A Computer Vision-Based Framework for Assessing Urban Walkability Using Multiscale Street Scene Features

Elke Kraushaar
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
kraushaar1979@ucf.edu

Sole Tedesco
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
tedescosole@colostate.edu

Clementina Portillo
School of Computing, Clemson University, Clemson, SC, USA.
clementinamail@clemson.edu

Abstract

Urban walkability assessment must reconcile coarse network-level indicators with fine-grained environmental qualities that pedestrians actually encounter. Conventional indices often reduce streetscapes to density, diversity, and design proxies, while image-based audits can capture visible conditions but frequently operate at a single analytical scale and provide limited treatment of uncertainty. This article develops a computer vision-based framework for assessing urban walkability from multiscale street scene features. The framework parses street-level imagery into semantic classes, extracts object-level cues such as sidewalk visibility, vegetation, sky openness, vehicles, and obstructions, aggregates these cues across viewpoints and spatial neighborhoods, and fuses the resulting object, view, and network descriptors into a latent walkability score. A scale-sensitive aggregation operator is specified, followed by a monotonic fusion model that is designed to preserve directional expectations about supportive and obstructive scene elements. The framework also formalizes scene diversity, error propagation from segmentation and viewpoint sampling, and a decision rule for identifying adequacy thresholds. Rather than treating computer vision as a direct substitute for manual audits, the approach positions visual evidence as one component within a structured measurement system that can be calibrated against audit or perception data when such evidence is available. The main contribution is a technically explicit account of how street scene features can be represented, aggregated, fused, and evaluated across scales without conflating visual detection with walkability itself. The framework offers a basis for more transparent, uncertainty-aware, and context-sensitive walkability assessment in planning research and practice.

Keywords

urban walkability; street scene analysis; semantic segmentation; multiscale aggregation; computer vision; uncertainty quantification

1. Introduction

Walkability research addresses a deceptively simple question: which urban environments support walking as a safe, comfortable, and useful activity? Empirical traditions in planning and public health have answered this question largely through observable dimensions of the built environment, including density, land-use mixing, street connectivity, sidewalk provision, and aesthetic quality [1], [2]. These dimensions are attractive because they can be measured from parcel data, road networks, and field audits. They also support comparative analysis across neighborhoods and policy scenarios. Yet the built environment that pedestrians experience is not fully described by such indicators. A connected street network may contain discontinuous sidewalks; a mixed-use district may present blank walls at eye level; and a green street may be visually screened by parked vehicles. Walkability is therefore both a network property and a scene property, and a credible assessment framework must relate the two without collapsing one into the other.

Public health research has established that neighborhood environments are associated with walking behavior and physical activity [3], [4]. The determinants of walking include functional access, safety, comfort, and aesthetic interest, and these determinants operate through different pathways [5], [6]. Functional access depends on destinations and connectivity; safety depends on traffic exposure and surveillance; comfort depends on sidewalk condition, shade, and enclosure; interest depends on visual variety and human-scale detail. Conventional walkability indices capture some of these pathways, but they tend to treat the street environment as an attribute of a geographic unit rather than as a sequence of scenes. Manual street audits can capture scene-level detail, and virtual audit methods have demonstrated that street-level imagery can support systematic observation of neighborhood conditions [7]. Reliability studies have further examined whether virtual streetscape audits can substitute for physical audits, with the implication that image-based observation is useful but not automatically equivalent to on-site assessment [8]. This tension motivates a framework that makes the visual measurement process explicit.

Computer vision offers a means to extract street scene features at scale. Semantic segmentation can assign each image pixel to classes such as road, sidewalk, building, vegetation, sky, person, and vehicle, enabling quantitative description of visible composition. Visual enclosure and street-level greenery have been measured from street view imagery using machine learning and image processing [9], [10]. Urban scene datasets and segmentation architectures provide the technical basis for robust parsing of complex street environments [11], [12], [13]. However, detection is not assessment. The presence of a sidewalk class in an image does not establish that the sidewalk is wide, continuous, unobstructed, or pleasant. Similarly, the proportion of vegetation is not equivalent to shade, enclosure, or perceived safety. The analytical gap lies in converting pixel-level and object-level observations into defensible multiscale constructs while preserving the distinction between visual evidence and walkability judgment.

