



## VISION WITHOUT SURVEILLANCE

### USING ZERO-SHOT CLIP FOR SCALABLE BUILT ENVIRONMENT FEATURE EXTRACTION

MALIKKA RAJSHAHI<sup>1,2</sup>, MARIELA ALFONZO<sup>2,3</sup>, COREY ASHBY<sup>2</sup>, EVAN TAYLOR<sup>2,4</sup>, BENJAMIN FEUER<sup>2,5</sup>

<sup>1</sup> CUNY GRADUATE CENTER, USA

<sup>2</sup> STATE OF PLACE, USA

<sup>3</sup> NYU TANDON, USA

<sup>4</sup> ARIZONA STATE UNIVERSITY, USA

<sup>5</sup> STANFORD, USA

#### KEYWORDS

*Built Environment  
Scaling  
Urban Design  
Computer-Aided  
Design  
Planning Support  
Systems*

#### ABSTRACT

*Municipalities need current, granular built-environment data to support effective planning. However, current collection methods are expensive, challenging to scale, and present privacy concerns. To address this gap, State of Place (SoP) collaborated with the US Department of Transportation (USDOT) through a Small Business Innovation Research (SBIR) initiative. We developed a scalable, privacy-preserving vision-language model (VLM) pipeline that leverages zero-shot CLIP to extract 150+ interpretable urban design features from street-level imagery. Applied to Philadelphia's bus network, the pipeline processed over two million images and generated block-level predictions for more than 8,000 blocks. This approach enables scalable, evidence-based urban planning across cities.*

## 1. Introduction

The USDOT initiated a three-phase SBIR grant after identifying gaps in data related to infrastructure, urban design, and travel behavior, selecting SoP as one of twelve recipients. Phase I focused on improving data collection and availability. This study introduces the research and development SoP implemented during Phase I to generate built environment data using a scalable, privacy-preserving vision-language model (VLM) pipeline for urban analysis, which employs zero-shot classification with CLIP.

Effective urban planning depends on detailed, diverse data that municipalities can use to evaluate, maintain, and redesign streets, infrastructure, and zoning priorities. Such data must reflect real-time conditions of urban landscapes to support improvements in pedestrian safety, public transit efficiency and maintenance, and climate resilience indicators (e.g., shade from tree coverage). Although detailed insights are essential for urban planning, collecting this data at scale remains a significant challenge.

Most cities currently rely on three primary data sources to address these gaps. The first is sensor networks, such as video-based traffic counters, multimodal detectors, and other fixed hardware, which provide measurements of travel behavior and street activity. These systems are costly to deploy and maintain, often requiring thousands of dollars per intersection, and their placement may reinforce inequities by favoring wealthier or more visible corridors (Du et al., 2019; U.S. Department of Transportation, ITS Joint Program Office, 2012).

The second source is mobile phone data obtained from commercial vendors, including location-based services and anonymized mobility datasets. Although these products provide broad geographic coverage, they present several limitations: high costs, limited interpretability, proprietary algorithms, and privacy risks, as individuals may still be identifiable even in supposedly anonymized datasets. Importantly, mobile data cannot capture physical street design elements (e.g., sidewalk width, tree canopies, crosswalk quality, etc.) which are critical for safety and multimodal performance (Toole et al., 2015; Barbeau et al., 2022).

The third source consists of manual audits or crowd sourced assessments. While informative, these methods require substantial labor, do not scale effectively, and rapidly become outdated. Consequently, many municipalities lack the street-level data necessary to diagnose issues, prioritize investments, or evaluate interventions (Ewing and Cervero, 2010; Haklay, 2010).

Recent research in computer vision demonstrates that deep learning models have the potential to extract built environment features from street-level imagery. However, most current approaches utilize supervised learning, in which models are trained on large datasets of manually annotated images with specific labels (e.g., the presence of sidewalks, crossings, street trees). Producing these labeled datasets is labor-intensive, expensive, and challenging to standardize across cities, as annotation protocols and visual interpretations often differ by local context. Furthermore, models trained in one city frequently underperform in others due to variations in architecture, climate, cultural context, or image conditions. These limitations constrain the practicality of computer vision for operational planning (Zhang et al., 2024; Biswas and Roy, 2023; Doiron et al., 2022).

Advances in vision-language models (VLMs), particularly CLIP, offer a promising alternative. CLIP, trained on hundreds of millions of image-text pairs, embeds images and natural language into a shared representation space, enabling zero-shot classification. Instead of training a new model for each task, practitioners can use natural-language prompts to describe features of interest (Radford et al., 2021; Zhang et al., 2024). This approach provides a powerful and flexible mechanism for extracting semantically meaningful indicators from street-level imagery without labeled training data. Notably, it scales to hundreds of features and can be rapidly adapted to new cities or planning questions.

Addressing these gaps in coverage and transferability, our work introduces a zero-shot VLM pipeline for citywide urban analysis that enables municipalities to extract interpretable, planning-relevant features at scale. Our contributions are threefold. First, we construct a flexible CLIP-based feature extraction framework that produces over 150 interpretable indicators

covering infrastructure, context, and travel behavior, without requiring training data or bespoke model development. Second, we develop a low-cost, open-source image capture and geospatial alignment methodology that gives cities autonomy over data collection while avoiding the privacy and surveillance concerns associated with sensors or mobile device traces. Third, we integrate these components into a block-level analytic system aligned with the SoP Index, a composite urban design metric, enabling interpretable assessments of urban design quality at the scale most relevant for planning decisions (Day et al., 2006).

This methodology is applied to more than two million images collected across Philadelphia's SEPTA bus network, demonstrating the feasibility, accuracy, and scalability of zero-shot VLMs for operational urban analytics. The results presented here establish technical feasibility and inform ongoing work focused on expanding model coverage and robustness, as well as deployment in additional urban contexts. By leveraging advances in AI and grounding the system in established planning frameworks, this work provides a new pathway for cities to automate data collection, address longstanding information gaps, and support evidence-based design and policymaking.

## **2. Background**

Municipalities increasingly depend on data-driven approaches to guide transportation, planning, and urban design decisions. Delivering safer streets, improving transit performance, and supporting resilient infrastructure all require timely and detailed information about the built environment. However, even with significant advances in digital tools and spatial data, cities continue to face persistent challenges in obtaining reliable street-level information at the scale necessary for operational decision-making. These challenges underscore the need for new analytic methods capable of producing comprehensive, interpretable, and up-to-date assessments of urban conditions.

