Mining Spatio-Temporal Attributes of Anomalies through Large Ego-Vehicle Dataset

Tiffany Ma

May 2022

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Srinivasa Narasimhan
Christoph Mertz
Stephen Smith

Submitted in partial fulfillment of the requirements
for the degree of Master of Science.

Abstract

In recent years, an increasing amount of urban visual big data has been collected through a diverse range of sources, such as taxi vehicle records, video from surveillance cameras, or images captured by mobile devices. The large collection of urban data contains rich implicit information that can help in numerous downstream tasks, such as monitoring for construction management companies, planning for government units, etc. However, it is challenging to efficiently extract the desired information from a dataset of such a large scale. In this work, we focus on developing methods for extracting the spatial attribute and the temporal attribute from these urban visual data. Specifically, we introduce a method of organizing large-scale urban visual data into a spatial-temporal data structure by mining attributes inherent in the data. We demonstrate the effectiveness of our method by using videos captured by the exterior camera of buses to detect and analyze work zones within the captured videos. The raw set of bus data needs to be further preprocessed into a spatial-temporal data structure. Next, we exploit the rich spatial and temporal attributes of bus data in the application of work zone detection and analysis. The goal of this work is to demonstrate the effectiveness of using spatial and temporal attributes to break down large-scale urban visual data and extract insights from large-scale unlabeled data.

Contents

Chapter 1

Introduction

In recent years, an increasing amount of urban visual data is collected through a diverse range of sources, such as vehicle recordings from taxis, videos from surveillance cameras, or images captured by mobile devices. The large collection of urban data contains rich implicit information that can help in numerous downstream tasks. For example, construction management companies can use videos near the construction site to monitor its progress. Government units can improve the infrastructure of the city by analyzing the traffic status of commuters. Autonomous vehicle teams can also use large-scale vehicle data to design algorithms that better adapt to realistic road environments. Although large-scale urban visual data contain rich information for many downstream tasks, analyzing and excavating the value of these big data is a significant challenge [14]. Sources such as surveillance cameras and vehicle recorders continuously collect data over long periods of time. This could accumulate up to terabytes or petabytes of unlabeled raw data. The amount of raw data makes it difficult to extract points of interest from the large pool of data. Another challenge is the lack of structure in the raw data forms. Each task needs to preprocess the raw data into structures that highlight the desired properties each task is analyzing.

Spatial and temporal properties are commonly observed in urban visual data. For example, given a video recording of a taxi that drove through some spatial region, we can infer the boundaries of the traffic region by analyzing the density of cars in each frame of the video. In another example, given a surveillance camera that looks at a parking lot, by analyzing changes at different hours, we can identify the hours at which a parking lot is busiest. Analysis of spatial and temporal attributes enables us to extract interesting events from large-scale urban visual data.

In this work, we demonstrate the effectiveness of using spatial and temporal attributes to extract interesting events. Specifically, we extract instances of work zones from the exterior bus recordings by exploiting the spatial and temporal properties of the bus. One of the commonly overseen sources of spatial-temporal data is bus data. For safety and liability, nowadays transit buses usually have cameras installed to observe the environment around the buses, together with some other sensors such as GPS. These sensors provide rich urban visual data in areas where public transport is widely available. Naturally, buses routinely traverse the same spatial region for a long period of time. Such data can be distilled to construct a dataset of rich spatial-temporal information. To mitigate the challenge of scale and lack of structure, we propose a method to map bus data to a spatial-temporal data structure.

Using this dataset, we want to detect and analyze work zones. Work zones are a common

source of traffic disruption, causing great inconvenience to commuters. Gaining a better understanding of the work zones can benefit various downstream applications. Construction groups can use this knowledge to improve the planning of future construction sites. Government units can also use the learned patterns of the work zones to help make decisions about modifying traffic flows near the work zones. The prediction teams in Autonomous Vehicles companies can apply spatial and temporal relations of the work zones to better react to roadside anomalies. The completion time of the work zones can range from hours to months. For instance, a highway infrastructure change may span up to months and across long regions of the highway; whereas a local road repainting work will be short and contained. To fully understand a work zone from start to finish, we should study patterns about a work zone in the temporal dimension. Most of the current works that study road construction and work zones are mainly in the domain of planning, safety, and transportation. Only a limited number of them utilize vision inputs. Part of this is due to the variability of the work zones and the lack of a concrete definition to detect and analyze these sites. In this work, we aggregate the definitions of road construction used in the transportation, vision, and safety community to find a common ground for understanding these sites through a visual modality.

In summary, we are interested in studying methods for mining spatial and temporal attributes in large-scale urban visual data. In this work, we work with a specific source of urban visual data: recordings collected from exterior cameras of the bus (bus data). The raw set of bus data needs to be further preprocessed into a spatial-temporal data structure. Next, we exploit the rich spatial and temporal attributes in bus data in the application of work zone detection and analysis. The main contributions of this thesis are explicitly stated as follows: We propose a structure for organizing bus data based on their spatial and temporal relations, which facilitates more accessible analysis performed on similar raw data forms. We demonstrate different types of spatial-temporal attributes that are present in work zones detected from bus data and show the potential of such findings.