A further complication is that objective scene measurements and subjective perceptions can align in some dimensions and diverge in others. Research comparing objective and subjective measurement of street perceptions has examined this coherence and divergence, indicating that visual indicators should not be treated as transparent proxies for experience [14]. Deep-learning studies of urban perception have shown that street images contain information associated with perceived safety, beauty, and liveliness, but these inferred qualities are model-dependent and context-sensitive [15], [16], [17]. The framework

developed here therefore treats computer vision as a measurement instrument rather than as an autonomous judge of walkability. It defines a multiscale feature hierarchy, a fusion model with directional constraints, an uncertainty account, and an evaluation logic. The central contribution is not a new segmentation architecture but a structured way to connect street scene features across object, view, and network scales so that visual evidence can be used transparently in walkability assessment.

2. Conceptual Foundations and Related Work

Walkability measurement has traditionally relied on three families of indicators: land-use and destination measures, network configuration measures, and streetscape quality measures. Density, diversity, and design summarize destination proximity and street layout, while audit instruments describe sidewalks, crossings, buffers, and micro-scale conditions [1], [2], [5]. These approaches have strong policy relevance because they connect to zoning, transportation investments, and street design standards. Their limitation is that they often aggregate over geographic units that are too coarse to represent the pedestrian's immediate environment. A block group may contain both a comfortable commercial street and a hostile arterial, and an average walkability score can conceal that contrast. Street-level imagery can recover some of this heterogeneity, but only if the sampling and aggregation logic are specified.

Virtual audit research has shown that street view imagery can support systematic observation of sidewalks, buildings, and public spaces [7]. The reliability of virtual audits relative to field audits depends on the feature being observed and on the protocol used [8]. This finding has two implications for a computer vision framework. First, the unit of observation should be the viewpoint or street segment rather than an entire neighborhood. Second, the framework should preserve information about observer location, viewing direction, and sampling density. A single panoramic image at a street intersection is not a random sample of the street environment; it is a situated observation shaped by occlusion, perspective, and field of view. Treating such observations as unbiased neighborhood samples would introduce a systematic error that no classifier can correct after the fact.

Image-based urban perception research has developed models that predict human ratings of safety, beauty, or liveliness from street scene images [15], [16], [17]. These models demonstrate the expressive capacity of convolutional and deep architectures for urban visual analysis. At the same time, they raise questions about construct validity. A model trained to predict perceived safety may learn cues associated with lighting, vegetation, signage, or architectural style, and those cues may not transfer across cities or cultural contexts. The framework proposed here does not replace perception modeling; it provides a feature representation and aggregation structure into which perception models can be inserted as calibrators or auxiliary evidence. The distinction matters because walkability is a normative and behavioral construct, whereas scene perception is an inferred visual attribute. A scene can be visually pleasant but functionally inaccessible, or visually ordinary but highly walkable because of destinations and crossings.

Multiscale representation is central to bridging these levels. Object-level features describe visible elements such as sidewalk width, curb presence, tree canopy, parked cars, and obstructions. View-level features describe composition and arrangement, including visual enclosure, openness, and scene diversity. Network-level features describe how these visible conditions are distributed along routes and whether supportive scenes are connected rather than isolated. The same semantic class can have different meaning at different scales. A high proportion of road pixels at an intersection may indicate exposure to traffic, while a high

proportion of road pixels along a commercial street may indicate a coherent pedestrian corridor with active frontage. The framework therefore does not assign fixed valence to raw pixel proportions. It treats scale, context, and arrangement as part of the feature definition.

3. Methodological Framework

The framework begins with a georeferenced sample of street-level images. Each observation is indexed by location i and viewpoint v. Images may be panoramic or perspective, but the framework requires that the camera height, orientation, and field of view be recorded or estimable, because these parameters affect visible composition. A semantic segmentation model assigns a class label to each pixel, producing a label map S_v(p) over the image domain. The model may be trained on an urban scene dataset and adapted to the target city through label harmonization and targeted review [11], [13]. The output is not treated as ground truth. It is a probabilistic or discrete observation of visible scene classes, and its uncertainty is carried forward into later stages.