### ***2.1. Existing Approaches to Urban Data Collection***

A common approach to collecting transportation and streetscape data is fixed sensor networks, such as video based traffic counters, radar detectors, and automated pedestrian and bicycle counters. Although these systems offer high temporal resolution and can capture multimodal activity, their installation and maintenance costs are substantial. Estimates from the U.S. Department of Transportation show that adaptive signal or video-based detection systems often cost between \$4,000 and \$16,000 per intersection for detector installation alone, while full adaptive traffic-signal systems can exceed \$65,000 per intersection (U.S. Department of Transportation, ITS Joint Program Office, 2012; Florida Department of Transportation, 2023). Federal Highway Administration reports similarly highlight that agencies continue to rely on expensive loop detectors and tube counters because lower cost, highly scalable alternatives remain limited.

In this context, (Ingles-Romero et al., 2018) proposed a low-cost alternative that uses existing surveillance cameras, combined with a BeagleBone Black embedded platform, to extract traffic metrics (Ingles-Romero et al., 2018). While their approach demonstrates that modest hardware can be repurposed to reduce deployment costs, it was tested on a single urban intersection and designed for a restricted set of traffic metrics (e.g., vehicle queue lengths). The system relies on scene-specific calibration, tailored camera angles, and manually defined regions of interest. These requirements complicate the system's broader adoption and scalability, as each deployment must undergo a labor intensive setup to ensure accuracy. Transferability is further hindered by the difficulty of adapting the system to diverse urban environments without substantial adjustments. Consequently, the method does not address the broader set of built-environment features that planners typically need, such as sidewalk conditions, vegetation, frontage activation, or street design quality (Ewing and Cervero, 2010). These limitations demonstrate how isolated attempts to lower costs do not replace fixed sensor networks, which remain expensive and difficult to scale, resulting in sparse coverage concentrated in downtown

corridors or priority areas, while many neighborhoods, particularly underserved communities, remain unmeasured.

Besides cost and scale, concerns about surveillance and data governance further limit the adoption of camera-based sensing systems. Research on smart city infrastructures highlights pervasive privacy and security risks associated with urban sensing networks, particularly when high resolution imagery or persistent monitoring is involved (Ismagilova et al., 2022). Participatory sensing studies similarly show that mobile device integrations often raise unresolved questions around re-identification and data ownership (Joy et al., 2016). These governance concerns constrict deployment and amplify equity concerns in data scarce urban contexts.

Typical supplements to sensor networks include commercial mobile device data, which enable broad geographic coverage by aggregating geolocation signals to infer travel flows and movement patterns (Gonzalez et al., 2008; Blondel et al., 2015). These datasets have been widely adopted for regional mobility analysis and travel demand modeling, particularly where fixed sensors are sparse. However, mobile device data present persistent challenges related to transparency, validation, and governance. Processing pipelines are typically proprietary, with limited access to raw data and few opportunities for independent audit or methodological replication (Toole et al., 2015; He et al., 2016). As a result, municipalities often have little visibility into how movement signals are filtered, aggregated, or weighted, constraining their ability to interpret results or assess uncertainty (Verhulst, 2021). Privacy concerns further complicate adoption: even aggregated or anonymized mobility datasets have been shown to carry significant re-identification risks, raising ongoing ethical and regulatory concerns de (Montjoye et al., 2013).

Furthermore, sensor networks most commonly deployed by municipalities (e.g., loop detectors, radar sensors, video counters) have a key limitation. They are designed to measure movement and activity rather than the physical design features of streets, such as sidewalk width, lighting conditions, or building frontage. While specialized sensing technologies such as LiDAR can capture aspects of urban geometry, they are rarely deployed at a citywide scale and do not directly represent the qualitative, human-scale design attributes that shape how streets are experienced (Wu et al., 2021). Extensive empirical research shows that built environment characteristics exert a measurable influence on travel behavior, pedestrian activity, and mode choice, with meta-analyses demonstrating that design qualities such as block structure, streetscape enclosure, and pedestrian infrastructure are central determinants of mobility Ewing and (Cervero, 2010, 2017; Sallis et al., 2016; Le et al., 2019; de Moraes Ferrari et al., 2020). Street-level imagery studies further demonstrate that design cues ranging from building form and façade quality to visible amenities and street lighting are detectable visually but remain invisible to sensor- or mobility-derived datasets (Geburu et al., 2017). The absence of these features is a fundamental limitation of sensor- and mobile-based approaches: they are designed to measure activity, not the built-environment conditions that produce that activity. As a result, even when these systems provide broad geographic coverage or detailed movement patterns, they cannot supply the qualitative design information planners rely on to diagnose safety concerns, assess accessibility, or guide public realm improvements.

Another method many cities rely on is manual audits or crowd-sourced assessments to collect detailed data on the built environment. Field surveys and community mapping yield valuable insights, but their labor-intensive nature limits scalability and makes regular updates challenging. A report on sidewalk inventory costs observed expenditures ranging from US \$86 to \$826 per mile, even in smaller municipalities (Florida Department of Transportation, 2023). Crowd-sourcing also faces coverage bias and update lag: Diop et al. (2022) identify exclusion of low-capacity communities and delays in data coverage as major limitations in global urban settings (Diop et al., 2022).

Consequent to these approaches' limitations, municipalities still lack an affordable, scalable, and privacy-preserving method for observing both street activity and physical design in sufficient detail. This gap restricts the ability to diagnose safety issues, target investments, monitor equity, and evaluate sites for policy intervention. Furthermore, the absence of

comprehensive built-environment data remains a significant barrier to addressing increasingly prevalent needs to mitigate pedestrian and vehicle fatalities, lower impacts of climate change, reduce crime, and focus on outcomes that improve quality of life and economic performance (Board et al., 2019).

To address this barrier, we develop a vision-language modeling framework that extracts rich, design-oriented features directly from street-level imagery without relying on sensors, mobile traces, or large annotated datasets. This approach directly addresses the shortcomings of previous methodologies and provides a pathway to scalable, privacy preserving urban analytics. Advances in the field and the slow but steady adoption of computer vision in urban design make such a framework possible.

## ***2.2. Computer Vision for Urban Analytics***

Recent advances in computer vision and the expansion of street-level imagery have opened new avenues for measuring the built environment. Platforms such as Google Street View, Mapillary, and agency-collected imagery provide high resolution street-level views. Researchers have used these sources to extract features including sidewalks, crosswalks, greenery, building setbacks, and perceived safety (Ning et al., 2021; Seiferling et al., 2017; Naik et al., 2014). Collectively, these efforts demonstrate that computer vision can automate components of the urban audit process that once required extensive manual fieldwork.

