standard CPU, with zero GPU requirement. Pipeline 2 requires approximately 7–9 hours on an NVIDIA A100 GPU for a 1 km² area (1,500 images), with segmentation alone consuming 4.5 hours per 1,000 panoramas. This renders Pipeline 2 impractical for city-scale deployment in its current form.

Synthesis: The inventory mapping pipeline (Pipeline 1) is the practical, scalable solution for city-wide sidewalk mapping. Regarding detection accuracy, geometric precision, spatial coverage, and computational cost, Pipeline 1 excels over Pipeline 2 precisely because it avoids geometry altogether. Instead of requiring complete geometric boundary reconstruction, even minimal visual evidence within the segmentation mask is sufficient for Pipeline 1 to successfully register sidewalk presence. It also delivers robust presence detection at minimal computational cost and can be deployed immediately wherever Mapillary coverage exists. Conversely, the geometry reconstruction pipeline (Pipeline 2) demonstrates the technical feasibility of deriving sidewalk geometry from crowdsourced imagery without LiDAR, but its current accuracy and computational demands position it as a research contribution rather than a production-ready tool.

6.2 Discussion

The results of this thesis reveal a fundamental tension in automated sidewalk mapping from crowdsourced imagery: the trade-off between information richness and reliability. Pipeline 1 extracts limited information (binary presence) but does so with high confidence, while Pipeline 2 attempts richer output (metric geometry) but at significantly reduced accuracy.

Strengths of the Approach: The theoretical and practical implications of this research are most evident when synthesized against the prevailing academic literature. The exclusive use of openly licensed data (Mapillary CC BY-SA 4.0) ensures that the outputs are legally compatible with open mapping initiatives, addressing a key limitation of prior work that relied on proprietary Google Street View (GSV) imagery. Furthermore, the research incorporates pedestrian-captured perspective imagery, which is inherently resilient to visual blockages caused by parked vehicles. Finally, the use of pre-trained foundation models (DINOv3, Depth Anything V3) without any fine-tuning demonstrates that zero-shot inference is highly viable for sidewalk extraction, eliminating the localized training requirements that have historically limited reproducibility.

The Hybrid Approach: Aerial and SVI Complementarity: One of the dominant sources of missed detections in the pipeline is visual occlusion by parked vehicles. The Boston study area demonstrates this most clearly: the per-image sidewalk detection rate is only approximately 70%, compared to 88–92% in the Amsterdam study areas (Table 5.3), due to sidewalks being blocked laterally by dense on-street parking. However, this limitation effectively positions the proposed Street View Imagery (SVI) methodology as highly complementary to aerial-based deep learning frameworks such as TILE2NET (Hosseini et al., 2023). While aerial segmentation performs exceptionally well on open streets and is completely unaffected by parked cars, it is often compromised by building shadows and dense tree canopies. Conversely, as demonstrated in the Hobbemakade study area along Heinzstraat, this ground-level SVI pipeline successfully maps sidewalks beneath heavy vegetation and urban canyons where overhead sensors are entirely occluded or ambiguous. A hybrid approach by combining aerial segmentation for broad urban coverage with SVI-based mapping for targeted infill in vegetated or shadowed zones maximizes the distinct geometric strengths of both data sources to outperform either system in isolation.

Limitations and Error Sources: The geometry reconstruction pipeline is subject to cascading error accumulation. Each processing stage introduces uncertainty: segmentation boundaries contain inherent pixel noise (DINOv3 mIoU of 63.0 on ADE20K); metric depth estimates (F1 of 87.1) project approximately 13% of 3D points with spatial errors exceeding 25 cm, and the SfM-corrected GPS positions lack any reported accuracy metric from Mapillary. When sidewalk masks from different image sequences, captured by different devices with varying GPS precision, are fused, the resulting polygons exhibit irregular boundaries and lateral displacement.

Also, during the testing phase, the multi-view reconstruction approach (SfM/MVS) was experimentally ruled out due to the inherently texture-less nature of sidewalk surfaces, which yield unreliable, scarce feature correspondences per sequence. This finding confirms that sidewalks present a fundamentally challenging case for correspondence-based methods, motivating the monocular depth approach despite its own limitations.