3.1 Scene Representation and Semantic Feature Extraction

For each viewpoint, the visible proportion of semantic class k is computed as the fraction of image pixels assigned to that class. This gives a simple but interpretable object-level representation of scene composition. The proportion is defined as

$$R(i, v, k) = (1/|\Omega_v|) \sum_\{p \in \Omega_v\} 1[S_v(p) = k]$$

where Ω_v is the set of pixels in viewpoint v, |Ω_v| is its cardinality, S_v(p) is the semantic label at pixel p, and 1[·] is an indicator function equal to one when its argument is true. The index i denotes the street location, v denotes the viewpoint, and k denotes the semantic class. Classes typically include road, sidewalk, building, vegetation, sky, vehicle, person, and bicycle. The proportion is useful because it is invariant to image resolution and easy to audit. It is also limited: it discards spatial arrangement, depth, and object identity. A sidewalk that occupies the lower center of an image and a sidewalk that appears as a thin sliver behind parked cars can have similar proportions but very different implications for walking.

To retain some information about arrangement and visual variety, the framework supplements class proportions with object detections and geometric cues. Object detectors provide counts and bounding boxes for pedestrians, vehicles, bicycles, street furniture, and trees. Monocular depth or vanishing-point estimation can approximate enclosure and openness, though such estimates are sensitive to camera pitch and lens distortion. Scene diversity is summarized by the entropy of the class distribution:

$$\mathrm{E}(i, v) = -\sum_{k=1}^{K} R(i, v, k) \log(R(i, v, k) + \varepsilon)$$

where K is the number of semantic classes and ε is a small positive stabilizer that prevents undefined logarithms when a class is absent. Higher entropy indicates a more balanced mix of visible classes; lower entropy indicates dominance by one class, such as sky, road, or wall. Entropy is not a walkability score. It is a descriptive feature that can support hypotheses about visual interest, monotony, or enclosure. The framework treats it as one input among several and does not assume that higher diversity is always better. A uniformly clean residential sidewalk may be more walkable than a visually diverse but cluttered service alley, and the fusion stage must be able to represent such context dependence.

3.2 Multiscale Feature Hierarchy

The multiscale hierarchy organizes features into three levels. The object level contains class proportions, object counts, and geometric descriptors for a single viewpoint. The view level summarizes composition across viewpoints at the same location or street segment, including maximum and minimum enclosure, visible sidewalk continuity, and the proportion of views with obstructions. The network level summarizes the distribution of view-level features along connected routes. This hierarchy is necessary because a location-level average can hide route-level discontinuities. A sequence of high-quality scenes does not constitute a walkable route if the sequence is interrupted by an un-crossable arterial. Conversely, a modest scene repeated along a safe, connected, and shaded corridor may support walking better than an isolated visually attractive pocket.

Spatial aggregation across the network is defined with a scale-specific kernel. For scale m, the aggregated feature at location i is

$$F(i,m) = \sum_{j \in N_m(i)} \omega_m(d_i j) \frac{z_j}{\sum_{j \in N_m(i)} \omega_m(d_i j)}$$

where N_m(i) is the set of locations within the neighborhood defined for scale m, d_ij is the network distance between locations i and j, ω_m(d_ij) is a distance-weighting kernel, and z_j is the view-level descriptor at location j. The kernel may be exponential, Gaussian, or distance-decay, but it must be fixed before calibration and reported with the results. The denominator normalizes the weights so that locations with different numbers of sampled neighbors remain comparable. The scale parameter m corresponds to object, view, and network neighborhoods; for example, a small neighborhood may capture the immediate street segment, while a larger neighborhood may capture a walking catchment. The equation makes explicit that multiscale features are not merely different variable sets. They are different spatial supports over which the same underlying visual evidence is summarized.

The choice of kernel and bandwidth is a substantive design decision. A narrow kernel emphasizes local conditions and is appropriate for sidewalk obstruction or curb presence. A broad kernel emphasizes contextual conditions and is appropriate for route continuity or access to destinations. Because the kernel determines how quickly evidence decays with distance, it should be selected with reference to pedestrian behavior and network topology rather than statistical convenience alone. If a pedestrian is unlikely to walk more than a few hundred meters to reach a destination, a kernel that assigns substantial weight to distant scenes will misrepresent the relevant environment. The framework therefore requires that scale definitions be justified in behavioral and network terms, not only in image-processing terms.

