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.

Abstract

Sidewalk width is an important attribute for pedestrian mobility and accessibility, yet large-scale measurements remain scarce. Existing approaches rely on costly field surveys, require high-resolution aerial imagery that is unavailable in many settings, or use simplified geometric assumptions that introduce systematic error. We present UrbanVGGT, a measurement pipeline for estimating metrically scaled sidewalk width from a single street-view image. The pipeline fits a local support plane to semantically labelled road and sidewalk points in reconstructed 3D geometry, recovers metric scale from the camera-to-plane distance and a known camera height, and measures width as a directional 3D quantity on that plane. The system combines SegFormer-B5 semantic segmentation, the Visual Geometry Grounded Transformer (VGGT) for 3D reconstruction, random sample consensus (RANSAC) ground-plane fitting, and camera-height scale calibration. On a Washington, D.C. ground-truth dataset, UrbanVGGT achieves a mean absolute error (MAE) of 0.25 m (median absolute error 0.22 m, bias -0.055 m) with 95.5% of estimates within 0.50 m. A controlled comparison against 13 alternative depth and reconstruction backbones using a shared downstream measurement protocol, together with ablation experiments, supports the design choices. As a feasibility demonstration, we apply the pipeline to construct SV-SideWidth, a preliminary layer spanning New York City, São Paulo, and Nairobi. The resulting layer adds sidewalk-width estimates to 527 OpenStreetMap segments; we present it as an unvalidated prototype to illustrate application potential rather than as a complete inventory.

1. INTRODUCTION

Sidewalk width is a fundamental micro-level attribute that directly affects pedestrian comfort, safety, and accessibility. Adequate sidewalk dimensions are essential for accommodating diverse users, including wheelchair users, caregivers with strollers, and visually impaired pedestrians, and are reflected in accessibility design standards; for example, the ADA Standards for Accessible Design specify minimum clear widths for accessible routes (U.S. Department of Justice, 2010). Despite its importance, sidewalk width data remain scarce in most cities worldwide, hindering evidence-based urban planning and equitable resource allocation.

Early efforts to build sidewalk and pedestrian-infrastructure inventories relied on field surveys and manual interpretation of imagery (Proulx et al., 2015). While these workflows can produce useful datasets, they require substantial human effort, making them costly, time-consuming, and difficult to scale across large areas. Recent advances in computer vision have enabled automatic sidewalk extraction from aerial and satellite imagery using semantic segmentation. For example, TILE2NET (Hosseini et al., 2023) generates sidewalk polygons across entire municipalities to derive geometric attributes and support large-scale dataset creation. Even so, these approaches depend heavily on high-resolution, orthorectified imagery, which limits their transferability to data-sparse regions. In addition, accuracy often deteriorates in the presence of occlusions from vegetation, canopies, or parked vehicles.

Street-view imagery provides ground-level detail that is often missed from overhead imagery (Biljecki and Ito, 2021). Platforms such as Google Street View (Anguelov et al., 2010) make large collections of street-level imagery available for analysis. Ning et al. (Ning et al., 2022) convert panoramic images to measurable land-cover maps by leveraging associated depth-map data. Lieu and Guhathakurta (Lieu and Guhathakurta, 2025) estimate 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 that can introduce systematic errors. More recently, vision and language models have shown potential for zero-shot, prompt-based streetscape assessment, though their measurement accuracy remains limited (Perez and Fusco, 2025).

Meanwhile, OpenStreetMap (Haklay and Weber, 2008) provides open, user-generated street-network data, yet sidewalk width tags are largely absent even in well-mapped cities. As Figure 1 illustrates, none of the 461 drivable street segments in Midtown Manhattan nor any of the 1 958 segments in Nairobi's central business district carry sidewalk-width attributes in OpenStreetMap. Street-view imagery therefore offers a complementary source that could help fill this gap.

To address these three challenges, namely the poor scalability of manual surveys, the data dependency of overhead methods, and the geometric simplifications of existing street-view approaches, we develop UrbanVGGT, a single-image measurement pipeline for metrically scaled sidewalk-width estimation from street-view imagery. The pipeline uses semantic road and sidewalk labels to recover a locally consistent support plane from arbitrary-scale 3D predictions, calibrates that geometry with camera height, and measures width as a directional 3D quantity on the plane. This paper makes three contributions:

1. We present this plane-based measurement formulation, which differs from prior work by performing width estimation in recovered 3D geometry rather than in image coordinates, and validate it on a Washington, D.C. benchmark.

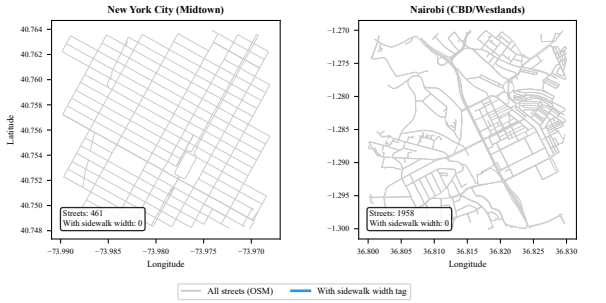

Figure 1. Sidewalk width data in OpenStreetMap. Grey lines denote all drivable streets; blue lines (if any) denote streets with a sidewalk-width tag. Both New York City (461 streets) and Nairobi (1 958 streets) have zero sidewalk-width tags, highlighting the data gap that UrbanVGGT aims to fill.

2. We conduct a controlled evaluation of 14 depth and reconstruction backbones, all sharing the same segmentation, plane-fitting, calibration, and outlier-filtering pipeline, so that performance differences isolate the effect of the 3D geometry backbone.

3. We apply the pipeline to three neighbourhood-scale study areas (New York City, São Paulo, Nairobi) to construct SV-SideWidth, a preliminary, unvalidated feasibility prototype that illustrates the potential for automated sidewalk-width generation from street-view imagery.

