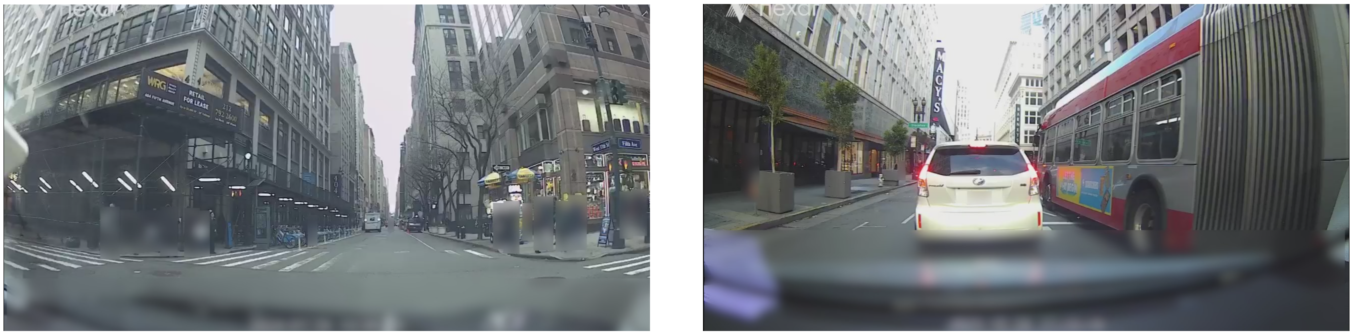
Towards Instrumented Fingerprinting of Urban Traffic: A Novel Methodology using Distributed Mobile Point-of-View Cameras

Matt Franchi
mwf62@cornell.edu
Cornell University
New York, United States

Debargha Dey
debargha.dey@cornell.edu
Cornell University
New York, United States

Wendy Ju
wendyju@cornell.edu
Cornell University
New York, United States



(a) 5th Avenue, Manhattan

(b) Union Square, San Francisco

Figure 1: Examples of POV Cam footage. Note obfuscation of license plate, dashboard and pedestrians for privacy protection.

ABSTRACT

Amidst the replication crisis, it is increasingly clear that we need to understand contextual factors that drive participant behavior, because those factors influence the applicability of study findings more broadly. For AutoUI, as we conduct interaction studies involving drivers, pedestrians, and other traffic participants, it is useful to characterize the traffic contexts that human participants are familiar with, because their prior experiences with traffic are likely to influence their behaviors within the context of a controlled study.

To address this, we propose a new method, 'POV (point-of-view) camera-driven urban fingerprinting,' which can be used to characterize differentiating features of urban environments. We introduce two approaches, Small-Scale Custom Instrumentation, and Large-Scale Collection and Aggregation, and show how they can be used to acquire a broad picture on the characteristics of any city. One key benefit of POV camera-based data collection is that it better captures the experiential aspects of traffic and road scenes than methods such as satellite imaging. This work is the first to formalize a specification for this data collection method by describing existing work and outlining standards to serve as a baseline for downstream research. This work posits the medium of crowd-sourced POVcam as a new and useful tool for transportation/ automotive interaction studies and infrastructure analysis. Subsequently, we provide future researchers with guidelines for characterizing urban traffic in cities for interaction study design.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; **Systems and tools for interaction design**; *Interactive systems and tools.*

KEYWORDS

Urban fingerprinting, POVCams, urban data, automotive

ACM Reference Format:
Matt Franchi, Debargha Dey, and Wendy Ju. 2024. Towards Instrumented Fingerprinting of Urban Traffic: A Novel Methodology using Distributed Mobile Point-of-View Cameras. In *16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '24), September 22–25, 2024, Stanford, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3640792.3675740

1 INTRODUCTION

In conducting interaction studies involving drivers, pedestrians, and other traffic participants, it is useful to characterize the traffic contexts that human participants are familiar with, because their prior experiences with traffic are likely to influence their behaviors within the context of a controlled study. Understandings of qualitative and quantitative similarities and differences between contexts can help to understand where study findings might be applicable; lab studies of driver stress in simulated urban driving scenarios in

AutomotiveUI '24, September 22–25, 2024, Stanford, CA, USA

ACM ISBN 979-8-4007-0510-6/24/09
https://doi.org/10.1145/3640792.3675740

City A, for example, are more likely to be applicable to drivers in City B if City A and City B were similar.

Open Data initiatives in cities worldwide can help provide aggregate data about the urban traffic conditions, but this data is often not fine-grained or temporally dense enough to capture what it is that urban denizens experience in their city every day [30]. To address this issue, we propose an instrumented approach towards urban traffic fingerprinting by collecting and analyzing data from distributed mobile cameras that capture video footage from the point of view of the agent (Point-of-View Cameras, or POV-Cams), such as action cameras or dashcams. Camera-based urban fingerprinting offers a methodological advancement in this regard, providing researchers with reliable data to effectively formalize the traffic characteristics of urban settings. By employing crowdsourced Point-of-View Camera (POVCam) data, this research offers a new tool for detailed data collection and analysis, establishing a framework that enhances the validity and reliability of traffic interaction studies. We outline at two approaches: (1) small-scale sampling by sending out individuals to get footage from the city (we term this *small-scale, custom instrumentation (SSCI)*, and (2) leveraging crowdsourced video footage from networked dashcams (or, *large-scale collection and aggregation (LSCA)*. We demonstrate the effectiveness of these approaches using New York City as a model, though the methodology is adaptable to any urban setting. This work is the first to formalize a specification for this data collection method by describing existing work and outlining standards to serve as a baseline for downstream research.

Contribution Statement. In this paper, we present POVCam footage as a novel, accessible, and low-effort resource to effectively get traffic characteristic information from a city, which can be foundational in setting up interaction studies. Our work is the first to formalize a rigorous and reproducible specification for this data collection method by describing existing work and outlining best practices. To help bootstrap replication of this work, we also provide a GitHub repository at https://github.com/FAR-Lab/urban-fingerprinting to analyze footage gathered in the two approaches we discuss in this paper.

2 RELATED WORK

This work outlines the use of POV cameras to characterize the traffic experienced by people in everyday driving or pedestrian activity. In many studies of cross-cultural differences, researchers focus on differences in the driving skills, risk-perception, driving behaviors, gaze patterns [32, 37, 38, 43, 47, 48]. Sometimes such research is conducted in driving simulators, but often they are based on self-reported responses from users on surveys. These measurement mechanisms do not assess the obvious role that the traffic environment plays on the driver actions, perceptions and skills in each of their cultures. This is clearly a shortcoming, but difficult to fault, because there is as yet no way to typify the driving environments that any of the participants in these studies normally experience.

In this section, we review the state of the art in traffic characterization and in the use of image datasets and datafeeds to analyze traffic.

