

UC Berkeley

Research Reports

Title

StreetTransformer: A Computer Vision-based Tool for Analyzing and Retrieving Longitudinal Changes in Urban Streetscapes

Permalink

<https://escholarship.org/uc/item/1bw139gk>

Author

Atkins, Jon

Publication Date

2025-12-01

DOI

10.7922/G26Q1VNN

StreetTransformer: A Computer Vision-based Tool for Analyzing and Retrieving Longitudinal Changes in Urban Streetscapes

Author Jon Atkins, University of California, Berkeley

Advisor Dr. Maryam Hosseini, University of California, Berkeley

December 2025

UCB-ITS-PSR-2024-08

doi:10.7922/G26Q1VNN



TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. UCB-ITS-PSR-2024-04	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A	
4. Title and Subtitle Transit to California's National Parks: An Assessment of Visitation and Sociodemographic Barriers		5. Report Date May 2024	
		6. Performing Organization Code: ITS Berkeley	
7. Author(s) Winnie Zhuang http://orcid.org/0009-0004-7655-2124		8. Performing Organization Report No. N/A	
9. Performing Organization Name and Address Institute of Transportation Studies (ITS Berkeley) University of California, Berkeley 109 McLaughlin Hall MC1720 Berkeley CA 94720-1720		10. Work Unit No. N/A	
		11. Contract or Grant No. USDOT Grant 69A3551747109	
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		13. Type of Report and Period Whitepaper (August 2023 - May 2024)	
		14. Sponsoring Agency Code USDOT OST-R	
15. Supplementary Notes DOI:10.7922/G2K64GDB			
16. Abstract California boasts nine national parks that are home to a plethora of recreational, cultural, and professional opportunities. It is no surprise that national parks draw in millions of recreational and non-recreational trips from across the country and internationally, however how visitors choose to arrive at these federal lands is a major consideration for land management, policy, and infrastructure decisions. According to the National Park Service's Visitor Use Statistics, an overwhelming majority of visitors arrive by automobile. Alternative transportation options to national parks in California may increase visitor diversity and encourage the modal shift away from car travel. Additionally, transportation infrastructure has the potential to take up more natural spaces in national parks as population growth and recreational popularity increases in the coming years. Roadway widening, repaving, and other disruptive events can have negative impacts on adjacent ecosystems and communities, as well as contribute to public health issues. In this report, the existing conditions of transit to national parks in California are examined in relation to census tract-level sociodemographic and origin-destination data to reveal the spatial distribution of visitor origins and gaps in national park accessibility. From an equity perspective, findings suggest that median household income is proportional to the network distance between a given trip origin and national park destination, and household vehicle access and race are strongly correlated with national park visitation. By examining the current transit network to national parks, select characteristics of park visitation, and demographics of these visitors, this research aims to reveal opportunities for expanded transit connectivity to California's national parks. Recommendations are presented to provide guidance to decision makers in the realm of transportation access to recreational spaces.			
17. Key Words National parks, accessibility, public transit, transportation equity, travel behavior, demographics		18. Distribution Statement No restrictions	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 117	22. Price N/A

About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes the following partners: California State University, Long Beach, Northern Arizona University, Pima Community College, University of California, Berkeley, University of California, Davis, University of California, Irvine, University of California, Los Angeles, University of Hawaii, and University of Nevada, Las Vegas.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at *improving the mobility of people and goods throughout the region*. The program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

US Department of Transportation Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the US Department of Transportation's University Transportation Centers Program. However, the US Government assumes no liability for the contents or use thereof.

Disclosure

This research took place at the University of California, Berkeley from August 2023 to May 2024 and was supported by the Pacific Southwest Region University Transportation Center research program in the amount of \$7,500.

StreetTransformer

A Computer Vision-based Tool for Analyzing and Retrieving Longitudinal
Changes in Urban Streetscapes

Jon Atkins

PROFESSIONAL REPORT

Submitted in partial satisfaction of the requirements for the degree
of

MASTER OF CITY PLANNING

in the
Department of City and Regional Planning
of the

UNIVERSITY OF CALIFORNIA, BERKELEY

APPROVED

Maryam Hosseini
Assistant Professor, DCRP

Marta Gonzalez
Professor, CEE, DCRP

Date: Fall 2025

Abstract

Urban street design plays a critical role in safety, mobility, community and public welfare. Yet, systematic, large analysis of streetscape change remains challenging due to the scale, heterogeneity and unstructured nature of publicly available data. We here introduce STREETTRANSFORMER, a unified, open-source, multi-modal visual analytics system and framework for gathering, comparing, retrieving and analyzing urban streetscape state and change over time.

We construct a novel spatiotemporally grounded corpus for New York City’s more than 47,000 intersections by integrating longitudinal aerial imagery, segmentation masks, civic design documents and capital reconstruction project data spanning nearly two decades of New York City history. Innovatively, this corpus includes a large cache of 33,000 unstructured pdf document pages scraped from the NYC DOT website, which have been parsed, categorized and spatially linked to intersections using OCR and LLM-integrations. Additionally, project metadata, collected from NYC Open Data, is standardized and filtered to safety-focused interventions, grounding the Documents, Images and Masks with the city’s objective records.

We use this corpus to benchmark recent iterations of LLMs and VLMs and find that they struggle on basic streetscape analysis tasks that require structured output. Motivated by this, we develop a custom retrieval pipeline using CLIP-derived embeddings, difference vectors, FAISS-accelerated HNSW indexing and whitening-based re-ranking, enabling fast, accurate and efficient sub-second similarity and change retrieval across millions of instances.

We encapsulate these methods in an interactive web-based tool aimed toward practitioners and researchers, allowing them to query, compare and explore streetscape similarity and transformation through image-, text- and change-based queries. STREETTRANSFORMER demonstrates the feasibility and value of large-scale, multi-modal data unification for urban design research, and provides a foundation for more robust machine and human understanding of built-environment change.

1 Introduction

As the arteries of urban systems and bedrock of urban form, city streets play an omnipresent role in civic life and serve as inputs into countless models of urban activity. Street Design is an important but understudied component of urban form and can dramatically change dynamics across the transportation networks the streets compose. Design choices at the individual street level can have an out-sized impact on various metrics of safety (NACTO, n.d.), economic vitality (Cortright, 2009), mode-choice (Ewing & Cervero, 2010), social-connection (Leslie et al., 2007), and general well-being (Office of the Surgeon General, 2023).