2. RELATED WORK

2.1 Sidewalk Measurement from Remote Sensing

Overhead imagery has been widely used to map sidewalk and pedestrian infrastructure. Proulx et al. (Proulx et al., 2015) developed an active transportation database through field and imagery-based collection of pedestrian-infrastructure attributes. Hosseini et al. (Hosseini et al., 2023) proposed TILE2NET, a scalable semantic segmentation approach that generates sidewalk network datasets from aerial imagery across entire municipalities. These datasets can in turn support the derivation of geometric attributes such as sidewalk width. While effective in well-mapped cities with high-resolution ortho-imagery, these methods face two inherent limitations: trees and building overhangs occlude sidewalks from a nadir perspective, and suitable high-resolution imagery is often unavailable in lower-data settings.

2.2 Sidewalk Measurement from Street-View Imagery

Street-view imagery provides an oblique, ground-level perspective that complements overhead views. Ning et al. (Ning et al., 2022) leverage depth maps associated with Google Street View panoramas to convert images into land-cover maps with metric coordinates, enabling sidewalk width extraction. Lieu and Guhathakurta (Lieu and Guhathakurta, 2025) estimate sidewalk width from a pair of street-view images captured at different pitch angles, applying trigonometric functions under a flat-ground assumption. However, this approach requires two images per measurement and is sensitive to camera field-of-view settings. Perez and Fusco (Perez and Fusco, 2025) explore vision and language models for streetscape assessment but report limited metric accuracy. The broader idea of recovering metric quantities from a single image dates back to Criminisi et al. (Criminisi et al., 2000), who showed that vanishing points and a known reference length suffice for single-view metrology. Our work builds on this line of research by replacing simplified geometric assumptions with learned 3D reconstruction and camera-height-based metric calibration, achieving higher accuracy from a single image while avoiding the need for manual vanishing-point annotation.

2.3 Monocular Depth and 3D Reconstruction

Monocular depth estimation has advanced rapidly. Metric-depth models such as ZoeDepth (Bhat et al., 2023), DepthPro (Bochkovskii et al., 2025), UniDepthV2 (Piccinelli et al., 2025), and Metric3D V2 (Hu et al., 2024) predict absolute depth from single images. Relative-depth models like Depth Anything V2 (Yang et al., 2024) and DPT (Ranftl et al., 2021) predict scale-ambiguous depth requiring external calibration. Feed-forward 3D geometry methods, including DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), π^3 (Wang et al., 2025c), and VGGT (Wang et al., 2025a), infer dense geometric representations from one or more images in a single forward pass. DUSt3R and MASt3R are most commonly used in pairwise or multi-view settings; accordingly, we exclude them from our single-image benchmark and focus that comparison on methods with a direct single-image evaluation path. In this work, we benchmark these model families for sidewalk-width measurement and adopt VGGT as our primary backbone because it supports single-image inference and performed most reliably in our preliminary experiments.

3. METHODOLOGY

Our framework integrates five stages into a unified pipeline: semantic segmentation, 3D geometry reconstruction, ground-plane fitting, metric calibration, and robust width estimation (Figure 2).

Rather than treating sidewalk width as a 2D pixel distance or approximating it with image-plane trigonometry, the pipeline measures width as a directional 3D quantity on a semantically recovered ground plane, requiring only one external scalar parameter: the camera mounting height. The following subsections detail each stage.

3.1 Semantic Segmentation

We employ SegFormer-B5 (Xie et al., 2021), fine-tuned on the Cityscapes dataset (Cordts et al., 2016), to produce pixel-level semantic labels. The model classifies each pixel into one of 19 urban categories; we retain the *sidewalk* and *road* classes for downstream processing. The segmentation map is used to (a) identify sidewalk boundary candidates via column-wise scanning and (b) gather ground-level 3D points for plane fitting. Small connected components smaller than 50 pixels or 0.1% of image area, whichever is larger, are removed, and a 3×3 morphological closing operation fills minor holes in the sidewalk mask. Before geometric measurement, we further filter out images with too little sidewalk or road support by requiring sidewalk pixels to cover at least 2% of the image, road pixels at least 1%, and sidewalk presence in at least 30% of columns inside the measurement band.

3.2 3D Geometry Reconstruction via VGGT

The Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025a) is a feed-forward neural network that, given