2.1 Characterizing Traffic

In transportation research, researchers commonly characterize traffic using aggregate planning-level analysis, keeping track of traffic volume, congestion, delays and accidents for the purposes of identifying targets for traffic or safety improvement. Boarnet et al., for example, developed a traffic congestion metric which relates the number of vehicles against the capacity of the roadway to handle it [5]. Another example is the annual Texas A&M Urban Mobility Report[1], which characterizes congestion (delay-hours per commuter, fuel wasted, total cost) of different US cities, using data from INRIX[2], which is private-sector provider of travel time information for commercial shipping and transportation. Other related research uses networks of probe vehicles to monitor traffic speed as it changes in key urban corridors [2].

More recent work addresses the dynamics of traffic, for example, identifying typical travel patterns in traffic networks [23], or accounting for heterogenetity of traffic in urban settings [39]. As image processing grew in the 20th century, so did applications rooted in traffic characterization; TITAN was an early example of this, capable of analyzing 4 images per second to ascertain traffic measurements like speeds, densities, flows, lane changes and queue lengths [4]. More recent projects take advantage of inexpensive web cameras and advances in computer vision to provide more fine-grained analysis of traffic. Synder and Do's STREET's project, for example, used a network of webcameras throughout the suburban Chicago area to characterize traffic differences in different communities. [45]. Now, emerging works use networked dashcams to analyze pedestrian behavior [34], detect anomalies [17], and to pursue road asset characterization [42].

2.2 Characterizing *the experience* of traffic

From the AutomotiveUI perspective, these types of traffic characterizations in the previous subsection not entirely relevant to the sense of traffic that study participants might have as drivers or pedestrians. In many ways, our phenomenological experience of a city is more like that captured by the Situationists in the 1960's [13]. The Situationists were an international organization of social revolutionaries made up of avant-garde artists, intellectuals, and political theorists engaged in the practice of the *dérive*, a method of drifting through space to explore how the city is constructed, as well as how it makes one feel. Kevin Lynch famously used the walking interview as a way to understand how people for mental maps of their environments [29]. During the pandemic, Yavo-Ayalon et al. employed Zoom over mobile phones to remotely interview people as they traversed through the city to reflect on changes the pandemic had wrought to their neighborhoods [51].

The advent of dash-cams, mobile phones and action cameras makes it possible to capture the first-person experience of a journey. For example, in AutoUI 2021, Bito, et al. used developed auto-summarization methods on user's dash-cams to give people video summaries of their road trips [3]. Ethnographers such as Barry Brown and Eric Laurier have drawn upon the dashcam videos that other people have uploaded to Youtube to analyze the experience people have with the Tesla Autopilot [7], with Autonomous Ubers

[1] https://mobility.tamu.edu/umr/
[2] https://inrix.com/

[6], with their everyday commutes [25] and even to investigate the meaning of different kinds of automobile beeps in India [26].

In our work, we strive to maintain the human point-of-view of these instruments in the way that the ethnographers and automotive UI researchers have, but to aggregate these so that they are able to typify key differences in urban traffic for the purpose helping to account for the role of context in automotive studies. We strive, in short, to characterize the experience of urban traffic at scale. Google Street View is arguably the most similar product to crowdsourced dashcam, with both products offering timestamped images of a geographic area. GSV has a number of advantages, but none of which make it more applicable to the task of generating behavioral driving simulation parameters; in fact, some characteristics of GSV make the task practically impossible.

The footage from GSV is higher resolution, and panoramic, making observation and characterization of smaller objects like street signs easier. Also, the general spatial coverage of GSV is presently higher than that of dashcam; GSV makes every attempt to gather footage for all roads in an area it offers coverage for. Despite this, recent research finds lacking coverage across the 'time machine' feature (ie. seeing all sampled images for a location, across GSV collection projects) [44]; spatio-temporal instability of imagery dates [11]; and only 44% of commuting routes being covered on average across 45 small and medium-sized American cities [22]. Spatio-temporal instability, in particular, means that images in the same collection project may be from different dates. And, further, granularity is not available at the *time of day*-level, as with dashcam data. This is important; traffic patterns at 9am are significantly different than those of 2pm [35].

3 METHODS

We present two POVCam approaches that represent data collection and analysis at different scales, but both aiming to fingerprint an urban environment.

3.1 SSCI: Small-Scale Custom Instrumentation, fine control using Action Cameras

In pursuit of gaining an impression on how pedestrians and vehicles interact in an urban environment and to ascertain traffic counts, we recorded 28 hours of footage on-foot in and around New York City over the Spring and Summer of 2023. We recruited 12 participants using convenience sampling from the university participant database to record video footage of their daily commute. This allowed us to get a wide variety of snapshots of the city at different locations and time of day. Recorders used GoPro Hero 8 camera[3] fixed to their body or backpack facing forward as a pedestrian, cyclist, or any other micromobility vehicle user, allowing for a very similar perspective to eyesight. Besides qualitatively coding the footage to categorize interactions, the goal is to understand the nature and composition of traffic as a snapshot of the city. However, the manually-intensive nature of qualitative coding [20] poses a time and resource problem when dealing with hundreds of hours of footage; hence we employed computational methods to isolate relevant events, while minimizing 'waste' of useful footage.

This was done by deploying an object detection model to detect intersections in our images, using proximity to zebra crossings, stop signs, and traffic lights as surrogate indicators of the likelihood of approaching an intersection. To that end, we incorporate a pre-trained YOLO model [40] that detects these three artifacts, namely: zebra crossings, stop signs and traffic lights. Then, we generate clips of each piece of footage that depict intersections. We pad the the starting and ending timestamp of each clip by 10 seconds, and merge overlapping intervals between clips. This gives us a set of temporally-disjoint clips, all predicted to depict intersections.

In the absence of GPS data (a limitation detailed in Section 4.1), the aggregation of traffic metrics at the road level is unfeasible. Nevertheless, it remains possible to assess traffic metrics at the level of individual footage clips, categorized over time. This analysis is facilitated by the application of a pre-trained YOLO object detection model, which identifies various traffic participants such as pedestrians, bicycles, and motorized vehicles from footage clips. This model, trained on the Common Objects in Context (COCO) dataset [27], can discern up to 80 different object classes. With this, we calculate metrics like the *average number of vehicles per second*[4] within a clip. These metrics serve as proxies for simulating traffic flows (NPCs: non-player characters) in virtual environments for interaction studies. Furthermore, for a more detailed analysis, traffic data can be segmented by vehicle type — differentiating among cars, SUVs, trucks, buses, etc. — to better represent the diverse vehicle mix typically seen in urban settings. This stratification allows for a more nuanced understanding and simulation of urban traffic patterns.

3.2 LSCA: Large-Scale Collection and Aggregation, less control but more reach using Nexar datasets

The alternative to individualized instrumentation is to build a data fusion approach that takes independent cameras and unifies collected footage over the dimensions of space and time. Several companies are in the business of creating such datasets from dash cams, and some have yielded data for early research explorations [15] and [41].