Federal and local governments in the United States spend over \$100 B annually in capital costs towards street reconstruction (Urban Institute, 2024). However, despite this, the country continues to suffer from rising traffic fatalities (National Highway Traffic Safety Administration (NHTSA), 2024), lagging alternative mode-share adoption (Ewing & Cervero, 2010), and what experts have described as an overly car-centric culture (Jackson, 1985). The National Association of City Transportation Officials (NACTO) has called city intersections “one of the most critical and most complicated elements in [roadway] design” (NACTO, n.d.). Urban Designers, Transportation Planners and Engineers devote careers to researching, proposing improvements and arguing over intersection designs with minimal consensus on best practices.

Without comprehensive, structured, and accessible data on the geometry and configuration of city intersections, professionals, researchers, and advocates cannot make fully-informed decisions. The underlying data exists, but it heterogeneous, dispersed, and highly unstructured. The challenge is compounded by the difficulty of defining human-scale urban form in the first place. An intersection exists in three physical dimensions, changes over time, and presents different features to different users

depending on how and when they encounter it. Any comprehensive representation of intersection-level street design must account for all of this.

This paper presents STREETTRANSFORMER, an urban visual analytics system and replicable framework that instantiates a structured, searchable database of built-environment streetscape features across urban areas. Interactive, multi-modal, and open-source, it fuses distinct modal data (e.g., images, text, documents, diagrams) across time to create a comprehensive corpus of street design that can prove useful for researchers and practitioners alike. In doing so, it solves for a number of novel data science questions, driving advancements in the fields of Urban Planning, Engineering, and Computer Science.

The research resulted in two peer-reviewed workshop publications at leading international venues in computer science and geospatial analytics: an oral presentation at Google’s Urban Mobility Foundation Models (UMFM) workshop at ACM SIGSPATIAL 2025, and a poster presentation at the NeurIPS 2025 Urban AI workshop. Work is ongoing on expanding the analytic tool’s interface and improving ways to leverage the diverse methodologies with a targeted paper submission in a top tier data visualization venue this Spring.

2 Literature Review

Urban Design & Form

The impact of built-environment design on urban travel patterns has been well-studied in Urban Planning literature by Cervero and Kockelman (1997) in their landmark “3D’s” paper. This affirmed prior work by Southworth and Ben-Joseph (1995) finding the layout of streets can predetermine modal choices and travel patterns. Beyond transportation, street design factors are shown to impact safety (NACTO, n.d.), public health (Saelens & Handy, 2008), economic development (Cortright, 2009) and community (Leslie et al., 2007). Despite this, urban design features have long been under-measured as Ewing and Handy (2009) show in “Measuring the Unmeasurable.”

Computer Vision and Cities

The field of Computer Vision has risen in prevalence in recent years as the capability for machine-understanding of visual input has grown exponentially. As it has grown in capability, it has increasingly been applied to the study of cities, producing various interesting approaches to streetscape analysis. Recent works on sidewalk extraction (Hosseini et al., 2023), street view segmentation, and change detection (Dong et al., 2024; Hosseini et al., 2022; Wang et al., 2025; Zhu et al., 2024) show the potential to evaluate infrastructure condition and assess neighborhood change. However, efforts like this generally draw conclusions from singular sources, often aerial imagery or Street View Images, or features, rather than using an ensemble-based approach to measuring change. Urban foundation models (W. Zhang et al., 2024) can handle city-scale prediction tasks, but treatment of the underlying built environment remains limited to encoding static networks or features.

Semantic Grounding

However, they found that grounding the models with formal and structured domain knowledge - e.g., ontology of common street-evaluation metrics - improved model performance and reliability considerably. Zhu et al. (2024), similarly, found that the performance of the Segment Anything Model (SAM) (Kirillov et al., 2023) on Remote Sensing Change detection was significantly improved by adding semantic guidance. While, Shu et al. (2025), found that traditional Remote Sensing Change Detection models struggled with some common tasks when applied to roads and bridges

without semantic understanding, and thus published a semantically-grounded dataset to augment such detection tasks in the future.

Urban Visual Analytic Systems

Recent academic inquiries at the intersection of Computer Science and Urban Studies have resulted in a wide array of standalone visual analytic systems - tools which "distill intricate urban analysis workflows into intuitive, interactive visual representations and interfaces" and allow users from a variety of fields to access and analyze complex interactions between geospatial, visual and temporal urban datasets in a reproducible, low-code manor. (Ferreira et al., 2024)

In *MobiSeg*, Wu et al. (2017) created a tool that automatically segments a city into different regions based on similar patterns found in mobility data. Garcia et al. (2021) worked with criminology experts to create *CrimAnalyzer*, a tool that allows users to easily see crime patterns across their city. And in *SensorAware*, Wei et al. (2021) uses pixel-based time-series visualization to propose a generic interactive visual system for analyzing readings from static and mobile sensors, with various applications for city government officials.

Looking closer at the built-environment Miranda et al. (2019) used foundational computer graphics techniques of ray tracing and shadow maps to develop *Shadow Profiler*, a tool that allows urban planners to simulate the effects of shadows throughout the day. At further granular level, *Urban Rhapsody*, (Rulff et al., 2022) leverages machine-learning models on top of sampled audio from New York City to create a large-scale, annotated database representation of the city’s soundscape. However, few tools focus specifically on street-level transportation and accessibility infrastructure. Hosseini et al. (2023) developed *Tile2Net*, an end-to-end tool to automatically detect roads, sidewalks, and crossings from aerial imagery and create the planimetric polygon and network representation of pedestrian infrastructure. This study makes use of these segmentation labels to detect infrastructure changes.

Another set of recent tools have leveraged recent advances in multi-modal large language models (MLLMs), to aid in analyzing the built environment by allowing coupling of textual and visual data. IF-City (Lyu et al., 2024) uses explainable AI to help Urban Planners simulate the effects of built-environment change at the city scale. At a more street-level, Froehlich et al. (2025) built a VLM-agent-powered tool to aid in street-level change detection for building-accessibility. In a benchmark of LLM capabilities with street-level understanding, C. Zhang and Wang (2024) found that GPT and other “zero-shot” foundation vision-language models are good when it comes to translating visual information into text (captioning) but bad when expecting exact measurements and objective answers. While Cai et al. (2025) found that MLLMs were able to provide assessments of specific streetscape factors related to walkability from Street View Imagery out-of-the-box based on general knowledge, but were generally too optimistic and prone to errors.