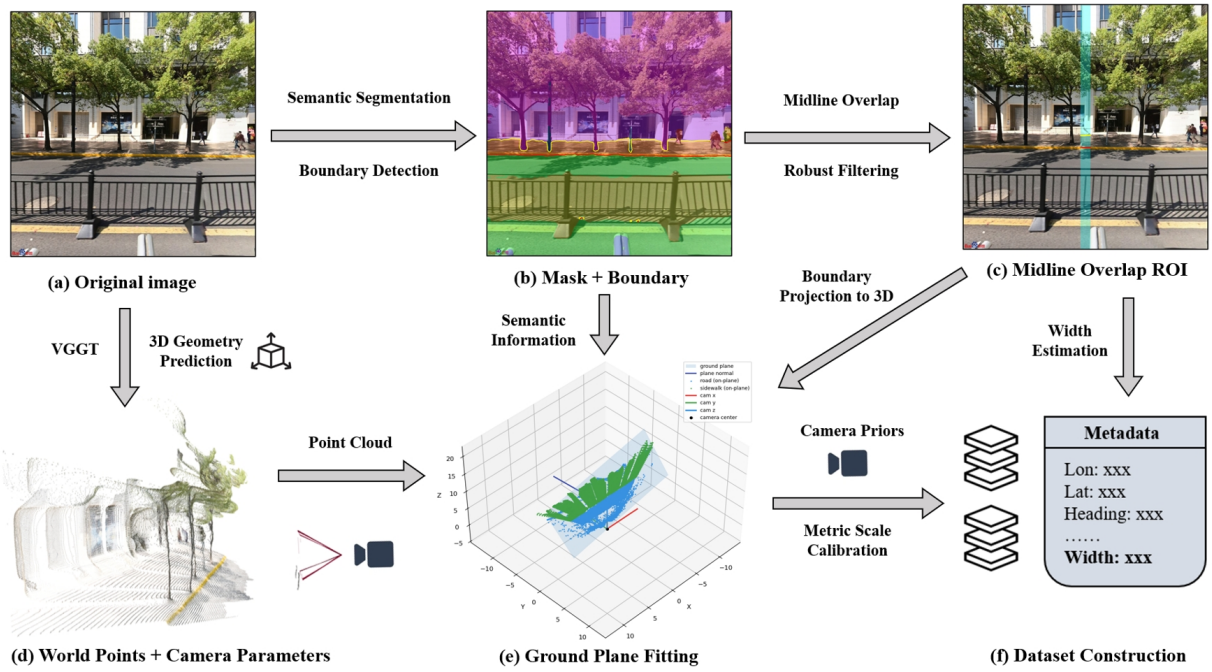

Figure 2. UrbanVGGT pipeline overview. (a) Input street-view image. (b) Semantic segmentation with inner (yellow) and outer (red) boundary detection. (c) Midline overlap region used to pair boundary points. (d) VGGT-based 3D reconstruction. (e) Ground-plane fitting with semantic point cloud. (f) Width estimation and preliminary layer construction.

one or more images, directly predicts per-pixel world-point coordinates and camera pose parameters in a single forward pass without iterative optimisation. In our implementation, each input image yields a dense world-point map of shape $(H_d, W_d, 3)$ and a 9-dimensional pose encoding comprising the camera centre (c_x, c_y, c_z), orientation quaternion (q_w, q_x, q_y, q_z), and field-of-view angles (fov_h, fov_w). Because VGGT produces geometry in an arbitrary coordinate system with unknown metric scale, an explicit scale calibration step is required.

3.3 Ground-Plane Fitting

We collect 3D points labelled as either *sidewalk* or *road* by the segmentation map and fit a ground plane $\pi : \mathbf{n} \cdot \mathbf{x} + d = 0$ using random sample consensus (RANSAC) (Fischler and Bolles, 1981), where $\mathbf{n}$ is a unit normal ($\|\mathbf{n}\| = 1$). To adapt to varying point-cloud noise levels, we first compute a coarse singular value decomposition based plane fit, then estimate the median absolute deviation of point-to-plane distances and set the inlier threshold adaptively as:

$$\tau = \mathrm{clip}(2.5 \times 1.4826 \times \mathrm{MAD},\ 0.005,\ 0.05)\,. \qquad (1)$$

After 1 000 RANSAC iterations, inlier points are refitted with singular value decomposition to obtain the final unit-normal plane parameters $(\mathbf{n}, d)$. This median-absolute-deviation adaptive strategy avoids the need for a manually tuned threshold and accommodates both smooth asphalt and rough unpaved surfaces.

3.4 Metric Scale Calibration

Because the VGGT output is expressed in an arbitrary coordinate frame, we recover metric scale from the known camera mounting height h_{cam}. The predicted camera centre $\mathbf{c}$ is extracted from the pose encoding, and the signed distance from $\mathbf{c}$ to the fitted ground plane gives $h_{\mathrm{pred}} = |\mathbf{n} \cdot \mathbf{c} + d|$. The scale factor is then:

$$s = h_{\mathrm{cam}} \,/\, h_{\mathrm{pred}}. \qquad (2)$$

All subsequent 3D measurements are multiplied by s to convert from VGGT units to metres. Because s is linear in h_{cam}, a relative error δ in the assumed camera height propagates directly to a relative error δ in the estimated width. In our experiments, $h_{\mathrm{cam}} = 2.5\,\mathrm{m}$ corresponds to the Google Street View camera mounting height reported by Anguelov et al. (Anguelov et al., 2010) for the standard acquisition vehicle; Section 4.3 presents a sensitivity analysis confirming that this value minimises error on the benchmark dataset. We note that the actual mounting height may vary by ± 0.1–$0.2\,\mathrm{m}$ across vehicle generations and countries, which would introduce a corresponding 4–8% systematic width bias.

3.5 Column-Wise Width Estimation

Rather than extracting morphological boundaries, which suffer from inner/outer confusion and stickiness artefacts, we scan the sidewalk mask column by column within the central 20% band of the image. For each column, we identify contiguous sidewalk pixel runs, discard runs shorter than 15 pixels, and select the most plausible run based on length and proximity to the global median y-centre. The top (inner/building-side) and bottom (outer/road-side) pixel coordinates of the selected run define the sidewalk edges. Columns whose edges touch the image border or deviate from the median by more than three times the median absolute deviation are rejected.

An over-segmentation detector compares the observed pixel gap with the geometrically expected gap at the estimated distance; when the ratio exceeds 1.3, gradient-based refinement searches for the true inner edge within the plausible region. The surviving edge coordinates are mapped to the VGGT world-point map resolution (which may differ from the input im-

age resolution), and the corresponding 3D points are projected onto the fitted ground plane. The across-sidewalk direction is determined via principal component analysis of the projected boundary points. Per-column widths are computed as the projection of top-to-bottom 3D vectors onto this direction, scaled by s.

A final median-absolute-deviation based outlier filter retains inlier widths, whose median yields the width estimate and whose standard deviation quantifies uncertainty. To suppress unstable cases, we require at least 10 valid columns, reject estimates outside $[0.50,\ 4.00]$ m, and discard results whose per-column coefficient of variation exceeds 0.40. A geometric cross-validation step compares the 3D-based width with a pinhole-geometry estimate and rejects images where their ratio falls outside $[0.33,\ 3.0]$, guarding against catastrophic reconstruction failures.