ADD THESIS OUTLINE

Chapter 2

Background

2.1 Urban Big Data

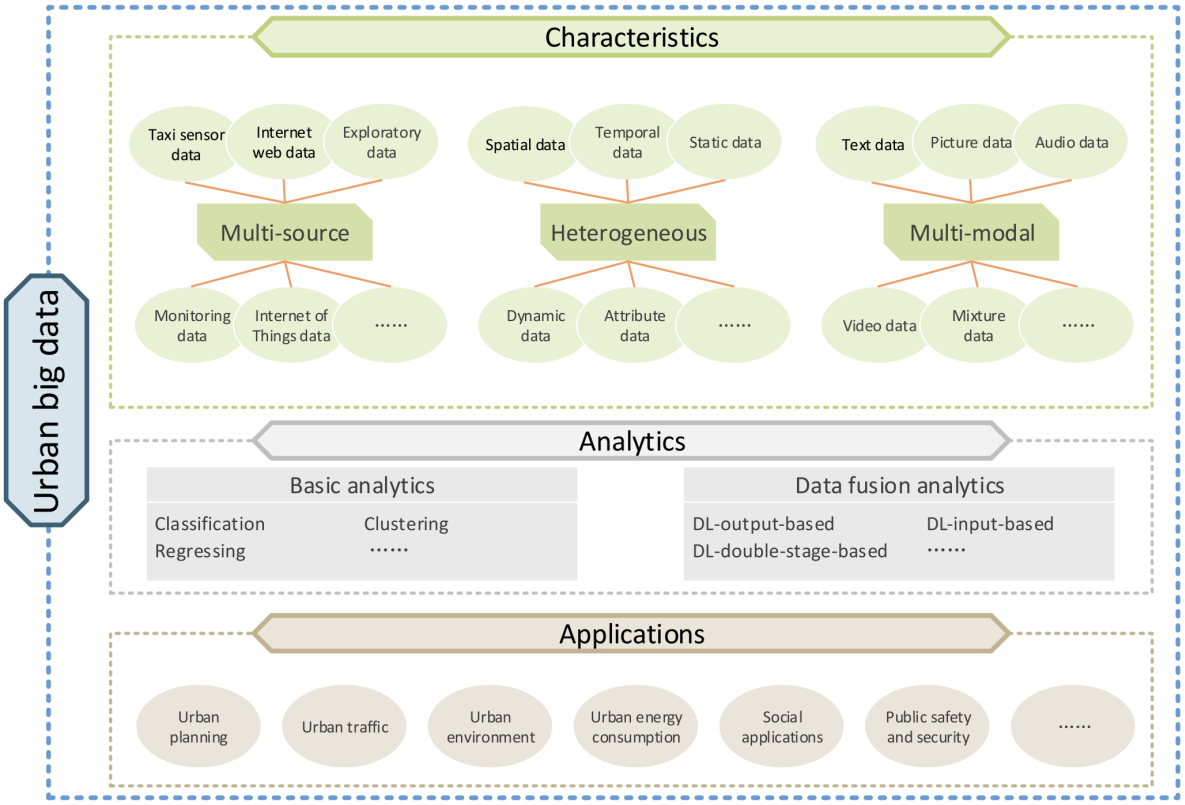


Figure 2.1: This diagram [14] breaks down the definition of urban big data by its characteristics, methods of analysis, and its downstream applications.

Urban big data refers to a combination of structured or unstructured data collected from various urban environments. Zhang et al. [28] summarized the five characteristics of big data as 5Vs, which refers to volume, velocity, variety, veracity and value. Volume refers to the amount

of data available. Velocity refers to the speed with which data is generated and collected. The latter three metrics may vary from task to task. Variety refers to the diversity of data types. Veracity refers to the quality of the collected data. Last but not least, value refers to the value this set of data can provide. These five features showcase the key characteristics that a set of big data should have. It also highlights common challenges in analyzing and analyzing the value of these big data. For example, a large volume of data may contain valuable data, but it is also computationally expensive to parse and analyze a large volume of data. As [16] notes, urban big data is very complex, and we only extract a small part of its knowledge.

From Figure 2.1, we see that [14] lists three broad characteristics that urban big data generally hold. First, we observe that urban big data is collected from a wide range of sources. Meng et al. [15] used real-time GPS readings of taxis, the road network from Internet Web data to infer the volume of urban traffic. Second, we see that urban big data exists in many domains, including spatial, temporal, static, and dynamic data [14]. Lastly, we see that urban big data is multimodal and may include visual, textual, or numeric data. Yi et al. [26] used three datasets to predict air quality, namely, air quality data, weather forecast data, and meteorological data. Weather is represented as textual input (sunny, cloudy, overcast, foggy, etc.), and wind speed is represented as numeric input. With appropriate analysis, urban big data can be used in a wide variety of applications, including urban planning, urban traffic, urban environment, social applications, and public safety and security.

Recently, there has been a rapid increase in the amount of visual data available in urban locations. However, there is still a lack of a structural approach to organize and understand such a vast amount of data. In our case, we are specifically interested in visual urban data. These data can be collected from cameras mounted on stationary traffic light poles, front cameras of moving vehicles, and even out-facing surveillance cameras.

2.2 Spatial-Temporal Data

According to [4], spatial-temporal data comprises spatial and temporal representations. Spatial-temporal data contains three distinct types of attributes, which are non-spatial-temporal, spatial, and temporal attributes. Spatial attributes are those related to the location, shape, and physical aspects of the object. Temporal attributes include timestamps and duration of the in-range data. Non-spatial-temporal attributes typically refer to other additional numeric evaluations of aspects that do not fall under spatial or temporal domains [4]. For example, [4] gave the example of air pollution measures. Air pollution levels and name of location are one example of non-spatial-temporal attributes. [12] used spatial-temporal data from 1,904 residential cars to generate a heat map of vehicle mobility in the city during COVID-19. Based on the heat map, they enforced flexible lockdown strategies to reduce population flow within the city. These examples demonstrate that with adequate analysis, spatial-temporal data can provide rich insights into the event at hand. One common example that uses spatial-temporal analysis is the field of traffic and transportation. Traffic data represents spatial-temporal trajectories that are used to discover periodic patterns. Rao et al. [19] observe that one challenge comes from the influence of nearby objects. Examples of such influence are spatial-temporal events, such as accidents, that may affect traffic patterns in irregular ways.