Analyzing a set of 130 similar tools for similar trends, Ferreira et al. (2024), found 46% of the systems allow for interactive queries, but only 1% allowed natural language queries and 2% allowed visual querying. Only 14% attempted to fuse multi-modal or multi-source datasets to improve the queries. (Ferreira et al., 2024) In a public-service focus field like Urban Planning, availability and cost are key considerations of any tool or data adoption. As such, it is important to note that Ferreira et al. (2024) found only 10% of those analyzed were open-source with data and source code publicly accessible.

This paper proposes a new urban visual analytics system, STREETTRANSFORMER, a multi-modal, open-source tool for built-environment change detection that allows users to query the history of streetscape design features using interactive selection, text, or image/document-based search. Few

studies or tools focus intently on intersection-level transportation infrastructure change detection and, crucially, no studies could be found that attempt to link multiple data modalities by combining open-source imagery with publicly available civic design **documents**. While the work of Stalder et al. (2024) and Dong et al. (2024) shows promise in improving visual change detection, they remain tied to the Image to Text comparison, while our research expands by incorporating additional modalities to create a fuller picture of streetscape change by linking longitudinal **Imagery** with the design **Documents** that mandated them and the **Project** data that catalogs them.

3 Data Processing

Understanding the complexity of Street Design requires diverse multi-modal data sources. STREET-TRANSFORMER utilizes a variety of publicly-available data, each representing a different aspect of streetscapes either at a given time or the difference between two times, and unifies them into a singular, structured, and queryable, spatiotemporally-grounded database.

3.1 Overview

The core unit of analysis in this paper is *city intersections*, defined as where any two car-accessible streets cross and hereafter referred to (for extensibility) as **Locations** (L). Sets of **Locations** (often with geographic or jurisdictional connections) are called **Universes**. Here, our universe consists of all intersections in New York City, analyzed over the **Timeframe** (T), 2006-2024.

Data is then gathered from a number of **Modalities**: Ortho-rectified aerial **Imagery**, (I) streetscape feature **Masks** (M), public official city **Documents** (D), and civic capital reconstruction **Project** data (P). Each of these goes through their own processing steps (detailed below) to standardize and ultimately assign them to **Locations** (L) and years (T).

3.2 Definition

State vs. Change Formally, for any $l \in L, t \in T \exists l_t$ that represents the *State* of the location l at the given time t . For this study, we specify that *State* as the existence (binary) and location (geometry) of a set of street-level transportation features (F), detailed in Table 1.

Streetscape Features (F)
Sidewalk/Curb
Median/Island
Car Lane
Crosswalk
Bike Lane
Bus Lane
Pedestrian Plaza

Table 1: The existence and location of this set of street-level design features (F) defines the *State* of a location (l)’s at any given time (t).

At any location, for any pair of years $\{l_t, l_{t'}\}$, we can define a *Change*: $\Delta_{l_t l_{t'}} = F_{l_{t'}} - F_{l_t}$ as the difference between the features observed at time t' and time t , and we can define a streetscape redesign as any set of $\{l, t, t'\}$ where $\Delta_{l_t t'} \neq \emptyset$. And we crucially know that this $\Delta_{l_t t'} \neq \emptyset$ for a countably finite amount of locations in the city (by our count, 6,000), which should be captured to some extent in modal data.



Figure 1: The redesign of Pearl St to include a pedestrian plaza was an early example of NYC’s new emphasis on streetscape initiative in 2007

The components of the set of modal data (I_l, M_l, D_l) assigned to each location and year (e.g. I_l is an aerial image of location l at time t), each describe a different, incomplete angle of the feature-scape F_l either at the time t or over a period relating to t (e.g. D_l is a document that defines a change prior to or after time t). Crucially, civic project data (P - detailed further below) in NYC allows us to know, for each location, if a change exists ($\Delta_l \neq \emptyset$ and more specifically, for which $f \in F, f_{l,t,t'} \neq \emptyset$).

Example Let us define l_1 as the Location at the intersection of Pearl Street and Anchorage Pl in Brooklyn Figure 1. We define *State* of l_1 for years 2006 and 2010 to include the set of streetscape infrastructure features at l_1 existent at each time. We also define the set of *Change* data ($\Delta_{l_1,2006,2010}$) between the years 2006 and 2010, which includes the difference between those two feature-sets. Most notably, between these years, a pedestrian plaza was added, so only looking at existence, $f_{pedplaza_{2010}} = 1$ yet $f_{pedplaza_{2006}} = 0$ and their geographic data is in turn defined as the shape of the new plaza.

3.3 Locations

As detailed in Figure 2, the data gathering process begins by gathering the canonical set of location data. For New York City, STREETTRANSFORMER gathers Intersection centroids (*nodes*) from the City’s open-source LION (New York City Department of City Planning, 2024) GeoDatabase. These points are then filtered to remove those nodes which do not have car-accessible streets as edges - this excludes nodes solely connected to footpaths, bike lanes, ferry stops, places where subways cross underground and no-longer-existent streets. Remaining points are then consolidated by merging those within 100 feet and those with degree two. And finally the remaining nodes are assigned street-names based on the edge-names of the preconsolidated nodes in order of frequency.

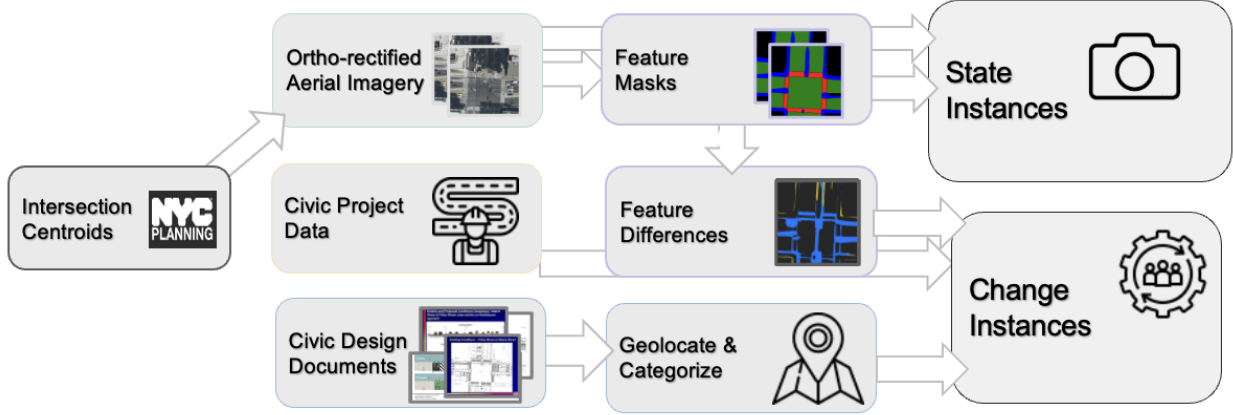


Figure 2: Beginning with data from the NYC’s LION dataset, intersection centroids are assigned to **Imagery**, **Feature Masks**, **Projects** and **Documents** to create instances of *State* and *Change* for all combinations of intersections and years.