4. EXPERIMENTS

Figure 3 shows representative measurement examples from the Washington, D.C. dataset, illustrating that the pipeline accurately localises sidewalk boundaries and recovers metric widths close to the ground truth.

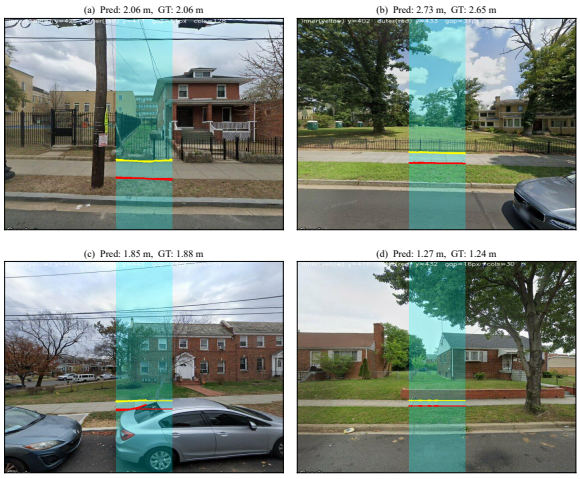

Figure 3. Qualitative measurement examples on the Washington, D.C. dataset. Each panel shows the segmentation overlay with detected inner (yellow) and outer (red) boundaries. Predicted width: model estimate; ground-truth width: reference measurement.

4.1 Study Area and Dataset

We evaluate UrbanVGGT on a ground-truth sidewalk-width dataset from Washington, D.C. published by Lieu and Guhathakurta (Lieu and Guhathakurta, 2025). From this dataset we sampled approximately 300 Google Street View images with associated reference width measurements. All images are 640×640 pixels at a 90° field of view. Ground-truth widths range from 0.56 m to 3.94 m.

4.2 Ablation Study

To quantify the contribution of each pipeline component, we conduct ablation experiments by selectively disabling key modules. Table 2 reports the results. The full pipeline achieves a MAE of 0.252 m, a median absolute error (MdAE) of 0.223 m, a RMSE of 0.293 m, and a bias of -0.055 m, with 95.5% of estimates within 0.50 m. The near-zero bias indicates that the pipeline does not systematically over- or under-estimate width. Removing scale calibration (i.e., setting $s = 1$) causes catastrophic failure with a MAE of 1.571 m, confirming that VGGT's raw coordinates are not metrically scaled and calibration is essential. Replacing 3D reconstruction with pinhole geometry alone increases the MAE to 1.096 m. Using the full image width instead of the central 20% band slightly increases the MAE to 0.265 m and reduces the within-0.50 m rate from 95.5% to 88.5%, indicating that peripheral columns introduce noise from perspective distortion.

Table 1. Ablation study on the Washington, D.C. dataset. Best values are shown in bold.

Variant	MAE (m)	RMSE (m)	<0.25 m (%)	<0.50 m (%)
Full pipeline	**0.252**	**0.293**	49.9	**95.5**
− Scale calibration	1.571	1.599	0.0	0.0
Pinhole only	1.096	1.203	4.5	11.0
Full image width	0.265	0.351	**54.2**	88.5

4.3 Camera Height Sensitivity

The camera mounting height h_{cam} is the sole external parameter. We fix $h_{\text{cam}} = 2.5$ m for Google Street View imagery throughout the study, following the acquisition-vehicle specification described by Anguelov et al. (Anguelov et al., 2010); this value is a fixed prior and is not tuned on the evaluation set. To characterise sensitivity, we sweep h_{cam} from 2.0 m to 3.0 m in 0.25 m increments (Table 3). At the adopted value of 2.5 m, the pipeline achieves the lowest MAE (0.252 m) and the highest within-0.50 m rate (95.5%), supporting the stability of this fixed assumption on the benchmark dataset. Performance degrades approximately symmetrically as h_{cam} departs from the adopted value: at 2.0 m (a -20% perturbation), the MAE rises to 0.384 m with strong negative bias (-0.380 m), while at 3.0 m ($+20$%), the MAE rises to 0.349 m with positive bias ($+0.263$ m). This behaviour is consistent with the linear scale-propagation property noted in Section 3: a ± 10% camera-height error produces a ± 10% width bias, corresponding to roughly ± 0.25 m for a typical 2.5 m-wide sidewalk. Figure 4 visualises this sensitivity curve. For non-Google street-view providers, h_{cam} should be set according to the respective platform specification; where the mounting height is uncertain, the sensitivity curve provides a basis for estimating the resulting width uncertainty.

Table 2. Camera height sensitivity analysis.

Camera height (m)	MAE (m)	Bias (m)	<0.25 m (%)	<0.50 m (%)
2.00	0.384	−0.380	29.7	65.5
2.25	0.283	−0.217	50.6	82.4
2.50	**0.252**	**−0.055**	49.9	**95.5**
2.75	0.276	+0.104	**50.0**	85.0
3.00	0.349	+0.263	46.5	69.1

4.4 Benchmark Against Alternative Backbones

We benchmark UrbanVGGT against 13 alternative depth and reconstruction models, organised into three categories that reflect distinct levels of geometric reasoning. To ensure a fair comparison, all methods share exactly the same downstream