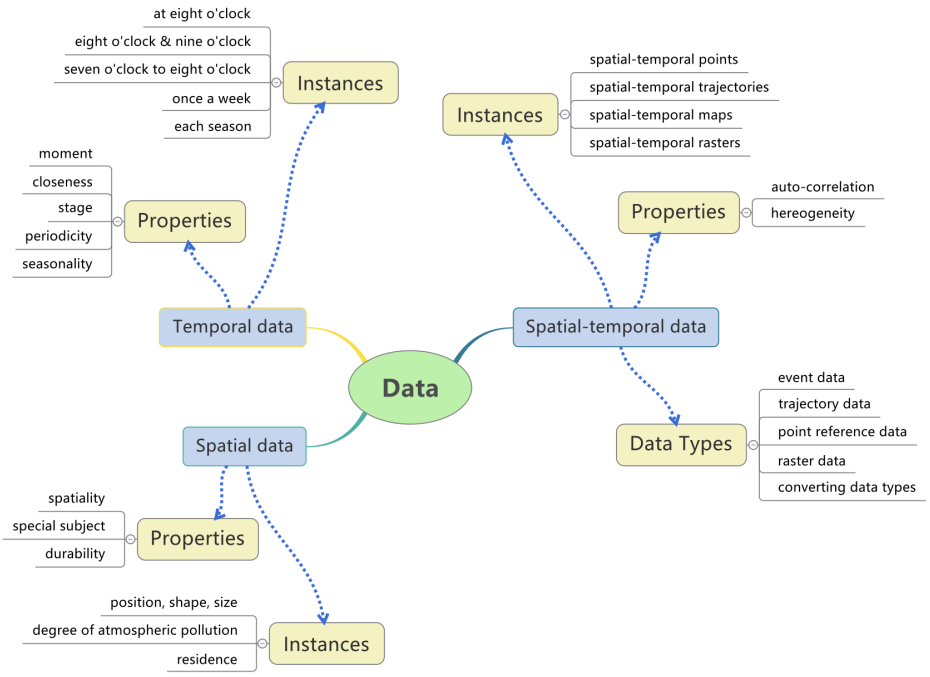


Figure 2.2: Based on the spatial and temporal dimensions, a given data can be divided into data with temporal attribute, data with spatial attribute, and data with both temporal and spatial attributes (spatial-temporal data). [14]

Unlike typical datasets, spatial-temporal data do not follow the assumption that data points are independently and identically distributed. Spatial-temporal objects that exist in neighboring spatial location or time period share similar characteristics that are often related. Spatial-temporal data can exploit relationships that are usually omitted in normal data distributions. Although spatial-temporal data have the potential of extracting rich insights and relationships, these patterns are challenging to mine for the following reasons. First, spatial-temporal relationships are typically high in complexity. Co-located objects in the spatial and temporal domain may influence each other, making detection of relationships difficult. Another difficulty occurs where these relationships are typically implicitly defined. Non-spatial-temporal data have explicit relationships represented through arithmetic relations, such as ordering, instance of, subclass of, and member of. Spatial relationships are built on the basis of qualities or features such as distance, volume, size, and time. These attributes are expressed in a continuous spectrum. These meanings or representations of these attributes can vary depending on interpretation and context, making it difficult to identify these relations [4].

Spatial-temporal data mining (STDM) [10] aims to tackle the above challenge. STDM discovers useful patterns from the dynamic interplay between space and time. STDM contains numerous tasks, such as prediction, clustering, hotspot detection, pattern discovery, outlier analysis, visualization, and visual analytics. These tasks are important in different applications, such as understanding the behavior of objects, scenes, and events. STDM pattern mining works on discovering hidden information (occurrences in space and time, such as movement patterns from

trajectories of spatial-temporal objects). Discovering spatial-temporal associations of trajectories is challenging due to long temporal duration, different moving directions, and lack of spatial accuracy.

Many traffic and transportation datasets contain correlated spatial and temporal attributes. Public transportation data fits the exact definition. We know that public transports traverse on a fixed trajectory. Therefore, for each image captured at a location, it is spatially related to the nearby images. In addition, public transits travel through the same region for a long period of time, adding rich temporal attributes to the collection of images. As [10] observes, modeling trajectory data in a spatial-temporal structure can be challenging. In the latter sections, we describe the design choices made to mitigate these challenges for the purpose of our application.

2.3 Work Zones

Work zones are a common source of traffic disruption, causing great inconvenience to commuters. Gaining a better understanding of work zones can benefit various downstream applications. According to [2], a work zone is an area where roadwork takes place and can involve lane closures, detours, and moving equipment. Highway work zones are set according to the type of road and the work to be done on the road. The work zone can be long or short term and can exist anytime of the year, but most commonly in the summer. Work zones are expected to follow a set of regularizations to ensure the safety of workers and nearby vehicles. There are official guidelines for how to set up a roadside work zone [1]. For example, temporary traffic control signs should be placed at some distance before the actual work zone site. Channelizing devices such as cones, vertical panels, and tubular markers should also be placed around actual construction sites.

Barricade	Used to block travel. Consists of horizontal strips often with orange and white striping (color may vary based on country, city, etc)	
Barrier	Used to block, separate, or channel traffic.	

Temporary Traffic Control (TTC) Sign	Temporarily placed during work period. Usually orange or yellow in the US	
TTC Message Board	Digital sign to provide info	
Arrowboard	Digital sign that uses arrows for directing traffic	
Work Vehicle	Vehicles with specific functions in road work zones, e.g., heavy machinery, vehicles with ttc message boards, bucket trucks, etc.	
Guide Sign	Signs used to direct traffic, e.g., detour signs	
Tubular Marker	Tube shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc.	
Vertical panel	Rectangular shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc	

Cone	Triangular shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc.	
Fence	Used as a barrier to restrict access to work area. Temporary meshed fence. Usually metal or plastic	
Worker	Any workers in the work zone. Usually wearing brightly colored vest and hard hat. Includes flaggers	
Drum	Barrel shaped traffic control device for channeling traffic through a work zone or as a warning of nearby road work	

Table 2.1: A table listing work zone related objects.

2.3.1 Building Roadside Work Zone Understanding

A work zone generally follows the structure in Figure 2.3 and uses objects in Table 2.1 as delineators. These structures are extremely helpful for understanding construction zones. [13] aims at classifying whether a given scene is a construction site or not using related indicators such as vehicle speed, speed limit in road segment, and presence of traffic signs. Each of these indicators is treated as a variable in the Bayesian model and aggregated to produce a probability score of how likely the given scene is a work zone. The goal of this work is to design an online detection pipeline that can handle uncertainty with efficient performance. However, this method does not take into account the spatial information of the indicators. For example, the location of indicator objects in the work zone scene and their relative spatial positioning could provide information on the structure of the understanding of the work zone.

[22] dives deeper into the understanding aspect of work zones by proposing a geometric

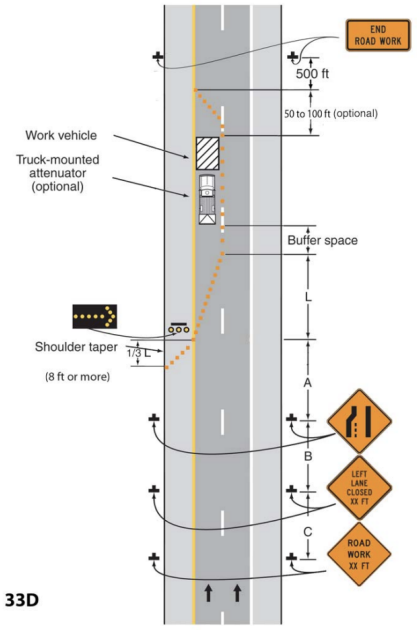


Figure 2.3: An illustration of a short-term stationary setup. The distances A, B, C should be proportional to the work zone speed limit. TTC Signs are placed some distance prior to the actual work zone. Delinators are placed surrounding the boundaries of the work zone. [1]

definition of the boundaries of the work zone. It does so by first considering all detected objects of interest as key points and then lifting the key points onto a bird's eye view plane of the scene. The work zone is mapped into a contour whose boundaries are defined by the key points on that bird's-eye view plane. The authors of [22] also noted that using the RGB input and the LiDAR input produces a more accurate contour. In our work, we will focus on using RGB cameras as our source of visual input since RGB cameras are more widely available.