3.4 Modalities

Imagery The aerial imagery process begins at the intersection level. For each **Location**, a bounding box of a fixed width is drawn around the intersection centroid, and translated into XYZ tiles, which are then retrieved from NYC’s Orthoimagery APINYC OpenData (2006-2024) for each even year (2006 - 2024) at zoom level 20 (15 cm resolution). These tiles are stitched into mosaics and cropped to the bounding box, producing ten temporal snapshots per location, resulting in a total of 470,000 images.

Feature Masks: Each **Image** is then fed into *Tile2Net* Hosseini et al. (2023), to create the segmentation labels, including roads, sidewalks & footpaths, and crosswalks for each intersection tile. This gives a representation of the intersection that is highly focused on the shapes of the features of study, theoretically improving the ability of models to detect and measure difference. Even for features not included in the model segmentation - medians are classified as background - changes in size and shape of these features are in most cases visible in the resulting masks, as can be seen in the change-calculation example in Figure 3.

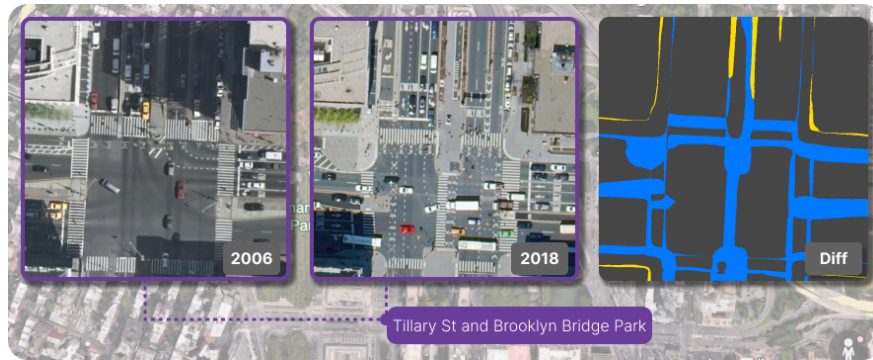


Figure 3: Images are passed into *Tile2net* to be segmented by into infrastructure masks. Then masks are subtracted to create difference masks

Documents: The inclusion of civic design documents is one of the key innovations of this project. Government-sponsored capital projects require a large paper trail of documentation, made available to the public through hearings or official releases. For the New York City Department of Transportation (NYC DOT), this takes the form of a large unstructured online repository of pdf documents (see Figure 4), which reflect all materials made available during necessary public hearings about capital reconstruction projects.



Figure 4: Four of 130 Publicly-available Document Pages, which were spatially-grounded to the intersection of Tillary St & Adams Rd in Brooklyn. These **Documents** can be used to augment and enrich understanding of streetscape change beyond traditional data sources.

These documents are intended for public consumption on an individual basis and thus are largely unstructured and un-indexed. Yet, as shown in Figure 4, they contain highly detailed, multi-modal data, including text descriptions of projects, checklists of project components, design schematics like detailed descriptions and labeled diagrams, and renderings. This data is feature-rich but difficult to standardize or localize without modern advances in document understanding, natural language processing and computer vision.

All publicly available "project document" PDFs were scraped from NYC DOT's website (2007-2024) (New York City Department of Transportation, 2025a) – totaling 1,000 documents organized by project name, borough and year. These documents were then split into 33,000 individual page files

and spatially grounded using a combination of regex and Google Gemini queries on OCR-extracted text linking each page to sets of cross-streets. In the process, the agent also assigns a type (**text**, **image**, or **diagram**) to each street-set based on how and where it found the street pair – this typology will be used later. The street name pairs are then normalized using a combination of regex and NYC-Geoclient-based Normalizer (NYC OTI, 2025), and joined to the **Locations** dataset by name so that 78% (with *gemini-2.5-flash*) of street pairs found within pages match to **Locations**.

Projects: From *NYC OpenData* (New York City Department of Transportation, 2025b) we extract a geocoded log of *Capital Reconstruction Projects* since 2001 and match them to **Locations** within 100 ft. This buffer is necessary to account for manual discrepancies and artifacts introduced for complex intersections during the consolidation process. Projects are then filtered to transportation related reconstruction projects (**n=2,799**) while filtering out minor changes like resurfacing or art installations (**n=1,266**), leaving the **safety_scopes** metadata column is consolidated and standardized to fit the following 6 categories:

Project Scopes – "Safety Interventions"
Lane Removal or Road Narrowing
Bus Bulb / Curb Extensions
New or Extended Median / Island
Raised Crosswalk / Intersection
Shared Street
Bike Enhancement

Table 2: NYC DOT’s Capital Reconstruction Project dataset defines a set of Safety-focused scopes for each project. Here we present a consolidated version.

3.5 Unification

These modalities all been cleaned and processed and are temporally (*Year*) and spatially (*location_id*) grounded and thus all modalities can be linked to a snapshot of a specific **Location** at any given time. Furthermore, *State*-based modalities (e.g. **Images** and **Masks**) can be paired and subtracted to compute *Change*. And *Change*-based modalities (e.g. **Documents**, and **Projects**) can be used to compute *State* features in their before and after years. Together the modalities creates a feature-rich, densely-populated spatiotemporally grounded corpus of multi-modal, mutually-supportive data.

4 Benchmarking Existing Tools

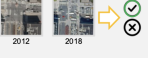
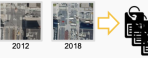
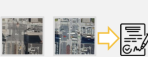
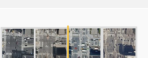
Recent advances in general-purpose Large Language Models (LLMs) and Vision-Language Models (VLMs) like ChatGPT and Gemini have shown immense promise in recent years across many fields. With that growth, researchers have increasingly applied their skills to urban streets, Cai et al. (2025) Huang et al. (2024) C. Zhang and Wang (2024) , to varying degrees of success. To test whether such "zero-shot" approaches might be useful as part of streetscape feature change search, we developed a benchmark of tasks upon which to test the tools.