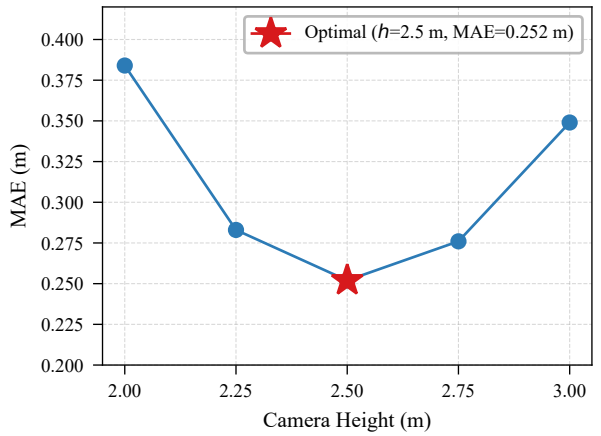

Figure 4. Camera height sensitivity: MAE as a function of assumed camera mounting height h_{cam}.

pipeline: SegFormer-B5 segmentation, column-wise boundary extraction within the central 20% band, RANSAC ground-plane fitting with the same adaptive threshold (Equation 1), identical outlier rejection criteria (minimum 10 valid columns, coefficient of variation < 0.40, width in $[0.50, 4.00]$ m, 3D-vs-pinhole ratio in $[0.33, 3.0]$), and median aggregation. The only difference across methods is the 3D geometry source. All benchmark images are fed to every method; the number of valid measurements varies because each backbone's reconstruction quality determines whether the downstream pipeline produces a result that passes quality control. Models that support multiple output modalities, such as UniDepthV2 and MapAnything, appear in more than one category, and each is evaluated through the corresponding measurement protocol (native scale, pinhole unprojection, or direct point cloud) so that the effect of the geometric representation can be separated from that of the backbone. Figure 5 provides a visual comparison of MAE across all methods.

4.4.1 Category 1: Metric Depth with Native Scale. These models produce metrically scaled depth maps and optionally 3D point clouds. We apply our column-wise width estimation directly on the native metric geometry without camera-height calibration. Table 4 reports the results. UniDepthV2 (Piccinelli et al., 2025) achieves the lowest MAE (0.289 m) in this category with a valid-estimate rate of 97.6%, closely matching UrbanVGGT. MapAnything (Keetha et al., 2026) also performs well (MAE = 0.463 m, 90.2% valid), while Depth Anything V2-metric (Yang et al., 2024) produces valid measurements for only 22.3% of images. The wide variance in both accuracy and yield highlights that metric-depth quality alone does not guarantee reliable width estimation.

Table 3. Category 1: metric depth models evaluated with native scale geometry. No camera-height calibration is applied.

Model	MAE (m)	<0.25 m (%)	<0.50 m (%)
DepthAnything V2-m	1.712	0.0	2.7
ZoeDepth	1.067	10.1	20.6
DepthPro	0.768	10.2	23.6
Metric3D V2	1.361	5.5	11.5
UniDepthV2	**0.289**	**54.7**	**83.3**
MapAnything	0.463	17.8	60.5

4.4.2 Category 2: Monocular Depth with Pinhole Unprojection. These models estimate per-pixel depth (metric or relative). We unproject depth maps into 3D point clouds using known or estimated camera intrinsics, fit a ground plane, and apply camera-height calibration using the same pipeline as UrbanVGGT but with monocular depth instead of VGGT 3D reconstruction. Table 5 reports the results. Metric3D V2 leads this category with a MAE of 0.417 m, followed by UniDepthV2 (0.436 m). Scale calibration substantially improves Depth Anything V2-relative (from 1.712 m with its metric variant in Category 1 to 0.628 m), confirming the value of our calibration pipeline even with scale-ambiguous depth models.

Table 4. Category 2: monocular depth models evaluated with pinhole unprojection and camera-height scale calibration.

Model	MAE (m)	<0.25 m (%)	<0.50 m (%)
DepthAnything V2-r	0.628	18.0	38.0
DepthPro	0.718	12.2	31.7
DPT	0.701	15.2	33.7
Metric3D V2	**0.417**	**39.9**	**67.8**
UniDepthV2	0.436	23.9	66.7
ZoeDepth	0.583	21.0	44.0

4.4.3 Category 3: Single-Image Point-Cloud Reconstruction. These models use their built-in point-cloud reconstruction capabilities to produce dense 3D geometry from a single image. We collect road and sidewalk points from the reconstructed point cloud, fit a ground plane, and apply camera-height calibration. Table 6 reports the results. UrbanVGGT achieves the lowest MAE of 0.252 m among all 14 methods across the three categories, with a 100% valid-estimate rate. π^3 (Wang et al., 2025c) (MAE = 0.324 m, 97.6% valid) and MapAnything (Keetha et al., 2026) (MAE = 0.334 m, 90.2% valid) are competitive alternatives. CUT3R (Wang et al., 2025b) shows higher error (0.672 m, 81.9% valid). UrbanVGGT also achieves the highest within-0.50 m rate (95.5%), compared with 79.6% for π^3. The combination of lowest error and highest yield distinguishes UrbanVGGT from the alternatives.

Table 5. Category 3: single-image point-cloud reconstruction models evaluated with camera-height scale calibration.

Model	MAE (m)	<0.25 m (%)	<0.50 m (%)
CUT3R	0.672	24.3	47.1
MapAnything	0.334	44.7	78.0
π^3 (Wang et al., 2025c)	0.324	49.2	79.6
UrbanVGGT (ours)	**0.252**	49.9	**95.5**

Statistical note. Because all 14 methods are evaluated on the same benchmark images with a shared downstream pipeline, differences in MAE reflect the 3D backbone rather than engineering variation. The gap between UrbanVGGT (0.252 m) and the next-best method, UniDepthV2 in Category 1 (0.289 m), is 0.037 m. We note that this margin, while consistent, is modest; UniDepthV2 further offers native metric scale without camera-height calibration, making it a practical alternative when camera height is unknown. The valid-estimate rate also differs meaningfully: UrbanVGGT produces measurements for all benchmark images, whereas Depth Anything V2-metric covers only 22.3%, illustrating that backbone reliability affects not just accuracy but coverage.