One of the foremost producers of large-scale dashcam image data is Nexar Inc.[5], which we used in this approach due to established coverage ([15], [41], and [12]). We developed software[6]. around a consumer-grade Nexar API[7] to download all of the image data and metadata in New York City uploaded to the company's servers daily, processing hundreds of thousands of images per day. Nexar uses proprietary algorithms to ensure their database of sampled images is distributed throughout the depicted area, which makes the data reliable for robust urban traffic characterization.

Nexar images are typically sourced from ridesharing vehicles, and so there is noted distribution shift [15] between a random trip taken by a random vehicle in the city and a randomly-sampled Nexar image. To help address this, Nexar stores multiple images per

[3] https://gopro.com/en/us

[4] This is done by counting the total number of vehicles in the entire clip of footage, divided by the total number of frames, and multiplied by frames per second

[5] https://data.getnexar.com/, last accessed April 12, 2024

[6] Software is made available on GitHub at https://github.com/FAR-Lab/urban-fingerprinting

[7] https://data.getnexar.com/product/citystream/

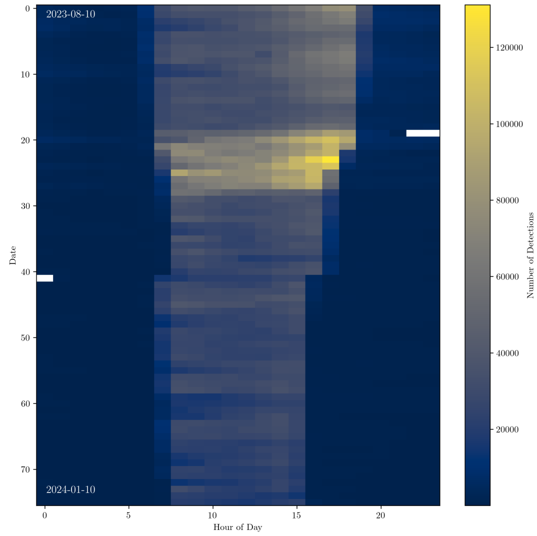


Figure 2: Sampling density in our LSCA image distribution over time. The horizontal axis goes over *time of day*, and the vertical axis goes from the first day of coverage at the top, to the most recent at the bottom. More yellow areas indicate denser coverage, and darker blues indicate lower coverage. A clear positive correlation of daytime and frame density is visible, caused by factors like commuting and human living patterns. White regions indicate a complete lack of coverage; as can be seen, it is very rare (only 3 cells) for us not to have a single image in any arbitrary one-hour length period.

map cell, implementing a first-in-first-out queue based on image recency.

We collected 29,852,687 images between August 2023 and January 2024. The distribution is visualized temporally in Figure 2. Our goal is to, at the road segment level, compute metrics of average pedestrian density and average vehicle density, stratified by time. Such metrics can allow a representative simulation of traffic conditions at a specific area in the city. We demonstrate proof-of-concept of road-level traffic parameter generation in Figure 3. We take all image data from three Thursdays in August 2023; the 17th, the 24th, and the 31st. Then, using the functionality of our toolkit, we generate road-by-road, averaged pedestrian counts across all three days. The Figure shows the distribution at four different times of day (all Eastern time): 9AM, 2PM, 6PM, and 10PM.

4 CHALLENGES AND BEST PRACTICES

In adopting these approaches effectively, there are certain considerations that can ensure the best results. We highlight the common issues and challenges we faced in setting up this methodology, and the subsequent suggested best practices.

4.1 Challenges with Geographic Localization

Knowing *where* a POVCam image or video was taken is imperative to doing any sort of downstream analysis in a direct fashion. We found that the Nexar dashcam imagery produced accurate GPS coordinates for all surveyed locations, including dense parts of New York City like the Wall Street neighborhood and Midtown South neighborhood. However, in cases where the imaging device does not contain built-in telemetry, or insufficiently-accurate location sensing, it is necessary to add an additional tool to the imaging setup. This is critically necessary for footage procured from the SSCI approach using GoPros, since – as described in Section 5.1.6 – the GPS information we found from the GoPro metadata, especially for New York City, was too imprecise to be of use. This is likely due to interference from tall buildings in dense, urban environments, by a factor that makes the signal obsolete, as seen in Figure 4.

Recommendation. To remedy this, we developed a simple apparatus that can be used to extend the localization of the GoPro, while still maintaining the portability and ease-of-use offered by the suite of cameras. We require two items: an ESP32[8] or similar microcontroller, and a portable battery. We flash the BBC with firmware provided by OpenHaystack [18] that allows location sampling from Apple's FindMy network, an Internet-of-Things (IoT) network that relies on nearby Apple devices (phones, smart watches, tablets, etc.) to triangulate location. Now, simply by plugging the microcontroller into a battery pack and keeping it in close proximity to the GoPro, we acquired accurate locations at a sufficient sampling frequency to infer routes [28]. We visualize a sample run of this tool in Figure 4b.

4.2 Storage Constraints

There is a direct relationship between temporal density of POVCam footage and storage cost. In extension, there is a direct relationship between footage resolution and storage cost. Compression remains an active research area in the realm of videos & images [53], and so it is still important to consider the maximum resolution footage we need and the footage 'density' (that is, how many times per period P we need N images in an area G).

Recommendation. We found that 720x1080 resolution images are sufficient for surveying the built environment, vehicles on the street, and other larger objects. For tasks like pothole detection or optical character recognition (OCR) of street signs, higher resolution footage might be required. However, these are special cases that require additional consideration, and are not typical use cases, in our experience. We recommend pre-defining the footage resolution prior to starting data collection; while it is possible to downsize footage through downsampling and frame skipping [21], this adds additional compute and time requirements at the post-processing stage of a project, which hinders efficiency.