Dependence on Platform Coverage: The most significant practical limitation is the dependency on Mapillary's spatial coverage. In Boston, only 18.2% (Table 5.4) of roadsides received any assessment, with the vast majority lacking imagery entirely. The gap between algorithmic recall and system recall is attributable to missing imagery rather than model error: where imagery exists, the pipeline achieves 97–99% precision (Table 5.5). This finding underscores that the quality of crowdsourced sidewalk mapping is ultimately bounded by

the density and diversity of contributed imagery. Consequently, rather than indicating a failure of the algorithm, the 18.2% Boston coverage highlights a clear directive for future data collection.

6.3 Future Work

Several directions emerge from the identified limitations:

Image Source Filtering: The current methodology processes all available images within a bounding box to maximise spatial coverage. An alternative strategy could filter sequences by capture source, retaining only pedestrian or cyclist captured imagery, which provides unobstructed lateral views of sidewalks. While this would reduce spatial coverage, it may improve geometric accuracy for Pipeline 2 by eliminating the vehicle-occlusion problem. However, such imagery is captured using consumer-grade smartphones with lower GPS precision, and the positional drift in urban canyon environments may offset any gain in segmentation quality.

Broader Geographic Testing: This thesis evaluated three study areas in two cities. Testing across diverse urban morphologies, including suburban areas with sparse sidewalk coverage and cities with different parking cultures, would clarify the generalizability of both pipelines and expose potential failure modes in non-Western contexts. Additionally, testing in areas with sparse ground-truth sidewalks would rigorously evaluate the false positive rate, which could not be fully assessed in the current near-100% sidewalk coverage urban environments.

Model Improvement: As foundation models for segmentation and depth estimation continue to improve, evidenced by the progression from DINOv2 (mIoU 49.5) to DINOv3 (mIoU 55.9) within a single model generation, the accuracy of both pipelines will benefit directly without requiring architectural changes. The pose-refined depth constraint in Depth Anything 3 is the current state-of-the-art; future models with even tighter multi-view consistency may substantially reduce the width estimation errors observed in this work.

Platform Evolution: Once Mapillary extends its pre-computed segmentation to panoramic imagery globally, the inventory mapping pipeline will become fully operational without any local computation in all regions, a scenario already demonstrated for Boston. Monitoring this platform capability and adapting the pipeline accordingly represents a low-effort, high-impact path to wider deployment.

Bibliography

Aanæs H, Jensen RR, Vogiatzis G, Tola E, and Dahl AB (2016). Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision (IJCV)*, 120(2):153–168. doi:10.1007/s11263-016-0902-9.

Abou-Senna H, Radwan E, and Mohamed A (2022). Investigating the correlation between sidewalks and pedestrian safety. *Accident Analysis & Prevention*, 166:106548. doi:10.1016/j.aap.2021.106548.

Aceves-González C, Ekambaram K, Rey-Galindo J, and Rizo-Corona L (2020). The role of perceived pedestrian safety on designing safer built environments. *Traffic Injury Prevention*, 21(sup1):S84–S89. doi:10.1080/15389588.2020.1812062. PMID: 32926653.

Ai C and Tsai Y (2016). An automated sidewalk assessment method for the americans with disabilities act compliance using 3-d mobile LiDAR. *Transportation Research Record*, 2542:25–32. doi:10.3141/2542-04.

Beuster L, van Selm M, Venverloo T, García-Sánchez C, Duarte F, and Ledoux H (2025). (slim) shady - the relative role of buildings and trees in urban shade provision for pedestrians. *Nature Cities (Preprint)*. doi:10.21203/rs.3.rs-6966874/v1. Preprint (Version 1) available at Research Square.

Biljecki F and Ito K (2021). Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning*, 215:104217. doi:10.1016/j.landurbplan.2021.104217.

Bochkovskii A, Delaunoy A, Germain H, Santos M, Zhou Y, Richter SR, and Koltun V (2025). Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*. doi:10.48550/arXiv.2410.02073.

Boeing G (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139. doi:10.1016/j.compenvurbsys.2017.05.004.

Brunken H and Gühmann C (2020). Road surface reconstruction by stereo vision. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88:433–448. doi:10.1007/s41064-020-00130-z.

Chen LC, Zhu Y, Papandreou G, Schroff F, and Adam H (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint*. doi:10.48550/arXiv.1802.02611. Published in ECCV 2018.

Cheng B, Misra I, Schwing AG, Kirillov A, and Girdhar R (2021). Masked-attention mask transformer for universal image segmentation. doi:10.48550/arXiv.2112.01527.

Delft High Performance Computing Centre (DHPC) (2024). DelftBlue Supercomputer (Phase 2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, and Houlsby N (2020). An image is worth 16x16 words: Transformers for image recognition at scale. doi:10.48550/arXiv.2010.11929.