Most computer vision techniques used in urban studies rely on supervised learning, which requires collecting and annotating large datasets to train specialized models for tasks such as tree detection, street furniture identification, or lane marking recognition. Seiferling et al. (2017) developed a supervised pipeline for quantifying urban tree cover from Google Street View imagery, which required substantial manual labeling to support segmentation and classification (Seiferling et al., 2017). Similarly, Hosseini et al., (2022) introduced a city-scale semantic segmentation model for sidewalk materials that relied on extensive annotated training data but demonstrated limited transferability across cities (Hosseini et al., 2022). These cases illustrate the high labor and cost associated with supervised methods, which often require thousands of examples to achieve reliable performance.

Geographic generalization poses an additional challenge. As Zhang et al. (2024) note in their review of urban visual intelligence, supervised detectors frequently perform poorly when applied to cities with different architectural styles, vegetation, signage conventions, or imaging conditions (Zhang et al., 2024). Updating such models to accommodate new design interventions, construction changes, or evolving policy priorities typically requires further annotation and retraining, as seen in façade-quality and streetscape assessment work such as that of Liu et al. (2016) (Liu et al., 2016). Finally, many supervised CV approaches focus on low-level pixel detections (e.g., texture classification, edge segmentation, or object bounding boxes) that do not align with the conceptual categories used in planning practice. As a result, their outputs often fail to capture higher level constructs such as street enclosures, active frontages, or human-scale design qualities, thereby limiting their interpretability and practical relevance to planners (Ibrahim and Cheng, 2020).

These limitations have restricted the adoption of computer vision in municipal planning. Effective deployment requires not only accuracy but also adaptability, straightforward maintenance, and alignment with planning concepts. Since supervised approaches depend on labeled data, they are not well-suited for situations where planning needs shift rapidly or data collection varies across cities and seasons. Our work addresses these constraints by introducing a zero-shot vision-language modeling framework that extracts planning relevant built environment features without requiring labeled data or retraining. Central to this capability is zero-shot classification, which supports the extraction of higher-level, concept-aligned urban design features.

### 2.3. Vision-Language Models and Zero-Shot Classification

The emergence of vision-language models (VLMs), particularly CLIP, represents a significant advancement in applying computer vision to urban analysis. CLIP learns joint representations of images and text from large collections of image-caption pairs, enabling the classification of new images based on natural-language descriptions rather than expert classification (Radford et al., 2021). This zero shot capability allows users to define features through text prompts, such as “a street with wide sidewalks,” and enables the model to assess how well an image matches the description.

Zero-shot VLMs differ from traditional supervised models in several key ways. They do not require labeled datasets to define new categories, enabling rapid expansion of the feature space without extra data collection. Their use of natural language improves interpretability and aligns with the conceptual constructs used in planning. Because CLIP is pretrained on a massive, diverse corpus of image-text pairs, it provides representations that generalize across different cities, seasons, and image sources, even without task-specific retraining. These models allow the extraction of many features from imagery using a single underlying model, eliminating the need to develop separate detectors for each task (Wang et al., 2024).

Despite these advantages, the adoption of VLMs in urban analytics remains limited as the application of CLIP and related models has focused primarily on narrow domains. Previous studies have demonstrated the potential of machine learning and multimodal models to support urban design analysis. Zhang et al. (2024) use multimodal large language models (LLMs) and street-level imagery to estimate perceived urban safety, showing that vision language models can capture subjective streetscape attributes relevant to planning (Zhang et al., 2024). Yan et al. (2024) develop UrbanCLIP, a text-enhanced contrastive model that aligns imagery and language to enable semantic profiling of urban regions, illustrating how CLIP-based approaches can support urban-scale representation learning (Yan et al., 2024). While these studies demonstrate the promise of VLMs for specific analytic objectives, they do not provide comprehensive or scalable pipelines that translate zero-shot outputs into actionable planning frameworks. In contrast, our method extracts a broad set of urban design features and aggregates them into interpretable, block-level metrics aligned with established built-environment indices.

### 2.4. The SoP Index

The SoP Index serves as the interpretive layer that links zero shot computer vision predictions to planning-relevant measures of the built environment. It is conceptually grounded in the Irvine-Minnesota Inventory (IMI), one of the most comprehensive and empirically validated instruments for assessing micro-scale urban design features associated with walk ability (Day et al., 2006). In addition to this conceptual foundation, the SoP Index is supported by evidence showing that composite, walkability-oriented measures of urban form consistent with IMI’s multi-dimensional framework are strongly associated with economic vitality, public health, and travel behavior in real-world planning contexts, including large U.S. metropolitan regions (Alfonzo and Leinberger, 2012).

Subsequent validation studies further demonstrate that IMI based dimensions exhibit high inter-rater reliability and construct validity when aggregated into higher-order constructs, reinforcing their utility as robust, planning-relevant indices rather than ad hoc feature collections (Alfonzo et al., 2014).

Drawing on this conceptual and empirical foundation, the SoP Index integrates numerous discrete built environment indicators, based on the IMI framework and subsequent refinements, into ten composite dimensions that represent key aspects of urban design, such as connectivity, density, and safety. Collectively, these ten dimensions comprise the SoP Index and are employed to summarize overall urban form and quality. Table 1 presents and defines these ten subindices.

**Table 1.** SoP Index Subdimensions

| SoP Subindex                | What It Represents                                                             | What It Measures                                                                                                                                           |
|-----------------------------|--------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Density                     | Intensity of development and activity within a block or its immediate context. | Features related to the concentration of buildings, land uses, and people.                                                                                 |
| Connectivity                | Potential of mobility through the street network.                              | Features capturing street network structure and permeability at the block level, including intersection presence, block structure, and network continuity. |
| Form                        | Physical configuration and enclosure of the street environment.                | Features related to built form characteristics such as building placement, street enclosure, and edge definition.                                          |
| Proximity                   | Closeness of destinations that support walking and daily activity.             | Features reflecting whether common destinations and land uses are present within a block.                                                                  |
| Parks & Public Spaces       | Availability of publicly accessible open spaces.                               | Features related to presence and visibility of parks, plazas, and civic open spaces along a block.                                                         |
| Recreational Facilities     | Opportunities for structured physical activity.                                | Features related to recreational facilities such as courts, fields, and play areas.                                                                        |
| Pedestrian & Bike Amenities | Infrastructure supporting non-motorized travel.                                | Features related to non-motorized accessibility such as sidewalks, crossings, bicycle infrastructure, and pedestrian amenities.                            |
| Traffic Safety              | Perceived protection from vehicle-related harm.                                | Features reflecting traffic calming, separation from vehicles, and pedestrian crossing support.                                                            |
| Aesthetics                  | Visual and experiential quality of the street.                                 | Features related to visual appeal, maintenance, greenery, and architectural character.                                                                     |
| Personal Safety             | Perceived safety from crime and disorder.                                      | Features related to lighting, visibility, upkeep, and signs of disorder.                                                                                   |

Source(s): Alfonzo, 2014

In our pipeline, VLM-derived feature predictions are aggregated at the block level, mapped into these ten dimensions, and further aggregated into our single SoP Index. This mapping grounds the outputs in a framework previously validated across multiple empirical studies. It ensures that raw VLM outputs can be translated into constructs widely used in transportation, planning, and urban design research. It also enables standardized cross-block and cross-city benchmarking.