This benchmark included a suite of five LLM tasks that utilize different aspects the corpus to answer questions that are key building blocks of STREETTRANSFORMER’s more complex queries. These benchmarks tested with various versions of Google’s Gemini model and with different media type inputs (**Images** and **Fusion (Image + Mask)**).

4.1 Benchmark Tests

Four tasks

1. *Change Classification*: Did a change occur (between image/mask pairs at the same location).
2. *Change Identification*: What specific streetscape features (pulling from **Safety Features** above) changed between these two **Images** or in this **Document**?
3. *Change Descriptor*: Write a short, 1-sentence description of the change that occurred between two **Images**. Do the same for a linked **Document** and compare them.
4. *Change Dater*: Find the image where in a sequence where a change first occurs.

Task	Description	Image Example	Data Types
Change Classifier	Given a "media pair" or document, classify if there are changes to Safety Features.		Image Image + Mask
Change Identifier	Given a media pair or document, identify what features in the image the change occurred.		Civic Documents
Change Descriptor	Given an media pair or document, describe in a sentence, the change that occurred between the two given a set vocabulary of Safety Features.		
Change Dater	Given a chronological set of images from the same location, identify where the change occurred.		

(a) Four benchmark tasks test important sub-capabilities across a variety of input .

Task	Accuracy	Macro-F1	Task-Specific Score (Δ)
Change Identifier	2.5 Flash (image) 0.563 2.5 Flash (+Mask) 0.543 2.5 F-lite (image) 0.539 2.5 F-lite (+Mask) 0.527	0.519 0.502 0.526 0.522	MCC 0.046 0.059 0.076 0.121
Change Classifier	2.5 Flash (image) 0.421 2.5 Flash (+Mask) 0.56 2.5 F-lite (image) 0.574 2.5 F-lite (+Mask) 0.774	0.008 0.007 0.004 0.003	MCC 0.013 0.019 0.031 0.033
Change Dater	0.238 0.178 0.071 0.021	2.5 Flash (image) 2.562 2.5 Flash (+Mask) 4.880 2.5 F-lite (image) 2.730 2.5 F-lite (+Mask) 5.099	MAD 0.013 0.019 0.031 0.033
Change Descriptor		2.5 Flash (image) 0.842 2.5 Flash (+Mask) 0.862 2.5 F-lite (image) 0.858 2.5 F-lite (+Mask) 0.838	RefCLIP 0.842 0.862 0.858 0.838
Tile2net (Baseline)			0.921 mIoU 0.037 frac

(b) Results of benchmark tasks were middling on all respective tests compared to Tile2net.

4.2 Findings & Discussion

These results show that all four models tested show minimal improvement over random chance at *Change Identification*, *Change Classification*, suggesting that the models are not very useful in this case. However, they show positively strong performance in the *Change Description* task, with each model showing *RefCLIP* scores are between 0.838 and 0.862 showing a high overlap between the descriptions generating from pre-linked images and documents.

These findings show that existing MLLM models might not be very helpful in change retrieval, lending more weight to the goal of developing our own algorithm is necessary.

Note: These steps were unfortunately undertaken just before Google Gemini 3 was released, which purports to excel at tasks like these. Initial small-scale tests were positive but there There are also additional, more complex tasks including Image to Document Linking and Change Localization, which were scrapped due to the cost and lack of success on initial tasks.

5 Indexing & Retrieval

Efficiently and accurately querying across a corpus of data this large and multi-modal is non-trivial. Some of STREETTRANSFORMER’s core contributions come from this methodology. With New York City’s 47,000 intersections, across 10 years, there are 470,000 *state* instances and 2.115 million *change* instances over which to search. Each of these instances contains a large [512,] embedding for each Media Type (**Image**, **Mask**, **Fusion (Image + Mask)**).

We developed this retrieval process to ensure accuracy across the space while still promising sub-second retrieval times on text- and image-based search, whether using local or cloud-based encoders.

1. Embed query (**Images**, **Masks**, **Text**, **Documents**) using OpenCLIP into a vector database (powered by PGVector)
 - (a) OpenCLIP (Cherti et al., 2023) is an open-source and active version of OpenAI’s CLIP (or Contrastive Language-Image Pre-training) (Radford et al., 2021) model which was accessed using huggingface. (MLFoundations et al., 2021)
2. Compute change vectors by subtracting pairwise embeddings and save in separate file.
3. Compute cosine distance between all embeddings and change vectors
4. Build retrieval engine using FAISS (a) on top of HNSW-indexes(b) and export to ‘faiss’ files.
 - (a) FAISS or "Facebook AI Similarity Search"(Douze et al., 2025) is a python library built by Facebook to optimize Vector Similarity Search.
 - (b) Alternative Index for IVFFlat provided for low-memory scenario if willing to sacrifice some accuracy.
5. For a given input, find closest N vectors to it using raw distance from FAISS.
6. Re-rank N results using pre-computed whitening transformations.
 - (a) Whitening is a vector-decomposition algorithm similar to Principal Component Analysis (PCA), which emphasizes differences in key features and filters out irrelevant features with similar variances across the universe of images, thus dramatically improving discretion and accuracy.

Because the indexes and whitening transformations can be precomputed, and the whitening only happens in re-ranking, we can store the vector changes for this large dataset in highly compressed local files.

6 StreetTransformerTool

This brings us to the core output of the project, the open source and customizable tool . Fashioned as a web application, STREETTRANSFORMERallows urban transportation researchers, professionals, users and advocates to efficiently access, compare, and ask questions about their urban built environment over time.

We demonstrate our approach using New York City as an initial test case. With Its size, Its quantity and quality of publicly available data, and two decades of streetscape redesign—resulting in

a plethora of examples across many semantic categories—make it an ideal first case. In this example, the **Universe** includes all 47,000 NYC intersections and the **Timeframe** is 18 years beginning in 2006, which is during the Bloomberg administration when Director of Transportation Jeannette Sadik-Khan first made street design an emphasis of the city government.

6.1 Use Cases

STREETTRANSFORMER is designed for non-technical practitioners and researchers in the urban design and transportation space who want to learn more about their city’s intersections. It allows users to search across the Universe of **Locations**, using either structured or open-ended text queries (detailed below) to retrieve Locations throughout the **Timeframe** based on a similarity score. This can allow designers to find precedents for new projects, traffic engineers to determine a control group for an impact study of a planned intervention, or allow researchers to gather a sample for a policy study.