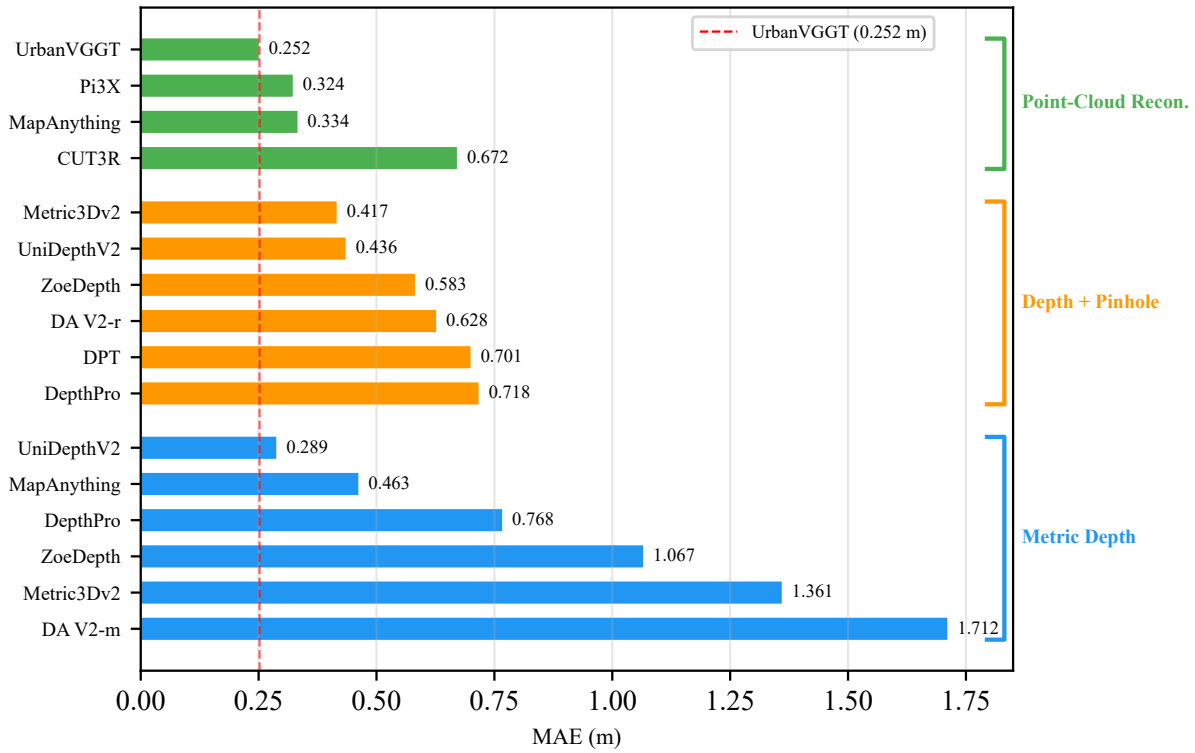

Figure 5. MAE comparison across all methods. Models are grouped by evaluation category: Category 1 (metric depth with native scale), Category 2 (monocular depth with pinhole unprojection and scale calibration), and Category 3 (single-image point-cloud reconstruction with scale calibration). All methods share the same segmentation, boundary extraction, plane fitting, and outlier filtering; only the 3D geometry backbone differs. The dashed red line indicates the UrbanVGGT MAE (0.252 m).

5. SV-SIDEWIDTH FEASIBILITY PROTOTYPE

To examine whether the method can help fill the data gaps identified in Figure 1, we construct SV-SideWidth, an automatically generated preliminary layer spanning three cities with different urban forms and levels of street-view availability. This deployment is a feasibility demonstration, not a validated inventory: we do not have ground-truth sidewalk widths in these cities and therefore cannot report quantitative accuracy. The goal is to test whether street-view imagery can provide candidate widths at neighbourhood scale and to identify practical bottlenecks.

5.1 Study Areas

We select three neighbourhood-scale study areas (Table 7): Midtown Manhattan, New York City (dense grid streets, abundant Google Street View coverage); Jardins/Paulista, São Paulo (high-quality sidewalks, dense Google Street View coverage); and the central business district and Westlands in Nairobi (sparser Google Street View coverage). These areas differ in urban morphology, sidewalk infrastructure quality, and street-view data availability.

Table 6. SV-SideWidth study areas.

City	Neighbourhood	Characteristics
New York City	Midtown	Dense grid, rich Google Street View coverage
São Paulo	Jardins/Paulista	Wide sidewalks
Nairobi	Central business district/Westlands	Sparse Google Street View coverage

5.2 Sampling and Measurement Pipeline

For each city, we query the OpenStreetMap (Haklay and Weber, 2008) street network via OSMnx (Boeing, 2017) and sample points at approximately 30 m intervals along each road segment. At each point, we compute the local road bearing and generate two perpendicular camera headings ($\pm 90^\circ$), yielding views of both sides of the street. A 20 m spatial grid deduplicates nearby points. Google Street View images are downloaded at 640×640 pixels with a 90° field of view. The UrbanVGGT pipeline processes each image, and successful measurements are aggregated to the OpenStreetMap way-segment level by taking the median width across all measurements assigned to the same segment.

5.3 Results and OpenStreetMap Coverage

Table 8 summarises the preliminary layer. Across the three cities, we generate 1 931 valid measurements covering 527 unique OpenStreetMap road segments. This output is partial: coverage ranges from 7.6% in Nairobi to 38.2% in New York City, reflecting both uneven Google Street View availability and the selectivity of the quality-control filters. In all study areas, OpenStreetMap contains zero sidewalk-width tags, so the measured segments add new candidate width attributes rather than replace an existing open inventory. Median widths are 2.58 m (New York City), 2.64 m (São Paulo), and 2.28 m (Nairobi), consistent with known differences in sidewalk infrastructure across these cities. Figure 6 shows the spatial distribution of measurements overlaid on the OpenStreetMap street network. Segment-level verification is still required before these outputs can be treated as authoritative citywide data.