The integration of VLM-based feature extraction with the SoP Index automates a traditionally expensive and labor intensive process while retaining theoretical and methodological continuity with established planning practice. This combination of scalable AI methods and an empirically grounded interpretive framework positions the pipeline to meet municipal needs for actionable, transparent, and cost effective urban form assessments.

### 3. Methodology

The Phase I research and development work integrates three core components: (1) a low-cost, open-source image capture system that enables cities to collect their own street-level imagery, (2) a zero-shot CLIP-based feature extraction pipeline that identifies more than 150 built environment features without requiring labeled training data, and (3) a block-level aggregation process aligned with the SoP Index. The methodology builds on SoP's existing urban analytics tools, including the SoP Index and software. Together, these components form a scalable workflow that replaces traditional supervised learning with a flexible zero-shot approach tailored to the complexities of urban imagery.

#### 3.1. Open-Source Image Capture and Preprocessing

To enable scalable and affordable data collection, an open-source image-capture system was developed using consumer-grade devices and free software. The deployment utilized Samsung Galaxy A15 smartphones running OpenCamera, mounted on vehicles traversing predetermined routes throughout the Philadelphia metropolitan area. Images were captured at fixed intervals and automatically geotagged. This approach facilitated the collection of more than 2 million images over 10 days of winter and spring fieldwork, covering over 1,000 linear miles of urban and suburban streets. The use of low-cost hardware and open-source software ensured that the image capture system remained accessible to municipalities and repeatable across various contexts.

After collection, images were processed using custom scripts to extract metadata, correct timestamps when necessary, and filter out corrupted or incomplete files. Each geotagged image was then mapped to the nearest street segment (hereafter referred to as a "block") using geodesic joins within a three-meter tolerance. Incorporating direction-of-travel metadata and duplicate image removal procedures ensured that block assignments were spatially consistent and minimally redundant.

#### 3.2. Evolution from Supervised Learning to Zero-Shot Classification

Our initial approach to extracting built environment features relied on conventional supervised learning with custom-trained convolutional neural networks. This process required curating labeled datasets for each feature of interest, which was slow, resource-intensive, and ultimately inadequate for the wide range of urban indicators required in planning practice. Despite significant annotation efforts, supervised models performed inconsistently across diverse visual contexts and lacked flexibility to accommodate new features without additional labeled data. These limitations motivated a methodological shift toward zero-shot classification using vision-language models.

CLIP, trained on a large corpus of image-text pairs, provided a promising alternative. Instead of developing and training feature-specific detectors, we defined new categories using natural-language prompts that described the presence or qualities of urban features. This approach enabled scaling from a limited set of supervised models to a library of over 150 features encompassing infrastructure, context, and travel behavior. More importantly, it facilitated rapid iteration on feature definitions through text adjustments rather than model retraining.

#### 3.3. Mitigating Privacy Risks

In contrast to previously used methods such as fixed sensors and passively collected mobile device data, our vision-language modeling workflow was designed to reduce privacy risk through data governance, technical safeguards, and aggregation choices. All imagery was collected in public rights-of-way and stored within SoP's secure, encrypted infrastructure with restricted access. While encrypted cloud storage is a viable option, this pilot deployment emphasized a conservative, privacy-first configuration and utilized local inferencing hardware (Apple M3 Mac Studio) which could be easily replicated by municipalities or other researchers at

low cost. Images were down-sampled to 336×336 pixels prior to CLIP-based feature extraction to standardize inputs and limit fine-grained detail, though resolution reduction alone does not eliminate personally identifiable information (PII).

### ***3.4. Iterative Development of Zero-Shot Feature Extraction: Prompt Engineering***

While CLIP supports zero-shot classification, its performance depends significantly on the phrasing of textual prompts. Initial experiments indicated that overly brief or generic prompts yielded inconsistent outcomes, motivating an iterative process to refine prompt wording for each feature. For example, a generic prompt such as “driveway” often returned false positives for alleys, curb cuts, or gaps between buildings. Slightly more precise prompts, such as “street driveway many” and “street driveway few,” reduced these errors and more accurately reflected distinctions relevant to urban design analysis. Prompts were systematically adjusted based on model output inspection and calibrated to correspond with SoP feature definitions.

For multi-class features, sets of prompts were developed to express ordered intensity levels. For example, to capture variation in parking lot scale, prompts such as “street parking lot small,” “street parking lot medium,” and “street parking lot large” were tested. Similarly, prompts such as “street parking garage” and “street parking garage shop restaurant” facilitated the distinction between parking structures with and without active ground-floor uses.

This iterative refinement enhanced semantic clarity and reduced misclassification, ensuring that CLIP’s zero-shot outputs aligned with the distinctions necessary for consistent block-level feature prediction.

### ***3.5. Threshold Calibration***

Zero-shot similarity scores produced by CLIP represent a continuous measure of how well an image matches a prompt. However, using raw similarity values for classification produced substantial overprediction in early iterations. In many cases, CLIP returned moderate similarity scores for features that were rare or absent in most images, leading to inflated block-level detection frequencies. To address this, we calibrated feature-specific thresholds based on a combination of empirical evaluation and domain expertise. In this context, CLIP similarity scores are treated as confidence measures indicating the strength of alignment between an image and a given textual prompt. Detection thresholds define the minimum similarity value required to classify a feature as present, effectively converting continuous confidence scores into urban indicators.

Threshold calibration involved reviewing model outputs across diverse street typologies and adjusting decision boundaries to reflect realistic prevalence rates. For example, features such as street vendors or bicycle parking are uncommon on many corridors; applying a uniform threshold would incorrectly mark them as present across large portions of the city. Conversely, features like paved sidewalks or building frontages are ubiquitous and require lower thresholds to avoid undercounting. This iterative process established a stable mapping between similarity scores and categorical feature values.

### ***3.6. From Image-Level Volumes to Block-Level Presence***

Our initial modeling framework aimed to estimate volumetric quantities, such as the number of pedestrians, trees, or lanes, directly from individual images. However, we determined that CLIP demonstrates greater reliability in detecting the presence of features than in estimating precise quantities (Radford et al., 2021; Yuksekgonul et al., 2023). This observation is consistent with established practices in urban design assessment, such as the Irvine-Minnesota Inventory framework, which emphasizes block-level characteristics over pixel-level object counts.