6.2 Available Queries

As of now, there are five implemented queries that together allow for a knowledgeable user to answer a wide range of domain-specific questions:

1. Image Similarity Retrieval

- **Type:** Image $\rightarrow$ Image, *State*
- **Use:** Find similar intersections
- **Detail:** "Given a location and a year, find the most similar locations to it"
- **Example:** Locations most similar to Times Square 2006.
- **Algorithm:** *FAISS(HNSW) + Whitening*
- **Options:**
 - Search using **Images**, **Masks** or both (**Fusion**)
 - Filter the "Target universe" using year, borough, etc.

2. Image Dissimilarity Retrieval

- **Type:** n/a, *Change*
- **Use:** Find *when* a change occurred.s
- **Detail:** "Given a single location across time, find which years are least similar to each other"
- **Example:** "When was this intersection redesigned?"
- **Algorithm:** *Cosine – distance*

3. Text to Image Retrieval

- **Type:** Text $\rightarrow$ Image, *State*
- **Use:** Search for streetscape features
- **Detail:** "Given a text-based description, find locations that match it."
- **Example:** "Find me intersections with a bus lane and pedestrian island."
- **Algorithm:** *OpenCLIP + FAISS(HNSW)*

- **Options:** Filter target universe.

4. Text to Change Retrieval

- **Type:** Text $\rightarrow$ Image Pair, *Change*
- **Use:** Search for streetscape change over time
- **Detail:** "Given a text-based description, find locations where the difference between two years matches it."
- **Example:** "Find me intersections where a pedestrian plaza was installed."
- **Algorithm:** *OpenCLIP* + *FAISS(HNSW)*
- **Options:**
 - Search using **Images**, **Masks** or both (**Fusion**)
 - Filter target universe.
 - Option for Sequential years only.

5. Image Pair Change Retrieval

- **Type:** Image Pair $\rightarrow$ Image Pair, *Change*
- **Use:** Find similar changes
- **Detail:** "Given a pair of images, find pairs of images (from the same location) that have similar semantic difference.
- **Example:** "Find me intersections that *also* had 3 crosswalks installed."
- **Algorithm:** *DifferenceVectors* + *FAISS(HNSW)* + *Whitening*
- **Options:**
 - Search using **Images**, **Masks** or both (**Fusion**)
 - Filter target universe.
 - Option for Sequential years only.

6.3 Interface

The interface for STREETTRANSFORMER focuses on simplicity and ease of use for users. Upon opening the app, the user is greeted by an interactive *Map* showing Street Reconstruction **Projects** across NYC. A user can hover over locations to view information about the project or click on one to select a *QueryLocation*. Above the *Map* is the *SearchBar*, with tabs for various Query Instances detailed above. In the first tab, Image State Search, a user can enter streets into a search bar and select a year from the set of available years. Once this temporally-grounded *QueryLocation* is selected, the *DetailsPanel* pops up over the *Map* on the left to reveal the relevant data about the *QueryLocation*. In different tabs, one can view **Imagery** and **Masks** from the location's bounding box through the years, **Document** pages which reference those cross streets, broken out by type, and tabular metadata for **Projects** that have affected the *QueryLocation* since 2001.

The *SearchBar* contains additional options, including the ability to select a *target year*, search methodology, or *Top N Results*, among others. With these options set, the user can click *Search*, which will cause the *ResultsPanel* to pop up over the *Map* on the right. Here, a structured representation of the *top N* most similar locations to the *QueryLocation* sorted based on the search criteria. These Locations also populate the background *Map* to show the geographic spread of the Results.

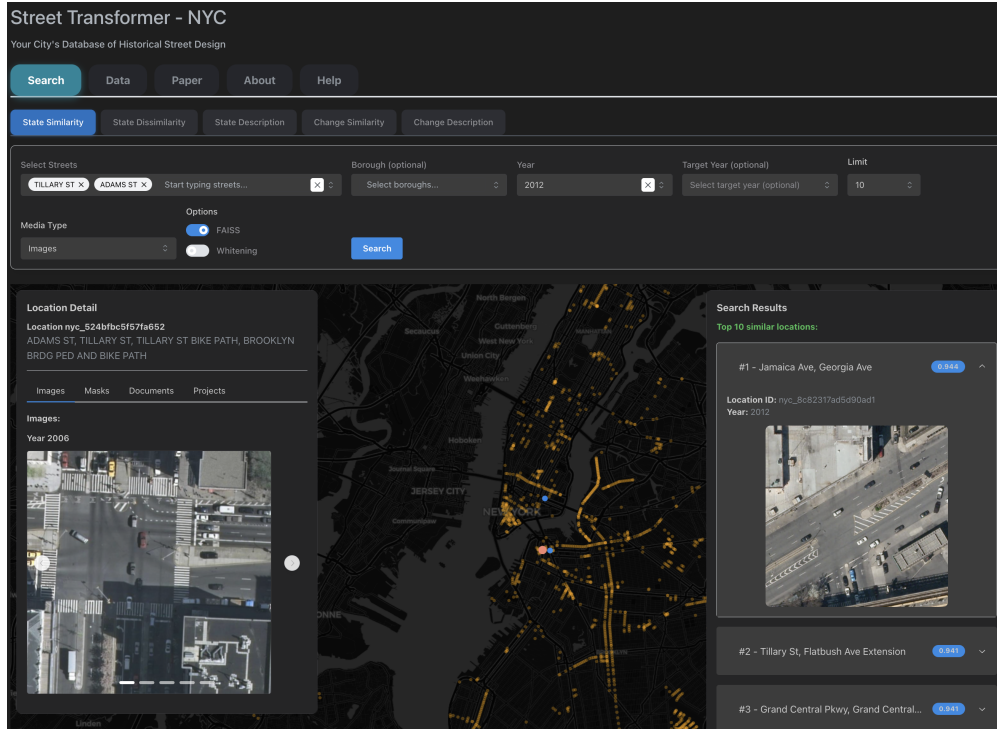


Figure 6: STREETTRANSFORMER provides a simple interface for accessing the multi-modal dataset and querying for streetscape similarity and change

This process is repeated throughout the application for other types of queries, with their respective criteria. By selecting the tabs above the *SearchBar*, a user can toggle between the available query types.

6.4 Extensibility

A key feature of STREETTRANSFORMER is its flexible design. On the backend, the application has been implemented with an eye toward extensibility and replicability in additional localities. Given the near universality of aerial imagery and Civic Design documents as well as the abundance of potentially available modalities beyond those used (e.g. *StreetView*, Planimetric features, Mobility or Land-use data), we believe this will allow the tool to grow both in breadth and adoption for maximal usefulness.