Previously, most work related to work zone detection has focused on using speed data and lane markers. There is a limited amount of work in the area of detecting vision-based roadside work zones. This is because there is a lack of publicly available work zone delineator datasets. Large ego-vehicle datasets such as BDD100K [27] and Cityscapes [8] contain minimal to no labeled instances of road construction object data. NuScenes [5] contains labels for some work zone delinators, such as barriers and traffic cones. However, training detectors for these objects faces yet other challenges. For example, compared to common instances such as vehicles or persons, NuScenes [5] contains a lower number of construction-related instances. This presents the challenge of class imbalance in the NuScenes dataset [5].

2.3.2 Challenges in Work Zones Understanding

Previous datasets focused more on common objects such as vehicles, pedestrians, and traffic signs. As we begin to build better models to capture these common objects, the ego-vehicle datasets begin to expand and include annotations for rarer instances. For example,example, NuScenes [5] included annotations of traffic cones and barriers in their most recent release

of the dataset. In the most recent release of the Argoverse 2.0 [24] dataset, the authors added the construction cone and the construction barrel to their annotation set. The growing amount of data related to road construction is crucial to developing an understanding of roadside work zones. However, to obtain a holistic understanding of these work zones, we need to take into account more related and rarer objects related to the work zone. into account more related and rarer objects related to the work zone.

2.3.3 Hierarchical Scene Understanding of Work Zones

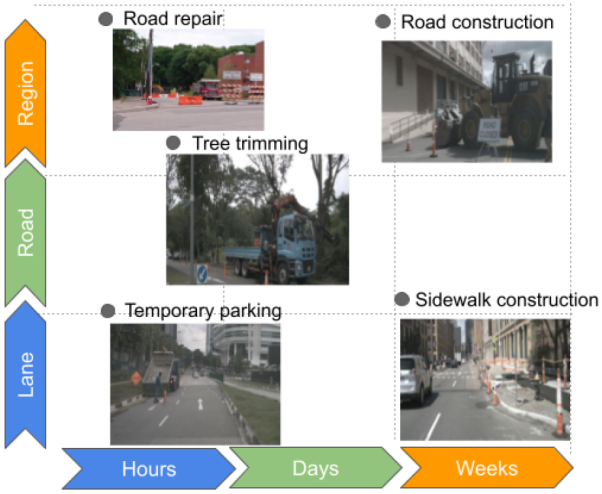


Figure 2.4: Based on the spatial and temporal dimensions, a given data can be divided into data with temporal attribute, data with spatial attribute, and data with both temporal and spatial attributes (spatial-temporal data). [22]

Part of the challenge in understanding road work zones comes from the amount of variance at construction sites. These work zones may differ in the delineator used on site, the scale of construction, and the environment in which they are located (urban, suburban, highway). Most scene understanding approaches are divided into two categories: top-down and bottom-up [17]. Top-down approaches look at the scene on a global scale without focusing on specific objects. Bottom-up approaches start with lower-level features, such as objects, and build understanding from the observed object categories and relationships. Understanding work zones can be built on many levels, such as scene level, frame level, and object level. The first two take a top-down approach, while the latter uses a bottom-up methodology. Below, we will cover different levels of understanding towards construction zones.

As mentioned, road construction zones vary on a variety of scales. Usually, the entire work zone is not fully contained in a single frame. To capture the entire work zone as a whole, we will need to analyze it across a series of neighboring frames that cover the entire work zone. In the work proposed by [22], each work zone is defined by the contoured region on the bird's-eye view plane. Key points detected from a series of neighboring frames are aggregated on the bird's-eye view plane to construct the work zone contour.

At frame level detection, the model loses context from neighboring frames and makes inferences based on a single image. [7] showed efforts to train an image-level classification model from a collection of queried images from the work zone on the road. The authors noticed that the model tends to identify construction zones based on the vibrant orange colors that are commonly seen in construction zones. However, this bias results in many false positives when the same color is visible on non-construction objects. The classification model falsely classified images with similar vibrant, orange color on non-construction objects as a roadside work zone image.

Object level scene understanding is built upon a bottom-up methodology. First, a detection model is trained to detect objects of interest. In this case, the objects of interest are delineators in the work zones. The detection results of each image are extracted and analyzed to build an understanding of the scene. The spatial and geometric relationships between the detected objects can provide information about the image category.

2.4 Object Detection

Object detection is a well-studied problem in computer vision. With the rise of deep learning, CNN-based approaches have become the dominant object detection solution and represent the state of the art. In the deep learning era, object detection can be further divided into two genres: two-stage detection and one-stage detection. The first line of work, pioneered by R-CNN [9], follows a coarse-to-fine detection process that will first generate class-agnostic region proposals of potential objects and then refine and classify them into different categories. Subsequently, Faster R-CNN [20] eliminates selective search by the introduction of the Region Proposal Network (RPN), making it the first end-to-end and near-realtime deep learning detector. Faster R-CNN with a well-performing backbone like ResNet-101 [11] can still be considered the state of the art for object detection. Similarly, we will use Faster R-CNN as the baseline detection model in the bottom-up approach to the detection of roadside work zones.

Chapter 3

Bus Data Organization

As previously mentioned, visual data from public transit systems is a good source for studying roadside work zones for many reasons. Public transports typically traverse on fixed, repeated routes. As such, this allows us to track changes across specific spatial regions on the route across time. This property is especially beneficial to roadside work zones since long-term work zones span across a length of time at a fixed location.