Figliozzi M and Jennings D (2020). Autonomous delivery robots and their potential impacts on urban freight energy consumption and emissions. *Transportation Research Procedia*. doi:10.1016/j.trpro.2020.03.159.

Forsyth A and Southworth M (2008). Cities afoot—pedestrians, walkability and urban design. *Journal of Urban Design*, 13(1):1–3. doi:10.1080/13574800701816896.

Google (2023). Street view imagery is not real time. https://www.google.com/streetview/policy/. Accessed: 2025-01-06.

Hamim OF, Kancharla SR, and Ukkusuri SV (2023). Mapping sidewalks on a neighborhood scale from street view images. *Environment and Planning B: Urban Analytics and City Science*, 51(7). doi:10.1177/23998083231200445.

Hosseini M, Sevtsuk A, Miranda F, Cesar Jr RM, and Silva CT (2023). Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101:101950. doi:10.1016/j.compenvurbsys.2023.101950.

Ito K, Zhu Y, Abdelrahman M, Liang X, Fan Z, Hou Y, Zhao T, Ma R, Fujiwara K, Ouyang J, Quintana M, and Biljecki F (2025). Zensvi: An open-source software for the integrated acquisition, processing and analysis of street view imagery towards scalable urban science. *Computers, Environment and Urban Systems*, 119. doi:10.48550/arXiv.2412.18641. ArXiv:2412.18641.

Iwasawa Y, Nagamine K, Yairi IE, and Matsuo Y (2015). Toward an automatic road accessibility information collecting and sharing based on human behavior sensing technologies of wheelchair users. *Procedia Computer Science*, 63:74–81. doi:10.1016/j.procs.2015.08.314.

Kang B, Lee S, and Zou S (2021). Developing sidewalk inventory data using street view images. *Sensors*, 21(9):3300. doi:10.3390/s21093300.

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, and Girshick R (2023). Segment anything. *arXiv preprint arXiv:2304.02643*. doi:10.48550/arXiv.2304.02643.

Lieu SJ and Guhathakurta S (2025). A novel approach for estimating sidewalk width from street view images and computer vision. *Environment and Planning B: Urban Analytics and City Science*. doi:10.1177/23998083251369602.

Lin H, Chen S, Liew JH, Chen DY, Li Z, Shi G, Feng J, and Kang B (2025). Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*. doi:10.48550/arXiv.2511.10647.

Liu S, Long Y, Zhang L, Wang L, and Yang L (2025). Accurate three-dimensional mapping for detecting shaded urban sidewalks and assessing tree canopy coverage. *International Journal of Geographical Information Science*. doi:10.1080/13658816.2025.2541196.

Ma B, Yang J, Di D, Zhang X, Cui J, Li H, Xie Y, and Chen W (2026). Metricanything: Scaling metric depth pretraining with noisy heterogeneous sources. doi:10.48550/arXiv.2601.22054.

Mapillary (2016). Semantic segmentation and object recognition. https://blog.mapillary.com/update/2016/09/27/semantic-segmentation-object-recognition.html. Mapillary Blog, accessed 21 May 2026.

Mapillary (2023). Opensfm: Open source structure-from-motion pipeline. https://opensfm.org/docs/.

Mapillary (2025a). Cc-by-sa license for open data. Accessed: 2025-01-01.

Mapillary (2025b). Images and data for openstreetmap. Accessed: 2025-01-01.

Mapillary (2025c). Mapillary api documentation. Accessed: 2025-01-07.

Merriam-Webster (2024). Sidewalk. https://www.merriam-webster.com/dictionary/sidewalk. Accessed: 2025-01-06.

Merry K and Bettinger P (2019). Smartphone GPS accuracy study in an urban environment. *PLOS ONE*, 14(7):e0219890. doi:10.1371/journal.pone.0219890.

Morra D, Zhu X, Liu C, Fu K, Duarte F, Mora S, He Z, and Ratti C (2024). Mapping sidewalk accessibility with smartphone imagery and visual ai: A participatory approach. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2285):20240106. doi:10.1098/rsta.2024.0106.

Neuhold G, Ollmann T, Bulò SR, and Kontschieder P (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4990–4999. doi:10.1109/ICCV.2017.534.

Ning H, Ye X, Chen Z, Liu T, and Cao T (2021). *Environment and Planning B: Urban Analytics and City Science*, 49(1):7–22. doi:10.1177/2399808321995817.