3.3 Fusion, Calibration, and Decision Logic

The multiscale features are fused into a latent walkability score. The score is not intended to be a universal measure of walkability across all cities and populations. It is a structured summary that can be calibrated to a specified planning question, such as route suitability for older adults, school access, or transit access. A logistic fusion form is

$$W(i) = \sigma\left(\beta_0 + \sum_{m=1}^{3} \beta_m^T F(i,m) + \gamma^T U(i)\right)$$

where W(i) is the latent walkability score bounded between zero and one, σ(·) is the logistic function, β_0 is an intercept, β_m is a vector of scale-specific coefficients, F(i,m) is the multiscale feature vector, U(i) is a vector of contextual covariates such as land-use mix or traffic volume, and γ is a coefficient vector for those covariates. The coefficients are not fixed by the framework. They are to be estimated only when suitable calibration evidence, such as audit ratings or behavioral observations, is available. Before calibration, the equation defines the model structure and the directional hypotheses that any calibration must respect.

Directional constraints are important because unconstrained regression can produce substantively implausible relationships. Sidewalk visibility, shade, and crossing provision are expected to be non-decreasing in walkability, while obstruction by vehicles or construction is expected to be non-increasing, all else equal. These expectations can be imposed as sign constraints or monotonicity constraints during calibration. The constraints do not guarantee validity, but they make the model's normative assumptions inspectable. They also reduce the risk that a learned coefficient will encode a spurious correlation, such as a positive association between parked cars and walkability that arises because both are more common in dense commercial districts. Contextual covariates and scale-specific features help separate such confounding, but they cannot eliminate it.

The fusion score supports a decision rule for identifying locations that meet a minimum adequacy standard. A simple rule is

$$A(i) = 1[W(i) \geq \tau and \mathrm{E}(i) \geq e_m in]$$

where A(i) is a binary adequacy indicator, τ is a walkability threshold, E(i) is a scene-diversity or composition criterion at the relevant scale, e_min is its minimum acceptable value, and 1[·] is an indicator function. The second condition prevents the rule from labeling a location adequate solely because it is visually uniform or dominated by one class. The threshold values are policy parameters, not properties of the image model. They should be set through deliberative planning processes and tested for sensitivity. The decision rule is deliberately transparent: it separates the continuous score from the binary judgment and allows planners to see how changes in threshold choices affect the classification of locations.

3.4 Uncertainty and Error Propagation

Uncertainty enters the framework at several points. Segmentation errors create class proportion errors; viewpoint sampling creates coverage errors; depth or enclosure estimates create geometric errors; and calibration error creates coefficient uncertainty. A first-order propagation approximation for the fused score is

$$\mathrm{Var}(W_i) \approx \sum_m J_m^T \Sigma_m J_m + \sigma_\varepsilon^2$$

where Var(W_i) is the approximate variance of the walkability score, J_m is the gradient of W with respect to the scale-m feature vector, Σ_m is the covariance matrix of the scale-m features, and σ_ε^2 is the residual variance from model misspecification or omitted contextual factors. The equation is not a substitute for empirical validation. It is a formal statement that uncertainty in visual features should not be ignored when visual features are used for decisions. A location with a high score but wide uncertainty may require audit review rather than automatic classification.

Segmentation error is often class-dependent. Road and sky are typically easier to separate than sidewalk and road at oblique angles, and vegetation may be confused with building

shadow in dense scenes. Viewpoint sampling error is spatial and directional: a single view toward a building facade may miss the sidewalk condition behind the camera. The framework addresses these errors by requiring multiple viewpoints per location where possible, by recording viewing geometry, and by reporting uncertainty alongside point estimates. It also recommends stratified review of high-stakes classifications, especially where the decision rule would change under plausible segmentation error. This is a practical form of error management that does not require the segmentation model to be perfect.

4. Analytical Assessment and Design Trade-offs

The framework's value depends on choices that cannot be resolved by computer vision alone. One trade-off concerns semantic granularity. A small class taxonomy is easier to segment consistently but may conflate sidewalk with road or vegetation with shade. A large taxonomy is more expressive but increases label imbalance and annotation cost. For walkability assessment, the most consequential distinctions are often between traversable and non-traversable surfaces, between static and moving obstructions, and between vegetation that provides shade and vegetation that blocks sight lines. The taxonomy should be designed around these distinctions rather than around the convenience of an existing scene dataset. Label harmonization is therefore not a technical afterthought; it is part of construct definition.