[8] https://www.espressif.com/en/products/socs/esp32

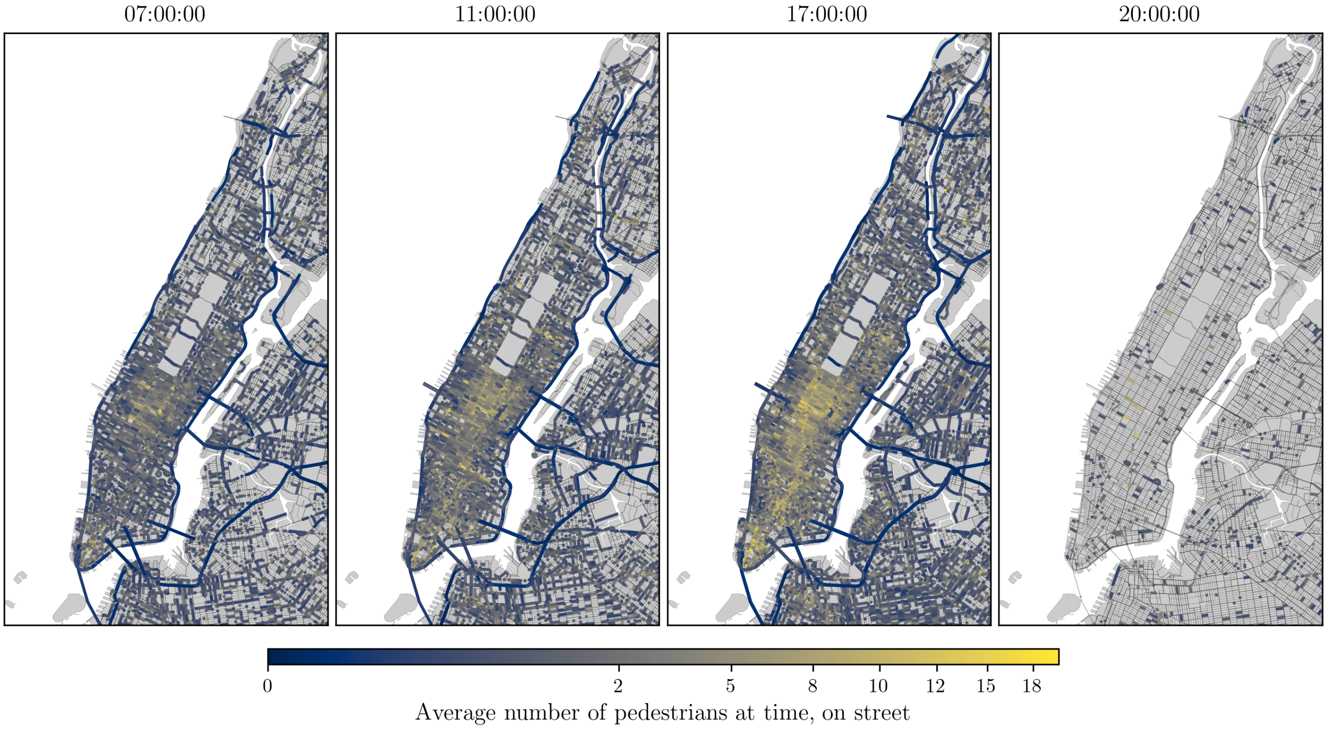


Figure 3: Pedestrian density changes over time. Average number of pedestrians per road-segment-with-coverage in New York City, aggregated across 3 Thursdays in the sampled data to get closer to a map of 'typical' Thursday traffic. A clear increase in pedestrian density is observable from morning to rush hour (around 5PM), and close-to-zero densities are visible on arterial highways going into the city. A power-law colormap is used to match the distribution of pedestrian density (mean 1.80, standard deviation 1.79).

4.3 Pre-processing Recommendations to counteract false positives

There are certain pre-processing steps that can make POV camera data more useful, including several image processing methods and computer vision techniques, which we discuss presently.

4.3.1 Shadow detection and de-weighting. Shadows in urban footage can confuse computer vision models and add false positives to the otherwise true distribution of pedestrian and vehicle counts [19]. To help reduce this error, we train a model to detect shadowed regions of images and reduce detection confidence from those predicted regions.

Recommendation. We originally sought to explore the possibility of addressing this with the implementation of the Meta Segment Anything Model (SAM) by fine-tuning it with two distinct datasets under the framework previously established by Kirillov [24], and Chen [10]. Initially, we employed the Image Shadow Triplets Dataset (ISTD) [46], comprising 1870 image triplets with dimensions 640x480, which includes original images, shadow masks, and shadow-free images. This dataset did not yield optimal results due to a mismatch between the dataset's simpler, lower resolution images and the complex, higher resolution footage from our action cameras. To address this issue, we explored alternatives and transitioned to the CUHK-Shadow dataset [19], which better matches the complexity and visual characteristics of our footage. CUHK-Shadow is derived from diverse sources including urban scenes, vehicles, and remote sensing images, offering a more relevant training environment for our application. This shift significantly improved the performance of our SAM model (now SAM-CUHK), as evidenced by enhanced accuracy and fewer false positives in shadow detection on the same footage. This enhancement led us to implement a de-weighting scheme in areas predicted to have shadows, further reducing the rate of false positives and refining our model's utility in real-world urban settings, which is our recommended approach to deal with shadows in collected footage.

4.3.2 Reflected Object Detection & Removal. POVCam footage can be complex, full of people, cars, buildings, and other street fixtures. In cities, we observe many reflective objects like windows and glass pillars; standard object detection models will falsely label reflected objects as true positives.

Recommendation. To remedy this, we recommend one of two approaches. Firstly, it is possible to train a more robust computer vision model (or new layers atop existing models) that detect and filter out reflected objects; this is a large overhead. Alternatively, there is work on specular removal ([50], [14], and [52]) that can be used to pre-process footage before analysis with a standard classifier like YOLO.

4.4 Other Considerations for Post-processing

When pursuing more complex analysis, it becomes essential to merge the results of an analysis with POVCam data with other spatio-temporal datasets.

In [15], the authors took detected distributions of police vehicles and merge them with American Community Survey (ACS) demographic data that produced correlations of police vehicle deployments with median household income, race, and other key demographic factors. In [41], the authors bootstraped the training of a sidewalk scaffolding detection model by filtering for training data near known scaffolding permits in the NYC Department of Buildings publicly-available permit dataset. As shown in these papers, merging POVCam data with public spatial and demographic datasets is one of its most valuable characteristics.

4.4.1 Geographic merges. When working with geographic data, it is important to keep the coordinates of the POVCam data and the coordinates of any merging datasets in the same projected coordinate plane. Cartesian (latitude/longitude) coordinates are **not** accurate enough for fine-grained analysis of proximity, i.e., matching POVCam images to corresponding crowdsourced citizen complaint (like the 311 system adopted in cities nationwide in the United States ([49], [33], and [9])), and will need to be projected [31]. Further, it is important to consider that an additional element of processing is needed to determine whether a POVCam image actually 'depicts' a sought-after object. This is only possible if the telemetry data of the POVCam includes information about camera heading; if so, it then becomes possible with basic trigonometry to determine not only if an image is close enough to a set of querying coordinates, but also if it faces in the right direction. This is necessary when looking for smaller, static objects like bus stops, scaffolding [41], or street signs.

4.4.2 Temporal merges. The same rule applies when working with longitudinal data and reporting results over a time grouping T; it is important to set granularity such that there is a non-zero amount of data for as many bins as possible in the distribution of T.

4.4.3 Distribution shift. In both SSCI and LSCA data collection approaches, we identify statistical problems with making claims about populations. For example, it is incorrect to draw conclusions about the entire bound of New York City when only having images from Manhattan commuters.

