UrbanVGGT: Scalable Sidewalk Width Estimation from Street View Images

Kaizhen Tan [1], Fan Zhang [2]

[1] Heinz College of Information Systems and Public Policy, Carnegie Mellon University, USA – kaizhent@cmu.edu
[2] Institute of Remote Sensing and Geographical Information System, Peking University, China – fanzhanggis@pku.edu.cn

Keywords: Sidewalk width; Street view images; 3D point cloud; Semantic segmentation; Ground-plane fitting; Metric calibration.

1. Introduction

Sidewalk width is a critical micro-level attribute for studying pedestrian mobility and related urban issues. However, reliable large-scale measurements and datasets remain scarce. Early efforts relied on field investigations and manual digitization from aerial imagery (Proulx et al., 2015). These workflows may produce useful datasets, but they require substantial human effort, making them costly, time-consuming, prone to human error, and difficult to scale across large areas. Recent advances in computer vision enable automatic sidewalk extraction from aerial and satellite imagery using semantic segmentation. TILE2NET extends this paradigm by generating polygons across entire municipalities to derive geometric attributes and support large-scale dataset creation (Hosseini et al., 2023). Even so, these approaches depend heavily on high-resolution, orthorectified imagery, which limits their transferability in data-sparse regions. In addition, accuracy often deteriorates in the presence of occlusions from vegetation, canopies, or parked vehicles.

Street-view imagery (SVI) offers a cost-effective alternative to remote sensing imagery, capturing detailed ground-level features that are often missed from above. Its increasing accessibility through online platforms has made it a valuable resource for sidewalk measurement. For instance, Ning et al. (2022) convert panoramic images to measurable land cover maps by leveraging associated depthmap data. Another work (Lieu and Guhathakurta, 2025) estimates sidewalk width from paired street view images captured at two pitch angles by applying trigonometric functions. Still, these analyses remain sensitive to camera field-of-view settings and rely on simplified geometric assumptions, which can introduce small but systematic variations. More recently, vision-language models (VLMs) have shown potential for zero-shot, prompt-based streetscape assessment, though their measurement accuracy remains limited (Perez and Fusco, 2025).

To address these gaps, we introduce UrbanVGGT, a unified neural framework that estimates metrically scaled sidewalk width from street-view imagery. At its core, UrbanVGGT infers world point coordinates and camera parameters from 2D inputs, fits a ground plane from semantically labeled point clouds, and recovers absolute scale to compute widths. Applying this pipeline, we construct and release SV-SideWidth, a city-scale dataset spanning both developed regions and the Global South. Across diverse cities, the framework delivers favorable accuracy with strong cost efficiency and stable generalization, filling data gaps in OpenStreetMap and in data-sparse areas. Overall, our work mitigates inequities in the availability of urban knowledge, providing a scalable foundation for more equitable assessments of street design, mobility, and safety.

2. Methodology

Our framework integrates semantic segmentation, 3D geometry reconstruction, ground-plane fitting, metric calibration, and robust width estimation into an end-to-end pipeline (Fig. 1). First, we use the SegFormer-B5 semantic segmentation model to extract sidewalks and roads, producing pixel-level masks and boundary candidates (Fig. 1b). Next, we leverage the Visual Geometry Grounded Transformer (Wang et al., 2025) to infer world point coordinates and camera parameters from 2D inputs (Fig. 1d). Then, ground points from the road/sidewalk regions are fitted with a RANSAC plane model to obtain the ground surface and calibrate metric scale from known camera height (Fig. 1e). The 2D boundaries are then projected onto the 3D ground plane, paired within a midline overlap region, and filtered to remove outliers (Fig. 1c). The median separation on the ground plane measured along the normal direction between paired boundaries provides the sidewalk width estimate, and dispersion across samples quantifies uncertainty. Finally, each image's sidewalk width estimate, uncertainty, and metadata (location, heading, scale) form the dataset (Fig. 1f).

3. Experiments

We evaluate UrbanVGGT on Washington, DC ground truth (Lieu and Guhathakurta, 2025) and ablate components to quantify their effect on width accuracy. We benchmark four depth backbones (DepthAnything, DepthPro, ZoeDepth, DPT) within a shared segmentation and geometry pipeline. We also compare single-view and multi-view reconstructions across VGGT, MapAnything, DUSt3R, MASt3R, and MonST3R to assess their impact on sidewalk width estimation accuracy. We further plan to create pilot datasets from Google Street View in sample cities, such as New York City, São Paulo, and Nairobi to validate scalability and generalization across regions.

4. Discussion

OpenStreetMap (OSM) provides globally consistent topological data, including building footprints and road networks, but micro-scale pedestrian attributes are often missing. Even in data-rich cities like New York, OSM contains some sidewalk width data, but these tags are inconsistent and scarce in other areas (Fig. 2a). The gap is more pronounced across Africa and other parts of the Global South. For example, in Nairobi, no sidewalk width data can be extracted from OSM (Fig. 2b). Complementing OSM's coverage, Google Street View (GSV) has broad and continually updated imagery worldwide (Fig. 3). UrbanVGGT converts widely available street-view imagery into metrically consistent sidewalk width layers that fill OSM gaps and extend coverage to data-scarce regions. This approach reduces geographic inequities in urban data and enables comparable assessments of street design, walkability, and safety in both well-mapped cities and underrepresented places. It also establishes a practical path for low-cost, incremental updates as new imagery becomes available.

5. Conclusion

UrbanVGGT estimates metric sidewalk width from street-view imagery using segmentation and 3D geometry. Paired with SV-SideWidth, it delivers accurate, transferable measurements that fill OSM gaps and enable scalable city-level analyses of walkability and street design.