Consequently, we transitioned from an image-level volumetric approach to a block-level presence-and-intensity model. CLIP was employed to classify each image as yes (present) or no (not present) for a given feature. These classifications were aggregated across all images associated with a block, with the number of affirmative detections serving as a proxy for feature

intensity. For example, benches were classified as present or absent at the image level, and aggregating affirmative detections across all images within a block provided insight into the presence and relative intensity of benches along blocks. This aggregation strategy leveraged CLIP's strengths and facilitated meaningful distinctions in urban design characteristics at the block level.

### ***3.7. Feature Library and Modular Architecture***

The system generates over 150 features encompassing infrastructure (sidewalk width, bike lanes, and lighting), context (vegetation, enclosure, and frontage), and behavior (pedestrian or vehicle presence). Each feature is specified by a set of prompts, classification thresholds, and aggregation rules. These configurations are managed via a no-code interface (Airtable), which stores the feature definitions, labels, and model selection parameters used by the pipeline, enabling rapid adjustments without requiring changes to the underlying code. The modular architecture permits practitioners to also do single feature extraction by inputting custom feature labels and prompt definitions for our zero-shot models to extract from images.

### ***3.8. Block-Level Aggregation and Integration with the SoP Index***

Following image-level classification, features are aggregated to their respective blocks. Aggregation rules vary by feature type but typically assign the block-level value based on either the most common (modal) classification or the maximum value observed across images. Blocks with a greater number of assigned images yield more robust feature estimates; however, the pipeline accommodates varying image densities.

Block-level predictions are subsequently coded according to the SoP Index algorithm and mapped to their corresponding SoP dimensions. The resulting dimension scores can then be combined to produce composite indices for block-level analysis. Integration with the SoP Index ensures that outputs are interpretable, comparable across neighborhoods, and actionable for planning and design purposes.

Collectively, these methodological components form a flexible, scalable, and interpretable pipeline for extracting urban design features from imagery at the citywide scale.

## **4. Results**

The primary focus of this study is the performance of the feature-extraction system, which serves as the foundation for analyses relevant to urban planning. The performance of the zero-shot vision-language model (VLM) pipeline was evaluated across three dimensions: (1) image-level feature classification accuracy, (2) block-level spatial coherence and patterning, and (3) large-scale deployment across Philadelphia.

### ***4.1. Image-Level Classification Performance***

To assess the accuracy of the CLIP-based feature extraction pipeline, we developed a manually labeled validation dataset, comprising more than 2,800 images from diverse street typologies across Philadelphia. Each image was annotated for the presence or absence of 114 features, excluding features that could not be classified with yes/no, such as "parking lot size," in the final evaluation set. Two accuracy metrics are reported: unbalanced accuracy, which measures the proportion of correct predictions across all classes, and value-balanced accuracy, which reweights performance based on feature prevalence to account for rare or infrequent conditions.

Across all evaluated features, the pipeline achieved an average unbalanced accuracy of approximately 90% and a value-balanced accuracy of 88%. These results indicate that our CLIP-based method can achieve strong performance while avoiding the computational and data-collection burdens of supervised approaches. Zero-shot classification, combined with prompt

and threshold calibration, enabled CLIP to produce reliable predictions across a wide range of built environment features without any feature-specific training data.

Performance varied by feature class. Structurally consistent features such as driveways, sidewalk paving, and buffers created by parked cars or trees exhibited the strongest improvements, with several features achieving accuracies above 90%. Features requiring finer distinctions among multiple categories, such as bike lane type, demonstrated more moderate but still substantial improvements following prompt refinement and model tuning. In contrast, features that are inherently block-level constructs, such as sidewalk presence and completeness, showed strong image-level recognition but require additional aggregation logic to translate image observations into reliable block-level classifications.

A key advantage of the zero-shot pipeline is its capacity for rapid, feature-specific improvement without retraining. Prompts, model ensembles, and classification thresholds can be adjusted independently for each feature, providing the advantage of iterative accuracy gains with minimal computational overhead. This flexibility allowed us to substantially improve performance for several infrastructure features that historically posed challenges for supervised models, particularly those involving small objects, partial occlusions, or fine-grained distinctions in streetscape conditions (e.g. pedestrian-activated sign). Table 2 summarizes improvements for a subset of twelve infrastructure features originally targeted for enhancement. These results highlight how zero-shot methods, combined with prompt refinement and threshold calibration, can yield large gains in predictive reliability while avoiding the costs associated with large-scale labeling and retraining.

**Table 2.** Balanced accuracy of legacy supervised models (block level) and new zero-shot CLIP models (image level) for 12 targeted infrastructure features.

| Feature                       | Old Pipeline Accuracy | New Pipeline Accuracy |
|-------------------------------|-----------------------|-----------------------|
| Bike lane type                | 33.3%                 | 62%                   |
| Parked cars as buffer         | 63%                   | 100%                  |
| Trees as buffer               | 67%                   | 100%                  |
| Driveways                     | 33.9%                 | 94.0%                 |
| Pedestrian-activated sign (A) | 50.0%                 | 100%                  |
| Pedestrian-activated sign (B) | 50.0%                 | 100%                  |
| Sidewalk presence*            | 74%                   | 72%                   |
| Sidewalk completeness         | 54.8%                 | 78%                   |
| Sidewalk paving               | 54.9%                 | 100%                  |
| Traffic calming (parked cars) | 76%                   | 90%                   |
| Zebra striping (A)            | 50.0%                 | 93%                   |
| Zebra striping (B)            | 50.0%                 | 93%                   |

\*The "Sidewalk presence" feature only had unbalanced accuracy recorded for the old pipeline; balanced accuracy would likely have been lower.

#### 4.2. Spatial Coherence and Block-Level Patterns

Block-level aggregation produces spatially coherent patterns that correspond with established urban morphologies and corridor characteristics (Naik et al., 2014). In our block-level aggregation, blocks along major commercial corridors exhibited dense built environment signatures, including high enclosure, frequent lighting, and active frontages. In contrast, residential neighborhoods display consistent features across adjacent blocks, including moderate tree canopies and reduced commercial intensity.

This spatial coherence offers indirect validation of the system's reliability, as features detected independently at the image level aggregate into block-level profiles that represent meaningful urban structure rather than noise. Correlation analyses further indicate that CLIP-derived features are logically related. For instance, blocks with high enclosure scores also tend

to exhibit reduced setbacks and greater continuity of building facades, a pattern that aligns with urban design theory. Likewise, blocks with high vegetation coverage frequently display lower built density, reflecting the spatial organization of Philadelphia's neighborhoods.

#### ***4.3. Large-Scale Deployment Across Philadelphia***

We applied the full pipeline to a geospatially filtered subset of approximately 80,000 images drawn from a larger corpus of two million collected across the SEPTA bus network. This subset produced feature predictions for up to 6000 blocks, representing all major street typologies in the region and a substantial share of the city's transportation corridors.