Recommendation. We recommend a statistical intuition to account for this. Logically, the process is to upweight images from areas that are under-represented in a dataset during the model training process. However, this requires that the dataset contains at least some footage in every grouping G elected to present analysis over. This can be harder to do in cities, such as San Francisco, which have less overall footage, compared to cities like New York, which have more (see Figure 5). As Figure 1a shows, there is clearly an inverse relationship between spatial area and the rate of overall penetration; a key element in analytical projects utilizing POVCam large-scale imagery is to select the right spatial granularity. To find the right granularity, it is critical to start small and decrease granularity (i.e. move from H3-6 to H3-7, and so on) until the overwhelming majority of areas contain footage.

4.5 Using the Toolkit

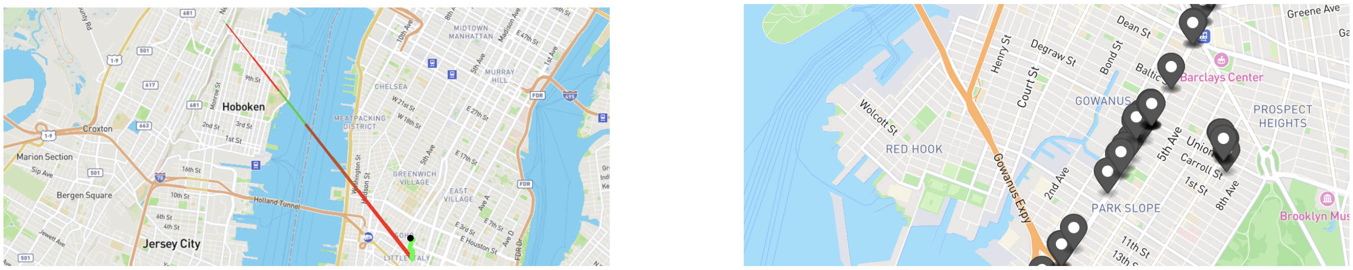
To help bootstrap replication of this work, we provide a GitHub repository at https://github.com/FAR-Lab/urban-fingerprinting to analyze footage gathered both with the SSCI and LSCA approach. For the SSCI approach, our script takes an individual MP4 video clip (or clips, if run in batch) and outputs the average number of pedestrians, bicycles, cycles, cars, trucks, trains, and buses, as described in Section 3.1. For the LSCA approach, the functionality described in the *Graph* class can be utilized, which reads in Nexar images from a folder, pairs the image data with matching metadata, and computes average traffic counts for each aforementioned medium over two dimensions: road edge and time. Note, there is an inverse relationship between time delta and frame density. This approach assumes access to Nexar data, however, with modifications, this can be adapted to any other source of POVCam data.

5 DISCUSSION

Now, we discuss tradeoffs between the small-scale custom instrumentation (SSCI) and large-scale collection and aggregation (LSCA) data collection approaches, potential applications of collected data, future work, and privacy & ethics considerations.

5.1 Trade-offs between SSCI and LSCA approaches

5.1.1 Choosing a data moat. How 'deep' (temporally) or 'wide' (spatially) of a data moat is required is a great factor in selecting an approach. SSCI, at the typical scale of data achievable from the efforts of a small academic team, is well suited for projects that require data from a small geographic area, over a small period of time. An additional advantage of SSCI is the availability of continuous footage; while theoretically available in the LSCA approach, storage constraints make it currently unavailable in the contemporary commerical market. We observe this possibility in our SSCI-compliant project; commuters record their commute every day through the same route, fixing a variable that is only at best approximated in the LSCA approach. LSCA, when using commercial products like that of Nexar, is well suited for projects that require dense data from a large geographic area, over any period of time. An important caveat is that historical data is often limited due to data replenishment, as described in Section 3.2. Even so, With proper tuning of frame collection algorithms, we can craft a more temporally-dense or more spatially-dense sample, depending on need. This also allows researchers to make representative conclusions about an entire area; when we have images from every neighborhood in a city, it becomes possible to not only characterize the city as a whole, but also characterize individual neighborhoods. Current offered products that comply with LSCA only offer non-sequential images; that is to say, analyzing a continuous interaction that occurs over say, 600 subsequent frames, is impossible. This limitation may diminish with storage & compute innovations, but at present, only applications that work with non-continuous footage are possible with this approach. Generally, objects of interest with low pixel 'half-lives'



(a) Native Body Camera GPS. Participant collecting footage was walking in Midtown Manhattan, and yet had coordinates logged across state lines in New Jersey.

(b) OpenHaystack-powered GPS. Participant collecting footage was taking a subway trip in North Brooklyn, and stopped to run errands.

Figure 4: GPS Traces

(or, objects that move nontrivial distances in short periods of time) are hard to detect with non-continuous footage.

5.1.2 Camera perspective. One important advantage of the SSCI paradigm is that the camera perspective is not fixed. In our study, researchers had participants strap action cameras to their chest, producing footage from the pedestrian point-of-view. Anywhere an action camera can be mounted can be depicted; so, off-road vehicles, aircraft, drones, and consumer vehicles are also all within scope. Inversely, with existing LSCA-compliant approaches, footage comes from the fixed perspective of vehicle dashboards; for most approaches requiring surveying of the street, this is not an important limitation.

5.1.3 Privacy. Another key privacy-oriented advantage of the SSCI approach is that consent is implicitly given via the user turning the camera on and revoked by turning the camera off. When having study participants record their commutes for our project, we saw this in action; some participants didn't care much about their individual privacy, leaving the cameras on sometimes for several minutes after arriving home. Others were sure to only turn the camera on after walking a street or two away from their home, ensuring that signal about their address was not embedded in the footage. However, in-house processing is required to properly obfuscate footage at a standard offered by comparable LSCA approaches, seen in Figure 1. We hope to make this overhead more manageable by including a YOLO-based pedestrian and license plate footage anonymization tool in our toolkit (described in Section 4.5). For LSCA-based approaches, additional privacy protections can be derived from the fact that these fast-moving objects are not traceable with current LSCA-compliant products. We expound on privacy considerations further in Section 5.2.

5.1.4 Cost models. A prominent disadvantage of the SSCI approach is cost; we pay participants \$15 / hour for footage, which is a steep price to pay with increasing scale. The LSCA approach is able to amortize these costs across thousands of vehicles, bringing prices down to a few hundred dollars per tens of thousands of images [1]. However, the LSCA approach is not easily replicable; the overhead required to set up such a system is often infeasible for any entity with marginal resources, including academics, non-profits, and smaller government agencies. As such, the most realistic route forward is to rely on private companies that are focused on crafting data collection in a way that maximizes revenue.

5.1.5 Standardization. A disadvantage of the SSCI approach, at least in terms of standardization, is the required human-in-the-loop component of data collection. Humans are required to turn the camera on and turn the camera off, which sometimes leads to accidents where footage is not recorded at all on a commute, or where footage is recorded far too long (i.e. continuing into a participant's place of residence). When using this approach, post-processing that filters out data close to a place of residence is highly recommended. On the other end, the LSCA approach is not as customizable, and more standardized. Coverage is limited to wherever and whenever the data source has cameras, and is only customizable if this is a part of the source's business model. This is useful for most research purposes, as it guarantees a quality of standardization necessary to make statistical conclusions about posed research questions.