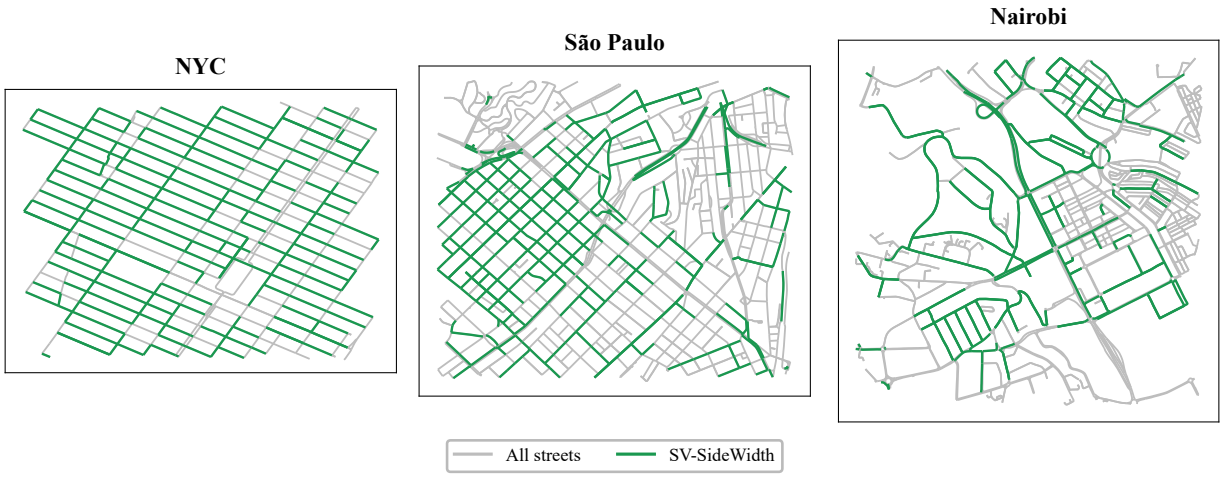

Figure 6. SV-SideWidth prototype layer coverage maps for New York City (left), São Paulo (centre), and Nairobi (right). Grey lines: OpenStreetMap street network; green lines: streets with SV-SideWidth measurements.

Table 7. SV-SideWidth prototype layer summary and OpenStreetMap coverage.

City	Valid measurements	Street segments covered	Total street segments	Coverage rate (%)	Median width (m)
New York City	502	176	461	38.2	2.58
São Paulo	866	203	1539	13.2	2.64
Nairobi	563	148	1958	7.6	2.28
Total	1931	527	3958	N/A	N/A

6. DISCUSSION

6.1 Complementing OpenStreetMap

OpenStreetMap provides globally consistent topological data, including building footprints and road networks, but micro-scale pedestrian attributes such as sidewalk width are largely absent. Even in our New York study area, sidewalk-width tags are sparse and inconsistent, and in Nairobi none are present at all (Table 8). As a result, many cities still lack openly available width data for pedestrian analysis.

Google Street View, by contrast, provides broad ground-level image coverage that can be processed in a consistent manner. UrbanVGGT converts those images into sidewalk-width estimates that can be attached to OpenStreetMap street segments without field surveys or high-resolution aerial imagery. Where no authoritative width inventory exists, even a partial layer can support screening, prioritisation, and comparative analysis. We stress, however, that SV-SideWidth is an unvalidated feasibility prototype: coverage is incomplete (7.6–38.2% of segments), no ground-truth verification has been performed in the three deployment cities, and the pipeline has been quantitatively validated only on the Washington, D.C. benchmark. Its current value is as a low-cost way to generate candidate attributes for downstream manual review, not as a substitute for authoritative survey data.

6.2 Strengths and Limitations

The primary strength of UrbanVGGT is that it estimates metric sidewalk width from a single street-view image without requiring stereo pairs, provider-supplied depth maps, or LiDAR data. The contribution is a measurement systems pipeline that assembles existing components (semantic segmentation, feed-forward 3D reconstruction, robust plane fitting) into a coherent workflow for a specific geospatial task, rather than introducing a new model architecture. In practice, the method is aimed at generating candidate widths from widely available imagery, not replacing survey-grade measurements on every segment. The Washington, D.C. benchmark results support this use case, but several limitations bound the current conclusions:

- **Single evaluation domain.** All quantitative ground-truth results come from a single Washington, D.C. dataset. Additional annotated datasets from cities with different street geometries, camera platforms, and sidewalk materials are needed to assess cross-domain generalisation.
- **Unvalidated prototype layer.** The SV-SideWidth output has not been audited against ground-truth data in New York City, São Paulo, or Nairobi, and should not be used for planning or policy decisions without such verification.
- **Coverage dependency.** Google Street View coverage is uneven, which reduces measurement yield in areas such as our Nairobi study area (7.6% segment coverage).
- **Fixed camera height.** The method assumes a single camera height for all images from a given provider. Camera-height variation across vehicle generations introduces systematic bias (Section 4.3).

6.3 Failure Modes

Failure cases arise when the visual evidence is insufficient or when the local scene violates the geometric assumptions of the measurement formulation. We identify four principal failure categories and note their expected effect:

- **Non-planar geometry.** Sloped streets, driveway ramps, curb cuts, and other non-planar transitions can bias the fitted support plane. The impact is a systematic width error proportional to the slope angle; for a 5° cross-slope, the geometric bias is approximately $\cos(5^\circ) \approx 0.4\%$, which is negligible, but steeper transitions (e.g., ramps at 8–10°) can produce larger deviations.
- **Occlusion.** Parked vehicles, pedestrians, utility poles, street furniture, or dense tree canopies can remove either the inner or outer sidewalk edge, causing boundary drift

or measurement rejection. In the Washington, D.C. benchmark, the pipeline's quality-control filters convert most occluded scenes into abstentions.