Previous works have used public transport data for statistical analysis [18], [21], [23]. However, most of these analysis focus on attributes directly related to the transportation task, such as number of stops visited, time elapsed at each stop, or number of pedestrians on board, etc. Limited amount of work place their focus on using visual data collected from outfacing cameras from the bus to perform analysis on its surroundings. In this work, we will be using data collected through the BusEdge [25] system. In the following sections, we will give a brief overview of the data collection process and the modalities of the data. Next, we will cover how the collection of bus data is organized into aligned trajectories.

3.1 Data Source and Collection Process

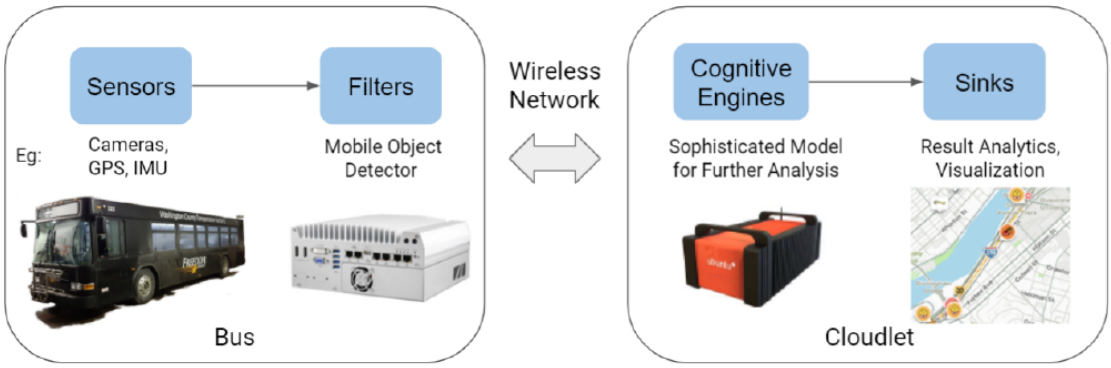
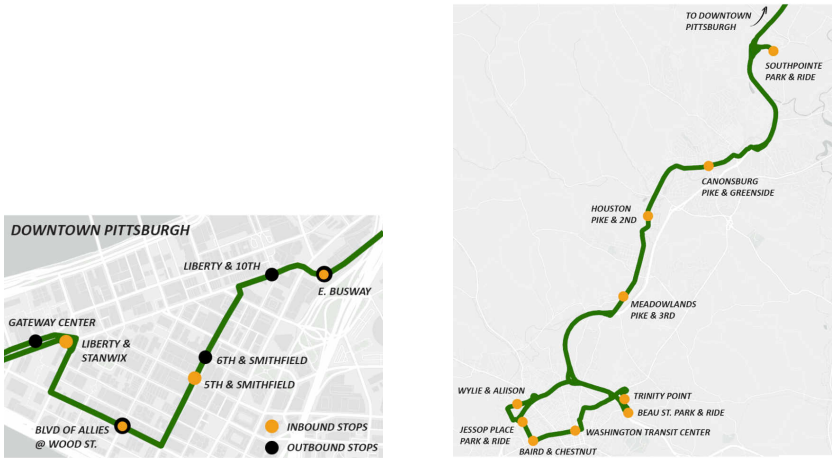


Figure 3.1: Pipeline of the BusEdge system. In this work, we are focusing on analyzing the data collected from the sensors.



(a) Downtown Pittsburgh

(b) Washington Country to Downtown Pittsburgh

Figure 3.2: Visualization of the trajectory taken by the bus

The data used in this work is collected from a metro commuter running between downtown Pittsburgh, PA and Washington County, PA. Two round trips are collected every workday. The traversed route is drawn at Figure 3.2. The inbound trip starts from the intersection of Wylie and Allison Avenue and ends at East Busway at downtown Pittsburgh. The outbound trip starts at East Busway, downtown Pittsburgh and ends at the Transit Center.

In total, there are five cameras installed on the bus. Four waterproof exterior cameras are placed at the top corners of the bus, with two front cameras looking backward and two rear cameras looking forward. The last camera is an outfacing, interior camera placed behind the windshield of the bus. The technical specifications of the cameras are listed in Table 4.5 and Table 3.2. In addition to the five RGB cameras, the bus is also equipped with a Mobile Mark's LTM501 Series Multiband MIMO antenna. This antenna has five build-in antenna, one of them being a GPS antenna. The collected raw data is uploaded to a cloud storage and can be accessed from there. Data is being continuously collected through the bus system. Currently, we have a collection of raw data from August 2019 to February 2022 (time of written).

Brand	Safety Vision
Model	43 series IP camera
Highest Resolution	1920 × 1080
Focal Length	2.8mm

Table 3.1: Specification of Interior Camera

Brand	Safety Vision
Model	37 series IP camera
Highest Resolution	1920 × 1080
Focal Length	2.8mm, 4.0mm

Table 3.2: Specifications of Exterior Cameras

3.2 Raw Data Definition

Our data source traverses between two endpoints: Easy Busway at downtown Pittsburgh and East Chestnut Street Transit Center. We define a run R of the bus as a sequence of data points that traverse between the end points. Each run is traveling in either the inbound or outbound direction. Next, we sort the collection of runs by the starting time of that run and enumerate the runs. Let R_i be the i-th run in the collection. Each run is a sequence of raw data instances, denoted as $R_i = [d_{i,0}, ..., d_{i,j}, ..., d_{i,n}]$. Let d_{ij} be the j-th raw data instance from the i-th run. $d_{i,j} = (c_{i,j}, t_{i,j}, I_{i,j})$, where $c_{i,j} \in \mathbb{R}^2$ is the latitude and longitude coordinate of the data instance, $t_{i,j} \in \mathbb{N}$ representing time in unix timestamp, and $I_{i,j}$ being the RGB image.

Our goal is to detect roadside work zones changes from the collection of runs. Specifically, we are looking for changes along the spatial or temporal axis. To observe changes in these axis, we should first define an ordering for the collecting of runs along the temporal and the spatial axis. Based on the above definition, we see that a natural derivation of the temporal axis is the indices of the runs, since the collection of runs are sorted by the starting time of that run. We will derive the spatial axis in the next few sections. In the following sections, we will formulate the input bus data as a series of trajectories, and cover an overview of algorithms available to cluster and sub-sample these points.

3.3 Data Cleaning

Based on prior knowledge about bus trajectories, we expect the collected data points to be continuous and the trajectories should be coarsely aligned. However, data collected in the wild does not always abide this assumption. There are rare instances when there are missing image or GPS coordinates in the sequence of data points. This could be caused by hardware malfunctions and should be filtered out.