5.1.6 Telemetry. Often embedded in the SSCI approach are imaging devices like GoPros with poor GPS localization abilities. Especially in large cities where the urban landscape comprises of tall buildings, poor GPS reception makes adequate localization impossible. As seen in Figure 4a, GPS traces from our body cameras provide unacceptably off-position coordinates; we found that it would log coordinates in Jersey City, across state lines, despite the participant being situated in Midtown Manhattan. This forces an additional overhead of either compute (i.e. detecting intersections with computer vision methods instead of simply aligning coordinates with street map data) or hardware (i.e., the Internet-of-Things-based telemetry solution we developed.) If this overhead is undesirable, leverage the LSCA approach for accurate telemetry data at the per-frame level, including latitude/longitude coordinates and camera heading (i.e. which direction the camera/vehicle is actually facing).

5.2 Privacy / Ethics

5.2.1 SSCI approach. We mention preliminary privacy concerns with the SSCI approach in Section 5.1.3. To discuss further, privacy is more at risk due to the small participant pool size typical of this approach. Whereas in the LSCA approach, data collection vehicles can be anonymized by proxy of being one of thousands and being

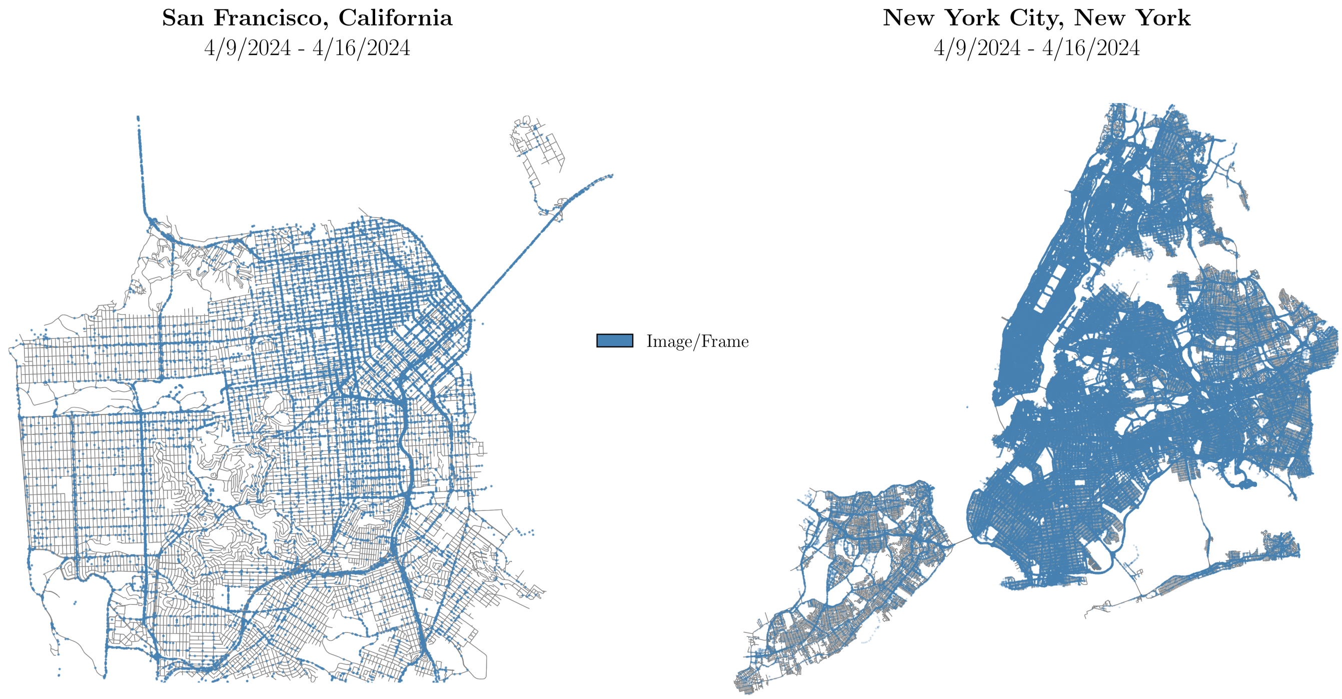


Figure 5: Varying image frame densities in different American cities. New York has much higher coverage than San Francisco over the same 7-day period. Even so, coverage on major roads and arterials remains sufficient for traffic characterization across both cities.

sampled & scrambled before storage, the research team is directly responsible for data storage, obfuscation, and handling in the SSCI approach. As this approach requires interfacing with human data collectors, Institutional Review Board (IRB) approval will likely be required before the start of data collection. Inversely, work using LSCA-compliant data [15] was IRB-exempt.

5.2.2 LSCA approach. There are serious privacy-rooted and ethical considerations at play when utilizing networked POVCams to collect data, with either the homegrown or large-scale approach. Geilser notes a cultural embedding in views towards individual privacy; when Google Street View was rolled out in 2007, American pushback was minimal compared to the general outrage expressed by Germany, despite evidence of equal legislative rights to an 'inviolate personality' [16].

5.2.3 On-edge processing. As edge computing devices become more powerful & smaller in form factor, rigorous analysis will become possible on-premises; companies like Hayden.AI are already able to utilize this, having on-board computers on city buses [8]. When image data never leaves the computing device, you have an embedded design assurance of privacy for those depicted in the footage (unless of course, the analysis being pursued requires individual-level identification).

5.2.4 Obfuscation. The alternate to on-edge processing is to stream footage off-premises of the recording device, and run post-processing on the data that obfuscates pre-determined regions of the image. Pre-determination might come from geometric bounds (for example, if you know the bottom 150 pixels of an image depict contents of a vehicle dashboard, crop that out), or computer-vision based inference. For research purposes, we have found the YOLO library of object detection models very useful for blurring pedestrians, automobiles, bicycles, and motorcycles in POVCam footage.

5.3 Limitations

We want to clarify that the fingerprint of urban traffic data provided by SSCI and LSCA provide representation but not "ground truth" about what is going on in the city, because the sampling methods are sparse and provide only some representation of the whole of urban traffic. In locations where the sampled data coincides with higher-temporal-resolution data, we can establish what percentage of vehicles, pedestrians and other road users are captured by the sporadic samples; what implications this has for the capture rate outside of these validation points is unclear. In future work, it would be useful to develop statistical methods to measure the sampling quality and the confidence interval for these fingerprint methods.