A second trade-off concerns the relative weight of visual and network evidence. A purely visual score may favor aesthetically rich but poorly connected streets, while a purely network-based score may ignore the conditions that make a connected route pleasant or safe. The multiscale fusion model allows both forms of evidence to enter, but it does not determine their relative importance. That determination belongs to the planning question and the calibration process. For a route-to-school assessment, crossing safety and sidewalk continuity may dominate; for a recreational walking assessment, shade, visual interest, and low traffic exposure may be more important. The same image features can therefore support different walkability constructs when the fusion weights and calibration targets differ.

A third trade-off concerns aggregation across viewpoints. Averaging class proportions over all viewpoints at a location reduces the influence of a single camera angle, but it can also cancel meaningful directional contrasts. A location may have a comfortable sidewalk on one side of the street and a narrow shoulder on the other. A simple average would report an intermediate condition that exists on neither side. The framework therefore recommends retaining directional or side-specific descriptors in addition to averaged features. This increases dimensionality, but it preserves information that is relevant to pedestrian experience. The appropriate level of aggregation depends on whether the planning question concerns a point, a street segment, or a route.

Sensitivity analysis is essential because the framework's outputs are model-based. The segmentation model's class proportions may vary with lighting, season, weather, and camera equipment. The spatial kernel may emphasize different neighborhoods depending on bandwidth. The fusion weights may change when calibration data are added or when the target population changes. A useful sensitivity analysis should vary these components one at a time and report how many locations change classification under plausible perturbations. It should also distinguish between rank stability and threshold stability. A framework may rank locations similarly even when absolute scores shift, and that distinction matters for different planning uses. If the purpose is to prioritize a small number of interventions, rank stability may be sufficient. If the purpose is to certify that a route meets a minimum standard, threshold stability is more important.

The framework also has boundary conditions. It is most defensible when street-level imagery is recent, georeferenced, and sampled at a density that reflects the relevant walking environment. It is less defensible when imagery is sparse, outdated, or captured only from vehicle-mounted cameras that systematically miss pedestrian-side conditions. It is not a substitute for accessibility assessment that requires information not visible in images, such as curb ramp gradients, surface friction, or crossing signal timing. It also cannot resolve questions of destination access without external land-use and network data. These limitations are not defects of the visual method; they are reasons to keep the visual method within a broader measurement system.

5. Discussion

The framework reframes computer vision as a disciplined measurement layer rather than a stand-alone walkability index. This reframing has several implications. First, it clarifies the object of measurement. Semantic class proportions describe visible scene composition; object detections describe the presence and location of elements; geometric cues describe enclosure and openness; network aggregation describes how these conditions are distributed along routes. These are different constructs, and treating them as interchangeable indicators of walkability weakens both interpretation and validation. The multiscale hierarchy keeps the distinctions visible and allows each construct to be evaluated separately before fusion.

Second, the framework makes uncertainty part of the assessment rather than an afterthought. In practice, visual audits are often presented as point estimates: a location is scored, ranked, or classified without an accompanying statement of confidence. The uncertainty equation and the decision rule provide a way to represent the fragility of those judgments. This is particularly important when visual evidence is used to allocate resources or to justify regulatory decisions. A location that is classified as inadequate under one segmentation model but adequate under another should be flagged for review, not silently averaged into a stable-looking score.

Third, the framework accommodates the coherence and divergence between objective scene measures and subjective perceptions [14]. If perception ratings are available, they can be used as calibration targets or as auxiliary outcomes, but they should not be confused with the visual features themselves. The same street scene may be perceived differently by different groups, and those differences may be systematically related to age, gender, disability, or prior experience. Research on neighborhood environments and walking among older adults, for example, indicates that built-environment associations with walking can vary across populations and contexts [18]. A walkability framework that claims universal applicability without population-specific calibration would therefore overstate its evidence.

The framework also has implications for data governance and reproducibility. Street-level imagery can raise privacy concerns, especially when pedestrians or vehicles are identifiable. Segmentation and object detection can be performed on de-identified or blurred imagery, but blurring may remove cues that are relevant to perceived safety or activity. The framework does not resolve this trade-off, but it makes it explicit. Reproducibility requires reporting the image source, capture dates, sampling protocol, segmentation model, label taxonomy, kernel bandwidths, and calibration procedure. Without such reporting, visual walkability scores are difficult to compare across studies or to defend in public decision-making.