From the bus, images are taken at 5 frames per second continuously. We perform preprocessing to remove low quality images. For example, at low illumination scenarios (before sunrise), the images collected tends to be overly blurry. When the bus is stationary (waiting for the red light), the collected images tends to be highly similar to its neighbors. We remove these instances by threshold at some blur and duplicate score. For each data instance $d_{i,j}$ from run R_i, we filter out images whose $blur\_score$ and dup_score is above some predefined threshold.

$$blur\_score = blur(I_{i,j})$$
$$dup\_score = dup(I_{i,j-1}, I_{i,j}) \text{ with } j > 0$$

3.4 Align Images using Spatial Coordinates

By nature, the trajectories contained by the bus data already traverse on a repeated path. However, we do not have a set of sample trajectory points at hand corresponding to the path. In addition, there are situations where on certain days the bus path deviates from the set routes. For instance, the weekday routes and weekend routes traverse between different points. Another example is when part of the road is blocked, then there may be a small detour at that section. Direct

(a) Blurred Image (b) Duplicate Image at time t (c) Duplicate Image at time $t+1$

Figure 3.3: (a) Images taken before sunrise tends to result in high-blur images. (b) and (c) is a pair of consecutive frames that are taken when the vehicle is stationary. This will generate a pair of images that are close duplicates of one another.

alignment of the bus data may fail to adapt to these changes. Thus, we generate a trajectory baseline by clustering and sampling points form the aggregation of trajectories at different start times and produce a mean representational path. Our goal is to extract a sequence of $K \subseteq \mathbb{R}^2$ coordinate as a representation of the average path the bus data traverses.

3.4.1 Generating a Baseline Trajectory

The work [3] organized a set of algorithms for map construction. They defined map construction as a task that automatically produces or update street map datasets using vehicle tracking data. This matches our purpose of constructing a street map from the bus tracking data. Map construction algorithms can be organized into three categories: point clustering, incremental track insertion, and intersection linking. Specifically, our data structure most closely aligns to the idea of incremental track insertion. Incremental track insertion uses ideas from map matching, where they cluster the tracks and refine them based on a rough baseline. This fits our use case since bus trajectories abide to the assumption that most of the trajectories are roughly aligned on majority of the segments.

We utilized Cao and Krumm [6] implementation of incremental track insertion algorithm. This incremental track insertion approach proceeds in two stages. In the first stage, simulation of physical attraction is used to modify the input tracks to group portions of the tracks that are similar together. This results in a cleaner data set in which track clusters are more pronounced and different lanes are more separated. Then, this much cleaner data is used as the input for a fairly simple incremental track insertion algorithm. This algorithm makes local decisions based on distance and direction to insert an edge or vertex and either merge the vertex into an existing edge, or add a new edge and vertex. From this algorithm, we now have a set of dense coordinates, K, that contains cluster centers of the original runs. In addition, we can impose an ordering to the coordinates in K into a sequence whose coordinate orders follow those of a bus trajectory ordering.

3.4.2 Downsampling the Baseline Trajectory

One challenge in working with Spatio-Temporal data comes from the vast amount of data. Unlike other independently, and identically distributed datasets. Time-series dataset lie on a continuous scale. One practical approach to working with these continuous data is to downsample the data to a coarse representation. Ego-vehicle datasets such as NuScenes [5] does so by extracting key-frames among the continuous stream of data. However, in our case, we are interested in the relationship between neighboring data both on the spatial and temporal axis. Thus, it is important to find a sampling rate that is able to extract an workable set of key points that retains interesting spatial and temporal attributes.

Given the densely clustered baseline trajectory, we want to downsample the points such that the sample clusters can retain spatial information. In the context of bus data, this means that neighboring clusters should share similar imagery features. For example, if two neighboring points are too far apart, then we would lose a lot of the spatial imagery feature since they are aggregated under a single cluster point.

3.4.3 Aligning to the Baseline Trajectory

Now that we have a sequence of baseline coordinates, we want to align the collection of runs onto the baseline coordinates. Let K be the sequence of baseline coordinates composed of the cluster centers from the map construction algorithms. Recall that a run is defined as $R_i = [d_{i,0}, ..., d_{i,j}, ..., d_{i,n}]$ and $d_{i,j} = (c_{i,j}, t_{i,j}, I_{i,j})$. For each data instance $d_{i,j}$, we assign it to the cluster $k = \text{argmin}_{k \in K}\, dist(c_{i,j}, k)$. As previously noted, we can order the coordinates in K such that they follow the trajectory order of the bus. The ordered sequence of coordinates in K is the spatial axis of the bus data.

From the above sections, we defined a structure to organize the vast collection of bus data. First, we split the data into collection of runs R_i. Each run consists of an ordered sequence of data instances. Each data instance stores an image $I_{i,j}$, the time $t_{i,j}$ when the image is taken, and the location $c_{i,j}$ at which the image is taken at. Each location $c_{i,j}$ is further mapped to one of coordinates $k \in K$, an averaged and downsampled representation of the bus trajectory. The ordering by starting time of each run formed the temporal axis, while the ordered sequence K formed the spatial axis.

Chapter 4

Work Zone Detection

Our goal is to identify roadside work zone within spatial-temporal data. Anomalies can occur on the spatial axis (changes across time) or on the temporal axis (changes across a single trajectory). As such, we are interested in answering the following questions:

- Is there a work zone at the current point (coordinate and time)?
- Is there an observable change when interpolating across time or space?

Previously, we have discussed several works that covered approaches of defining a work zone. For instance, [13] utilized a probabilistic model to model a binary classifier for work zones. However, this approach does not consider these objects on a sequential scale when observed by an incoming vehicle. [22] analyzes work zone from a ego-vehicle perspective, but the authors treats all related objects equally as key point indicators, thereby losing information each specific object may indicate.

In this work, we want to dive deeper into work zones from a ego-vehicle perspective. Specifically, we first identify these objects that are correlated to roadside work zones. Next, we want to further analyze their spatial-temporal attributes present in a collection of time series ego-vehicle data (bus data).

4.1 Related Objects at Roadside Work Zones

In most place around the world, these is a written definition for work zones. For instance, in United States, there is written regulations for how to setup temporary traffic control (TTC) zone devices. It is expected that all of the traffic control devices used should conform to the setup written on the manual whenever possible. These include instructions on what objects to place on site and where to place them. Below we will list a few examples of the commonly seen at roadside work zones:

- Barricade: Used to block travel. Consists of horizontal strips often with orange/white striping (color may vary based on country, city, etc)
- Barrier: Used to block, separate, or channel traffic
- Temporary Traffic Control (TTC) Sign: Temporarily placed during work period. Usually orange or yellow in the US.