5.4 Future Work

As transportation research with large-scale POVCam expands, it will be interesting to see the new ways in which researchers preprocess, manipulate, and visualize this data. With the two contemporary approaches being at separate poles (either homegrown and small-scale or purchased and large-scale), we envision a third option that strikes a sort of compromise. City fleets are often quite large; for example, New York City operates over 30,000 vehicles

[36], much larger than the 6,000 vehicle fleet private companies like Nexar are able to assemble. A city or publicly-owned fleet of vehicles offers several advantages to that of solely private vehicles. To get a better coverage — and subsequently, understanding — of the characteristics of a city, it is important to have access to several different *types* (in a sense of, typical routing) of vehicles. For example, buses offer a fixed traversal of an identical route. In contrast, roving vehicles like police cars offer a more sparse, yet far-reaching sample of vehicles (similar to the pool of ride-sharing vehicles utilized by Nexar). This would help create a much more robust characterization of a city to set up interaction studies.

6 CONCLUSION

In conclusion, we have introduced POVCam footage as a way to extract traffic-related data from urban environments, which can help illuminate the influence of contextual factors in interaction studies for the AutomotiveUI community. We have illustrated two approaches for data collection and analysis, Small-scale Custom Instrumentation, and Large-scale Collection and Aggregation, and show how these were used to characterize daily traffic. By applying computer vision to captured data, it is possible to gain insight on the urban traffic environment at scale. This method improves upon traditional traffic counts, providing more precise data for enhanced urban planning and traffic management. A GitHub repository containing resources for analyzing footage collected through the two approaches elucidated in this paper is provided at https://github.com/FAR-Lab/urban-fingerprinting.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under awards #2028009 and #2212431 and conducted under Cornell University IRB protocol #IRB0147258. Thanks to Ilan Mandel, Stacey Li, Maureen Hsu, Yuhen Ye and Tsung-yin Hsieh for their advice and assistance throughout the project.

REFERENCES

[1] [n. d.]. AWS Marketplace: Nexar Inc. https://aws.amazon.com/marketplace/seller-profile?id=680456a2-6bcb-4481-ae3f-edd4e0fe6712

[2] Javed Aslam, Sejoon Lim, Xinghao Pan, and Daniela Rus. 2012. City-scale traffic estimation from a roving sensor network. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. 141–154.

[3] Kana Bito, Itiro Siio, Yoshio Ishiguro, and Kazuya Takeda. 2021. Automatic Generation of Road Trip Summary Video for Reminiscence and Entertainment using Dashcam Video. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 181–190.

[4] J. M. Blosseville, C. Krafft, F. Lenoir, V. Motyka, and S. Beucher. 1990. TITAN: NEW TRAFFIC MEASUREMENTS BY IMAGE PROCESSING. In *Control, Computers, Communications in Transportation*, J. P. Perrin (Ed.). Pergamon, Oxford, 35–42. https://doi.org/10.1016/B978-0-08-037025-5.50011-6

[5] Marlon G Boarnet, Eugene Jae Kim, and Emily Parkany. 1998. Measuring traffic congestion. *Transportation Research Record* 1634, 1 (1998), 93–99.

[6] Barry Brown. 2017. The social life of autonomous cars. *Computer* 50, 2 (2017), 92–96.

[7] Barry Brown and Eric Laurier. 2017. The trouble with autopilots: Assisted and autonomous driving on the social road. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 416–429.

[8] Charles Territo. 2022. Hayden AI Completes First Phase of New York MTA Automated Bus Lane Enforcement System Expansion Hayden AI. https://hayden.ai/press

[9] Akemi Takeoka Chatfield and Christopher G. Reddick. 2018. Customer agility and responsiveness through big data analytics for public value creation: A case study of Houston 311 on-demand services. *Government Information Quarterly* 35, 2 (April 2018), 336–347. https://doi.org/10.1016/j.giq.2017.11.002

[10] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. 2023. SAM Fails to Segment Anything? – SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. https://doi.org/10.48550/arXiv.2304.09148 arXiv:2304.09148 [cs].

[11] Jacqueline W Curtis, Andrew Curtis, Jennifer Mapes, Andrea B Szell, and Adam Cinderich. 2013. Using google street view for systematic observation of the built environment: analysis of spatio-temporal instability of imagery dates. *International Journal of Health Geographics* 12, 1 (2013), 53. https://doi.org/10.1186/1476-072X-12-53

[12] Bahar Dadashova, Chiara Silvestri Dobrovolny, and Mahmood Tabesh. 2021. *Detecting Pavement Distresses Using Crowdsourced Dashcam Camera Images*. Report. SAFE-D: Safety Through Disruption National University Transportation Center. https://vtechworks.lib.vt.edu/handle/10919/111515 Accepted: 2022-08-12T19:13:43Z.

[13] Guy Debord. 1967. *The society of the spectacle*. Editions Buchet-Chastel (Paris).

[14] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W. H. Lau. 2021. Location-aware Single Image Reflection Removal. http://arxiv.org/abs/2012.07131 arXiv:2012.07131 [cs].

[15] Matt Franchi, J.D. Zamfirescu-Pereira, Wendy Ju, and Emma Pierson. 2023. Detecting disparities in police deployments using dashcam data. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 534–544. https://doi.org/10.1145/3593013.3594020

[16] Roger C. Geissler. 2011. Private Eyes Watching You: Google Street View and the Right to an Inviolate Personality Note. *Hastings Law Journal* 63, 3 (2011), 897–926. https://heinonline.org/HOL/P?h=hein.journals/hastlj63&i=905

[17] Sanjay Haresh, Sateesh Kumar, M. Zeeshan Zia, and Quoc-Huy Tran. 2020. Towards Anomaly Detection in Dashcam Videos. In *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, Las Vegas, NV, USA, 1407–1414. https://doi.org/10.1109/IV47402.2020.9304576

[18] Alexander Heinrich, Milan Stute, Tim Kornhuber, and Matthias Hollick. 2021. Who Can Find My Devices? Security and Privacy of Apple's Crowd-Sourced Bluetooth Location Tracking System. *Proceedings on Privacy Enhancing Technologies* (2021). https://petsymposium.org/popets/2021/popets-2021-0045.php

[19] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. 2021. Revisiting Shadow Detection: A New Benchmark Dataset for Complex World. *IEEE Transactions on Image Processing* 30 (2021), 1925–1934. https://doi.org/10.1109/TIP.2021.3049331

[20] Matthias Huber. 2020. Video-based Content Analysis. 37–48. https://doi.org/10.4324/9780367321116-5

[21] Jenq-Neng Hwang, Tzong-Der Wu, and Chia-Wen Lin. 1998. Dynamic frame-skipping in video transcoding. In *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175)*. 616–621. https://doi.org/10.1109/MMSP.1998.739049

[22] Junghwan Kim and Kee Moon Jang. 2023. An examination of the spatial coverage and temporal variability of Google Street View (GSV) images in small- and medium-sized cities: A people-based approach. *Computers, Environment and Urban Systems* 102 (June 2023), 101956. https://doi.org/10.1016/j.compenvurbsys.2023.101956

[23] Jiwon Kim and Hani S Mahmassani. 2015. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia* 9 (2015), 164–184.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. https://doi.org/10.48550/arXiv.2304.02643 arXiv:2304.02643 [cs].