Runtime analysis indicated that the pipeline could process thousands of predictions per minute on a single GPU-equipped consumer machine. These results demonstrate the feasibility of applying the methodology not only to large cities but also to multi-city or regional deployments.

Our pipeline outputs include a structured dataset of 150 block-level features, along with a GeoJSON dataset suitable for spatial analysis and integration into planning workflows. This infrastructure enables agencies to generate up-to-date assessments of urban design quality and to evaluate conditions along current or proposed transportation corridors.

### **5. Discussion**

The development of this zero-shot vision-language model pipeline required navigating a series of methodological, conceptual, and practical challenges. The implementation process that led to our final system involved iterative refinement of image capture methods, prompt design, model ensembles, and aggregation techniques to ensure that zero-shot predictions aligned with planning-relevant feature definitions and performed reliably across diverse urban contexts. This pipeline also introduces broader implications for urban analytics by demonstrating how general-purpose vision-language models can support scalable, interpretable, and resource-efficient assessments of the built environment. Finally, we identified limitations that emerged during development. Particularly, those related to feature rarity, semantic ambiguity, and block-level inference, and outlined several directions for future research to enhance robustness, generalization, and application across cities.

#### ***5.1. From Supervised Constraints to Zero-Shot Flexibility***

The transition from supervised learning to zero-shot classification marked a turning point in our methodology. Early prototypes using supervised models yielded modest success in detecting a handful of well-defined features. However, the approach proved too rigid and resource-intensive to support the diversity of indicators needed for comprehensive urban design assessment. Developing a separate, labeled dataset for each feature, combined with the need to retrain models whenever planning priorities shifted, introduced delays incompatible with real-world municipal workflows.

With the adoption of CLIP, we scaled the system through prompt engineering rather than additional labeling, enabling a significantly faster and more adaptable workflow. The zero-shot paradigm enabled us to rapidly explore new built environment features, refine definitions, and improve interpretability without retraining. This shift highlights a broader methodological insight: for many urban analytics tasks, the bottleneck is not model architecture but the availability of flexible, semantically meaningful representations of planning concepts. Zero-shot VLMs alleviate this constraint by enabling features to be "programmed" using natural language rather than annotated examples.

## 5.2. Iterative Refinements and Practical Discoveries

The final design of our pipeline emerged through iterative experimentation with CLIP outputs, informed by urban design expertise and empirical evaluation. Several key insights shaped the system.

First, we learned that CLIP excels at identifying the presence of features but is less reliable at estimating precise counts or volumetric measures. This realization prompted a shift from image-level volume estimation to evaluating intensity and presence of features via block-level aggregation, aligning more closely with established urban auditing practices such as the Irvine-Minnesota Inventory.

Second, prompt engineering proved essential. While CLIP can interpret natural language descriptions, the specificity and clarity of prompts significantly influence performance. Vague or overly narrow prompts yielded inconsistent predictions, while carefully crafted prompts captured the essential semantics of planning-relevant features. Through repeated refinements, we developed prompt sets that balanced descriptive richness with conceptual clarity.

Third, the project demonstrated that threshold calibration is not merely a technical detail but a conceptual task requiring domain knowledge. Determining where to draw the line between “present” and “absent” depends on both similarity scores and the expected prevalence of features in real urban environments. Thresholds that made statistical sense were sometimes misaligned with urban design realities, leading to implausible block-level patterns. The calibration process thus demanded a continual interplay between quantitative evaluation and qualitative judgment.

These iterative lessons underscore the importance of interdisciplinary collaboration in developing AI systems for planning applications. The pipeline did not emerge solely from algorithmic optimization but from a close engagement between modeling techniques and the conceptual frameworks used in urban design.

## 5.3. Implications for Urban Planning Practice

The resulting system carries implications for municipal planning, urban analytics, and the commercialization of such tools. Because classification is zero-shot, the marginal cost of a new feature is a prompt and a validation pass rather than an annotation campaign and a training run. By leveraging open-source image capture and zero-shot classification, municipalities gain a privacy-preserving, cost-effective alternative for monitoring street conditions. The cost efficiency also means this pipeline can be delivered at a price public agencies can budget. State of Place has already commercialized this path: the Index and recommendation platform has served nearly 50 citymakers, and the automated VML pipeline has supported analysis across 50,000+ blocks nationwide.

Moreover, the modular design allows users to customize feature sets or develop new indicators in response to local priorities. For example, a project focused on heat resilience might require additional prompts on shade structures or surface albedo (e.g., reflective materials), while a project focused on public realm activation might emphasize street furniture or storefront transparency. Our pipeline’s flexibility reduces reliance on vendor-defined, fixed metrics and empowers agencies to control their analytical frameworks.

This methodology is also the first step to greater refinement and increased accuracy. Keyword and prompt iteration, feature-specific confidence thresholds, and binary image-level detection with block-level aggregation are all inference-time components. Instead of retraining models for better accuracy, this pipeline asks only to re-run. Precision can therefore focus on the features a given customer actually depends on and improve after deployment in response to observed failures.

Feature extraction represents only the initial stage. Because the features are human-readable and aligned with the SoP Index, they aggregate into a validated metric spanning 144 features across ten urban design dimensions. The forecasting models connect the index to outcomes cities are accountable for, including travel behavior, collisions, chronic disease, vehicle miles

traveled (VMT), transit ridership, etc. The recommendation engine then systematically identifies which design changes will most effectively advance a user-specified target. Together, the VML pipeline, the SoP Index, the forecasting models, and the recommendation engine give cities a granular understanding of the built environment and, more importantly, insight into how to improve it.

#### **5.4. Limitations**

While our methodology is promising, it also has several limitations. Consumer-grade equipment introduced challenges such as inconsistent write speeds, weather sensitivity, and the need for daily hardware swaps. Despite this, the overall system demonstrates sufficient robustness to support citywide data collection. Insights from this proof-of-concept phase informed the design of a more durable Raspberry Pi-based capture platform for future deployments.

In addition to equipment, VLMs also introduced limitations. Like all models trained on large-scale internet imagery, CLIP inherits biases from its training corpus. These biases may influence predictions in ways that correlate with geographic, cultural, or socioeconomic differences across neighborhoods. Additional work is needed to systematically evaluate and mitigate these effects in urban analytics contexts.

The reliance on threshold calibration introduces a degree of subjectivity, as thresholds reflect domain-informed judgments about feature prevalence and semantic boundaries. Although this approach improves alignment with planning concepts, it may reduce reproducibility across cities unless standardized procedures are developed further.