Ning H, Ye X, Li Z, Wang C, Hodgson ME, Huang X, and Li X (2022). Converting street view images to land cover maps for metric mapping: A case study on sidewalk network extraction for the wheelchair users. *Computers, Environment and Urban Systems*, 95:101808. doi:10.1016/j.compenvurbsys.2022.101808.

OpenStreetMap contributors (2024). Openstreetmap. *OpenStreetMap*.

OpenStreetMap Wiki (n.d.a). Automated edits code of conduct. Accessed 21 May 2026.

OpenStreetMap Wiki (n.d.b). Import/guidelines. Accessed 21 May 2026.

OpenStreetMap Wiki (n.d.c). Seattle, washington/sidewalk import. Accessed 21 May 2026.

OpenStreetMap Wiki (n.d.d). Sidewalks. Accessed 21 May 2026.

OpenStreetMap Wiki (n.d.e). Tag:area:highway=footway. Accessed 21 May 2026.

Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jégou H, Mairal J, Labatut P, Joulin A, and Bojanowski P (2024). Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*. doi:10.48550/arXiv.2304.07193.

Saha M, Saugstad M, Maddali HT, Zeng A, Holland R, Bower S, Dash A, Chen S, Li A, Hara K, and Froehlich J (2019). Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14. doi:10.1145/3290605.3300292.

Schönberger JL and Frahm JM (2016). Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2016.445.

Senlet T and Elgammal A (2012). Segmentation of occluded sidewalks in satellite images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 805–808. ISBN 978-1-4673-2216-4.

Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C, Khalidov V, Szafraniec M, Yi S, Ramamonjisoa M, Massa F, Haziza D, Wehrstedt L, Wang J, Darcet T, Moutakanni T, Sentana L, Roberts C, Vedaldi A, Tolan J, Brandt J, Couprie C, Mairal J, Jégou H, Labatut P, and Bojanowski P (2025). Dinov3. doi:10.48550/arXiv.2508.10104.

Smith V, Malik J, and Culler D (2013). Classification of sidewalks in street view images. In *Proceedings of the 2013 IEEE International Green Computing Conference (IGCC)*, pages 1–6. doi:10.1109/IGCC.2013.6604476.

Teed Z and Deng J (2020). Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*. doi:10.48550/arXiv.2003.12039.

Weld GC, Jang E, Li A, Zeng A, Heimerl K, and Froehlich JE (2019). Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*, pages 196–209. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308561.3353798.

Xia J, Gong G, Liu J, Zhu Z, and Tang H (2023). Pedestrian-accessible infrastructure inventory: Enabling and assessing zero-shot segmentation on multi-mode geospatial data for all pedestrian types. *arXiv preprint*. doi:10.48550/arXiv.2310.09918. Version 4, last revised 28 Nov 2023.

Xu J, Liu S, Vahdat A, Byeon W, Wang X, and Mello SD (2023). ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation with text-to-image diffusion models. *arXiv preprint*. doi:10.48550/arXiv.2303.04803. Published in CVPR 2023 (Highlight).

Yang L, Kang B, Huang Z, Xu X, Feng J, and Zhao H (2024). Depth anything v2. *arXiv preprint arXiv:2406.09414*. doi:10.48550/arXiv.2406.09414.

Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, and Torralba A (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321. doi:10.1007/s11263-018-1140-0.

Zhou Q, Duan J, Qiao J, Liu Z, Yang H, et al. (2025). A large crowdsourced street view dataset for mapping road surface types in africa. *Scientific Data*, 12(1003). doi:10.1038/s41597-025-05153-y.

Zuniga-Teran AA, Orr BJ, Gimblett RH, Chuang WC, and Marsh SE (2017). Designing healthy communities: Testing the walkability model. *Frontiers of Architectural Research*, 6(1):63–73. doi:10.1016/j.foar.2016.11.005.

Appendix : Declaration of AI/LLM usage

This appendix provides a declaration of the use of Artificial Intelligence (AI) and Large Language Models (LLMs) throughout the entire research process for this thesis.

What AI/LLM tool(s) are used?

The tools used in this study are Claude (Anthropic) and Gemini (Google).

How are they used?

Claude : Used for assistance with debugging Python code, particularly when encountering confusing error messages or unexpected behaviour that could not be resolved through conventional search engines. As the computational work was performed on the DelftBlue HPC cluster, Claude was also utilized to troubleshoot Linux command-line environments and refine batch job submission scripts. Additionally, it was used as secondary help to understand and verify technical concepts.