- Work vehicle: Vehicles with specific functions in road work zones, e.g., heavy machinery, vehicles with ttc message boards, bucket trucks, etc.
- Tubular marker: Tube shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc.
- Verticle panel: Rectangular shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc.
- Cone: Rectangular shaped markers used to divide traffic, mark road edges, divert traffic, restrict turns, etc.
- Drum: Barrel shaped traffic control device for channeling traffic through a work zone or as a warning of nearby road work

Even though all of these objects are categorized as delineators, as noted in the descriptions, they have different spatial implications. These prior knowledge provide us clear characteristics to search for when trying to detect work zones.

4.2 Challenges in Work Zone Object Detection

4.2.1 Lack of Dataset

As previously mentioned, one of the largest challenge towards training a work zone object detector comes from the lack of dataset in this domain. To the best of our knowledge, there is no dataset that contains a majority of the objects above. There are efforts in expanding the annotations towards some of the work zone labels. For instance, in the most recent release of Argoverse 2.0 [24], the authors added annotations for construction cone and construction barrel. Similarly, [5] also contains annotations for construction cones and barricade. However, annotations across datasets cannot be easily combined into training. [29] notes that joining two datasets may introduce conflicts in their joined annotation. This calls for a stronger need for a specialized dataset that focuses on work zone related objects.

4.2.2 Class Imbalance at Work Zones

We noticed most objects appear at roadside work zones at different frequency and counts. For example, channelizing devices such as construction cones and vertical panels appear at most construction zones and in large quantities. On the other hand, work vehicles are generally only present at larger scale work zones. This introduces an imbalance between quantities of each object, which introduces class imbalance to the detection model. Class imbalance has been a long time challenge for detection models. A few ways to combat class imbalance is to apply weighted sampling or weighted loss that is inversely proportional to the quantity of the class label in the training set.

4.2.3 Inter-Class Similarities

Within the domain of work zone objects lies many challenges that may be overseen. In order for these objects to be salient, many of these object share a similar feature of having bright orange color and having reflective stripes. Since most of these objects share similar salient features, it becomes easy for the detection model to get confused about labels between work zones.

4.2.4 Hierarchical Relationships to Common Class Labels

Another common example is when our objects of interests are subclasses of common objects. For instance, the class ”worker” is a specific instance of ”person”. Some of the instances of ”work vehicle” also fall under the broader category of ”cars”. This makes categorizing these challenging because ”person” and ”cars” also appear frequently in most ego-vehicle datasets. This requires the model to learn to differentiate between these two instances under the situation where there is less of the instances of our interest.

4.3 Detection Methods

Our goal is to identify work zones among a continuous stream of data. As previously covered, some of the possible approaches include top-down classification or bottom-up detection. Top-down approaches tends to suffer from the high variance in the distribution of work zone domain. Based on [7], we notice that the classification model ends up falsely relying on salient features such as orange colors. In this work, we will focus on applying a bottom-up approach. Specifically, we want to look for the prior knowledge indicated above to mine information about the bus data.

4.3.1 Dataset

	Traffic Cone	Vertical Panel	Tubular Marker
NuScenes	87,603	-	-
Google Images	1,044	471	283
Bus Data	1,042	344	30

Table 4.1: Number of instances in each class label for each dataset

	Number of Images
NuScenes	
Google Images	302
Bus Data	158

Table 4.2: Number of images for each dataset

As previously mentioned, there yet to be a publicly available dataset that focuses on roadside work zone objects in the vision community. The most relevant dataset is NuScenes [5], which contains 87,603 instances of traffic cones. In an effort to supplement this field of study, we begin gathering roadside work zone images and annotating the instances listed in Section 4.1. The two sources of image data are Google searched images and bus data. In the table 4.1, we can see that different instances of roadside construction objects appear at different frequencies. Among all the work zone related instances, channeling devices, including traffic cones, vertical panel, and tubular marker, are most commonly seen in the construction images. At this time, there is an insufficient amount of training data for all work zone related objects, except for traffic cones. As such, we use NuScenes traffic cone instances as training data to build a single-class detection model, and use traffic cones as an bottom-up indicator of existence of roadside work zones.

4.3.2 Model Architecture

For the purpose of this work, we are interested in detecting roadside work zones and analyzing their spatial-temporal relation. As such, we are more interested in obtaining accurate bottom-up detection than efficient real-time performance. Two-stage proposal generation network generally outperforms one-stage detection model architecture. Thus, for our baseline model, we employ the Faster R-CNN model architecture with ResNet-50 and Feature Pyramid network as backbone. In addition, we also experiment with the current state of the art, which uses a transformer model Swin-T as backbone architecture.

Roadside work zone sites tends to be cluttered and frequently contains cluttered objects. As such, we are interested in seeing if training through instance segmentation would result in better performance. An extension of this work would be performing more accurate analysis and localization of the scene through instance segmentation of the objects.

	Backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN	ResNet-50 FPN	46.8	83.8	45.5	35.6	57.1	65.2
Faster R-CNN	Swin-T FPN	55.7	90.8	59.0	44.7	65.1	75.0
Mask R-CNN	ResNet-50 FPN	54.2	88.9	56.3	42.5	64.4	74.3
Mask R-CNN	Swin-T FPN	53.4	89.2	56.5	39.5	65.8	79.8

Table 4.3: Object detection results on dev set of NuScenes on the Traffic Cones class.

	Backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN	ResNet-50 FPN	11.7	31.6	5.1	11.4	23.3	-
Faster R-CNN	Swin-T FPN	31.8	76.6	19.6	29.8	52.2	-
Mask R-CNN	ResNet-50 FPN	35.8	81.4	25.6	32.5	64.7	-
Mask R-CNN	Swin-T FPN	41.0	77.4	36.4	36.6	73.1	-

Table 4.4: Object detection results on the bus dataset on the Traffic Cones class.

Data	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
NuScenes [Train]	54.2	88.9	56.3	42.5	64.4	74.3
Bus Data [Evaluation]	35.8	81.4	25.6	32.5	64.7	-

Table 4.5: Object detection results using Mask R-CNN with ResNet-50 FPN backbone.