[25] Eric Laurier and Hayden Lorimer. 2012. Other ways: Landscapes of commuting. *Landscape Research* 37, 2 (2012), 207–224.

[26] Eric Laurier, Daniel Muñoz, Rebekah Miller, and Barry Brown. 2020. A bip, a beeeep, and a beep beep: How horns are sounded in Chennai traffic. *Research on Language and Social Interaction* 53, 3 (2020), 341–356.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[28] Linbo Luo, Xiangting Hou, Wentong Cai, and Bin Guo. 2020. Incremental route inference from low-sampling GPS data: An opportunistic approach to online map matching. *Information Sciences* 512 (Feb. 2020), 1407–1423. https://doi.org/10.1016/j.ins.2019.10.060

[29] Kevin Lynch. 1964. *The image of the city*. MIT press.

[30] Brendan J. Lyons. 2023. 'Poor communication' led to waste of equipment during pandemic. https://www.timesunion.com/state/article/poor-communication-led-waste-equipment-pandemic-18506732.php

[31] D. H. Maling. 2013. *Coordinate Systems and Map Projections*. Elsevier.

[32] Karl A Miller, Peter Chapman, and Elizabeth Sheppard. 2021. A cross-cultural comparison of where drivers choose to look when viewing driving scenes. *Transportation research part F: traffic psychology and behaviour* 81 (2021), 639–649.

[33] Scott L. Minkoff. 2016. NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City. *Urban Affairs Review* 52, 2 (March 2016), 211–246. https://doi.org/10.1177/1078087415577796 Publisher: SAGE Publications Inc.

[34] Prerana Mukherjee and Brejesh Lall. 2021. Pedestrian Behavior Analytics on Dashcam Videos in Chaotic Environments. *IEEE Sensors Journal* 21, 14 (July 2021), 15660–15669. https://doi.org/10.1109/JSEN.2021.3062762

[35] National Research Council (Ed.). 1997. *Advanced transportation management systems and transportation system management: papers contained in this volume were included on the program for the 76th TRB Annual Meeting held in January 1997.* Number 1603 in Transportation research record Highway operations, capacity and traffic control. National Acad. Press, Washington, DC.

[36] NYC Department of Citywide Administrative Services. 2024. Fleet Management - Department of Citywide Administrative Services. https://www.nyc.gov/site/dcas/agencies/fleet-services.page

[37] Türker Özkan, Timo Lajunen, Joannes El Chliaoutakis, Dianne Parker, and Heikki Summala. 2006. Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation research part F: traffic psychology and behaviour* 9, 3 (2006), 227–242.

[38] Türker Özkan, Timo Lajunen, Joannes El Chliaoutakis, Dianne Parker, and Heikki Summala. 2006. Cross-cultural differences in driving skills: A comparison of six countries. *Accident Analysis & Prevention* 38, 5 (2006), 1011–1018.

[39] Vincenzo Punzo and Marcello Montanino. 2020. A two-level probabilistic approach for validation of stochastic traffic simulations: Impact of drivers' heterogeneity models. *Transportation research part C: emerging technologies* 121 (2020), 102843.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-Decem. IEEE, 779–788. https://doi.org/10.1109/CVPR.2016.91 arXiv:1506.02640

[41] Dorin Shapira, Matt Franchi, and Wendy Ju. 2024. Fingerprinting New York City's Scaffolding Problem with Longitudinal Dashcam Data. https://doi.org/10.48550/arXiv.2402.06801 arXiv:2402.06801 [cs].

[42] Michael Sieverts, Yoshihiro Obata, Mohammad Farhadmanesh, David Sacharny, and Thomas C. Henderson. 2022. Automated Road Asset Data Collection and Classification using Consumer Dashcams. In *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, Bedford, United Kingdom, 1–6. https://doi.org/10.1109/MFI55806.2022.9913859

[43] Michael Sivak, José Soler, Ulrich Tränkle, and Jane Maria Spagnhol. 1989. Cross-cultural differences in driver risk-perception. *Accident Analysis & Prevention* 21, 4 (1989), 355–362.

[44] Cara M. Smith, Joel D. Kaufman, and Stephen J. Mooney. 2021. Google street view image availability in the Bronx and San Diego, 2007–2020: Understanding potential biases in virtual audits of urban built environments. *Health & Place* 72 (Nov. 2021), 102701. https://doi.org/10.1016/j.healthplace.2021.102701

[45] Corey Snyder and Minh Do. 2019. Streets: A novel camera network dataset for traffic flow. *Advances in Neural Information Processing Systems* 32 (2019).

[46] Jifeng Wang, Xiang Li, Le Hui, and Jian Yang. 2017. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. http://arxiv.org/abs/1712.02478 arXiv:1712.02478 [cs] version: 1.

[47] Wuhong Wang, Qian Cheng, Chenggang Li, Dietrich André, and Xiaobei Jiang. 2019. A cross-cultural analysis of driving behavior under critical situations: A driving simulator study. *Transportation research part F: traffic psychology and behaviour* 62 (2019), 483–493.

[48] Henriette Wallén Warner, Türker Özkan, and Timo Lajunen. 2009. Cross-cultural differences in drivers' speed choice. *Accident Analysis & Prevention* 41, 4 (2009), 816–819.

[49] Ariel White and Kris-Stella Trump. 2018. The Promises and Pitfalls of 311 Data. *Urban Affairs Review* 54, 4 (July 2018), 794–823. https://doi.org/10.1177/1078087416673202 Publisher: SAGE Publications Inc.

[50] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. 2018. Seeing Deeply and Bidirectionally: A Deep Learning Approach for Single Image Reflection Removal. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Vol. 11207. Springer International Publishing, Cham, 675–691. https://doi.org/10.1007/978-3-030-01219-9_40 Series Title: Lecture Notes in Computer Science.

[51] Sharon Yavo-Ayalon, Cheng Gong, Harrison Yu, Ilan Mandel, and Wendy Ju. 2022. Walkie-talkie maps–a novel method to conduct and visualize remote ethnography. *International Journal of Qualitative Methods* 21 (2022), 16094069221115519.

[52] Xuaner Zhang, Ren Ng, and Qifeng Chen. 2018. Single Image Reflection Separation with Perceptual Losses. https://doi.org/10.48550/arXiv.1806.05376 arXiv:1806.05376 [cs].

[53] Yun Zhang, Linwei Zhu, Gangyi Jiang, Sam Kwong, and C.-C. Jay Kuo. 2023. A Survey on Perceptually Optimized Video Coding. *Comput. Surveys* 55, 12 (Dec. 2023), 1–37. https://doi.org/10.1145/3571727