Finally, the framework suggests a research agenda organized around validation rather than novelty. The most useful next step is not another segmentation architecture but a

systematic comparison of multiscale fusion against simpler baselines: raw class proportions, manual audit composites, and network-only walkability indices. Such comparisons should evaluate not only predictive accuracy but also rank stability, uncertainty calibration, and sensitivity to viewpoint sampling. They should also examine whether the framework's outputs are interpretable to planners and whether they support decisions that are defensible to residents. These are questions about measurement validity, not merely model performance.

6. Conclusion

This article has developed a computer vision-based framework for assessing urban walkability using multiscale street scene features. The framework treats street-level imagery as situated evidence rather than as a direct readout of walkability. It parses scenes into semantic and object-level features, summarizes them across viewpoints and spatial neighborhoods, fuses them into a latent walkability score with directional constraints, and formalizes uncertainty and decision thresholds. The central contribution is the multiscale structure: object-level cues, view-level composition, and network-level distribution are kept analytically distinct and then related through explicit aggregation and fusion equations. This structure addresses a persistent weakness in image-based walkability research, namely the tendency to move from pixel proportions to neighborhood judgments without accounting for viewpoint, scale, or context.

The framework is deliberately calibration-dependent. It does not claim that a particular set of weights or thresholds is universally valid, and it does not treat semantic segmentation as a substitute for field observation. Instead, it specifies what must be measured, how features are aggregated, how uncertainty is represented, and how a decision rule can be made transparent. This makes the approach suitable for planning contexts in which walkability is not a fixed technical quantity but a policy-relevant construct that must be defined and defended. The framework can support comparative assessment, route screening, and audit prioritization, provided that its assumptions are reported and its outputs are tested against independent evidence. Its scientific value lies in making the pathway from street scene to walkability judgment inspectable, rather than in hiding that pathway inside an opaque model.

References

1. Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. Journal of the American Planning Association, 76(3), 265–294.

2. Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. Transportation Research Part D: Transport and Environment, 2(3), 199–219.

3. Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., & Hess, P. M. (2010). The development of a walkability index: Application to the Neighborhood Quality of Life Study. British Journal of Sports Medicine, 44(13), 924–933.

4. Saelens, B. E., Sallis, J. F., Black, J. B., & Chen, D. (2003). Neighborhood-based differences in physical activity: An environment scale evaluation. American Journal of Public Health, 93(9), 1552–1558.

5. Pikora, T., Giles-Corti, B., Bull, F., Jamrozik, K., & Donovan, R. (2003). Developing a framework for assessment of the environmental determinants of walking and cycling. Social Science & Medicine, 56(8), 1693–1703.

6. Handy, S. L., Boarnet, M. G., Ewing, R., & Killingsworth, R. E. (2002). How the built environment affects physical activity: Views from urban planning. American Journal of Preventive Medicine, 23(2), 64–73.

7. Rundle, A. G., Bader, M. D., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to audit neighborhood environments. American Journal of Preventive Medicine, 40(1), 94–100.

8. Badland, H. M., Opit, S., Witten, K., Kearns, R. A., & Mavoa, S. (2010). Can virtual streetscape audits reliably replace physical streetscape audits? Journal of Urban Health, 87(6), 1007–1016.

9. Yin, L., & Wang, Z. (2016). Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. Applied Geography, 76, 147–153.

10. Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forestry & Urban Greening, 14(3), 675–685.

11. 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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3213–3223).

12. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848.

13. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ADE20K dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 633–641).

14. Song, Q., Li, M., Qiu, W., Li, W., & Luo, D. (2022, November). The coherence and divergence between the objective and subjective measurement of street perceptions for Shanghai. In International Conference on Advanced Data Mining and Applications (pp. 244-256). Cham: Springer Nature Switzerland.

15. Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore: Predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 779–785).

16. Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In European Conference on Computer Vision (pp. 196–212). Springer.

17. Porzi, L., Bulò, S. R., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. In Proceedings of the 23rd ACM International Conference on Multimedia (pp. 139–148).

18. Nagel, C. L., Carlson, N. E., Bosworth, M., & Michael, Y. L. (2008). The relation between neighborhood built environment and walking activity among older adults. American Journal of Epidemiology, 168(4), 461–468.