Gemini : Used primarily for writing-related tasks, including editing and structuring thesis text, improving grammar and clarity of written sections, formatting tables, etc.

To what extent are they used?

Claude was used occasionally during the implementation phase when bugs could not be resolved through conventional search engines and for verifying understanding of technical concepts encountered in the literature. Gemini was used more regularly during the writing phase, particularly towards the end of the thesis when the text had to be structured and refined.

Ethics: I confirm adherence to academic integrity and that AI/LLM tools did not generate falsified, fictional, or plagiarised content. AI/LLM tools were used exclusively to support tasks such as code debugging, text editing, structural guidance, and concept clarification, and not to generate original research findings or introduce unverified information.

Reproducibility self-assessment

This appendix includes a self-assessment of the reproducibility of the research presented in this MSc thesis, addressing data availability, code availability, documentation, and an overall reproducibility rating.

Data Availability

The data utilized and generated in this study consists of both publicly available datasets and restricted data:

- **Input Imagery:** The street-level panoramas used as input for both pipelines are publicly accessible and can be queried and downloaded via the Mapillary API Documentation. A free developer API token can be obtained by registering through the Mapillary Developer Portal. It is worth noting that Mapillary is a continuously growing platform, with new imagery being contributed by users globally. As a result, future researchers applying this pipeline to the same study areas may have access to a denser and more recent set of panoramas, which could yield improved results compared to those reported in this thesis.
- **Boston Ground Truth Dataset:** The sidewalk inventory data used to validate the pipeline in Boston is openly available through the Analyze Boston Sidewalk Inventory.
- **Amsterdam Ground Truth Dataset:** The ground truth dataset used for the Amsterdam case study is not publicly accessible. This dataset was provided by my thesis supervisor for the purpose of this research. Due to municipal licensing constraints, it cannot be shared publicly.

Model Weights

The pipelines rely on two pre-trained deep learning models. Their availability differs:

- **DINOv3** (semantic segmentation): The model code is available through the DINOv3 GitHub Repository. However, the pre-trained model weights are gated and must be requested through Meta's DINOv3 Access Portal. Alternatively, the weights can be

requested via the DINOv3 Hugging Face Model Page. For this thesis, access was granted for research purposes.

- **Depth-Anything-V3** (metric depth estimation): Both the model code and pre-trained weights are freely and publicly available through the Depth Anything V3 GitHub Repository.

Code Availability

The complete codebase and analysis workflow developed for this thesis are openly available on GitHub through the Pixel2Sidewalk Repository. The repository contains both pipelines developed during this research: the Sidewalk Geometry Reconstruction Pipeline and the Sidewalk Inventory Pipeline. Each pipeline is self-contained in its own directory with its own documentation and environment specification.

Documentation

Each pipeline directory includes a detailed `README.md` file explaining the pipeline phases, required API keys, model weight placement, and step-by-step run instructions. An `environment.yml` file is provided so that the exact Conda environment and all dependencies can be recreated. Generalized SLURM batch scripts are also included for replication on high-performance computing clusters.

Reproducibility Rating

Based on the provided scale, I rate the overall reproducibility of this research as **High**.

While the Amsterdam validation dataset is restricted, the Boston ground truth dataset remains entirely open, allowing other researchers to validate the methodology independently. All software developed for this thesis is publicly available and documented. The only potential barrier is obtaining access to the DINOv3 model weights through Meta's gated download process, which requires a brief research access request. A researcher can clone the repository, recreate the Python environment, obtain the necessary API keys and model weights, and execute the end-to-end pipeline to reproduce the core results of this thesis.

Current State of OpenStreetMap Sidewalk Coverage



Figure 1: Existing OpenStreetMap (OSM) sidewalk data for Utrecht. The map illustrates the sparsity of explicitly mapped sidewalk centerlines (`footway=sidewalk`) and confirms the complete absence of 2D polygon representations (area:highway=footway).



Figure 2: Existing OpenStreetMap (OSM) pedestrian network data for Amsterdam. Similar to Utrecht, sidewalk polygon representations are entirely missing. The current dataset is only a centerline and represents only a fraction of the city's actual pedestrian infrastructure.

Colophon

This document was typeset using LaTeX, using a modified the KOMA-Script class `scrbook` available at https://github.com/tudelft3d/msc_geomatics_thesis_template. The main font is Palatino.