Some features remain challenging for zero-shot models, particularly those involving small objects, partially occluded items, or complex spatial relationships. Although block-level aggregation mitigates some of these issues, certain urban indicators, such as fine-grained accessibility features or subtle design elements, may require alternative modeling techniques or multimodal data sources.

Finally, although the proposed pipeline avoids many privacy risks associated with mobile device data and fixed sensors, privacy remains a key limitation. The pipeline includes an optional open-source blurring module for obscuring faces and other identifiable elements (not activated during this work due to the controlled storage environment and access). The system is designed for built-environment feature extraction rather than identity-level analysis, with no facial recognition, license plate detection, or tracking components implemented, and all outputs are aggregated at the block level prior to reporting. Street-level imagery can capture identifiable people, vehicles, or property, raising concerns about consent and secondary use even for analytics. Our analysis relies on aggregated, block-level data without storing or examining individual identities. However, upstream data collection still demands strict governance, robust anonymization, and compliance with local laws. The use of on-device filtering, automated blurring, and clear retention policies can reinforce privacy and uphold ethical and legal standards.

#### **5.5. Future Directions**

The methodological advances presented here have already begun to inform the next stage of development beyond Phase I. In particular, the zero-shot VLM pipeline developed in this study provides the foundation for ongoing Phase II work to extend automated built-environment measurement into a broader multimodal urban data and decision-support system. A key component of this effort is IRIS (Intelligent Remote Imaging & Sensing), an open source mobile, multi-sensor data collection platform designed to support scalable street-level data acquisition. An initial IRIS prototype has already been deployed in a real-world customer environment, collecting data across more than 1,000 city blocks and demonstrating end-to-end integration between field data collection, the VLM processing pipeline, and downstream visualization tools.

Building on these results, Phase II expands the framework beyond RGB imagery by incorporating additional sensing modalities, including LiDAR, inertial, and environmental inputs.

These modalities can complement visual inference for features that are difficult to estimate from imagery alone, including depth, clearance, and dynamic conditions. Phase II also extends the analytical pipeline through optimized QA/QC, forecasting models, and a scenario-based recommendation engine that predicts outcomes and benefits.

The framework may ultimately support more flexible forms of interaction with urban data as well. One example is integrating our pipeline with LLMs to enable natural-language querying of street-level imagery. Rather than relying solely on predefined feature sets, planners could pose open-ended questions (e.g., “identify streets lacking shade near transit stops”).

Importantly, these directions are no longer solely theoretical extensions of the Phase I research. The development and deployment of IRIS, ongoing pipeline improvements, and Phase II expansion can translate the zero-shot methodology into an operational, commercially scalable platform. Future work focuses on reduced reliance on third-party imagery, supporting recurring and geographically adaptable data collection, and integrating these capabilities with an existing commercial decision-support platform. The current framework serves not only as a methodological contribution, but also as the technical foundation for a deployable system through which cities can continuously observe, interpret, and design the built environment.

## **6. Conclusion**

Developed through State of Place’s USDOT SBIR Phase I project, this methodology introduces a scalable, interpretable, and privacy-preserving framework for extracting built environment information from street-level imagery using zero-shot vision-language models. By combining a low-cost image-capture system with a flexible CLIP-based feature-extraction pipeline and block-level aggregation aligned with the SoP Index, our pipeline enables municipalities to generate detailed assessments of urban design conditions without relying on expensive sensor networks, proprietary mobile device data, or labor-intensive manual audits.

Our results demonstrate that zero-shot classification can produce high-quality, planning-relevant features across a wide range of infrastructure and contextual categories. The system achieves strong image-level accuracy, generates spatially coherent block-level patterns, and can be deployed at a citywide scale using consumer-grade hardware. These findings underscore the potential for VLMs to serve as general-purpose urban analytics tools that support diverse planning tasks, from corridor evaluations to equity analyses to long-term monitoring.

Beyond its technical contributions, the framework advances a broader vision for urban data systems: one in which municipalities retain autonomy over data collection, privacy is preserved without sacrificing analytical richness, and AI serves as a practical tool to improve the equity, safety, and livability of streets. As cities confront increasingly complex challenges from climate adaptation to multimodal transportation, the ability to obtain detailed, interpretable, and up-to-date data on the built environment will become indispensable.

Building on these Phase I results, ongoing Phase II work is extending the framework through IRIS, a mobile multi-sensor data-capture system, while integrating multimodal sensing, forecasting, and decision-support capabilities with State of Place’s existing commercial platform. Future work will also focus on expanding the feature library and developing interactive tools that allow planners to query and interpret streetscape conditions through natural language. By continuing to bridge advances in machine learning with the practical needs of urban planning, this line of research offers a promising pathway toward more responsive, data-informed cities.

## **7. Acknowledgements**

The present text arises within the framework of a Small Business Innovation Research grant funded by the US Department of Transportation.

## References

- Alfonzo, M., Guo, Z., Lin, L., & Day, K. (2014). Walking, obesity and urban design in Chinese neighborhoods. *Preventive Medicine*, 69(Suppl. 1), S79–S85. <https://doi.org/10.1016/j.ypmed.2014.10.002>
- Alfonzo, M., & Leinberger, C. B. (2012). *Walk this way: The economic promise of walkable places in metropolitan Washington, D.C.* Brookings Institution. <https://www.brookings.edu/articles/walk-this-waythe-economic-promise-of-walkable-places-in-metropolitan-washington-d-c/>
- Barbeau, S. J., Lozano, W., Turner, S., Sener, I. N., Maness, M., & Luong, T. (2022). *Evaluating the accuracy of smartphone-based travel behavior data* (Research Report No. 14). National Institute for Congestion Reduction. <https://doi.org/10.5038/CUTR-NICR-Y1-3-1>
- Biswas, G., & Roy, T. K. (2023). Measuring objective walkability from pedestrian-level visual perception using machine learning and GSV in Khulna, Bangladesh. *Geomatics and Environmental Engineering*, 17(6), 5–34. <https://doi.org/10.7494/geom.2023.17.6.5>
- Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4, 10. <https://doi.org/10.1140/epjds/s13688-015-0046-0>
- Day, K., Boarnet, M., Alfonzo, M., & Forsyth, A. (2006). The Irvine–Minnesota Inventory to measure built environments: Development. *American Journal of Preventive Medicine*, 30(2), 144–152. <https://doi.org/10.1016/j.amepre.2005.09.017>
- de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 1376. <https://doi.org/10.1038/srep01376>
- de Moraes Ferrari, G. L., Kovalskys, I., Fisberg, M., Gómez, G., & Rigotti, A. (2020). Socio-demographic patterns of public, private and active travel in Latin America: Cross-sectional findings from the ELANS study. *Journal of Transport & Health*, 16, 100788. <https://doi.org/10.1016/j.jth.2019.100788>
- Diop, E. B., Chenal, J., Tekouabou, S. C. K., & Azmi, R. (2022). Crowdsourcing public engagement for urban planning in the Global South: Methods, challenges and suggestions for future research. *Sustainability*, 14(18), 11461. <https://doi.org/10.3390/su141811461>
- Doiron, D., Setton, E. M., Brook, J. R., Kestens, Y., McCormack, G. R., Winters, M., Shoostari, M., Azami, S., & Fuller, D. (2022). Predicting walking-to-work using street-level imagery and deep learning in seven Canadian cities. *Scientific Reports*, 12, 18380. <https://doi.org/10.1038/s41598-022-22630-1>
- Du, R., Santi, P., Xiao, M., Vasilakos, A. V., & Fischione, C. (2019). The Sensable City: A survey on the deployment and management for smart city monitoring. *IEEE Communications Surveys & Tutorials*, 21(2), 1533–1560. <https://doi.org/10.1109/COMST.2018.2881008>
- Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, 76(3), 265–294. <https://doi.org/10.1080/01944361003766766>
- Ewing, R., & Cervero, R. (2017). “Does compact development make people drive less?” The answer is yes. *Journal of the American Planning Association*, 83(1), 19–25. <https://doi.org/10.1080/01944363.2016.1245112>
- Florida Department of Transportation. (n.d.). *Traffic signal budgeting and costs – tables*. <https://wbt.dot.state.fl.us/ois/tsmo/TrafficSignalBudgetingCostsAccessibleTables.htm>

- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50), 13108–13113. <https://doi.org/10.1073/pnas.1700035114>
- González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779–782. <https://doi.org/10.1038/nature06958>
- Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4), 682–703. <https://doi.org/10.1068/b35097>
- He, H., Guler, S. I., & Menendez, M. (2016). Adaptive control algorithm to provide bus priority with a pre-signal. *Transportation Research Part C: Emerging Technologies*, 64, 28–44. <https://doi.org/10.1016/j.trc.2016.01.009>
- Hosseini, M., Miranda, F., Lin, J., & Silva, C. T. (2022). CitySurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 79, 103630. <https://doi.org/10.1016/j.scs.2021.103630>
- Ibrahim, M. R., Haworth, J., & Cheng, T. (2020). Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities*, 96, 102481. <https://doi.org/10.1016/j.cities.2019.102481>
- Inglés-Romero, J. F., Doménech-Asensi, G., Cano-Baños, M.-D., & Morales-Esteras, V. (2018). Traffic metrics at urban intersections using a low-cost image processing approach. In *Proceedings of the 12th International Conference on Distributed Smart Cameras* (Article 9, pp. 1–7). Association for Computing Machinery. <https://doi.org/10.1145/3243394.3243693>
- Ismagilova, E., Hughes, L., Rana, N. P., & Dwivedi, Y. K. (2022). Security, privacy and risks within smart cities: Literature review and development of a smart city interaction framework. *Information Systems Frontiers*, 24(2), 393–414. <https://doi.org/10.1007/s10796-020-10044-1>
- Joy, J., McGoldrick, C., & Gerla, M. (2016). *Mobile privacy-preserving crowdsourced data collection in the smart city* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1607.02805>
- Le, H. T. K., Buehler, R., & Hankey, S. (2019). Have walking and bicycling increased in the US? A 13-year longitudinal analysis of traffic counts from 13 metropolitan areas. *Transportation Research Part D: Transport and Environment*, 69, 329–345. <https://doi.org/10.1016/j.trd.2019.02.006>
- Liu, L., Wang, H., & Wu, C. (2016). *A machine learning method for the large-scale evaluation of urban visual environment* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1608.03396>
- Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. A. (2014). Streetscore—Predicting the perceived safety of one million streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 793–799). IEEE. <https://doi.org/10.1109/CVPRW.2014.121>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Benchmarking and comparative measurement for effective performance management by transportation agencies*. The National Academies Press. <https://doi.org/10.17226/25365>
- Ning, H., Ye, X., Chen, Z., Liu, T., & Cao, T. (2022). Sidewalk extraction using aerial and street view images. *Environment and Planning B: Urban Analytics and City Science*, 49(1), 7–22. <https://doi.org/10.1177/2399808321995817>

- Otero, B. G., & Verhulst, S. G. (2025, September 13). *Why data governance and collaboration are essential for the future of urban digital twins*. The GovLab Blog. <https://blog.thegovlab.org/why-data-governance-and-collaboration-are-essential-for-the-future-of-urban-digital-twins>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Aspell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR. <https://proceedings.mlr.press/v139/radford21a.html>
- Sallis, J. F., Cerin, E., Conway, T. L., Adams, M. A., Frank, L. D., Pratt, M., Salvo, D., Schipperijn, J., Smith, G., Cain, K. L., Davey, R., Kerr, J., Lai, P.-C., Mitáš, J., Reis, R., Sarmiento, O. L., Schofield, G., Troelsen, J., Van Dyck, D., . . . Owen, N. (2016). Physical activity in relation to urban environments in 14 cities worldwide: A cross-sectional study. *The Lancet*, 387(10034), 2207–2217. [https://doi.org/10.1016/S0140-6736\(15\)01284-2](https://doi.org/10.1016/S0140-6736(15)01284-2)
- Seiferling, I., Naik, N., Ratti, C., & Proulx, R. (2017). Green streets—Quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, 165, 93–101. <https://doi.org/10.1016/j.landurbplan.2017.05.010>
- Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58, 162–177. <https://doi.org/10.1016/j.trc.2015.04.022>
- U.S. Department of Transportation, ITS Joint Program Office. (2012, January 17). *The average installation cost per intersection of an adaptive traffic control system (ATCS) is \$65,000*. <https://www.itskrs.its.dot.gov/2012-sc00249>
- Wang, C., Jia, R., Liu, X., & Song, D. (2024). *Benchmarking zero-shot robustness of multimodal foundation models: A pilot study* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2403.10499>
- Wu, B., Yu, B., Shu, S., Liang, H., Zhao, Y., & Wu, J. (2021). Mapping fine-scale visual quality distribution inside urban streets using mobile LiDAR data. *Building and Environment*, 206, 108323. <https://doi.org/10.1016/j.buildenv.2021.108323>
- Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., & Liang, Y. (2024). UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024* (pp. 4006–4017). Association for Computing Machinery. <https://doi.org/10.1145/3589334.3645378>
- Yüksekgönül, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2023). When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. <https://openreview.net/forum?id=KRLUvxh8uaX>
- Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., & Ratti, C. (2024). Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, 114(5), 876–897. <https://doi.org/10.1080/24694452.2024.2313515>