7 Discussion & Future Steps

STREETTRANSFORMER is an urban visual analytics system designed to serve practitioners and researchers in analyzing urban design and infrastructure. Development is still ongoing on additional features, functionality and user interface improvements to best serve this purpose and aid its potential userbase. With its existing core functionalities, users are able to access, explore and compare data about individual street design projects in New York City over the past two decades. Looking forward, this prototype also serves as a proof of concept and template for what is possible in open-source decomposition, indexing and machine understanding of the Urban built environment, and provides a good base from which to improve.

In non-trivially linking **Images**, **Documents** and **Project Data** in a spatiotemporally-aware

dataset that can be queried with a simple user-interface, STREETTRANSFORMER allows those interested access to this multi-modal data like never before. This can have broad implications for human understanding of built-environment change, allowing residents and policy makers to easily access the history of their local intersections or serving as a human-accessible layer on top of the open-source corpus for practitioners and researchers to conduct additional analysis.

As for machine understanding of built environment change, the findings so far are promising but modest. There is some definitive signal for similarity search across state-based **Images** and in change similarity search for differences that are either non-semantic (e.g. "zebra-striped" crosswalk additions) and/or are unmistakable on Segmentation **masks** (e.g. median expansion). However, the algorithms struggle to link across locations where semantic differences might be more necessary (e.g., classification of bike lanes and turn lane configurations).

The team plans to continue working on the tool and the underlying algorithm, which will expand the project along multiple vectors.

Additional semantic data The team plans to improve access to semantic grounding with both human expert interviews and by and by incorporating additionally structured data like through an ontology-based deconstruction of the documents.

Fine Tuning & RAG These additional data could be combined with fine-tuning and incorporated into a RAG-enhanced grounding, which has shown promise in improving results within the domain.

Additional Modalities Expanding the modal set to include additional layers can help gain a broader understanding of state and change. In particular, Street View Images (SVI) or other street-level views for their universality and the wealth of research already conducted with them. These additional visual angles on street infrastructure, while less consistent and harder to standardize than the ortho-rectified aerial imagery, can undoubtedly provide a more complete picture of urban streetscape at various times. Additionally, due to its sparse and multi-faceted nature it can help fill in for literal and figurative blind-spots of the aerial imagery such as tree-cover or intervening time between snapshots to better pinpoint and categorize change occurrence.

Additional Geographic Support Expanding beyond just New York City is another priority. While the application and database structure has been designed to be extensible already, doing so would require gathering additional modal data from new jurisdictions and writing custom source loaders to translate the raw data into the STREETTRANSFORMERbackend format.

User Interviews Finally, we plan to continue to expand the functionality of the tool with functional, ergonomic and aesthetic enhancements. There are a number of items on this to do list already but near future next step will involve user interviews with practitioners and researchers to get a better sense of what will be most helpful for their needs.

8 Conclusion

The realm of urban street design is important for public welfare, highly complex and challenging to ascertain objective truth and representative examples. However, in this paper we also demonstrated that it is flush with feature-rich if highly unstructured data from many public sources which have

been heretofore underutilized at scale despite their usefulness for practitioners. We argue it is thus ripe for intervention and that the type of data unification and retrieval that recent advances in Computer Vision, Vector Search, Vision Language Models can help empower urban planners, designers and researchers with increased knowledge.

We further showcased STREETTRANSFORMER, a visual analytics system built to do just that, by first walking through its process for unifying publicly available aerial **imagery**, public design **documents** and civic **project data** across New York City to create a novel corpus of multi-modal, spatiotemporally-grounded streetscape data over which a retrieval engine could be built. We then showed how current LLMs struggle with even basic sub-tasks necessary to reliably operate such a retrieval process using existing foundation models. And then detailed our custom algorithms for such retrieval using combination of CLIP-based encoders, FAISS-enabled HNSW indexing and whitening for re-ranking to . We finally showcased how this dataset and engine have been built into an interactive tool, for which we walk through use-cases and features, and mapped out the design process implemented with an eye toward future extensibility, and discussed such future plans in the discussion.

STREETTRANSFORMER represents a major advance toward better machine and human understanding of the dynamic built-environment of urban streetscapes. The combination of novel dataset and algorithm packaged into a simple user interface shows promise for a tool that can empower urban designers and other practitioners with design data they need and offers an opportunity for Computer Science and Machine Learning to play a role in improving cities through design.

References

- Cai, C., Kuriyama, K., Gu, Y., Biljecki, F., & Herthogs, P. (2025). Can a large language model assess urban design quality? evaluating walkability metrics across expertise levels. *arXiv preprint arXiv:2504.21040*. <https://arxiv.org/abs/2504.21040>
- 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. [https://doi.org/10.1016/S1361-9209\(97\)00009-6](https://doi.org/10.1016/S1361-9209(97)00009-6)
- Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., & Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2829. <https://doi.org/10.1109/CVPR52729.2023.00276>
- Cortright, J. (2009, August). *Walking the walk: How walkability raises home values in u.s. cities* (Research Report) (Accessed via CEOs for Cities website; original PDF available through affiliated sources). CEOs for Cities. <https://ceosforcities.org/research/walking-the-walk/>
- Dong, S., Wang, L., Du, B., & Meng, X. (2024). Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 53–69.
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). The faiss library. <https://arxiv.org/abs/2401.08281>
- 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., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, 14(1), 65–84. <https://doi.org/10.1080/13574800802451155>
- Ferreira, L., Moreira, G., Hosseini, M., Lage, M., Ferreira, N., & Miranda, F. (2024). Assessing the landscape of toolkits, frameworks, and authoring tools for urban visual analytics systems.