Chapter 5

Results

Our goal is to identify instances of work zones from a set of spatio-temporal data collected from fixed public transport. To do so, we first aggregate the collection of bus data into a base line trajectory and anchor all the bus trajectories onto the sample points from the baseline trajectory. Next, we use a bottom approach to evaluate the work zone score at each point. Finally, we visualize the outputs into a spatial and temporally varying 2D graph to observe how work zones change across the spatial and temporal dimension.

5.1 Sample Outputs

5.1.1 Single Image Results

For each of the images, we generate bounding box predictions for traffic cones. Below we will show a few examples of scenarios where there is a high number of instances predicted and where there are low number of instances predicted.

5.1.2 Spatial Changes

In this section, we want to focus on changes that occurred across the spatial dimension.

5.1.3 Temporal Changes

5.1.4 Spatial-Temporal Visualization

5.2 Analysis

5.3 Spatio-Temporal Attributes

We observe that many of the roadside work zone related scenes consists of spatio-temporal attributes. Below we will show a few examples that we aim to detect from the bus data.

5.3.1 Variations across Time

From the figures above, we see that this is a temporary, small-scale roadwork zone at a fixed location. Across time, there are changes to the number of tubular markers present and changes in the blocked-off zones. A few questions that are of interested are: can we identify the starting and ending timestamps where the workzone started and ended?

2) Variations across spatial location on a fixed run In this example, we use images that belong to the same bus run. The presented construction zone is a large-scale work zone on the side of a highway subsegment. First, we see a TTC sign approximately xxx meters before the start of the actual construction zone to warn the drivers. As the vehicle gets closer to the work zone, there is an increasing number of vertical panels residing on the side of the road. Similarly, we notice a gradual decrease in the number of vertical panels towards the end of this work zone segment. This is one example where it is possible to use prior knowledge to draw insights and predictions about the parts of the work zone.

Chapter 6

Conclusion

Bibliography

[1] 2.3, 2.3

[2] Work/construction zones. URL https://www.nhtsa.gov/. 2.3

[3] Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. A comparison and evaluation of map construction algorithms using vehicle tracking data. *GeoInformatica*, 19(3):601–632, 2015. 3.4.1

[4] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41, 2018. 2.2

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2.3.1, 2.3.2, 3.4.2, 4.2.1, 4.3.1

[6] Lili Cao and John Krumm. From gps traces to a routable road map. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 3–12, 2009. 3.4.1

[7] Brian Chen, Robert Tamburo, and Srinivas Narasimhan. Automatic detection of road work and construction with deep learning model and a novel dataset. URL https://www.youtube.com/watch?v=pPjIS1rz5SU. 2.3.3, 4.3

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2.3.1

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2.4

[10] Ali Hamdi, Khaled Shaban, Abdelkarim Erradi, Amr Mohamed, Shakila Khan Rumi, and Flora D Salim. Spatiotemporal data mining: a survey on challenges and open problems. *Artificial Intelligence Review*, pages 1–48, 2021. 2.2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2.4

[12] Peng Jiang, Xiuju Fu, Yee Fan, Jiri Klemeš, Piao Chen, Stefan Ma, and Wanbing Zhang. Spatial-temporal potential exposure risk analytics and urban sustainability impacts related

to covid-19 mitigation: A perspective from car mobility behaviour. *Journal of Cleaner Production*, 279:123673, 08 2020. doi: 10.1016/j.jclepro.2020.123673. 2.2

[13] Philipp Kunz and Matthias Schreier. Automated detection of construction sites on motorways. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1378–1385, 2017. doi: 10.1109/IVS.2017.7995903. 2.3.1, 4

[14] Jia Liu, Tianrui Li, Peng Xie, Shengdong Du, Fei Teng, and Xin Yang. Urban big data fusion based on deep learning: An overview. *Information Fusion*, 53:123–133, 2020. 1, 2.1, 2.2

[15] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. City-wide traffic volume inference with loop detector data and taxi trajectories. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2017. 2.1

[16] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Multimodal data fusion for sensitive scene localization. *Information Fusion*, 45:307–323, 2019. ISSN 1566-2535. 2.1

[17] Prajakta Ganesh Pawar and V Devendran. Scene understanding: A survey to see the world at a single glance. In *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, pages 182–186, 2019. doi: 10.1109/ICCT46177.2019.8969051. 2.3.3

[18] Simone Porru, Francesco Edoardo Misso, Filippo Eros Pani, and Cino Repetto. Smart mobility and public transport: Opportunities and challenges in rural and urban areas. *Journal of traffic and transportation engineering (English edition)*, 7(1):88–97, 2020. 3

[19] Amudapuram Mohan Rao and Kalaga Ramachandra Rao. Measuring urban traffic congestion-a review. *International Journal for Traffic & Transport Engineering*, 2(4), 2012. 2.2

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2.4

[21] Prasanta K Sahu, Babak Mehran, Surya P Mahapatra, and Satish Sharma. Spatial data analysis approach for network-wide consolidation of bus stop locations. *Public Transport*, 13(2):375–394, 2021. 3

[22] Weijing Shi and Ragunathan Raj Rajkumar. Work zone detection for autonomous vehicles. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1585–1591, 2021. doi: 10.1109/ITSC48978.2021.9565073. 2.3.1, 2.4, 2.3.3, 4

[23] Shijie Sun, Naveed Akhtar, Huansheng Song, Chaoyang Zhang, Jianxin Li, and Ajmal Mian. Benchmark data and method for real-time people counting in cluttered scenes using depth sensors. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3599–3612, 2019. 3

[24] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes,

Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 2.3.2, 4.2.1

[25] Canbo Ye. Busedge: Efficient live video analytics for transit buses via edge computing. Master's thesis, Pittsburgh, PA, July 2021. 3

[26] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, page 965–973, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219822. URL https://doi.org/10.1145/3219819.3219822. 2.1

[27] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2.3.1

[28] Lili Zhang, Yuxiang Xie, Luan Xidao, and Xin Zhang. Multi-source heterogeneous data fusion. In *2018 International conference on artificial intelligence and big data (ICAIBD)*, pages 47–51. IEEE, 2018. 2.1

[29] Bowen Zhao, Chen Chen, Wanpeng Xiao, Xi Xiao, Qi Ju, and Shutao Xia. Towards a category-extended object detector without relabeling or conflicts, 12 2020. 4.2.1