- **Narrow sidewalks.** When ground-truth width is below approximately 1.0 m, only a small number of usable columns remain inside the central measurement band, so small segmentation errors produce large relative width errors.
- **Complex curb geometry.** Bulb-outs, parking bays, slip lanes, bus stops, shared-space streets, or intersections where the road-side edge is not approximately parallel to the building-side edge make the across-sidewalk direction ambiguous.

These failure modes are consistent with the safeguards built into the pipeline. On the Washington, D.C. benchmark, only 4.5% of estimates exceed 0.50 m error, and the pipeline's quality-control filters convert many difficult scenes into abstentions rather than grossly wrong estimates: the full pipeline produces valid measurements for all benchmark images, but the pinhole-only variant yields only 91.7%, illustrating how 3D reconstruction quality affects measurement yield. The coefficient-of-variation filter and 3D-versus-pinhole consistency check are the most active rejection mechanisms, accounting for the majority of discarded images in the ablation variants. This abstention-over-error behaviour is desirable for large-scale automatic generation, but it also reinforces why SV-SideWidth should be understood as a feasibility prototype whose outputs require downstream verification, particularly in streetscape conditions not represented in the Washington, D.C. benchmark.

6.4 Comparison with Prior Work

Lieu and Guhathakurta (Lieu and Guhathakurta, 2025) estimate sidewalk width from paired street-view images captured at two pitch angles using trigonometric functions. Their method requires two images per measurement and precise camera field-of-view metadata. In contrast, UrbanVGGT operates on a single image and requires only camera mounting height while also producing additional outputs including a 3D point cloud, ground-plane model, and per-measurement uncertainty estimate. This simpler acquisition requirement makes batch processing easier when the objective is to generate width data for an entire neighbourhood or city.

6.5 Future Directions

Future work should prioritise systematic quality auditing of automatically generated widths, including uncertainty calibration and targeted manual review of difficult segments. Multi-view fusion could aggregate measurements from overlapping images along the same street segment, reducing per-measurement variance and increasing coverage. Temporal analysis could track sidewalk-width changes using historical street-view imagery. Integration with lightweight depth backbones could enable deployment on resource-constrained devices for real-time field surveys. Extending the segmentation model to distinguish sidewalk subtypes (raised vs. flush, paved vs. unpaved) would enrich the generated layer with qualitative attributes beyond width.

7. CONCLUSION

We presented UrbanVGGT, a measurement pipeline for estimating metrically scaled sidewalk width from single street-view images. The pipeline assembles semantic segmentation, feed-forward 3D reconstruction, robust plane fitting, and camera-height calibration into a coherent workflow that measures width as a directional 3D quantity on a recovered ground plane. On the Washington, D.C. benchmark, the method achieves a MAE of 0.252 m with 95.5% of estimates within 0.50 m and attains the lowest MAE among the 14 evaluated methods under a controlled protocol. Ablation studies show that scale calibration is the most critical component. As a feasibility demonstration, the method produces SV-SideWidth, an unvalidated prototype layer covering 527 road segments across three cities. These results suggest that street-view imagery can support automated generation of candidate sidewalk-width attributes, while underscoring the need for ground-truth validation in deployment cities and broader cross-domain evaluation before such outputs can inform planning decisions.

ACKNOWLEDGEMENTS

The Google Street View API costs for this study were self-funded by the first author. We thank OpenStreetMap contributors for maintaining the open street-network data used in this study.

References

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google Street View: Capturing the world at street level. *IEEE Computer*, 43(6), 32–38.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., Müller, M., 2023. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, 215, 104217.

Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S. R., Koltun, V., 2025. Depth Pro: Sharp monocular metric depth in less than a second. *International Conference on Learning Representations (ICLR)*.

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Criminisi, A., Reid, I., Zisserman, A., 2000. Single View Metrology. *International Journal of Computer Vision*, 40(2), 123–148.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Haklay, M., Weber, P., 2008. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4), 12–18.

Hosseini, M., Sevtsuk, A., Miranda, F., Cesar Jr, R. M., 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.

Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S., 2024. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10579–10596.

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Rota Bulò, S., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P., 2026. MapAnything: Universal feed-forward metric 3D reconstruction. *International Conference on 3D Vision (3DV)*.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3D with MASt3R. *European Conference on Computer Vision (ECCV)*, 71–91.

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*.

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.

Perez, J., Fusco, G., 2025. Streetscape Analysis with Generative AI (SAGAI): Vision–language assessment and mapping of urban scenes. *Geomatica*, 77(2), 100063.

Piccinelli, L., Sakaridis, C., Yang, Y.-H., Segu, M., Li, S., Abbeloos, W., Van Gool, L., 2025. UniDepthV2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*.

Proulx, F. R., Zhang, Y., Grembek, O., 2015. Database for active transportation infrastructure and volume. *Transportation Research Record*, 2527, 99–106.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12179–12188.

U.S. Department of Justice, 2010. ADA standards for accessible design. `https://www.ada.gov/law-and-regs/design-standards/2010-stds/`.

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

Wang, Q., Zhang, Y., Holynski, A., Efros, A. A., Kanazawa, A., 2025b. Continuous 3D perception model with persistent state. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10510–10522.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. DUSt3R: Geometric 3D vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20697–20709.

Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2025c. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 12077–12090.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth Anything V2. *Advances in Neural Information Processing Systems (NeurIPS)*.