- Computers Graphics*, 123, 104013. <https://doi.org/https://doi.org/10.1016/j.cag.2024.104013>
- Froehlich, J. E., Hwang, J., Wang, Z., O'Meara, J. S., Su, X., Huang, W., Zhang, Y., Fiannaca, A., Nelson, P., & Kane, S. (2025). Does the cafe entrance look accessible? where is the door? towards geospatial ai agents for visual inquiries. *arXiv preprint arXiv:2508.15752*. <https://doi.org/10.48550/arXiv.2508.15752>
- Garcia, G., Silveira, J., Poco, J., Paiva, A., Nery, M. B., Silva, C. T., Adorno, S., & Nonato, L. G. (2021). Crimanalyzer: Understanding crime patterns in são paulo. *IEEE Transactions on Visualization and Computer Graphics*, 27(4), 2313–2328. <https://doi.org/10.1109/TVCG.2019.2947515>
- Hosseini, M., Miranda, F., Lin, J., & Silva, C. (2022). Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 79, 103630.
- Hosseini, M., Sevtuk, A., Miranda, F., Cesar Jr, R. M., & Silva, C. T. (2023). Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101, 101950.
- Huang, T., Wu, Z., Wu, J., Hwang, J., & Rajagopal, R. (2024, July). *Using ai to map urban change* (HAI Policy Brief) (Accessed 2025-12-12). Stanford University. <https://hai.stanford.edu/assets/files/2024-07/HAI-Policy-Brief-Using-AI-Map-Urban-Change.pdf>
- Jackson, K. T. (1985). *Crabgrass frontier: The suburbanization of the united states*. Oxford University Press.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. <https://arxiv.org/abs/2304.02643>
- Leslie, E., Coffee, N., Frank, L., Owen, N., Bauman, A., & Hugo, G. (2007). Walkability of local communities: Using geographic information systems to objectively assess relevant environmental attributes [Part Special Issue: Environmental Justice, Population Health, Critical Theory and GIS]. *Health Place*, 13(1), 111–122. <https://doi.org/https://doi.org/10.1016/j.healthplace.2005.11.001>
- Lyu, Y., Lu, H., Lee, M. K., Schmitt, G., & Lim, B. Y. (2024). If-city: Intelligible fair city planning to measure, explain and mitigate inequality. *IEEE Transactions on Visualization and Computer Graphics*, 30(7), 3749–3766. <https://doi.org/10.1109/TVCG.2023.3239909>
- Miranda, F., Doraiswamy, H., Lage, M., Wilson, L., Hsieh, M., & Silva, C. T. (2019). Shadow accrual maps: Efficient accumulation of city-scale shadows over time. *IEEE Transactions on Visualization and Computer Graphics*, 25(3), 1559–1574. <https://doi.org/10.1109/TVCG.2018.2802945>
- MLFoundations et al. (2021). OpenCLIP: An open source implementation of CLIP [Accessed: January 8, 2026].
- NACTO. (n.d.). Intersection design principles [Design guidelines published to shape design projects by practitioners].
- National Highway Traffic Safety Administration (NHTSA). (2024, June). *Overview of motor vehicle traffic crashes in 2022* (Traffic Safety Facts Research Note No. DOT HS 813 560) (Revised June 2024; accessed 2025-12-12). U.S. Department of Transportation. Washington, DC. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813560>
- New York City Department of City Planning. (2024). *Lion: Linear integrated ordered network* [Geographic street centerline dataset for New York City; accessed 2025-12-12]. NYC Department of City Planning. <https://www.nyc.gov/content/planning/pages/resources/datasets/lion>

- New York City Department of Transportation. (2025a). *Capital projects* [Overview of NYC DOT capital infrastructure programs and project types. Accessed December 12, 2025]. <https://www.nyc.gov/html/dot/html/infrastructure/capital-projects.shtml>
- New York City Department of Transportation. (2025b). *Street and highway capital reconstruction projects* [Capital project-level dataset describing street and highway reconstruction projects in New York City.]. NYC Open Data. New York City Department of Transportation. <https://data.cityofnewyork.us/Transportation/Street-and-Highway-Capital-Reconstruction-Projects/97nd-ff3i>
- NYC OpenData. (2006-2024). *Nyc orthoimagery tile index* [Accessed via Data.gov. Last modified September 23, 2022. Metadata updated September 2, 2023.]. City of New York. <https://catalog.data.gov/dataset/nyc-orthoimagery-tile-index>
- NYC OTI. (2025). Nyc geoclient api [Accessed: 2025-08-29].
- Office of the Surgeon General. (2023). *Our epidemic of loneliness and isolation: The u.s. surgeon general’s advisory on the healing effects of social connection and community* (Surgeon General’s Advisory) (Accessed: 2025-12-12). U.S. Department of Health and Human Services. <https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 8748–8763. <https://doi.org/10.48550/arXiv.2103.00020>
- Rulff, J., Miranda, F., Hosseini, M., Lage, M., Cartwright, M., Dove, G., Bello, J., & Silva, C. T. (2022). Urban rhapsody: Large-scale exploration of urban soundscapes. *Computer Graphics Forum*, 41(3), 209–221. <https://doi.org/10.1111/cgf.14534>
- Saelens, B. E., & Handy, S. L. (2008). Built environment correlates of walking: A review. *Medicine & Science in Sports & Exercise*, 40(7 Suppl), S550–S566. <https://doi.org/10.1249/MSS.0b013e31817c67a4>
- Shu, Q., Chen, S., You, Z., Lu, W., Tang, J., & Luo, B. (2025). Rb-scd: A new benchmark for semantic change detection of roads and bridges in traffic scenes. *arXiv preprint arXiv:2505.13212*.
- Southworth, M., & Ben-Joseph, E. (1995). Street standards and the shaping of suburbia. *Journal of the American Planning Association*, 61(1), 65–81. <https://doi.org/10.1080/01944369508975620>
- Stalder, S., Volpi, M., Büttner, N., Law, S., Harttgen, K., & Suel, E. (2024). Self-supervised learning unveils urban change from street-level images. *Computers, Environment and Urban Systems*, 112, 102156. <https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2024.102156>
- Urban Institute. (2024). Highway and road expenditures [Accessed: 2025-11-23].
- Wang, Y., Yu, W., & Ghamisi, P. (2025). Change captioning in remote sensing: Evolution to sat-cap—a single-stage transformer approach. *arXiv preprint arXiv:2501.08114*.
- Wei, D., Li, C., Shao, H., Tan, Z., Lin, Z., Dong, X., & Yuan, X. (2021). Sensoraware: Visual analysis of both static and mobile sensor information. *Journal of Visualization*, 24(3), 597–613. <https://doi.org/10.1007/s12650-020-00717-z>
- Wu, W., Zheng, Y., Cao, N., Zeng, H., Ni, B., Qu, H., & Ni, L. M. (2017). Mobiseg: Interactive region segmentation using heterogeneous mobility data. *2017 IEEE Pacific Visualization Symposium (PacificVis)*, 91–100. <https://doi.org/10.1109/PACIFICVIS.2017.8031583>
- Zhang, C., & Wang, S. (2024). Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7839–7849.

- Zhang, W., Han, J., Xu, Z., Ni, H., Liu, H., & Xiong, H. (2024). Urban foundation models: A survey. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6633–6643.
- Zhu, Y., Li, L., Chen, K., Liu, C., Zhou, F., & Shi, Z. (2024). Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*.