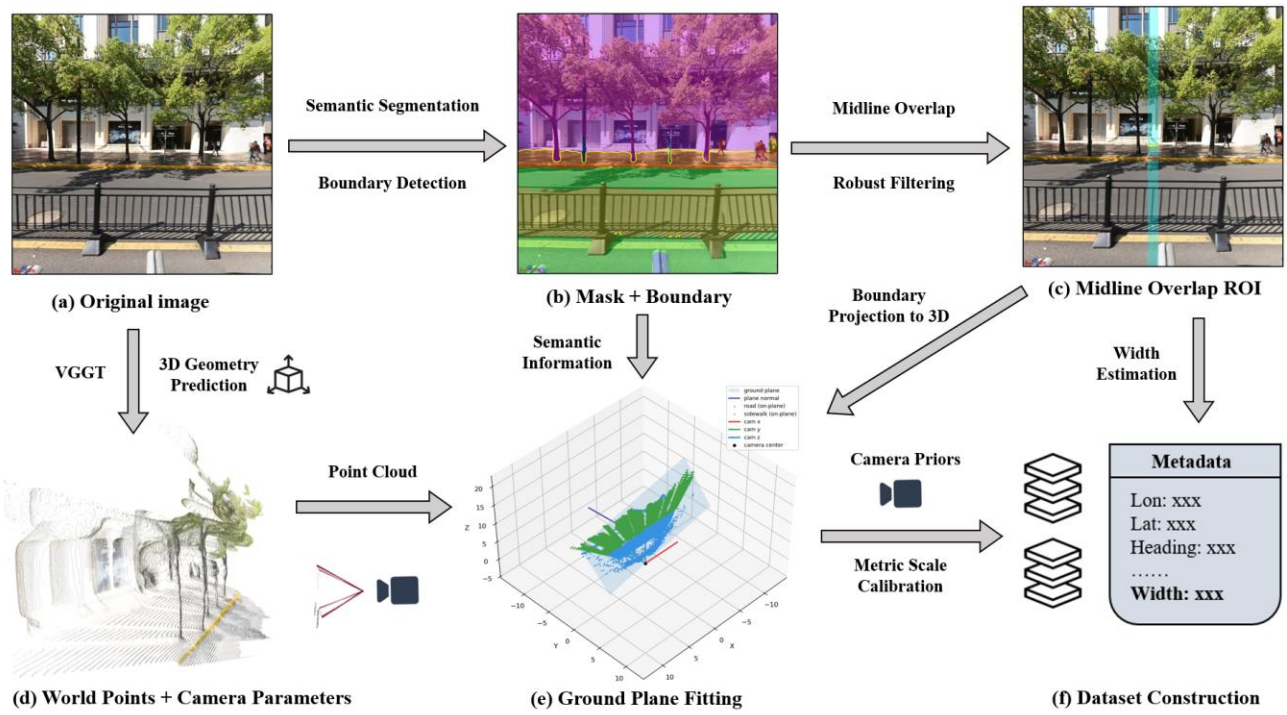

Figure 1. Technical framework. (a) Original image. (b) Segmentation of sidewalk with outer (red) and inner (yellow) boundary detection. (c) Midline overlap region (cyan band) used to pair boundary points for width computation. (d) VGGT-based 3D reconstruction. (e) Ground plane fitting with semantic point cloud (green points indicate sidewalk classes, blue points indicate road classes, and the fitted plane is shown as a translucent sheet). (f) Dataset construction.

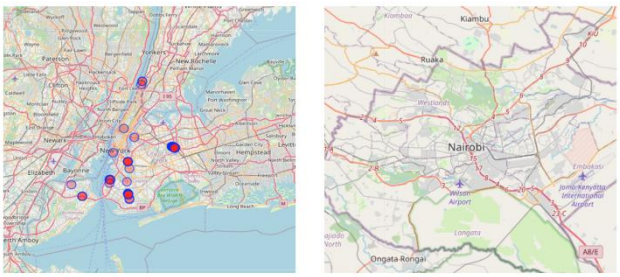

Figure 2. Sidewalk Width Data from OpenStreetMap in New York City (left) and Nairobi (right).

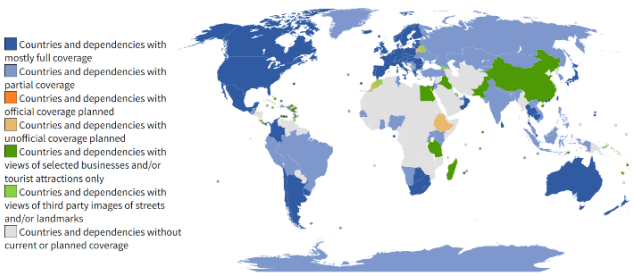

Figure 3. Global Coverage of Google Street View.

References

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

Lieu, S. J., 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*. https://doi.org/10.1177/23998083251369602.

Ning, H., Li, Z., Wang, C., Hodgson, M. E., Huang, X., 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. https://doi.org/10.1016/j.compenvurbsys.2022.101808.

Perez, J., Fusco, G., 2025: Streetscape Analysis with Generative AI (SAGAI): vision–language assessment and mapping of urban scenes. *Geomatica*, 77(2), 100063. https://doi.org/10.1016/j.geomat.2025.100063.

Proulx, F. R., Zhang, Y., Grembek, O., 2015: Database for active transportation infrastructure and volume. *Transportation Research Record*, 2527, 99–106. https://doi.org/10.3141/2527-11.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025: VGGT: visual geometry grounded transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, 5294–5306. https://doi.org/10.1109/CVPR52734.2025.00499.